Unrestricted Bridging Resolution

Dissertation

zur

Erlangung der Doktorwürde

der Neuphilologischen Fakultät

der Ruprecht-Karls-Universität Heidelberg

vorgelegt

von

Yufang Hou

aus China

Referent: Prof. Dr. Michael Strube
Korreferent: Prof. Dr. Katja Markert
Einreichung: 02.02.2015

Abstract

Anaphora plays a major role in discourse comprehension and accounts for the coherence of a text. In contrast to *identity anaphora* which indicates that a noun phrase refers back to the same entity introduced by previous descriptions in the discourse, *bridging anaphora* or *associative anaphora* links anaphors and antecedents via lexico-semantic, frame or encyclopedic relations.

In recent years, various computational approaches have been developed for bridging resolution. However, most of them only consider antecedent selection, assuming that bridging anaphora recognition has been performed. Moreover, they often focus on subproblems, e.g., only part-of bridging or definite noun phrase anaphora. This thesis addresses the problem of **unrestricted** bridging resolution, i.e., recognizing bridging anaphora and finding links to antecedents where bridging anaphors are not limited to definite noun phrases and semantic relations between anaphors and their antecedents are not restricted to meronymic relations.

In this thesis, we solve the problem using a two-stage statistical model. Given all mentions in a document, the first stage predicts bridging anaphors by exploring a **cascading collective classification** model. We cast *bridging anaphora recognition* as a subtask of learning fine-grained information status (IS). Each mention in a text gets assigned one IS class, bridging being one possible class. The model combines the binary classifiers for minority categories and a collective classifier for all categories in a cascaded way. It addresses the multi-class imbalance problem (e.g., the wide variation of bridging anaphora and their relative rarity compared to many other IS classes) within a multi-class setting while still keeping the strength of the collective classifier by investigating relational autocorrelation among several IS classes. The second stage finds the antecedents for all predicted bridging anaphors at the same time by exploring a **joint inference** model. The approach models two mutually supportive tasks (i.e., *bridging anaphora resolution* and *sibling anaphors clustering*) jointly, on the basis of the observation that semantically/syntactically related anaphors are likely to be *sibling anaphors*, and hence share the same antecedent. Both components are based on rich linguistically-motivated features and discriminatively trained on a corpus

(ISNotes) where bridging is reliably annotated. Our approaches achieve substantial improvements over the reimplementations of previous systems for all three tasks, i.e., *bridging anaphora recognition*, *bridging anaphora resolution* and *full bridging resolution*.

The work is – to our knowledge – the first bridging resolution system that handles the unrestricted phenomenon in a realistic setting. The methods in this dissertation were originally presented in Markert et al. (2012) and Hou et al. (2013a; 2013b; 2014). The thesis gives a detailed exposition, carrying out a thorough corpus analysis of bridging and conducting a detailed comparison of our models to others in the literature, and also presents several extensions of the aforementioned papers.

Zusammenfassung

Anaphorik spielt eine große Rolle im Textverstehen und ist grundlegend für die Kohärenz eines Textes. Im Unterschied zur direkten Anaphorik, bei welcher eine Nominalphrase eine bereits vorher im Diskurs erwähnte Entität referenziert, verknüpft indirekte Anaphorik (auch Bridging genannt) Anapher und Antezedent vermöge lexico-semantischer, Frame- oder enzyklopädischer Relationen.

In den letzten Jahren wurden verschiedene informatische Verfahren für die Auflösung von Bridgingrelationen entwickelt. Die meisten dieser Ansätze setzen jedoch voraus, dass Bridginganaphern bereits erkannt wurden, und beschäftigen sich nur mit der Auswahl von Antezedenten. Darüber hinaus konzentrieren sich diese Ansätze oft auf Teilaufgaben wie zum Beispiel Meronymie-Bridging oder Bridgingauflösung von definiten Nominalphrasen. Die vorliegende Arbeit beschäftigt sich mit unbeschränkter Bridgingauflösung, das heißt es werden sowohl Bridginganaphern entdeckt als auch Verknüpfungen zu Antezedenten gefunden. Hierbei sind die Anaphern weder auf definite Nominalphrasen beschränkt, noch werden nur bestimmte Relationen zwischen Anapher und Antezedent untersucht, wie zum Beispiel Meronymie-Relationen.

In dieser Arbeit lösen wir das Problem durch ein zweistufiges statistisches Modell. Gegeben alle Erwähnungen in einem Dokument, sagt die erste Stufe Bridginganaphern durch ein kaskadierendes Modell zur kollektiven Klassifikation voraus. Wir fassen die Erkennung von Bridginganaphern als eine Unteraufgabe der Erkennung von Informationsstatus (IS) auf. Jeder Erwähnung in einem Text wird eine Informationsstatusklasse zugewiesen, wobei Bridging eine solche Klasse ist. Das Modell kombiniert Binärklassifikatoren für Minoritätskategorien und einen kollektiven Klassifikator für alle Kategorien durch Kaskadierung. Es befasst sich mit der Klassenunausgewogenheit (d.h. der Seltenheit von Bridginganaphern im Vergleich zu vielen anderen IS-Klassen) in einem Multinomialklassifizierungsszenario. Gleichzeit behält es die Stärke des kollektiven Klassifikators bei, indem relationale Autokorrelation zwischen verschiedenen IS-Klassen berücksichtigt wird. In der zweiten Stufe werden Antezedenten für alle im ersten Schritt erkannten Bridginganaphern gefunden. Hierbei benutzen wir ein Modell, welches die Antezedenten für alle Anaphern gleichzeitig voraussagt. Der von uns entwickelte

Prozess modelliert zwei Aufgaben gemeinsam, die sich gegenseitig beeinflussen: zum einen die Auflösung von Bridginganaphern, zum anderen das Clustering von Anaphern. Hierbei wird die Beobachtung ausgenutzt, dass semantisch oder syntaktisch verwandte Anaphern wahrscheinlich den gleichen Antezedenten haben. Die Modelle für beide Stufen benutzen vielfältige linguistisch motivierte Features und werden diskriminativ auf einem Korpus trainiert, welcher zuverlässig für Bridging annotiert ist. Unser Ansatz erzielt substantielle Verbesserungen im Vergleich zu Reimplementierungen früherer Systeme für Erkennung und Auflösung von Bridginganaphern und der Kombination beider Aufgaben (die sogenannte Bridgingauflösung).

Unseres Wissens nach präsentiert die vorliegende Arbeit das erste System für Bridgingauflösung welches das unbeschränkte Problem in einem realistischen Szenario behandelt. Die in dieser Dissertation vorgestellten Methoden wurden ursprünglich in Markert et al. (2012) und Hou et al. (2013a; 2013b; 2014) behandelt. Die vorliegende Arbeit stellt verschiedene Erweitungen zu diesen Artikeln vor. Insbesondere präsentieren wir eine gründliche Korpusanalyse eines für Bridging annotierten Korpus und führen einen detaillierten Vergleich der von uns entwickelten Modelle zu vorangegangenen Arbeiten aus der Literatur durch.

Acknowledgments

I am so fortunate to have both Michael Strube and Katja Markert as my supervisors. Over the years, they have helped me to shape my taste in research. Michael taught me patiently all things about academic research: how to perform interesting and impactful work, how to write clear yet exciting papers, how to give an impressive talk, and how to be brave to ask questions at conferences. Katja was always helpful during these years. The problem addressed in this thesis is based on her work during her visiting stay at HITS as a Humboldt scholarship holder. She closely guided me in my first year at HITS. In that year, it was a great experience for me to walk into her office and discuss problems every two or three days. I am always inspired by her enthusiasm, passion and curiosity about research. I still remember that I can hear the sound of her typing even we sat in different rooms. After she finished her visiting stay at HITS, we still continued to discuss problems through emails or telephone calls. It is also a memorable experience for me to receive five or six emails from her within ten minutes and decide to which one to reply first.

HITS and the department of computational linguistic department at Heidelberg University offer a wonderful environment to do research. The administrative team at HITS is supportive and helpful. The PhD student colloquium provides me with a platform to get to know various research topics, to discuss new ideas and to get feedback about my own research. I thank the Research Training Group *Coherence in Language Processing* at Heidelberg University, which supported my research and organized the tutorials for PhD students once per year. I also thank the researchers who gave excellent tutorials at the university over the past three years: Ido Dagan, Dan Roth and Noah Smith, for the inspirations I draw from their fantastic works and for their generous advices for my research.

I would like to thank also Coursera and its two founders Andrew Ng and Daphne Koller. The courses I have taken on Coursera since 2011 helped me in various ways. Some memorable courses I took are: Machine Learning by Andrew Ng, Probabilistic Graphical Models by Daphne Koller, and Natural Language Processing by Michael Collins.

Contents

Chapter 1

Introduction

Bridging resolution is the task to recover various non-identity but necessary relations between certain noun phrases (bridging anaphora) and their referring expressions (antecedents) in a text. In Example 1.1, the bridging anaphors, i.e., **The windows**, **The carpets** and **walls** can be felicitously used because they are semantically related via a part-of relation to their antecedent *the Polish center*[1].

(1.1)   If Mr. McDonough's plans get executed, as much as possible of *the Polish center* will be made from aluminum, steel and glass recycled from Warsaw's abundant rubble. **The windows** will open. **The carpets** won't be glued down and **walls** will be coated with non-toxic finishes.

However, bridging resolution is an extremely challenging task. First of all, there are no clear syntactic or other surface clues to indicate the existence of bridging anaphora. In Example 1.2, the bridging anaphor **low-interest disaster loans** associates to the antecedent *the Carolinas and Caribbean*, whereas in Example 1.3 the noun phrase (NP) <u>loans</u> is a generic use. Furthermore, the semantic relations between anaphor and antecedent are usually context-specific. In Example 1.4, the bridging anaphor **The opening show** represents a non-identity relation with its antecedent *Mancuso FBI*, whereas the NP <u>the show</u> represents an identity relation with the expression *Mancuso FBI*.

(1.2)   The $2.85 billion measure comes on top of $1.1 billion appropriated after Hugo stuck *the Carolinas and Caribbean* last month, and these totals don't reflect the additional benefit of **low-interest disaster loans**.

(1.3)   Many states already have Enterprise Zones and legislation that combines tax incentives, <u>loans</u>, and grants to encourage investment in depressed areas.

---

[1]All Examples, if not specified otherwise, are from OntoNotes (Weischedel et al., 2011). Bridging anaphors are typed in boldface; antecedents in italics. This convention is used throughout the thesis.

(1.4)   Over the first few weeks, *Mancuso FBI* has sprung straight from the headlines. **The opening show** featured a secretary of defense designate accused of womanizing (a la John Tower).

...

Most of all though, $\boxed{\textbf{the show}}$ is redeemed by the character of Mancuso.

While the phenomenon illustrated in Example 1.4 (*Mancuso FBI* – $\boxed{\textbf{the show}}$) has attracted a lot of interest in the field of natural language processing under the heading of *coreference resolution*, the bridging relations illustrated in Example 1.1, Example 1.2 and Example 1.4 (*Mancuso FBI* – **the opening show**) have been largely ignored.

In this thesis, we explore the problem of bridging resolution. In order to characterize the nature of the phenomenon in a real scenario, we first carry out a statistical study in a corpus where bridging is reliably annotated. Secondly, we investigate suitable computational methods to resolve bridging automatically. We apply joint inference models to address two subtasks of bridging resolution separately, i.e., bridging anaphora recognition and bridging anaphora resolution. We then model bridging resolution in a pipeline fashion: (1) recognizing bridging anaphors and (2) finding the correct antecedent among candidates.

In this chapter, we start with discussing the linguistic background of the definition of bridging in Section 1.1, drawing distinctions between bridging in this thesis and other related linguistic phenomena. Section 1.2 motivates bridging resolution by discussing practical applications. Section 1.3 lists a variety of research questions that we aim to explore. We summarize the main contributions of this thesis in Section 1.4. Finally, Section 1.5 gives an overview of the dissertation.

1.1   Problem Definition

Before defining the problem that we explore in this thesis (Section 1.1.2), we first review the main linguistic studies about the definition of bridging (Section 1.1.1). A further, more complete survey of various studies of bridging can be found in Chapter 2. We discuss the relations between bridging and other related linguistic phenomena in Section 1.1.3.

1.1.1   Linguistic Studies of Bridging Definition

Clark's taxonomy of bridging relations.   The term *bridging* was originally introduced by Clark (1975) to refer to the construction of implicatures, which is an obligatory part of the process of comprehension. Clark (1975) presents a broad classification of bridging relations (see Table 1.1). In the two categories of indirect reference in Table 1.1, a bridging implicature is needed when the listener cannot find a direct antecedent for a reference in his memory. The

direct antecedent is an object/event explicitly mentioned in the previous discourse, while the indirect antecedent means that the existence of an antecedent has to be inferred.

However, Clark does not offer an adequate definition of bridging. Moreover, Clark's taxonomy of bridging relations seems too broad. It even includes anaphoric use of NPs in which anaphoric NPs have an identity relation with their antecedents (Table 1.1, Example (1) – (3)). We will narrow the definition of bridging in Section 1.1.2, leaving out such cases as well as the rhetorical relation cases (Table 1.1, Example (10) – (13)).

Type	Example
1. Direct reference	
Identity	(1) I met *a man* yesterday. **The man** told me a story.
Pronominalization	(2) I met *a man* yesterday. **He** told me a story.
Epithets	(3) I met *a man* yesterday. **The bastard** stole all my money.
Set membership	(4) I met *two people* yesterday. **The woman** told me a story.
2. Indirect reference by association	
Necessary parts	(5) I looked into *the room*. **The ceiling** was very high.
Probable parts	(6) I walked into *the room*. **The windows** looked out to the bay.
Inducible parts	(7) I walked into *the room*. **The chandeliers** sparkled brightly.
3. Indirect reference by characterization	
Necessary roles	(8) John was *murdered* yesterday. **The murderer** got away.
Optional roles	(9) John was *murdered* yesterday. **The knife** lay nearby.
4. Reasons, causes, consequences and concurrences	
Reasons	(10) John fell, what he wanted to do was scare Mary.
Causes	(11) John fell. What he did was trip on a rock.
Consequences	(12) John fell. What he did was break his arm.
Concurrences	(13) John is a Republican. Mary is slightly daft too.

Table 1.1: Clark's taxonomy of bridging relations. Examples are from Clark (1975).

Associative anaphora. The term *associative anaphora* is usually used for definite NPs. Hawkins (1978) analyzes the associative anaphoric use of definite NPs. According to Hawkins, speaker and hearer may have shared knowledge of the relations between certain objects (the *triggers*) and their components or attributes (the *associates*). In Example 1.5 (borrowed from Hawkins (1978)), the mention of *a wedding*, can trigger off the hearer's associations which permit that **The bride** and **the cake** can be used felicitously.

(1.5)   I went to *a wedding* last weekend. **The bride** was a friend of mine. She baked **the cake** herself.

Although work on associative anaphora has typically concentrated on definite descriptions, Löbner (1998) points out that associative anaphora does not bind to definiteness. For instance, in Example 1.6 (modified from Example (7) in Table 1.1), **a window** is clearly interpreted as one of the room's windows. In Section 1.1.2, we will follow Hawkins' notion of "*associative anaphora*" but do not restrict it to definite NPs.

(1.6)   I walked into *the room*. **The chandeliers** sparkled brightly and **a window** was broken.

Inferrables.   Prince (1981; 1992) proposes a taxonomy of *Assumed Familiarity* for discourse entities represented by NPs in a text. A discourse entity is *inferrable* if the speaker assumes that the hearer can infer it from certain other discourse entities already mentioned. Furthermore, the connections between an inferrable NP and the previous discourse entities should not be specified as part of the NP itself, e.g., The bride of the wedding shown in Example 1.7 (modified from Example 1.5) is not an inferrable NP. Indeed, Prince calls such NPs "*Containing Inferrables*".

(1.7)   I went to *a wedding* last weekend. The bride of the wedding was a friend of mine.

Prince assesses the age of an entity as old or new from the hearer's head and from the discourse model respectively. She then discusses *Inferrables* in terms of Hearer-new/Hearer-old and Discourse-new/Discourse-old. On the one hand, *Inferrables* are technically Hearer-new and Discourse-new in the sense that the hearer is not expected to already have in his/her head the entity in question. On the other hand, *Inferrables* are like Hearer-old and Discourse-old entities: they rely on certain assumptions about what the hearer does know (e.g., a wedding typically has a bride), and they rely on entities which are already in the discourse model to trigger the inference, e.g., *a wedding* – **The bride** in Example 1.5.

Prince (1992) also claims that both indefinite and definite NPs may represent *Inferrable* entities. She further characterizes the form of *Inferrables* with regard to definite/indefinite. *Inferrables* are indefinite when the inference for a trigger entity T is something like "a T typically has Es (entities) associated with it" and "the Inferrable refers to a proper subset of the set of Es" (Prince, 1992, p.20). Thus the *Inferrable* in Example 1.8 is represented by an indefinite NP, i.e., **a page**.

(1.8)   I picked up *that book I bought* and **a page** fell out.

Prince's *Inferrables* are close to our bridging anaphora definition which will be discussed in Section 1.1.2.

Mediated/bridging. Building on Prince's work (1981; 1992), Nissim et al. (2004) define the Information Status (IS) of an entity as the degree to which the entity is available to the hearer with regard to the speaker's assumptions about the hearer's knowledge and beliefs. They present a scheme for IS in dialogue. `Old` entities are known to the hearer and have been mentioned previously. `Mediated` entities have not been mentioned before but are also not autonomous, i.e., they can only be correctly interpreted by reference to another entity or to prior world knowledge. All other entities are `new`, i.e., they have not yet been introduced into the discourse and the hearer cannot infer them from previously mentioned entities or their prior world knowledge.

Following Nissim et al. (2004) in distinguishing three major IS categories (`old`, `new` and `mediated`), Markert et al. (2012) propose an annotation scheme for IS in written text. Markert et al. (2012) distinguish six subcategories for `mediated`. A detailed description of this scheme can be found in Section 3.1 of Chapter 3. Among all subcategories of `mediated`, `Mediated/bridging` corresponds to Prince's *Inferrables*. We discuss this subcategory in detail in the following section.

1.1.2   Bridging in This Thesis

This thesis focuses on **unrestricted bridging** where bridging anaphors are NPs that refer back to discourse expressions (antecedents) in a manner that is neither *coreferential* nor *comparative*. We characterize **unrestricted bridging** from the following three perspectives.

1. Bridging anaphors are not restricted to definite NPs. Our understanding of bridging anaphora in this thesis is close to Hawkins's *associative anaphora* and Prince's class of *Inferrables* discussed in the previous section in the sense of "indirect anaphoricity". Therefore, unlike previous empirical studies on bridging which include cases where antecedent and anaphor are coreferent but do not share the same head noun (Vieira & Poesio, 2000; Bunescu, 2003), we restrict bridging to non-coreferential cases. We also exclude *comparative anaphora* (Modjeska et al., 2003) from bridging. Such referential NPs often have clear surface indicators (e.g., NPs with modifiers "other" or "another", such as **three other cities** in Example 1.9).

(1.9)   About 200,000 East Germans marched in *Leipzig* and thousands more staged protests in **three other cities**.

Furthermore, unlike most previous empirical work (Vieira & Poesio, 2000; Lassalle & Denis, 2011; Cahill & Riester, 2012) which limit bridging anaphora to definite NPs, we do not restrict bridging anaphora to any specific type of NPs. We argue that anaphoricity of bridging anaphora is a pragmatic tendency that exists independently of syntactic properties of

NPs. In Example 1.10, the bridging anaphor **The five astronauts** is a definite NP modified by the definite article "the", whereas the bridging anaphor **touchdown** is a bare NP (NPs without determiners). In Example 1.11, the bridging anaphor **A food caterer** is an indefinite NP modified by the indefinite article "a".

(1.10) *The space shuttle Atlantis* landed at a desert air strip at Edwards Air Force Base, Calif., ending a five-day mission that dispatched the Jupiter-bound Galileo space probe. **The five astronauts** returned to Earth about three hours early because high winds had been predicted at the landing site. Fog shrouded the base before **touchdown**.

(1.11) Still, *employees* do occasionally try to smuggle out a gem or two. **One man** wrapped several diamonds in the knot of his tie. [...] **A food caterer** stashed stones in the false bottom of a milk pail.

2. Bridging antecedents are not restricted to entities and are necessary to interpret bridging anaphors in a given situation. First, the bridging antecedent can be an NP (Example 1.12), a verb (Example 1.13), or a speech act denoted by a clause or in some cases a sentence (Example 1.14).

(1.12) I bought *a bicycle*. **A tyre** was already flat.

(1.13) I *traveled* to Frankfurt. **The train** was very full.

(1.14) *Why do humans collaborate?* **The answer** lies in ...

Second, the bridging antecedent is **necessary** for the hearer to interpret the bridging anaphor in question in a given situation. In Example 1.15, the bridging anaphor **any loosening this year** is interpreted as "any loosening this year of the high interest rates". Here, *the high interest rates* is crucial to understand the anaphor **any loosening this year** in its context. One may argue that according to the whole context, **any loosening this year** could be further inferred as "any loosening this year of the high interest rates (loosen what?) by the British government or the government officials (who loosen?)". However, we only consider discourse expressions which are **necessary** to interpret the bridging anaphor – on the basis of common sense knowledge and the context – as the antecedents. Such "necessary discourse expressions" are usually implicit core semantic roles of the anaphor (Example 1.15). Yet not all bridging antecedents are semantic roles of the bridging anaphor. In *set/membership* bridging in which the bridging anaphor is a subset or a member of the antecedent, the antecedent could be any NP which makes the anaphor accommodate into the context so that the text is coherent (e.g., *employees* – **One man** in Example 1.11). The relation between bridging and implicit semantic roles is discussed further in Section 1.1.3.

(1.15)   Britain's current account deficit dropped to [. . .] Chancellor of the Exchequer Nigel Lawson views *the high interest rates* as his chief weapon against inflation, [. . .] Officials fear that **any loosening this year** could rekindle inflation or further weaken the pound against other major currencies.

3. Bridging relations are not restricted to meronymic relations.   Unlike previous empirical works (Poesio et al., 2004a; Markert et al., 2003) which only concentrate on mereological bridging relation, we consider all possible bridging relations in running text. Bridging relations are tremendously diverse from a semantic perspective. It could be a meronymic relation between antecedent and anaphor (e.g., *the Polish center* – **The windows** in Example 1.1), a set membership relation (e.g., *employees* – **One man** in Example 1.11), a functional relation between organization or country and people (e.g., *Japan* – **the chief cabinet secretary** in Example 1.16), or an attribute-of relation (e.g., *meat, milk and grain* – **prices** in Example 1.17), inter alia.

(1.16)   Yet another political scandal is racking *Japan*. [. . .] On Friday, **the chief cabinet secretary** announced that **eight cabinet ministers** had received five million yen from the industry.

(1.17)   In June, farmers held onto *meat, milk and grain*, waiting for July's usual state-directed price rises. The communists froze **prices** instead.

1.1.3   Bridging Resolution and Implicit Semantic Role Labeling

Recently, Ruppenhofer et al. (2010) presented a task at SemEval-2010 that includes an implicit (core) role linking challenge based on FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005). They cover a wide variety of nominal and verbal predicates. Another vein of research on implicit semantic role labeling only focuses on a small set of ten nominal predicates derived from verbs (Gerber & Chai, 2012; Laparra & Rigau, 2013), such as *investor*, *loan* and *plan*.

There is a partial overlap between bridging resolution and implicit semantic role labeling, i.e., in some bridging cases, antecedents are implicit semantic roles of bridging anaphors. The main differences between bridging resolution and implicit semantic role labeling lie in the following aspects.

First, bridging resolution considers all possible nominal bridging anaphors in running text. Some bridging anaphors are not considered as "nominal predicates" in (implicit) semantic role labeling, e.g., **One man** in Example 1.11.

Second, implicit semantic role labeling for nominal predicates tries to link all possible implicit core roles for the nominal predicate in question. Yet not every nominal predicate

under consideration is a bridging anaphor from the discourse entity's perspective. In Example 1.18[2], the nominal predicate **losses** in the first sentence has three explicit roles: the role *arg0*, the entity losing something; *arg1*, the thing lost; and *arg3*, the source of that loss. These three arguments are considered as implicit roles for the nominal predicate **losses** in the second sentence. However, from a discourse entity's point of view, the second occurrence of **losses** in this example is not a bridging anaphor. Instead, it is an `old` entity which is coreferential with the NP **losses** introduced in the previous sentence.

(1.18) [The network]$_{arg0}$ has been expected to have [**losses**]$_{predicate}$ [of as much as \$20 million]$_{arg1}$ [on baseball]$_{arg3}$ this year. It isn't clear how much those [**losses**]$_{predicate}$ may widen because of the short Series.

1.2 Motivations for Bridging Resolution

Bridging resolution plays an important role to establish entity coherence in a text. Barzilay & Lapata (2008) model local coherence by using entity grid which is based on (approximate) coreference. Table 1.2 shows the corresponding the entity grid for Example 1.1. The rows of the grid correspond to sentences, the columns correspond to discourse entities. Ideally, a discourse entity is the set of coreferent noun phrases in a text. Each cell of the grid represents the presence or absence of the corresponding entity in the corresponding sentence. The absence of an entity in a sentence is signaled by gaps ($-$). The presence of an entity in a sentence is signaled by the entity's grammatical function in this sentence: S stands for subject, O stands for object, X stands for others. Example 1.1 does not exhibit any coreferential entity coherence, i.e., the three sentences do not share entities. In this example, entity coherence can only be established when bridging is resolved, i.e., *center* is (implicitly) realized in sentence 2 and sentence 3 in Table 1.2. Therefore entity-based coherence models can profit from bridging resolution.

	Mr. McDonough	plans	*center*	aluminum	Warsaw	rubble	**windows**	**carpets**	**walls**
S1	X	S	X	X	X	X	-	-	-
S2	-	-	-	-	-	-	S	-	-
S3	-	-	-	-	-	-	-	S	S

Table 1.2: The entity grid for Example 1.1

Furthermore, a number of text understanding applications can benefit from bridging resolution. We describe some of them below.

---

[2]The example is from Laparra & Rigau (2013).

Textual entailment. Textual entailment (TE) determines whether the meaning of a given textual assertion, i.e., Hypothesis (H), can be inferred from the meaning of certain text (T) (Dagan et al., 2006). Mirkin et al. (2010) show that bridging has the potential of improving textual entailment recognition. The Fifth Recognizing Textual Entailment (RTE-5) challenge introduces a search task, where the sentence under consideration is interpreted in the context of the full discourse. In Example 1.19[3], resolving bridging in the discourse (*China* – **A recent accident**) is necessary for determining entailment.

(1.19)   **T:** *China* seeks solutions to its coal mine safety. **A recent accident** has cost more than a dozen miners their lives.
   H: A mining accident in China has killed several miners.

Mirkin et al. (2010) analyzed 120 sentence-hypothesis pairs of the RTE-5 development set. They found that 44% of the pairs contain reference relations whose resolution is mandatory for inference. Among those reference relations, 27% are bridging relations.

Question answering. Question answering (QA) systems take users' natural language questions and automatically locate answers from large collections of documents. In a real-world setting, questions are not asked in isolation, but rather in a cohesive manner that involves a sequence of related questions to meet the user's information needs. Therefore, the tenth Text REtrieval Conference (TREC 2001) initiated a context QA task (Voorhees, 2001). In this task, the questions are grouped into different series, such as the one shown in Example 1.20.

(1.20)   **Q1:** Where is *Hawaii* located?
   Q2: What is **the state fish**?
   Q3: Is it endangered?
   Q4: Any **other endangered species**?

To answer these kinds of questions, a QA system needs to track the discourse entities across the individual questions of a series. For example, resolving the bridging relation between question Q2 and question Q1 (i.e., *Hawaii* – **the state fish**) is mandatory to answer question Q2. In this example, question Q4 relates to other questions in a more complex way. In Q4, **other endangered species** is a comparative anaphor with *it* in Q3 or *the state fish* in Q2 as the antecedent[4]. It is also a bridging anaphor in this context with *Hawaii* in Q1 as the antecedent. In order to answer question Q4, we need to restate the question with the discourse processing results from the context: what are endangered species other than fish in Hawaii?

Among all systems participating the TREC 2001 QA track, Harabagiu et al. (2001) integrate a meronymic reference resolution module to resolve meronymic bridging in context

---

[3]The example is from Mirkin et al. (2010).
[4]*it* in Q3 and *the state fish* in Q2 are coreferent.

questions. For example, **galleries** from "Which galleries were involved?" are referenced as a part of the *museum* from the preceding question "Which museum in Florence was damaged by a major bomb explosion in 1993?".

Opinion mining.   Opinion mining or sentiment analysis refers to the computational treatment of *opinion*, *sentiment*, and *subjectivity* in a text (Pang & Lee, 2008). Among various approaches, information extraction-oriented opinion mining focuses on extracting and analyzing opinions of multiple aspects or features of a single item from unstructured text data, i.e., who feels how on which aspect of which subject. This task includes two sub-tasks: (1) the identification of item features; and (2) extraction of opinions associated with these features.

Kobayashi et al. (2007) collected a corpus of weblog posts which consists of around 2,800 articles in four domains: restaurant, automobile, cell phone and video game. Four constituents are annotated in the corpus: opinion holder (e.g., the author of the weblog), subject (e.g., a car model name), aspect (a part, member or related object, or an attribute of the subject on which the evaluation is made, e.g., *engine* or *size*) and evaluation (e.g., *good* or *poor*). Aspects of a subject may have a hierarchical structure, such as "*the leather cover of the seats of a car*". Kobayashi et al. (2007) then explore bridging resolution for aspect-of relation extraction. They show that resolving bridging by combining contextual clues and distributional semantic features yields the best result on aspect-of relation extraction.

1.3   Research Questions

Given the bridging definition in Section 1.1.2, we aim to investigate a variety of research questions in this thesis. We organize them into two groups.

Characterizing bridging on the basis of a corpus analysis.   Prior corpus-linguistic studies on bridging are beset by three main problems. First, reliability is not measured or low (Fraurud, 1990; Poesio, 2003; Gardent & Manuélian, 2005; Riester et al., 2010). Second, annotated corpora are small (Poesio et al., 2004a; Korzen & Buch-Kromann, 2011). Third, they are often based on strong untested assumptions about bridging anaphora types, antecedent types or bridging relations, such as limiting it to definite NP anaphora (Poesio & Vieira, 1998; Poesio et al., 2004a; Gardent & Manuélian, 2005; Caselli & Prodanof, 2006; Riester et al., 2010; Lassalle & Denis, 2011), to NP antecedents (all prior work) or to part-of relations between anaphor and antecedent (Markert et al., 2003; Poesio et al., 2004a).

To gain a better view of the nature of the phenomenon, we carry out a statistical study in a corpus where bridging is reliably annotated and is not limited to specific anaphora/antecedent types or specific relations (Chapter 3). We wish to explore: How often does bridging appear

in texts? Among all bridging anaphors, how many of them are definite NPs? Is bridging a local coherence phenomenon like previous work claimed (Poesio et al., 2004a)? How do the typical bridging relations (i.e., part-Of relation and set membership relation) distribute in the corpus? Are there any common lexical or syntactic patterns indicating bridging? How does bridging interact with other discourse phenomena (i.e., discourses relations)? What are the most frequent bridging relations?

Resolving bridging automatically. The linguistic knowledge of bridging from the above empirical work as well as from various linguistic studies have been brought to bear in designing the model for bridging resolution. Bridging resolution involves two subtasks: (1) bridging anaphora recognition; and (2) bridging anaphora resolution. In this thesis, we wish to explore: Can we apply computational methods to solve these tasks automatically? Which computational model suits each task separately? What kind of features are crucial to recognize and to resolve bridging anaphora? Is there a way of incorporating world knowledge? How does entity salience (`global` and `local`) affect bridging anaphora resolution?

1.4 Contributions of the Thesis

The main contributions presented in this thesis are summarized below.

1. We model *bridging anaphora recognition* as a subtask of learning fine-grained information status. We design **discourse structure**, **lexico-semantic** and **genericity detection features** and integrate these features into a **cascading collective classification algorithm**. The model combines the binary classifiers for minority categories and a collective classifier for all categories in a cascaded way. It addresses the multi-class imbalance problem (e.g., the wide variation of bridging anaphora and their relative rarity compared to many other IS classes) within a multi-class setting while still keeping the strength of the collective classifier by investigating relational autocorrelation among several IS classes. Our model achieves state-of-the-art performances both for the overall IS classification accuracy as well as for bridging anaphora recognition.

2. We model *bridging anaphora resolution* by exploring a joint inference framework. We develop **semantic**, **syntactic** and **salience features** based on linguistic insights for this task. Our **joint inference model** expresses the interesting topological property of bridging, i.e., *semantically/syntactically related anaphors are likely to be **sibling anaphors**, and hence share the same antecedent*. Sibling anaphors are bridging anaphors which share the same antecedent with other bridging anaphors, e.g., **The windows** and **walls** in Example 1.1 are sibling anaphors. This joint inference model significantly outperforms other local models and baselines.

3. In the task of *bridging anaphora resolution*, we propose **a novel method to select antecedent candidates** for bridging anaphors by exploring the discourse relation `Expansion` and modeling salience from different perspectives. The method reflects different interpretive preferences (*local* or *global* focus) of bridging anaphors. This is the first empirical work to show the positive impact of discourse relations on bridging resolution.

4. We model **unrestricted bridging resolution** using a two-stage statistical model: (1) recognizing bridging anaphors and (2) finding the correct antecedent among candidates. Both components explore rich linguistic features and joint inference models as described above. This is the first full bridging resolution system that tackles the **unrestricted phenomenon** (i.e., bridging anaphors are not limited to definite NPs and semantic relations between anaphors and their antecedents are not restricted to meronymic relations) in a real setting.

1.5   Thesis Overview

The remainder of the thesis is organized as follows.

    Chapter 2 reviews relevant work on bridging. We first focus on research work on bridging from three perspectives: pragmatic theories, corpus studies and computational approaches. We then discuss two problems which are closely related to bridging resolution, i.e., implicit semantic role labeling and relation extraction. We review the main trends and representative work in these two areas. We also analyze the similarities as well as the differences between these problems and bridging resolution.

    Chapter 3 details the corpus used throughout the thesis. After describing the annotation scheme, we focus on analyzing bridging from different perspectives: the syntactic and the topological character of bridging anaphora and of bridging antecedents, the distance between bridging anaphors and antecedents, and the interaction between bridging and discourse relations. The above analysis and results are important as they provide us with prior knowledge of linguistic structure when we design computational models to resolve bridging automatically.

    Chapter 4 describes computational methods as well as lexical semantic resources used in this thesis. Examples of computational methods are Markov logic networks (MLNs) and support vector machines (SVMs), both of which are widely used in this dissertation.

    Chapter 5 investigates our model for bridging anaphora recognition. We describe why and how we model this problem in a cascading collective classification scheme. Moreover, we design local features as well as relational features for this task based on linguistic insights. We also compare and analyze our results in comparison to related approaches.

Chapter 6 explores our model for bridging anaphora resolution. We use a joint inference approach to model two mutually supportive tasks (i.e., *bridging anaphora resolution* and *sibling anaphors clustering*) jointly. The approach is based on the observation that semantically/syntactically related anaphors are likely to be sibling anaphors, and hence share the same antecedent. In addition, we propose a method to select antecedent candidates for bridging anaphors by exploring the discourse relation `Expansion` and modeling salience from different perspectives. Finally, we give a closer comparison of our model to other approaches.

Chapter 7 proposes two approaches for unrestricted bridging resolution, i.e., recognizing bridging anaphora and finding links to antecedents. One approach is a learning-based system combining the two models for *bridging anaphora recognition* (described in Chapter 5) and *bridging anaphora resolution* (described in Chapter 6) in a two-stage framework. The other is a rule-based system consisting of eight rules which target different relations based on linguistic insights. Both systems considerably outperform a reimplementation of a previous rule-based system and a learning-based pairwise baseline with regard to bridging resolution.

Chapter 8 revisits the questions raised previously (Section 1.3). We summarize the findings of this thesis and discuss potential future directions of research.

1.6   Generated Resources and Publications

Most contents of this thesis have been published before. The material presented in Chapter 5 is an extension of Markert et al. (2012) and Hou et al. (2013a). Chapter 6 is an extension of Hou et al. (2013b). Part of the work described in Chapter 7 has been originally presented in Hou et al. (2014).

The corpus[5] used throughout the thesis can be downloaded from: `http://www.h-its.org/english/research/nlp/download/isnotes.php`.

---

[5]The development of the annotation scheme and of the framework, the annotation study and the agreement study were carried out by Katja Markert.

Chapter 2

Literature Review

This chapter reviews previous scientific work relevant to the problem that we investigate in this thesis. We first look at research on bridging, focusing on theoretical studies, corpus studies as well as computational approaches (Section 2.1). We then discuss two problems which are closely related to bridging resolution, i.e., implicit semantic role labeling (Section 2.2) and relation extraction (Section 2.3). We analyze the similarities as well as the differences between these two problems and bridging resolution.

2.1 Bridging

In linguistics, an *anaphor* is an expression whose interpretation depends upon another expression (known as the *antecedent*) that appears previously in the discourse. Figure 2.1 shows an excerpt of a news article with three anaphoric references: "its" is a pronominal anaphor referring back to the antecedent "The business" (denoted as (1) in Figure 2.1), and "The business" is a nominal anaphor referring back to the antecedent "The Bakersfield Supermarket" (denoted as (2) in Figure 2.1). Both of these two anaphors have an identity relation with their antecedents. Differently, the bridging anaphor "friends" represents an non-identity relation with its antecedent "its owner" (denoted as (3) in Figure 2.1). The phenomena illustrated in (1) and (2) have attracted a lot of interest in the field of natural language processing under the heading of *coreference resolution* (Hobbs, 1978; Hirschman & Chinchor, 1997; Soon et al., 2001; Bengtson & Roth, 2008; Ng, 2010; Lee et al., 2013). This thesis, however, focuses on the phenomenon illustrated in (3) known as "*bridging*" (Clark, 1975). Other terms used for this phenomenon in the literature are "*associative anaphora*" (Hawkins, 1978), "*Inferrables*" (Prince, 1981; 1992), and "*indirect anaphora*" (Schwarz, 2000).

In this section, we first review theoretical studies related to bridging inference in Section 2.1.1. We then discuss empirical corpus studies on bridging in Section 2.1.2. Section 2.1.3 reviews automatic algorithms for bridging resolution.

[The Bakersfield Supermarket]$_{new}$ went bankrupt [last May]$_{mediated/knowledge}$
                    (2)   (1)
[The business]$_{old}$ closed when [[its]$_{old}$ owner]$_{mediated/syntactic}$ was
murdered by [robbers with [guns]$_{new}$]$_{new}$
                                                                        (3)
[The murder]$_{old}$ saddened [friends]$_{mediated/bridging}$

—————— coreference anaphora
————→ bridging anaphora
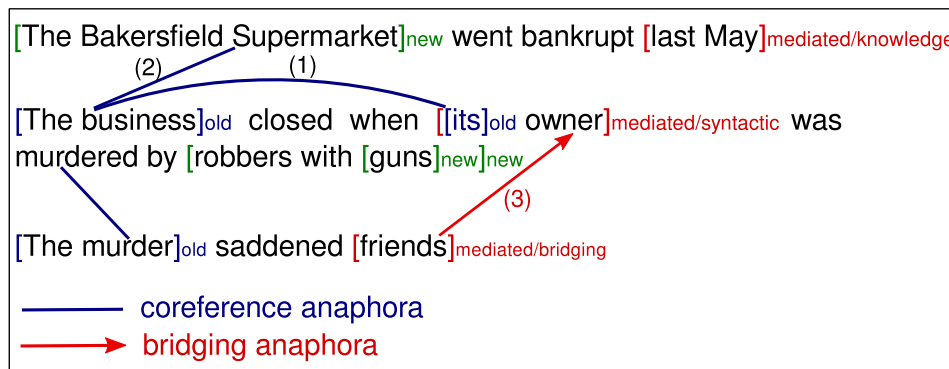
Figure 2.1: Coreference anaphora and bridging anaphora.

2.1.1   Theoretical Studies

Quite a few theoretical studies have discussed bridging inference from different perspectives, including psycholinguistic studies (Clark, 1975; Clark & Haviland, 1977; Garrod & Sanford, 1982), pragmatic and cognitive studies (Erkü & Gundel, 1987; Gundel et al., 2000; Schwarz, 2000; Matsui, 2000), and formal accounts (Hobbs et al., 1993; Bos et al., 1995; Asher & Lascarides, 1998; Löbner, 1998; Cimiano, 2006; Irmer, 2009). In the following, we describe some of the most influential works.

Given-new contrast.   Clark & Haviland (1977) proposed the given-new contrast to account for how a listener comprehends the utterances they hear. Based on Grice's cooperative principle for successful communication[1] (Grice, 1975), the given-new contrast states that speakers and listeners have an implicit agreement about how (a) information that is known to the listener, and (b) information that is novel to the listener are to appear in sentences. The heart of the given-new contrast is the *maxim of antecedence*:

> Maxim of Antecedence: Try to construct your utterance such that the listener has one and only one direct antecedent for any given information and that it is the intended antecedent.   (Clark & Haviland, 1977, p.4)

---

[1]Grice's cooperative principle states, "make your conversational contribution such as is required, at the state at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (Grice, 1975, p.307). This cooperative principle is divided into four maxims: (1) Quantity (informativeness): Make your contribution as informative as is required, and do not make your contribution more informative than is required; (2) Quality (truthfulness): Say only that which you both believe and have adequate evidence for; (3) Relation (relevance): Be relevant; and (4) Manner (clarity): Be perspicuous, avoid ambiguity, obscurity, and prolixity.

According to Clark & Haviland (1977), when the principle of *maxim of antecedence* is deliberately violated by the speaker[2], the listener is induced to draw implicatures in comprehending certain sentences in context. Under this theory, bridging can be interpreted as the following process: when the listener cannot find a direct antecedent, they then form an indirect antecedent by building an inferential bridge from something they already know. Clark & Haviland (1977) further found psychological evidence that the comprehension time for sentences containing direct anaphora (e.g., Example 2.1) was less than for those containing indirect anaphora (e.g., Example 2.2)[3].

(2.1)   Horace got *some beer* out of the car. **The beer** was warm.

(2.2)   Horace got *some picnic supplies* out of the car. **The beer** was warm.

Bridging as coercive accommodation.   Bos et al. (1995) interpreted bridging by extending Van der Sandt's theory of *presupposition projection*[4] (Van der Sandt, 1992) with lexical knowledge. They used the *generative lexicon* (Pustejovsky, 1995) as lexical knowledge. The *generative lexicon* is based on the theory of *qualia structure*, where Pustejovsky defines four types of prototypical information associated with lexical entries which denote entities or objects:

- formal: distinctive features of entities, such as *size* or *color*

- agentive: factors involved in the origin or creation of the entity, such as *creator*

- constitutive: the relation between an object and its constituents or proper parts

- telic: the purpose and function of an object

On the basis of Discourse Representation Structures (DRS) (Kamp & Reyle, 1993), Bos et al. (1995) resolved bridging by accommodating a missing antecedent inferred from *qualia* information. In Example 2.3[5], the bridging anaphor **the barkeeper** is resolved to the missing antecedent *barkeeper*, which is a *qualia role* of the discourse marker *a bar*.

(2.3)   When I go to *a bar*$_{barkeeper}$, **the barkeeper** always throws me out.

---

[2]In practice, the speaker tries to avoid tedious and repetitive sentences, therefore he "leaves gaping holes between his sentences that he expects the listener to fill in with the intended implicatures" (Clark & Haviland, 1977, p.19).

[3]Example 2.1 and Example 2.2 are from Clark & Haviland (1977).

[4]Van der Sandt (1992) considers presuppositions as anaphora with more descriptive content. Therefore the same mechanism that handles anaphoric expressions can be applied to account for presuppositions. Also the descriptive content of a presupposition enables it to accommodate an antecedent in case discourse does not provide one.

[5]The example is from Bos et al. (1995).

Bridging in SDRT.   In Segmented Discourse Representation Theory (SDRT), bridging infer-
ences are seen as "a byproduct of computing how the current sentence connects to the previous
ones in the discourse" (Asher & Lascarides, 1998, p.23). SDRT is a theory of discourse se-
mantics designed to explore systematically the interface between semantics, pragmatics and
discourse structure. The resolution of bridging anaphora in SDRT relies on the following four
meta-rules:

1. If possible use identity.

2. Bridges must be plausible.

3. Discourse structure determines bridging.

4. Maximize discourse coherence.

These four rules are applied in the indicated order. The first rule reflects the preference of
resolving an anaphor to an identical antecedent. The second rule specifies that world knowl-
edge provides plausible clues to resolve bridging anaphora. In the third rule, rhetorical re-
lations between the involved discourse segments are explored to resolve bridging anaphora.
The last rule prefers resolving bridging anaphora in a way that maximizes discourse coherence
(*maximise discourse coherence*)[6].

In essence, *maximise discourse coherence* guarantees that maintaining discourse coher-
ence takes priority over default world knowledge. This could be reflected by resolving the
bridging anaphor **The rent** in Example 2.4[7].

(2.4)   a. Alice moved from Eppelheim to *Heidelberg*.
             b. **The rent** was less expensive.

The first meta-rule for resolving bridging cannot be applied because there is no available
antecedent in the first sentence that could be coreferential to the description **The rent**. The
second rule considers world knowledge. According to common sense world knowledge, on the
one hand, the rent in Eppelheim is cheaper than in Heidelberg; on the other hand, a lower rent
in the destination is a cause for moving. Therefore world knowledge provides clues for both
interpretations, that the rent was either "the rent in Eppelheim" or "the rent in Heidelberg". We
then proceed to the third rule, which explores the discourse structure to determine bridging.
There are two discourse relations which are plausible to hold between the two sentences in

---

[6]In SDRT, *maximise discourse coherence* means: (1) the minimum number of labels, (2) no inconsistencies,
(3) the maximum number and the highest quality of rhetorical connections, and (4) the fewest unresolved seman-
tic ambiguities (including anaphoric conditions). See Asher & Lascarides (2003) for a detailed explanation.

[7]The example is adapted from Asher & Lascarides (1998). However, the origin of the example is from the
questionnaire compiled by Matsui (2000).

Example 2.4. In the *Background* relation, sentence (b) provides background information, i.e., the rent Alice paid in Eppelheim is less expensive than the rent of the place she moved to, which is Heidelberg. In the *Explanation* relation, sentence (b) provides an explanation to account for the event described in sentence (a), i.e., the rent of the place that Alice moved to, which is in Heidelberg, is less expensive than the rent she paid in Eppelheim. Again, the third rule leads to conflicting interpretations. Finally, the fourth rule - maximize discourse coherence - ensures only one reading is chosen: the *Explanation* relation is preferred over the *Background* relation since *Background* relations only convey little thematic continuity. With this, the bridging anaphor **The rent** in Example 2.4 is resolved to refer to the rent in Heidelberg, which is in contrast to world knowledge.

Bridging in an abductive inference framework. Hobbs et al. (1993) proposed a framework of abductive inference to interpret texts. The process of interpreting sentences in discourse can be viewed as a process of providing the best explanation of why the sentences could be true. In order to interpret a sentence, the following steps are taken:

> Prove the logical form of the sentence,
> — together with the constraints that predicates impose on their arguments,
> — allowing for coercions,
> Merging redundancies where possible,
> Making assumptions where necessary.
>
> (Hobbs et al., 1993, p.2)

The logical form of a sentence has been produced by syntactic analysis and semantic translation of the sentence. The interpretation of the sentence is then proved abductively by combining its logical form together with the facts and rules in the presupposed knowledge base. According to Hobbs et al. (1993), in a discourse situation, the speaker and hearer have their sets of private beliefs as well as a large overlapping set of mutual beliefs. The process of abductive proof anchors some information in mutual beliefs (hence the given information) and makes assumptions for some information from the speaker's private beliefs (hence the new information). In abductive inference, merging redundancies is a way of getting a minimal, and hence best interpretation.

Hobbs et al. (1993) showed that the idea of "interpretation as abduction" can be used to solve local pragmatics problems, including "*reference resolution*", "*compound nominal interpretation*", "*syntactic ambiguity resolution*", and "*metonymy resolution*". Here we use an example from Clark (1975) to explain how bridging resolution is modelled in the framework of Hobbs et al. (1993).

(2.5)　　a. John walked into *the room*.

　　　　b. **The chandelier** shone brightly.

In Example 2.5, in order to infer the antecedent for the bridging anaphor **The chandelier**, Hobbs et al. (1993) suppose there are some facts in the knowledge base:

- f1: rooms have lights.
  $\forall r\ room(r) \supset \exists l\ light(l) \wedge in(l, r)$

- f2: lighting fixtures with several branches are chandeliers.
  $\forall l\ light(l) \wedge hasBranches(l) \supset chandelier(l)$

In Example 2.5, the first sentence indicates the existence of a room, i.e., $room(R)$. To solve the definite reference **The chandelier** in the second sentence, one needs to prove the existence of a chandelier. Backward-chaining[8] on axiom $f2$, one needs to prove the existence of a light with branches. Backward-chaining from $light(l)$ in axiom $f1$, one needs to prove the existence of a room. This is already given by the first sentence, i.e., $room(R)$. To complete the derivation, one needs to assume that light l in room R has branches. Therefore the chandelier is the light in the room mentioned in the first sentence. The above interpretation is illustrated in Figure 2.2.

Logical Form:

$$\ldots \wedge chandelier(x) \wedge \ldots$$

Knowledge Base:

$$light(l) \wedge \boxed{hasBranches(l)} \supset chandelier(l)$$

$$room(r) \supset light(l) \wedge in(l, r)$$

$$room(R)$$
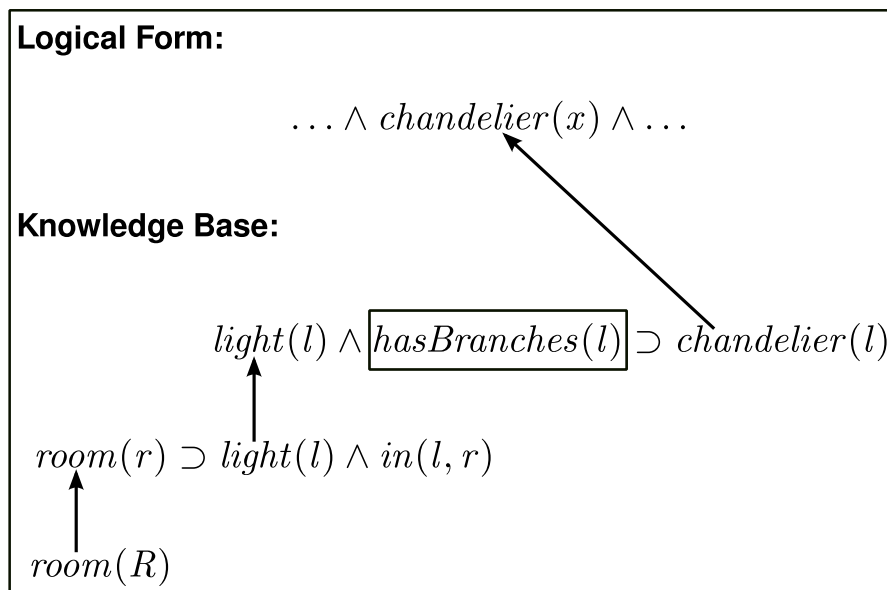
Figure 2.2: Interpretation of "**The chandelier**" in Example 2.5 (Hobbs et al., 1993, p.20).

---

[8]Backward-chaining is an inference method used in artificial intelligence applications. It starts with a goal and works backwards from consequences to causes to see if there is data available that will support any of the consequences.

Bridging inference based on focus theory.   Sidner (1979) proposed a framework to interpret definite anaphoric expressions on the basis of focus theory. She formalizes *focussing* as a process in which a speaker centers attention on a particular aspect of the discourse. This process supports comprehension of definite anaphora: "definite anaphora are signals which the speaker uses to tell the hearer what element in the discourse is the current discourse focus; at the same time, the element in focus constraints which anaphoric expressions can be used to signal the focus" (Sidner, 1979, p.234). According to Sidner (1979, p.68), the discourse expected focus of a sentence is predicted as follows:

1. The subject of a sentence if the sentence is an *is-a* or *there-insertion* sentence.
   This step presumes information from a parse tree about what the subject, and verb are and about whether the sentence is *there-insertion*.

2. The first member of the default expected focus list (DEF list), computed from the thematic relations of the verb, as follows:
   order the set of phrases in the sentence using the following preference schema:
   – theme, unless the theme is a verb complement in which case theme from the complement is used
   – all other thematic positions with the agent last
   – the verb phrase

By applying this algorithm, the expected focus for the sentence "I went to a new restaurant today" is *a new restaurant* and the corresponding DEF list is:

> 1. (theme) a new restaurant
> 2. today
> 3. (agent) I
> 4. (full verb phrase) went to (I, a new restaurant, today)

On the basis of the focus formalization, Sidner proposed *implicit backwards specification* to interpret bridging anaphora. She further analyzed four kinds of implicit specifications: *associated, inferential, set-element*, and *computed*.

In an *associated specification*, the bridging anaphor specifies an element of the association network surrounding the focus. Such association involves common sense inferences which are true about the world. In Example 2.6[9], the association between the focus of the first sentence *a new restaurant* and the bridging anaphor **The waitress** involves a hearer's common sense knowledge which is generally true. In contrast, an *inferred specification* involves a supposition

---

[9]Example 2.6 – Example 2.9 are from Sidner (1979).

of the speaker which is not necessarily true. For instance, in Example 2.7, the assumption of the speaker "the heiress was murdered" may not be true. In a *set-element specification*, the bridging anaphor specifies one member of the set represented by the focus (e.g., *clowns* – **The clown with the unicycle** in Example 2.8). Finally, a *computed specification* is indicated by modifiers such as *first, last, next*, and *second* (e.g., *a meeting* – **The last meeting** in Example 2.9).

(2.6)   a. I went to *a new restaurant* today.
         b. **The waitress** was nasty.

(2.7)   a. *The heiress* lived the life of a recluse.
         b. She died under mysterious circumstances, but **the murderer** was never found.

(2.8)   a. I went downtown today, and there were *clowns* performing in the square.
         b. **The clown with the unicycle** did this really fantastic stunt.

(2.9)   a. Aunt Het's Sewing Bee wants to have a meeting this week.
         b. *The meeting* should be on Tuesday.
         c. **The last meeting**, which was at 5, was too late, so schedule this one earlier.

Bridging inference based on relevance theory.   Matsui (2000) proposed an account of bridging on the basis of relevance theory (Sperber & Wilson, 1986; Wilson & Sperber, 2002). According to Wilson & Sperber (2002), "any external stimulus or internal representation which provides an input to cognitive processes may be relevant to an individual at some time" and "an input is relevant to an individual when its processing in a context of available assumptions yields a *Positive Cognitive Effect*" (Wilson & Sperber, 2002, p.251). These two quoted statements explain what sort of things are relevant and when they are relevant. But how do humans decide which input is more relevant than another? Wilson & Sperber (2002) first claim that "human cognition tends to be geared to the maximisation of relevance (*cognitive principle of relevance*)" (Wilson & Sperber, 2002, p.254), then they measure relevance in terms of *cognitive effects* and *processing effort*:

> **Relevance of an input to an individual**
> (a) Other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of the input to the individual at that time.
> (b) Other things being equal, the greater the processing effort expended, the lower the relevance of the input to the individual at that time.
>
>   (Wilson & Sperber, 2002, p.252)

Under *the principle of relevance*, in verbal communication, the hearer and the speaker co-operate to fulfill the communicative intention: on the one hand, the hearer will follow a path of least effort to compute cognitive effects to test interpretive hypotheses (e.g., reference resolution, disambiguation) till their expectations of relevance are satisfied. On the other hand, the speaker should make their utterance to be as easy as possible to understand. An utterance with two apparently satisfactory competing interpretations would cause the hearer the unnecessary extra effort of choosing between them.

Matsui (2000) applied relevance theory to explain bridging reference assignment. The relevance-theoretic comprehension process involves a mutual adjustment of content, context and cognitive effects. Therefore, the hearer's expectations of cognitive effect may alter the accessibility of candidate interpretations. In Example 2.10[10], *the town* is the antecedent of the bridging anaphor **The traffic**. From the processing effort side, first, *the town* is the most accessible antecedent because it is the expected focus of the first sentence; second, there is an encyclopaedic relation between *town* and **traffic**. Under this interpretation, the second sentence is expected to achieve adequate cognitive effects by answering questions implicitly raised by the first one, such as "Isn't the town too noisy"? In Example 2.11, however, the antecedent for the bridging anaphor **The traffic** is *the country*: choosing *the town* as the antecedent of the bridging anaphor **The traffic** in this example will put the hearer to unjustifiable processing effort. Therefore, the expectation of the coherence between the two sentences will guide the hearer to search for relevance by constraining the choice of contexts and cognitive effects.

(2.10)   a. I prefer *the town* to the country.
          b. **The traffic** doesn't bother me.

(2.11)   a. I prefer the town to *the country*.
          b. **The traffic** really bothers me.

Summary.   The aforementioned theoretical studies provide us with insights about bridging from different perspectives. Although these approaches are rooted in different theories, they all agree that common sense knowledge plays an important role in bridging resolution. In addition, Asher & Lascarides (1998) point out that discourse structure also has an influence on bridging interpretation. Indeed, both Asher & Lascarides (1998) and Hobbs et al. (1993) model bridging from a discourse coherence perspective. These linguistic theories on bridging later inspire us to design features for bridging resolution in our computational models (Chapter 5, Chapter 6 and Chapter 7). Specifically, in order to capture common sense world knowledge, we create a distributional semantic resource by mining big corpora (see Section

---

[10]Example 2.10 and Example 2.11 are from Matsui (2000).

4.2.1 for a detailed description). We also utilize the lexical resources such as WordNet (Fellbaum, 1998) and the General Inquirer (Stone et al., 1966). In Section 3.5, we empirically analyze the interaction between bridging and discourse relations. Furthermore, we propose a novel method to select antecedent candidates for bridging anaphors by exploring the discourse relation `Expansion` (Section 6.4).

2.1.2   Corpus Studies

Since the 1990's, empirical corpus studies related to bridging have been carried out on various genres and different languages. In this section we first review main works in this area, we then compare the resulting corpora on different dimensions.

Fraurud's study on the use of referring expressions.   Fraurud (1990) carried out a corpus study concerning the usage of NPs in a corpus, which consists of eleven professional written Swedish proses from four different sources: brochures, newspapers, textbooks and debate books. Perhaps the most important result of Fraurud's study is that the case of "first-mentioned" is by no means restricted to indefinite NPs. The author found that about 61% of all definite NPs and 85% of the indefinite NPs were "first-mentioned", i.e., without a coreferential NP antecedent. Among those first-mention definite NPs, Fraurud reported that 36% of them have interpretations which "appear to involve a relation to contextual elements outside the definite NP itself" (Fraurud, 1990, p.406). Therefore bridging anaphora triggered by definite NPs are not a marginal phenomenon.

The Vieira/Poesio dataset.   Poesio & Vieira (1998) focused on the use of definite descriptions in written texts. The resulting dataset (hence *the Vieira/Poesio dataset*) contains 33 articles from the Penn Treebank corpus (Marcus et al., 1993). Two annotation experiments were conducted on this dataset. In the first experiment conducted on 1,040 definite descriptions (DDs) in 20 articles, Poesio & Vieira (1998) classified DDs into five categories: *anaphoric s.h.* (DDs and antecedents are coreferent and have the same head noun), *associative anaphora*, *larger situation/unfamiliar*, *idiom* and *doubt*. Here *associative anaphora* includes DDs which co-refer with antecedents but do not have the same head (*different-head coreference*, e.g., *three bills – the legislation*) as well as anaphoric DDs which have non-identity relations with antecedents (e.g., *house – the chimney*). In the second experiment conducted on 464 DDs in 14 articles, they merged the first case of *associative anaphora* (i.e., *different-head coreference*) with the *anaphoric s.h.* category and separated *larger situation/unfamiliar* into two categories. This led to a taxonomy that was closer to Hawkins' and Prince's classification schemes for the usage of noun phrases (Hawkins, 1978; Prince, 1981; 1992): *coreferential, bridging, larger situation, unfamiliar*, and *doubt*. Here "*bridging*" coin-

cides with Hawkins' *associative anaphora* and Prince's class of *Inferrables*[11]. It is also close to the definition of bridging in this thesis.

This work found that a great number of definite descriptions in text are discourse-new (i.e., larger situation and unfamiliar), with the average percentage of discourse-new DDs being 50% in the first experiment and 46% in the second experiment. This coincides with our corpus study in Chapter 3 on the same genre. Moreover, Poesio & Vieira (1998) found that the agreement among annotators in the second experiment ($\kappa = 0.63$) was worse than in the first experiment ($\kappa = 0.73$). Particularly, the per-class agreement for associative definite descriptions in the first experiment is 59% while only 31% on bridging descriptions in the second experiment. On the contrary, the corpus used in this thesis (detailed in Chapter 3) achieves a reasonable agreement level for the most difficult category, i.e., bridging: κ is over 70 for two expert annotators and agreement is around 80% for all annotator pairings for selecting bridging antecedents.

On the basis of this study, later Vieira (1998) developed a standard annotation on the whole data set using the taxonomy in the first experiment. However, the original names of some categories were changed: "anaphoric s.h." was replaced by "direct anaphora", "associative anaphora" was replaced by "bridging", and "larger situation/unfamiliar" was replaced by "discourse new". The use of bridging here is rather confusing, since it contains the "real" bridging cases as well as the hard coreferent cases as explained before. We call it "lenient bridging" in this thesis. The corpus was used later to develop computational models to resolve lenient bridging (see Section 2.1.3).

The GNOME corpus (the bridging reference part). The GNOME corpus (bridging reference part) (Poesio, 2004) contains about 500 sentences and 3,000 NPs from two domains. The museum subcorpus consists of texts describing museum objects (with associated pictures), and the pharmaceutical subcorpus are from a selection of leaflets providing patients with legally required information about drugs. In this corpus, only three types of bridging relations were annotated: set membership (*ELEMENT*), subset (*SUBSET*) and generalized possession (*POSS*), which includes both part-of relations and ownership relations. By limiting bridging relations to these three types, a higher agreement among two annotators was observed on associative relations: only 4.8% of the relations were actually marked differently. However, only 22% of bridging references were annotated in the same way by both annotators and 73.17% of relations were marked by only one or the other annotator. In total, the GNOME corpus contains 153 mereological bridging references, 58 of them are realized by definite descriptions.

---

[11]See Section 1.1.1 for a detailed description of Hawkins' *associative anaphora* and Prince's *Inferrables*.

The Switchboard corpus (the IS part).   Nissim et al. (2004) defined the information status (IS) of an entity as "reflecting the speaker's assumptions about the hearer's knowledge/belief" (Nissim et al., 2004, p.1023). On the basis of the works in Prince (1992) and Eckert & Strube (2000), they proposed a scheme for IS annotation for all NP types. The scheme contains three main categories: an *old* entity is known to the hearer and has been mentioned in the conversation; a *new* entity is unknown to the hearer and has not been previously referred to; a *mediated* entity is newly mentioned in the dialogue but is inferrable from previously mentioned entities, or generally known to the hearer. There are nine subtypes of the *mediated* category: *general, bound, part, situation, event, set, poss, func_value* and *aggregation*. Four of them (*part, set, situation* and *event*) are specifically used to mark instances of bridging. Table 2.1 provides detailed explanations and examples for these four subtypes.

Subtype	Description
med/part	used to mark part-whole relations for physical objects
	e.g. *the car – the engine* or *the house's door*
med/set	used to indicate any kind of set relation (subset/superset/co-set-member)
	e.g. *rap – British rap* or *your life – somebody else's life*
med/situation	used to indicate the anaphor is part of a situation set up by the antecedent
	e.g. *wedding – the bride* or *murder – the killer*
med/event	used to link an entity back to an event (rather than an object)
	e.g. *traveling around – the bus*

Table 2.1: Bridging subtypes in the Switchboard corpus.

Nissim et al. (2004) then applied the above scheme to a portion of the Switchboard corpus (LDC, 1993), resulting in an IS corpus containing 147 dialogues. They reported a high agreement for the overall fine-grained IS annotation (with $\kappa = 0.788$). The κ scores for the four bridging subtypes are lower than other subtypes, but are still reasonable ($\kappa = 0.594$ for *part*, $\kappa = 0.696$ for *set*, $\kappa = 0.719$ for *situation*, $\kappa = 0.794$ for *event*).

It is interesting to notice that in this dialogue corpus, only 14.3% of NPs are *new*. This is different from the findings on the written newspaper corpora mentioned previously: Poesio & Vieira (1998) reported that around 50% of definite descriptions are discourse-new. We also found that the largest proportion (36.7%) of NPs are *new* in the ISNotes corpus (see Chapter 3). We believe this discrepancy comes from different genres since the news genre tends to introduce more new information compared to the dialogue genre.

Although this corpus is reliably annotated, we claim that it has two main flaws with regard to the bridging annotation. First, antecedent information for bridging NPs is not annotated. Second, the constraints on bridging annotations make the definition and the distribution of bridging deviate from the ground truth. The reason lies in the following aspects: (1) *med/part*

includes non-anaphoric, syntactically linked part-of and set-member relations, such as *the house's door*. Also the part-of relation must be encoded in WordNet (Fellbaum, 1998). (2) The *med/set* category includes comparative anaphora which are indicated by surface markers such as *different, another*. Furthermore, this category requests that the anaphor and the antecedent should have the same head or synonymic heads[12], or the anaphor is a hyponym or hypernym of the antecedent. Such synonymy, hyponymy and hypernymy relations must be encoded in WordNet. (3) The *med/situation* category depends heavily on the knowledge encoded in FrameNet (Baker et al., 1998), i.e., the anaphor exemplifies a role in the frame that the antecedent evokes. The above mentioned constraints on annotations make bridging in this corpus biased on the relations encoded in WordNet and FrameNet[13].

The DIRNDL corpus (the referential IS part). The DIRNDL corpus (Eckart et al., 2012; Björkelund et al., 2014) is a spoken corpus of German radio news. It contains various layers of annotation, i.e., morpho-syntactic annotations, prosodic annotations, semantic annotations, and pragmatic annotations. The pragmatic annotation layer previously contained annotations of *referential information status* following Riester et al. (2010). Later they were extended to two-dimensional information status annotations following the *RefLex* scheme (Baumann & Riester, 2013), which distinguishes between *referential* and *lexical* information status.

Here we briefly describe the annotation scheme for *referential information status* since bridging belongs to the discourse reference level. Riester et al. (2010) assumed that (referential) information status strongly depends on (in)definiteness and therefore provided different categories for each class indicated by overt in(definiteness) markers. The whole scheme includes four coarse categories for the definite expressions (i.e., *given, situative, bridging* and *unused*), four categories for the indefinite expressions (i.e., *new, generic, resumptive* and *partitive*), and several other categories (i.e., *cataphor, expletive, null* and *relative*). Although their definition of bridging is close to ours, they limited bridging to definite expressions.

By applying this scheme to annotate the DIRNDL corpus, a kappa score of .78 was achieved for six top-level categories. However, the confusion matrix in Riester et al. (2010) shows that the anaphoric bridging category (BRI) is frequently confused with other categories so that the two annotators agreed on only less than a third of bridging anaphors.

The DEDE corpus. DEDE (Gardent & Manuélian, 2005) is a French corpus built from *Le Monde* articles in 1987. It contains 4,910 definite descriptions classified into five main

---

[12]In these two cases, the anaphor and the antecedent should differ in modification and therefore are not coreferential.

[13]"Selective annotation" is common in the field of NLP. It saves the annotation cost. Moreover, it makes annotation work applicable by balancing between the difficulty of the annotation task and the agreement among the annotators.

categories (i.e., *autonomous, coreferential, associative, situational* and *non-referential*). The *associative* (bridging) category is further divided into the following four sub categories:

- *mero*: meronymy relation, e.g., *a tree* – **the trunk**

- *circ*: modifier-modified relation, e.g., *Besancon* – **the region**[14]

- *rel*: predicate-argument relation, e.g., *the attack* – **two robbers**

- *mod*: relation introduced by a modifier, e.g., *in July* – **the next month**

The DEDE corpus contains 530 bridging anaphors and 60% of them are mereological bridging references.

The Caselli/Prodanof dataset. Caselli & Prodanof (2006) carried out a corpus study of identifying bridging anaphors which have the form of "definite article + (possessive) + N"[15] in Italian. They proposed a scheme consisting of six main classes (i.e., *first mention, possessives, direct anaphora, bridging, idiom* and *doubt*) for classifying definite NPs. In this scheme, a *direct anaphor* refers back to an entity (antecedent) mentioned previously in the text, and the anaphor and the antecedent may or may not share the same head. A *bridging anaphor* is not mentioned before, but its interpretation depends on an entity already present in the text. In a dataset of 17 articles chosen from the Italian newspaper "il Sole-24 Ore", Caselli & Prodanof (2006) found that among 1,412 definite NPs, only 12.03% of them are direct anaphors, whereas the class of bridging represents 63.88% of all anaphoric definite NPs and 21.17% of all definite NPs. They assumed that bridging is a more productive cohesive strategy in Italian compared to other languages, i.e., English. In addition, Caselli & Prodanof (2006) reported a relative high agreement among annotators with regard to bridging anaphora identification: $\kappa = .58$ among three annotators and $\kappa = .71$ between two annotators who received training before completing the task. The agreement for selecting bridging antecedents was 78% among all annotators[16].

The CDT corpus (the associative anaphora part). Korzen & Buch-Kromann (2011) proposed a hierarchical scheme to classify bridging relations, on the basis of the following two parameters: (1) the anaphor is associated with the antecedent with regard to the *qualia structure*[17] of the antecedent; and (2) the antecedent is predicative and the anaphor is a semantic role.

---

[14]The original example is: … les moyens mis en oeuvre depuis près d'une semaine à *Besancon* et dans **la région** (Gardent & Manuélian, 2005).

[15]The definite NPs with this form represent 31.54% of all definite NPs in the corpus.

[16]The calculation was based on all cases of definite NPs that were classified by all three annotators as bridging.

[17]Pustejovsky's *qualia structure* theory (Pustejovsky, 1995) defines four types of prototypical information associated with entities denoted by words, i.e., *formal* encodes distinctive features of entities (e.g., *size* or *color*);

Korzen & Buch-Kromann (2011) then applied the scheme to annotate associative anaphora (bridging) and their relations in the Copenhagen Dependency Treebanks (CDT) corpus. The corpus consists of five parallel open-source treebanks for Danish, English, German, Italian and Spanish. It contains 100,000 words compiled from 200-250 word excerpts from Danish mixed-genre texts, which have been translated into the other languages by native translators. In 25 texts where 85 bridging relations were annotated, Korzen & Buch-Kromann (2011) reported that acceptable agreement results were achieved between two annotators for some relations (i.e., *assoc-const* and *assoc-telic*). The agreement was measured by calculating the probability that another annotator assigns the same label and/or out-node to the relation. However, the total number of bridging references in the CDT corpus is unknown[18].

The Prague dependency treebank (the bridging anaphora part). Nedoluzhko et al. (2009) developed a scheme for annotating nominal coreference and bridging anaphora in the Prague dependency treebank. The scheme contains six subtypes of bridging relations:

- *part*: part-of relation, e.g., *room* – **ceiling**

- *subset*: set-subset or element of the set, e.g., *participants* – **some participants**

- *func*: individual function of an object, e.g., *government* – **prime minister**

- *contrast*: coherent relevant discourse opposites, e.g., *People* – **cow** in the context "*People don't chew, it's cows who chew.*"

- *anaf*: explicitly anaphoric relations without coreference or one of the above mentioned semantic relations, e.g., *Rainbow* – **this word** in the context "'*Rainbow'? The priest put the finger on this word, so that he didn't forget, where he stopped.*"

- *rest*: further underspecified relations, e.g., *location* – **resident**

This annotation scheme was applied to the whole PDT corpus by two instructed annotators. In the inter-annotator agreement study, Nedoluzhko et al. (2009) showed that recognizing bridging was difficult, with F_1-measure ranging from 0.42 to 0.59 among four times of agreement measurement[19]. By 2009, roughly 8,000 bridging anaphora nodes[20] in 755 documents from the PDT corpus had been annotated.

---

constitutive encodes the relation between an object and its constituents or proper parts; *agentive* encodes the factors involved in the origin or creation of the entity (e.g., *creator*); *telic* describes the purpose and function of an object.

[18]According to the CDT website (*https://code.google.com/p/copenhagen-dependency-treebank*) and private email communication with Prof. Korzen, the anaphora annotation in CDT has not been developed any further since 2011.

[19]The agreement was measured on arguments of bridging relations, i.e., bridging anaphors and antecedents.

[20]The number is roughly calculated according to the chart reported in Nedoluzhko et al. (2009).

Discussion.   Table 2.2 compares the above mentioned corpora in different dimensions[21]. These previous corpus-linguistic studies on bridging have some limitations. First, the definition of bridging is extended improperly to include coreferential NPs with lexical variety (Vieira, 1998) or other non-anaphoric NPs (Nissim et al., 2004). Second, reliability is not measured or low (Poesio & Vieira, 1998; Gardent & Manuélian, 2005; Nedoluzhko et al., 2009; Riester et al., 2010). Third, annotated corpora are small (Poesio, 2004; Caselli & Prodanof, 2006). Fourth, they are often based on strong untested assumptions about bridging anaphora types, antecedent types or bridging relations, such as limiting it to definite NP anaphora (Poesio & Vieira, 1998; Gardent & Manuélian, 2005; Caselli & Prodanof, 2006; Riester et al., 2010), to NP antecedents (all prior work) or to part-of relations between anaphor and antecedent (Poesio, 2004). On the contrary, the corpus used in this thesis (ISNotes) circumvents these problems, i.e., human bridging recognition is reliable, it contains a reasonable number of bridging cases, also bridging anaphors/antecedents/relations are not limited to certain types (see Chapter 3 for a detailed description and a thorough analysis for the ISNotes corpus).

---

[21]We do not include the corpus study carried out by Fraurud (1990) in Table 2.2 since it did not count bridging as a separate category.

Corpus	Size	Language	Genre	Definition	Anaphora	Ante.	Relation
Vieira/Poesio dataset	33 articles (285 bridging NPs)	English	written news	lenient bridging	definite NPs	entity event	(1) identity (2) compoundNoun (3) meronymy
GNOME	500 sentences 3,000 NPs (581 bridging NPs)	English	written descriptions for museum objects and medicines	strict bridging	all NPs	entity	(1) *ELEMENT* (2) *SUBSET* (3) *POSS*
Switchboard	147 dialogues 69,004 NPs (14,829 bridging NPs)	English	speaking dialogues	strict bridging plus some non-anaphoric NPs	all NPs	–	(1) event (2) part-of/set-member in WordNet, (3) other relations in FrameNet
DIRNDL	3,221 sentences (449 bridging NPs)	German	radio news	strict bridging	definite NPs	entity	undefined
DEDE	4,910 definite NPs (530 bridging NPs)	French	written news	strict bridging	definite NPs	entity	(1) *mero*, (2) *circ* (3) *rel*, (4) *mod*
Caselli/Prod-anof dataset	17 articles (299 bridging NPs)	Italian	written news	strict bridging	definite NPs	entity	undefined
CDT	410,000 words (the size of bridging NPs is unclear)	Danish English Italian German Spanish	excerpts from mixed-genre texts	strict bridging	all NPs	entity	16 relations under two coarse categories *assoc-QUALIA* and *assoc-SEMROLE*
PDT	755 documents (8,000 bridging NPs)	Czech	written news	strict bridging	all NPs	entity	(1) *part*, (2) *subset* (3) *func*, (4) *anaf*, (5) *rest*

Table 2.2: Comparison of different corpora with regard to bridging annotation.

2.1.3 Computational Approaches

One branch of study on bridging focuses on incorporating knowledge from various theoretical studies into computational models, and therefore automatically resolving bridging in documents. Research works in this vein often build computational models based on a corpus where bridging is annotated. Such models are constructed either by applying machine learning techniques or by designing hand-crafted rules. Normally, a new unseen dataset or cross-validation is used to evaluate the performance of the computational model.

In recent years, various computational approaches have been developed for resolving bridging. However, most of them only concentrate on antecedent selection, assuming that bridging anaphora recognition has already been performed. Some of them handle bridging anaphora recognition under a fine-grained information status (IS) classification scheme. Only a few works try to tackle the problem of full bridging resolution, i.e., recognizing bridging anaphora and finding links to antecedents. In the following, we detail these approaches according to the (sub)tasks they aim to resolve, as well as highlighting the particular methods that inspire our work in this thesis.

Bridging anaphora recognition. Recent work on bridging anaphora recognition models it as a fine-grained information status (IS) classification problem. Given a text, the system assigns one IS class to each mention according to its accessibility to the reader at a given point in the text, bridging being one possible class.

Rahman & Ng (2012) studied the fine-grained IS classification problem on the Switchboard dialogue corpus (Nissim et al., 2004). They first designed a rule-based system to assign IS classes to mentions on the basis of Nissim's IS annotation guidelines (Nissim et al., 2004). They then applied an $\text{SVM}^{multiclass}$ algorithm for this task by combining the prediction from the rule-based system, the ordering of the rules as well as two lexical features.

Rahman & Ng (2012) reported high results for the four subcategories of bridging (i.e., *med/part, med/situation, med/event, med/set* with F-scores ranging from 63.3 to 87.2), but we argue that this does not reflect the real difficulty of the problem. The reason comes from the annotation flaws for bridging in the Switchboard corpus that we pointed out in the previous section: bridging in this corpus includes non-anaphoric, syntactically linked part-of and set-member relations, as well as comparative anaphora which are marked by surface indicators such as *other* or *different*. Moreover, the "selective annotation" for bridging biases the distribution of bridging in this corpus to the relation types encoded in certain knowledge resources.

Another work on bridging anaphora recognition was carried out by Cahill & Riester (2012). They assumed that the distribution of IS classes within sentences tends to have certain linear patterns, e.g., *old > mediated > new*. Under this assumption, they trained a CRF model with syntactic and surface features for fine-grained IS classification on the German DIRNDL

radio news corpus (Riester et al., 2010). Although Cahill & Riester's definition for bridging is similar to ours, they did not report the result for the *bridging* subcategory. Moreover, we claim that the embedded structure of IS classes and the flexible position of spatial or temporal scene-setting elements (e.g., *In New York*) weaken the linear pattern assumption. This may explain that Cahill & Riester (2012) found that the CRF model only performed slightly better than a simple multiclass logistic regression model. In contrast, our approach on fine-grained IS classification (Chapter 5) considers syntactic structural patterns among IS classes for embedded mentions. For instance, in prepositional phrases, the IS class of the (syntactic) parent is dependent on the IS class of the (syntactic) child (e.g., [professors at [Cambridge]$_{mediated/worldKnowledge}$]$_{mediated/syntactic}$ vs. [professors at [a college]$_{new}$]$_{new}$). We explore Markov logic networks (Domingos & Lowd, 2009) to represent such structures and find that our collective classification model outperforms other models (Nissim, 2006; Rahman & Ng, 2011) significantly. This model, however, performs unsatisfactory for the bridging category due to the wide variation of bridging anaphora and their relative rarity compared to many other IS categories. We then propose a cascading collective classification model which combines the binary classifiers for minority categories and our collective model in a cascaded way. The new system addresses the multi-class imbalance problem (for rare categories without strong indicators, such as *bridging*) while still keeping the strength of the collective classification.

Bridging anaphora resolution. Bridging anaphora resolution is the main focus of most previous computational research on bridging. The task is to find the correct antecedents for bridging anaphors.

Based on the corpus created by Poesio & Vieira (1998) (the Vieira/Poesio dataset described in the previous section), various studies have been conducted to resolve "lenient" definite bridging references in this dataset[22]: Vieira & Teufel (1997) discussed how many antecedent/anaphor pairs can be correctly resolved by exploring WordNet, with a focus on 38 cases of synonymy/hyponymy/meronymy relations. Poesio et al. (1997) proposed several heuristics (including exploring WordNet) to resolve different types of bridging descriptions. Schulte im Walde (1998) resolved bridging descriptions by using word clustering algorithms. The bridging anaphors were resolved by finding the closest antecedent candidate in a high-dimensional space. The space was constructed by exploring the British National Corpus containing 30 million words, with all nouns and verbs from the Vieira/Poesio dataset (training part) as target words and 2,061 most frequent words from the BNC corpus as context words.

---

[22]We call the bridging category in Poesio & Vieira (1998) *lenient bridging* since it includes coreferential cases where anaphor and antecedent do not have the same head. The full Vieira/Poesio dataset contains 285 lenient bridging anaphors in total with 204 in the training data and 81 in the test data. Most works only focus on resolving bridging references in the training set.

The best result (accuracy of 22.7%) was obtained by using the cosine metric as distance measure. Poesio et al. (2002) explored resolving bridging references by acquiring lexical knowledge, with a focus on part-of relations. They utilized three syntactic patterns (i.e. *the NP of NP, NP of NP, NP's NP*) to mine meronymic information from the BNC corpus. Markert et al. (2003) also tried to resolve mereological bridging references (12 cases) by searching the Web with shallow patterns, such as "*floors of (the OR all) ⋆ apartments*". Overall, the above work was done on a small dataset where bridging was extended to include coreferential NPs with lexical variety.

Following the definition of bridging in the Vieira/Poesio dataset (Vieira, 1998), Fan et al. (2005) casted the task of bridging anaphora resolution as finding semantic paths between concepts in knowledge bases. They used both inductive reasoning and abductive reasoning to infer the semantic relation between anaphor and antecedent. Fan et al. (2005) evaluated their framework on the revised Vieira/Poesio dataset[23] as well as the revised Brown dataset (Bunescu, 2003)[24]. They found that their approach boosted recall without loss of precision compared to the system proposed by Vieira (1998).

Another line of work has sought to apply machine learning techniques to find the best combination of various heuristics. Poesio et al. (2004a) applied a pairwise model combining lexical semantic features as well as salience features to perform mereological bridging resolution in the GNOME corpus. They used a five sentence window to choose the antecedent candidates on the basis of a previous study which claimed that local focus is preferred for the interpretation of a bridging reference in the same corpus (Poesio, 2003). To address the data sparseness problem (e.g., some part-of relations are not covered by WordNet), they used the Web to estimate the part-of relations expressed by certain syntactic constructions. For instance, big hit counts for the query *the door of the house* against Google could indicate that *door* and *house* stand in a part-of relation. In Chapter 6, we extend this idea to a more general preposition pattern and normalize the hit counts of the pattern queries for each anaphor respectively. However, the high results reported in Poesio et al. (2004a) were not based on a real evaluation setting: in the first two experiments they distinguished only between the correct antecedent and *one* or *three* false candidates; in the "hard" test they tried to find the correct antecedents for six bridging anaphors among all possible candidates. We suggest cross-validation would be more appropriate considering that a test set containing only six cases is

---

[23]Fan et al. (2005) claimed that the dataset described in Vieira & Poesio (2000) was partially annotated. They then completed the annotation. However, Fan et al. (2005) did not report the details of their further annotation. In addition, they did not consider named entities in their experiments because their knowledge base converted from WordNet only contains little knowledge about named entities.

[24]As we will discuss in the next part, this dataset only includes definite anaphors which bear the syntactic pattern "*the + head*". Fan et al. (2005) then remove all named entities as well as direct anaphors (anaphors and antecedents are coreferent and have the same head noun). This yields to a test dataset containing 196 instances of "lenient" bridging anaphors.

rather small.

On the basis of the method proposed by Poesio et al. (2004a), Lassalle & Denis (2011) developed a system that resolves mereological bridging anaphora in French. The system was enriched with meronymic information extracted from raw texts. Such information was extracted in a bootstrapping fashion by iteratively collecting meronymic pairs and the corresponding syntactic patterns. Lassalle & Denis (2011) evaluated their system on mereological bridging anaphors annotated in the DEDE corpus and reported an accuracy of 23%.

In contrast to all the aforementioned work on bridging anaphora resolution, we model this problem on the discourse level without limiting it to the specific type of anaphora (e.g., definite NPs) or to certain bridging relations (e.g., part-of relation). Our model resolves all bridging anaphors in one document by taking the phenomenon of "sibling anaphors" into account. Sibling anaphors are bridging anaphors that share the same antecedent and are often syntactically or semantically related (see Example 2.12). We then model the two tasks, i.e., *bridging anaphora resolution* and *sibling anaphors clustering* jointly since they mutually support each other. Furthermore, unlike previous work which simply uses a sentence window to form the set of antecedent candidates, we explore an advanced strategy to choose antecedent candidates for bridging anaphors on the basis of their discourse scopes (i.e., *local* or *non-local*). We evaluate our model on the ISNotes corpus (Chapter 3) and show that our system outperforms a reimplementation of a previous system (Poesio et al., 2004a) by a large margin (Chapter 6).

(2.12)   If Mr. McDonough's plans get executed, as much as possible of *the Polish center* will be made from aluminum, steel and glass recycled from Warsaw's abundant rubble. **The windows** will open. **The carpets** won't be glued down and **walls** will be coated with non-toxic finishes.

Full bridging resolution.   Full bridging resolution is a challenging task. It requires recognizing that a bridging anaphor is present and finding the correct antecedent among a list of candidates.

A line of work in this field attempted to resolve all anaphoric definite NPs (including bridging) together. However, it is not clear how well these systems perform on the task of "strict" bridging resolution. Vieira & Poesio (2000) proposed a rule-based system for processing definite NPs[25]. However, their definition of bridging (*lenient bridging*) includes cases where anaphors and antecedents are coreferent but do not share the same head[26]. In addition,

---

[25]See Appendix A for the detailed information of this system.

[26]Such cases are called "*coreferent bridging*" in Versley (2011), i.e., a definite description is a subsequent mention of an entity which has already been mentioned, but none of the prior mentions has the same head noun as the current mention.

Vieira & Poesio (2000) reported results for the whole anaphora resolution but did not report results for *lenient bridging* resolution only.

Bunescu (2003) discussed the differences between identity anaphora and associative anaphora. An identity anaphor and its antecedent refer to the same entity while an associative anaphor has a non-identity relation with its antecedent. The author then created a subset of definite NPs from 32 documents taken from the Penn Treebank corpus (Marcus et al., 1993). The initial list of definite NPs was extracted by only considering definite descriptions containing only one noun, with the additional constraint that the NPs do not have prepositional phrases or relative clauses attached to the heads. Bunescu further excluded definite NPs whose head noun occurs previously in the document from the above list. The resulting list contains 686 definite NPs which then were annotated by hand. 324 of them are anaphoric (identity anaphoric or associative anaphoric) and have entity (noun) antecedents. Bunescu (2003) attempted to resolve anaphoric NPs in this predefined subset of definite NPs by searching the Web using particular lexico-syntactic patterns (e.g., "*restaurant*. **The service** is/are/would"). Although Bunescu's associative anaphora is close to our bridging definition, he did not distinguish associative anaphora from identity anaphora both in the dataset creation and in the experimental evaluation. Another limitation of this work is that the dataset is biased to a certain type of definite NPs (i.e., *the + (adj) + head*). Therefore it does not reflect the real distribution of anaphoric definite NPs in texts.

Among works which distinguish bridging resolution from other anaphora resolution, Hahn et al. (1996) presented a rule-based framework which integrates language-independent conceptual criteria and language-dependent functional constraints. Their conceptual criteria were based on a knowledge base from the information technology domain, which consists of 449 concepts and 334 relations. They focused on definite bridging anaphora only and considered certain types of relations (e.g., *has-property, has-physical-part*). Hahn et al. (1996) evaluated the system on a small-scale technical domain dataset (5 texts in German with 109 bridging anaphors) and reported a recall score of 55.0% and a precision score of 73.2% . Although the results seem satisfactory, the system is heavily dependent on the domain knowledge resources. Therefore it is hard to apply it to other domains where such knowledge resources are not available.

Sasano & Kurohashi (2009) applied a probabilistic model for associative anaphora (bridging) resolution in Japanese. They regarded associative anaphora as a kind of zero anaphora and resolved associative anaphora and zero anaphora together by using automatically acquired (verb and nominal) case frames. The nominal case frames (lexical knowledge) used for associative anaphora resolution is constructed by conducting an analysis of the "N_h of N_m" (e.g., *price of car*) phrases from large corpora on the basis of an ordinary language dictionary. The language dictionary and large corpora are used in a complementary way: the dictionary is applied to generalize the case slots for nouns (e.g., "country" is a case slot for "prime minister"

where country is the generalized form for "Japan", "Germany", "China", etc.), whereas the examples from the large corpora provide evidence to predict obligatory case slots for nouns. For instance, given that the statistics distribution is 0.82 for "prime minister of country" and 0.13 for "prime minister of vehicle", "country" is chosen as an obligatory case slot for "prime minister" according to some thresholds set manually. Their approach can be understood as a generative process from right to left (i.e., generating antecedents from anaphors): given a potential bridging anaphor and discourse entities appearing before it, the system considers how likely it is that this bridging anaphor generates a case frame, together with the direct case assignment and the indirect case assignment. The direct case assignment finds the explicit realizations of the obligatory case slots, whereas the indirect case assignment finds the implicit realizations of the obligatory case slots (i.e., finding the antecedents for a potential bridging anaphor). Sasano & Kurohashi (2009) evaluated their approach on a set of 62 web documents (5.23 sentences per document) containing 110 associative anaphoric relations and reported an F-measure of 42.7% (a salience-based baseline achieved an F-measure of 20.2%). They further showed that associative anaphora resolution can benefit from resolving associative anaphora and zero anaphora simultaneously. Although it is not clear how associative anaphora are distributed in the web corpus and whether this approach can be effectively applied in other languages (e.g., English), the structured lexical knowledge resource constructed by the authors is general and therefore can capture diverse bridging relations.

Recently, Rösiger & Teufel (2014) applied a coreference resolution system with several additional semantic features from WordNet (e.g., *hypernymy, meronymy, topic*) to find bridging links in scientific text where bridging anaphors are limited to definite NPs. They reported preliminary results for bridging resolution using the CoNLL scorer. However, we argue that the coreference resolution system and the evaluation metric for coreference resolution are not suitable for bridging resolution since bridging is not a set problem.

Our study departs from the aforementioned related work in that we strive to resolve **unrestricted** bridging (both bridging anaphora types and bridging relations are unrestricted) in English. To tackle this challenging task, we propose two approaches: a rule-based system consisting of eight rules which target different relations based on linguistic insights, as well as a learning-based system combining the two models for *bridging anaphora recognition* and *bridging anaphora resolution* in a pipeline way (Chapter 7). Although the result is still undesirable, we believe that our research helps to clarify the problem (i.e., bridging resolution) since the concept of bridging or associative anaphora has been used in a mixed way in the related literature. Furthermore, by conducting a thorough analysis of bridging in the ISNotes corpus (Chapter 3) as well as error analyses of different systems in three tasks (i.e., *bridging anaphora recognize* in Chapter 5, *bridging anaphora resolution* in Chapter 6 and *bridging resolution* in Chapter 7), our study opens doors for future research on bridging resolution (Chapter 8).

2.2   Implicit Semantic Role Labeling

Semantic role labeling (SRL) or shallow semantic parsing is the task of assigning semantic roles to the semantic arguments associated with a predicate (e.g., a verb or a noun). For instance, in Example 2.13[27], the verb "borrowed" triggers the borrowing frame, the constituent "the girl" serves as the semantic role *borrower* and the constituent "the car" serves as the semantic role *theme*.

(2.13)   [The girl$_{borrower}$] **borrowed**$_{frame\_borrowing}$ [the car$_{theme}$] [from her sister$_{lender}$].

Semantic role labeling has attracted a large amount of interests since Gildea & Jurafsky (2002) first explored statistical methods to automatically label semantic roles on the basis of the FrameNet lexicon (Baker et al., 1998). The development of PropBank (Palmer et al., 2005) and Nombank (Meyers et al., 2004), together with CoNLL shared tasks on semantic role labeling (Carreras & Màrquez, 2004; 2005; Surdeanu et al., 2008) boosted research in this area.

However, the majority of work on semantic role labeling only focuses on the local context, i.e., the system only recognizes semantic arguments from the sentence where the predicate is present and thus ignores arguments from the wider discourse context. In Example 2.14[28], "the tenants" evokes the frame *residence* and serves as one core semantic role of this frame (i.e., *resident*). However, the filler of another core semantic role (i.e., *location*) of this frame appears two sentences before (i.e., *the house*).

(2.14)   "Now, Mr. Holmes, with your permission, I will show you round [*the house$_{location}$*]."
            The various bedrooms and sitting-rooms had yielded nothing to a careful search.
            Apparently [**the tenants**$_{resident}$]$_{frame\_residence}$ have brought little or nothing with them.
            (DNI)

According to the frame semantics paradigm, the non-local realized core semantic roles (also called *Core Frame Elements*) are considered *Null Instantiations* (NI). NIs are divided into three categories: *definite null instantiations* (DNI) are NIs whose fillers are accessible from the context (such as *the house$_{localtion}$* in Example 2.14), while for *indefinite null instantiations* (INI) and *constructional null instantiations* (CNI), the fillers are inaccessible. For instance, in Example 2.15[29], the filler of the core semantic role *creator* for the frame *intentionally_create* is not accessible from the context. Also in Example 2.16, the filler of the core semantic role *speaker* for the frame *statement* is not accessible from the context because the grammar of the sentence allows or requires an omission.

---

[27]The example is taken from FrameNet (Baker et al., 1998).

[28]The example is from the SemEval-2010 shared task on "Linking Events and Their Participants in Discourse (Ruppenhofer et al., 2010).

[29]Example 2.15 and 2.16 are from FrameNet (Baker et al., 1998).

(2.15)   Each legislator is only one out of 535 when it comes to [national policy$_{created\_entity}$] **making**$_{frame\_intentionallyCreate}$. (INI)

(2.16)   It is no longer possible to make **claims**$_{frame\_statement}$ to understand a culture simply by classifying it in terms of its relations to a present western culture. (CNI)

Recognizing DNIs and finding the corresponding fillers from the context is out of the scope of the traditional SRL systems which model the problem on a sentence level. The reason is partially due to the fact that such information was not annotated in large-scale corpora (e.g., FrameNet, PropBank). To address the issue of non-local (implicit) argument identification for predicates, Ruppenhofer et al. (2010) organized a shared task "Linking Events and Their Participants in Discourse" in SemEval-2010. Since then, there has been a growing interest in developing algorithms for resolving locally unrealized (implicit) semantic roles. In the remainder of this section, we first review works in this field and then discuss its connection to bridging resolution.

Implicit semantic role labeling: corpora and algorithms.   Task 10 in Semeval-2010 (Ruppenhofer et al., 2010) distributed a corpus that includes a manual annotation of implicit semantic roles. The corpus consists of two texts by Arthur Conan Doyle. The annotation consists of frame-semantic argument structures, coreference chains, information about null instantiations, i.e., the NI type (DNI vs. INI) and the fillers of DNIs. Table 2.3 shows basic statistics about this corpus.

	sentences	tokens	frame instances	frame types	overt frame elements	DNIs (resolved)	INIs
training set	438	7,941	1,370	317	2,526	303 (245)	277
test set	525	9,131	1,703	452	3,143	349 (259)	361

Table 2.3: Statistics for the SemEval-2010 Task-10 corpus.

Although the SemEval-2010 Task 10 includes a *full task* track and a *NIs only* track, its main focus is "linking locally uninstantiated roles to their coreferents in the wider discourse context" (Ruppenhofer et al., 2010, p.106). In the *full task* track, given the target words and their corresponding frames, the participants need to: (1) recognize the overt arguments of the target and label them; (2) recognize DNIs of the target and find links to antecedents from the context. In the *NIs only* track, the task is restricted to the second part of the full task while the gold standard local semantic argument structure (i.e., target words, frames, locally realized semantic roles and their fillers) is provided. This NIs linking task (DNI resolution) involves

three subtasks: identifying NIs, classifying NIs as either definite or indefinite (INI/DNI recognition), and finding the correct fillers for the DNIs (DNI linking).

Two systems participated in the NIs linking task. Chen et al. (2010) participated with their SRL system SEMAFOR (Das et al., 2010). They extended the SEMAFOR 1.0 frame-semantic parser by allowing fillers from a wider context (previous three sentences) and replacing the syntactic path features with two new semantic features derived from FrameNet's lexicographic exemplar annotations. The second system VENSES++ participating in the task (Tonelli & Delmonte, 2010) was built on an existing text understanding system. NIs for verbal and nominal predicates were resolved differently by exploring heuristics. Compared to the results for the full task, the results of both systems for the NIs linking task are far from satisfactory: the first ranked system achieved an F-measure of 1.40%. However, these initial works lead to a benchmark for further research in this field. Later, a few algorithms were developed for the NIs linking task on the basis of the same corpus. They focus on different subtasks (e.g., DNI linking or NIs recognition). We briefly describe these algorithms below.

Tonelli & Delmonte (2011) revisited the task by exploring a heuristic approach for INI/DNI identification and for DNI binding. The lexical statistics of the INIs and DNIs in the training set (i.e., how frequently an argument is annotated as INI and DNI) is used to decide whether a potential NI is a DNI. They then devised a relevance score to bind a DNI to its discourse referent. The relevance score is based on how often a lexical item is assigned to a frame element (FE) label in the training set and the distance between the candidate and the DNI. Tonelli & Delmonte (2011) reported an F-measure of 41% for DNI recognition and an F-measure of 8% for DNI resolution.

Silberer & Frank (2012) modeled this problem as an anaphora resolution task and explored an entity-based coreference resolution system for DNI linking. To address the data sparsity problem, the authors extended the training set by automatically acquiring heuristically labeled data. They showed that features adapted from coreference resolution (e.g., sentence distance, prominence score of the entity) have strong effects on DNI linking. Silberer & Frank (2012) achieved an F-measure of 27.7% for DNI linking and an F-measure of 10.1% for DNI resolution.

Laparra & Rigau (2012) explored the explicit frame element (FE) annotation for DNI resolution. They recognized DNIs by collecting the most common FE patterns of the corresponding frame of the target under consideration. For instance, *Resident Location* is the most common FE pattern of the frame "*residence*" (see Example 2.14). Therefore the core FE (i.e. *Location*) in this pattern is likely to be the DNI if it is omitted for the target which is being processed. The authors showed that this approach can help to improve the result for DNI recognition (with an F-measure of 57%). For the DNI linking task, the closest candidate which belongs to the same semantic type as the DNI is chosen. Laparra & Rigau (2012) reported an F-measure of 19% for DNI resolution.

Another vein of research on implicit semantic role labeling focuses on a small set of ten nominal predicates derived from verbs (Gerber & Chai, 2012), such as *investor*, *loan*, and *plan*. In contrast to the SemEval-2010 Task 10 which considers all target predicates for running texts (i.e., 3,073 instances covering 769 frame types), the corpus created by Gerber & Chai (2012) only covers ten nominal predicates (i.e., 1,247 instances covering 10 predicates). Another difference is that the SemEval-2010 Task 10 only links *core* implicit arguments (DNIs), while Gerber & Chai (2012) consider all implicit arguments. Gerber & Chai (2012) then formed the task as finding the fillers of all implicit arguments for the predicate given the gold-standard PropBank (Palmer et al., 2005) and NomBank (Meyers et al., 2004) information. This roughly corresponds to the DNI linking subtask in the SemEval-2010 Task 10 but limited to fewer predicates. Gerber & Chai (2012) applied a supervised learning method for this problem. They trained a logistic regression model by combining a set of syntactic, semantic and discourse features. The authors reported an average F-measure scores of 50.3%. Recently, Laparra & Rigau (2013) proposed a deterministic algorithm for the same problem and achieved competitive results compared to the supervised learning approach explored by Gerber & Chai (2012). They assumed that "in a coherent document the different occurrences of a predicate tend to be mentions of the same event, and thus, they share the same argument fillers" (Laparra & Rigau, 2013, p.1181). We integrate this idea into our models for bridging resolution (Chapter 6 and Chapter 7).

Implicit semantic role labeling and bridging resolution. There is a partial overlap between bridging resolution and implicit semantic role labeling for nominal predicates. Indeed, most bridging antecedents can be understood as the implicit core semantic roles of the anaphors. However, the main differences between implicit semantic role labeling and bridging resolution lie in the following aspects.

First, bridging resolution considers all possible nominal bridging anaphors in running text. Some bridging anaphors are not considered as "nominal predicates" in (implicit) semantic role labeling, e.g., **One man** in Example 2.17. Indeed, in *set/membership* bridging in which the bridging anaphor is a subset or a member of the antecedent (e.g., *employees* – **One man**), the antecedent could be any NP which makes the anaphor accommodate into the context so that the text is coherent.

(2.17)   Still, *employees* do occasionally try to smuggle out a gem or two. **One man** wrapped several diamonds in the knot of his tie.

Second, implicit semantic role labeling for nominal predicates tries to link all possible implicit core roles for the nominal predicate in question. Yet not every nominal predicate under consideration is a bridging anaphor from the discourse entity's perspective. In Example

2.18[30], the nominal predicate **losses** in the first sentence has three explicit roles: the role *arg0*, the entity losing something; *arg1*, the thing lost; and *arg3*, the source of that loss. These three arguments are considered as implicit roles for the nominal predicate **losses** in the second sentence. However, from a discourse entity's point of view, the second occurrence of **losses** in this example is not a bridging anaphor. Instead, it is an `old` entity which is coreferential with the first occurrence of **losses** introduced in the first sentence.

(2.18)   [The network]*arg0* has been expected to have [**losses**]*predicate* [of as much as \$20 million]*arg1* [on base ball]*arg3* this year. It isn't clear how much those [**losses**]*predicate* may widen because of the short Series.

Despite of the differences between implicit semantic role labeling and bridging resolution as analyzed above, we suggest that these two tasks should benefit from each other. In our work, we explore statistics from NomBank (Meyers et al., 2004), a corpus with annotations of nominal predicates and their semantic roles, to predict bridging anaphors. In addition, inspired by the work of Laparra & Rigau (2013) on implicit semantic role labeling, we design features for bridging resolution. In chapter 8, we further discuss that the *null instantiation* annotation in FrameNet could be explored for bridging anaphora recognition.

2.3   Relation Extraction

Relation extraction is the task of identifying semantic relations between entities in text. For instance, in Example 2.19, the entity *Jack Kuehler* bears a relation *employee-of* with the entity *IBM*.

(2.19)   [Jack Kuehler], [IBM's] president, said IBM is also considering letting other companies participate in additional semiconductor work.

In this section, we briefly review the main trends and representative works in this area. We then discuss the connection between relation extraction and bridging resolution.

2.3.1   Relation Extraction With Predefined Relation Types

Traditional relation extraction focuses on a small number of predefined relation types. Research in this area has been catalyzed by a series of tasks at the Message Understanding Conferences (MUC) and the NIST Automatic Content Extraction (ACE) meetings. The Template Relation task introduced at MUC-7 only focuses on the extraction of three binary relations (i.e., *product-of, employee-of*, and *location-of*) between named entities in text. After the

---

[30]The example is from Laparra & Rigau (2013).

MUC programs, The Relation Detection and Characterization (RDC) task organized by ACE includes more relation types, e.g., The RDC task at ACE 2005 contains 18 relation types, with 10,650 relation instances from 754 documents.

Various approaches have been developed for the ACE RDC task. Most of them explore *supervised machine learning* techniques, using a wide variety of "flat" features (i.e., lexical, syntactic and semantic features, such as unigrams, POS tags, entity types) (Roth & Yih, 2004; Kambhatla, 2004) or/and structural features (i.e., different kinds of paths between two entities in constituent/dependency trees) (Zelenko et al., 2003; Culotta & Sorensen, 2004; Zhao & Grishman, 2005).

Although standard supervised machine learning methods yield high performance for relation extraction in a series of RDC tasks, they rely on training data where relation instances for the target relation types are richly annotated. However, labelling textual relations is time-consuming and can only cover a small set of relation types. This leads to considerable interest in weakly supervised learning and distantly supervised learning. In the following, we discuss works in these two areas respectively.

Weakly supervised approaches. Since Hearst (1992) pioneered harvesting hyponymy relations automatically by using a number of lexical patterns designed manually, much work has explored automatic methods for discovering new patterns. Brin (1998) used a **weakly supervised** learning approach to identify the *author-book* relation from the Web. Brin collected patterns and examples for this specific relation by bootstrapping from a small set of labeled data. Research following this line includes Agichtein & Gravano (2000), Cederberg & Widdows (2003) and Pantel & Pennacchiotti (2006). Agichtein & Gravano (2000) focused on the *company-location* relation extraction. They improved Brin's approach by using a Named Entity tagger and by evaluating the quality of the new patterns and instances generated at each step of the extraction process. Cederberg & Widdows (2003) assumed that a hyponym and its hypernym should be semantically similar and that NPs occurring together in lists are usually semantically similar in some way. For instance, a coordinated noun phrase *cinnamon, cloves or coriander* indicates that "cinnamon", "cloves" and "coriander" are semantically similar. They then improved the results of hyponymy relation extraction by applying latent semantic analysis to filter false positive instances and by using noun coordination information to find more true positive instances. Pantel & Pennacchiotti (2006) used the same formalism to learn regular expressions over words and POS tags for a variety of relations, including *is-a, part-of, succession, reaction* and *production*.

Distantly supervised approaches. To reduce human annotation efforts, recently a significant amount of work has explored distantly supervised approaches to extract relations. Mintz et al. (2009) used Freebase (Bollacker et al., 2008) to provide distant supervision for relation

extraction. The intuition was that any sentence that contains a pair of entities which partic-ipate in a known Freebase relation is likely to express that relation in some way. However, some sentences may give incorrect clues. In order to alleviate the negative influences from these noisy training instances, Mintz et al. (2009) aggregated the features from all sentences containing the two entities into a single feature vector. The authors claimed that combining information from many different surface variations for the same entity pair can help to pre-dict its relation type. The intuition can be illustrated by the following example. Given two sentences which both contain the entity pair {*Steven Spielberg, Saving Private Ryan*}, the phrase in one sentence "[Steven Spielberg]'s film [Saving Private Ryan]" can indicate the re-lation type *film-director* or *film-producer*; whereas the phrase in another sentence "[Saving Private Ryan], directed by [Steven Spielberg]" can indicate the relation type *film-director* or *company-CEO*. The combination of these two phrases (features) precisely predicts the correct relation type *film-director* for the entity pair under consideration. In total, Mintz et al. (2009) collected a dataset which contains 1.2 million Wikipedia articles and 1.8 million instances of 102 relation types connecting 940,000 entities. They showed that a multi-class logistic classi-fier trained over these "noisy" instances by aggregating features from different sentences can extract 10,000 instances of 102 relation types at a precision of 67.6%.

Although the approach proposed by Mintz et al. (2009) works well for the *aggregating re-lation extraction* task (i.e., extracting ground facts from a corpus where the facts are expressed somewhere in the corpus) when the textual corpus is tightly aligned to the database, it leads to more noisy data and the poor extraction performance when applied more broadly. Riedel et al. (2010) observed that 31% of false positives when matching Freebase relations with New York Times articles. Indeed, the assumption of "one relation per one entity pair in all sentences" in Mintz et al. (2009) is often violated: first, if two entities participate in a relation, only some sentences that mention these two entities might express the relation. Second, the same pair of entities may have multiple relations. For instance, in Example 2.20, the entity pair {Barack Obama, United States} has two relations expressed in different sentences.

(2.20)   s1: [Obama] was born in the [United States] just as he has always said. (*bornIn*)
           s2: [Barack Obama] is the 44th and current President of the [United States]. (*employedBy*)

In order to overcome the problems mentioned above when applying distantly supervised learning to extract relations between entity pairs, several works have explored different ways to combine sentence-level relation extraction models with corpus-level relation extraction mod-els. Riedel et al. (2010) modeled the task as a *multi-instance single-label* problem. They relaxed Mintz et al.'s assumption, saying that "at least one sentence which mentions two re-lated entities expresses the corresponding relation". They then proposed an undirected graph-ical model to solve the following two tasks jointly: (1) predicting the relation for an entity

pair; and (2) predicting which sentences express this relation. Hoffmann et al. (2011) and Surdeanu et al. (2012) relaxed Mintz et al.'s assumption even further by modeling the task as a *multi-instance multi-label* problem, i.e., two related entities can have multiple relations in the corpus-level and each relation can be expressed in different sentences. Both approaches explored probabilistic graphical models and modeled the labels (relations) assigned to the instances of an entity pair as hidden variables. Hoffmann et al. (2011) used a deterministic model that aggregates latent instance labels into a set of labels for the corresponding entity pair. In contrast, Surdeanu et al. (2012) jointly inferred relation extraction decisions for the aggregated level (i.e., labels for an entity pair) and for the sentence level (i.e., labels for the instances of this entity pair in text) simultaneously. Moreover, this framework also learns dependencies between candidate labels for the same entity pair. For instance, it can learn that two relation labels (e.g., *bornIn* and *spouseOf*) cannot be generated jointly for the same entity pair.

2.3.2 Open-domain Relation Extraction

Unlike traditional relation extraction where target relation types are predefined, open-domain relation extraction aims to extract relation instances from the text where neither relation types nor labelled relation instances are available. Research in this direction has explored various *unsupervised approaches* to extract relation instances.

Open IE (Banko et al., 2007) aims to discover relations which are independent of specific domains. Banko et al. (2007) employed a self-learner to extract relation instances from the Web. They then assessed the reliability for each extracted relation instance on the basis of the Web redundancy: the more times a relation instance appears in different sentences, the more likely it is a valid one. Banko et al. (2007) reported that their Open IE system *TextRunner* extracted over 1,000,000 concrete facts among 11,000,000 highest probability relation instances on experiments over a corpus containing 9 million web pages. Although Open IE is highly scalable and efficient, it lacks the ability to generalize. For instance, Open IE does not know that "is headquartered in" and "is based in" are different ways of expressing "*Headquarters (X, Y)*".

To gain generalization, another line of work has explored inducing relation instances along with relation types via clustering. The general idea is based on Harris's *distributional hypothesis* (Harris, 1954), which states that words which occur in the same contexts tend to have similar meanings. Following this idea, Hasegawa et al. (2004) applied hierarchical clustering to group pairs of named entities according to the similarity of context words intervening between the named entities. Unlike Hasegawa et al. (2004), Lin & Pantel (2001) represented relations as triples: *Slot X - dependency path - Slot Y*. They then defined the context of a path as its fillers, i.e., "government" and "problem" are fillers for the path "*X finds a solu-*

tion to Y". Therefore, two paths have similar meanings if they tend to connect similar fillers, such as "*X is manufacturer of Y*" and "*Y is product made by X*". Indeed, the method proposed by Lin & Pantel (2001) is a general framework to identify paraphrases, which can be used in wide applications such as textual entailment, relation extraction and query expansion for question answering. On the basis of the triple representation developed by Lin & Pantel (2001), Yao et al. (2011) applied an unsupervised probabilistic generative approach (i.e., topic models) to extract the relations between named entities. The model also infers the underlying relation types along with the fine-grained entity types simultaneously. The aforementioned approaches, however, assume "one sense per each path". This is not always true, e.g., "A beats B" can mean that athlete A beats athlete B in a sports match or that person A wins over person B in competing for a political position. To deal with the polysemy problem, Yao et al. (2012) first employed a topic model to partition entity pairs of a path (e.g., *A play B*) into different sense clusters (e.g., *sportsTeam play sportsTeam* and *musician play musicPiece*), they then used Hierarchical Agglomerative Clustering (HAC) to merge senses into relation types.

Unlike the above mentioned approaches which mainly focus on clustering similar relations, the RESOLVER system (Yates & Etzioni, 2007) applies clustering to both relations and entities. The system takes relation instances (i.e., *object1- relation string - object2*, such as *Washington - is capital city of - U.S.*) extracted from the Web as input. It then applies a probabilistic model to cluster relation strings into sets of synonyms and to cluster objects into sets of coreferential names in an iterative manner.

Recently, Riedel et al. (2013) proposed a *universal schema* approach for open-domain relation extraction. The *universal schema* is the union of all source schemas, including original input forms (e.g., variants of surface patterns similar to OpenIE) and relations in the schemas of many pre-existing databases. They then applied matrix factorization models to operate simultaneously on relations observed in text and in structured databases. Riedel et al. (2013) showed that such models can learn asymmetric implicature among relations, and therefore can predict relations that do not appear explicitly in text. For instance, after observing the relation instance *FERGUSON – historian at – HARVARD* the system infers the relation instance *FERGUSON – professor at – HARVARD*, but not vice versa.

2.3.3 Relation Extraction and Bridging Resolution

Relation extraction and bridging resolution share some common properties. For instance, both tasks aim to extract relations between entities in text. From the perspective of relation extraction, bridging resolution can be seen as a task of extracting non-identity, implicit relations between discourse entities, with the constraint that the second entity of a relation is anaphoric.

However, there are two major differences between these two tasks. First, most works

on relation extraction only consider extracting relations within sentences[31]. On the contrary, bridging resolution often needs to extract relations across sentence boundaries.

Second, most works on relation extraction only focus on extracting semantic relations between named entities or extracting lexical relations between simple nouns. Therefore, the output of these approaches could be aggregated as context independent resources, such as facts or ontologies. In contrast, in bridging resolution, we need to consider extracting context dependent relations between all noun phrases or between noun phrase and verb/clause. More importantly, the second entity in a bridging relation should be "anaphoric".

There is a partial overlap between set/membership bridging and entity instantiation (McKinlay & Markert, 2013). Entity instantiation considers member-of or subset-of relations between entities in text, such as "several EU countries" and "the UK" in Example 2.21[32] and "employees" and "one man" in Example 2.22. Although both set/membership bridging and entity instantiation consider similar relations between all noun phrases in the whole discourse level, not all entity instantiations are set/membership bridging relations. In Example 2.21, the set member, "the UK", is not anaphoric.

(2.21)   Inflation has increased sharply in [several EU countries].
         In [the UK], this has accompanied a drop in interest rates.

(2.22)   Still, [*employees*] do occasionally try to smuggle out a gem or two. [**One man**] wrapped several diamonds in the knot of his tie.

Overall, we suggest that research on bridging resolution and information extraction can benefit from each other. In Chapter 3, inspired by Hasegawa et al.'s work on unsupervised relation extraction (Hasegawa et al., 2004), we apply hierarchical clustering to analyze bridging relations in the ISNotes corpus. In Chapter 8, we assume that bridging resolution can be applied to extract relations across sentence boundaries.

---

[31]Swampillai & Stevenson (2011) are an exception with regard to the scope of their relations. They applied an SVM model to extract relations within and across sentences by exploring structural features, i.e., joining parse trees for pairs of entities under a new root node.

[32]The example is from McKinlay (2013).

Chapter 3

ISNotes: A Corpus for Information Status

This chapter focuses on the corpus ISNotes that we use throughout the thesis. Section 3.1 provides an overview of the corpus. On the basis of this corpus, we then analyze bridging from five different perspectives: the syntactic and the topological character of bridging anaphora and bridging antecedents respectively (Section 3.2 and Section 3.3), the distance between bridging anaphors and antecedents (Section 3.4), and the interaction between bridging and discourse relations (Section 3.5), Finally, we summarize the chapter in Section 3.6.

3.1 An Overview of ISNotes

Data. ISNotes contains 10,980 mentions[1] annotated for information status in 50 texts taken from the Wall Street Journal portion of the OntoNotes corpus (Weischedel et al., 2011). The corpus[2] can be downloaded from: `http://www.h-its.org/english/research/-nlp/download/isnotes.php`.

Information status in ISNotes. Information status (IS henceforth) describes the degree to which a discourse entity is available to the hearer regarding the speaker's assumption about the hearer's knowledge and beliefs. We explain the eight IS categories annotated in ISNotes in the following. More detailed discussion of annotation decisions can be found in the annotation scheme (Markert, 2013).

A mention is `old` if it is either coreferent with an already introduced entity, or if it is a generic or deictic pronoun. ISNotes integrates the OntoNotes coreference annotation into its `old` IS annotation.

---

[1] In ISNotes, *mentions* are noun phrases (NPs) which carry information statuses.

[2] The development of the annotation scheme and of the framework, the annotation study and the agreement study were carried out by Katja Markert.

`Mediated` mentions have not been mentioned before but are not autonomous, i.e., they can only be correctly interpreted by reference to another mention or to prior world knowledge. ISNotes distinguishes six subcategories of `mediated` mentions:

- `mediated/worldKnowledge` mentions are generally known to the hearer. This category includes many proper names, such as *Poland*.

- `mediated/syntactic` mentions are syntactically linked via a possessive relation, a proper name premodification or a PP (prepositional phrase) postmodification to other `old` or `mediated` mentions, such as:

  [[their]$_{old}$ liquor store]$_{mediated/syntactic}$,

  [the [Federal Reserve]$_{mediated}$ boss]$_{mediated/syntactic}$, and

  [the main artery into [San Francisco]$_{mediated}$]$_{mediated/syntactic}$.

- `mediated/bridging` mentions are inferable because a related entity or event (antecedent) has been previously introduced in the discourse, such as **the streets** in Example 3.1 and **The reason** in Example 3.2.

- `mediated/comparative` mentions usually include a premodifier that makes clear that this entity is compared to a previous one (antecedent), such as **others** in Example 3.3. Normally the `mediated/comparative` mention and its antecedent (the previous mention it compares to) are not identical but belong to the same semantic type.

- `mediated/aggregate` mentions are coordinated mentions where at least one element in the conjunction is `old` or `mediated`, such as *[Not only [George Bush]$_{mediated}$ but also [Barack Obama]$_{mediated}$]$_{mediated/aggregate}$.*

- `mediated/function` mentions refer to a value of a previously explicitly mentioned function (e.g., *3 points* in Example 3.4). The function needs to be able to rise and fall.

(3.1)   *Oranjemund, the mine headquarters*, is a lonely corporate oasis of 9,000 residents. Jackals roam **the streets** at night . . .

(3.2)   The Bakersfield supermarket *went out of business* last May. **The reason** was . . .

(3.3)   As the death toll from last week's temblor climbed to 61, the condition of *freeway survivor Buch Helm*, who spent four days trapped under rubble, improved, hospital officials said. Rescue crews, however, gave up hope that **others** would be found.

(3.4)   IBM shares <u>were down</u>$_{function}$ *3 points*.

`New` mentions are entities that have not yet been introduced in the discourse and that the hearer cannot infer from either previously mentioned entities/events or general world knowledge.

Antecedents for mediated/bridging and mediated/comparative. In ISNotes, antecedents for both `mediated/bridging` and `mediated/comparative` categories are annotated[3]. The antecedents can be noun phrases (e.g., *Oranjemund, the mine headquarters* in Example 3.1), verb phrases (e.g., *went out of business* in Example 3.2) or even clauses. For the NP antecedent, if it has several instantiations within the text, ISNotes chooses the one which is the closest to the `bridging` or `comparative` mention. However, other instantiations can be inferred from the coreference annotation and are counted as correct antecedents as well.

Sometimes a `mediated/bridging` mention could have several antecedents when all antecedents are its missing mandatory roles. In Example 3.5, both antecedents (i.e., *Japan's* and *cars, trucks and buses*) are necessary to interpret the bridging anaphor **Domestic demand**.

(3.5)   *Japan's* production of *cars, trucks and buses* in September fell 4.4% from a year ago. [. . .] **Domestic demand** continues to grow, . . .

Agreement study. ISNotes is reliably annotated for most categories. The agreement study was carried out among three annotators. Annotator A is the scheme developer and a computational linguist. Annotator B and C have no linguistic training or education. Annotator A and B are fluent English speakers, living in English-speaking countries, but are not native speakers. Annotator C is a native speaker of English. The annotation task consisted of marking all mentions for their information status (`old`, `mediated` or `new`). Annotators also had to mark all subcategories of the three main types as well as the antecedents for comparative and bridging anaphora[4].

The scheme was developed on nine texts, which were also used for training the annotators. Inter-annotator agreements was measured on 26 new texts, which included 5,905 pre-marked potential mentions. The annotations of 1,499 of these were carried over from OntoNotes[5], leaving 4,406 potential mentions for annotation and agreement measurement.

Table 3.1 shows agreement results (percentage agreement as well as Cohen's κ (Artstein & Poesio, 2008) between all three possible annotator pairings) for the overall scheme at the coarse-grained (four categories: not a mention, old, new, mediated) and the fine-grained version (nine categories: not a mention, old, new and the six mediated subtypes)[6]. The results

---

[3]Antecedents for `old` mentions are from the OntoNotes coreference annotation.

[4]In ISNotes, a comparative anaphor is a mention whose information status is `mediated/comparative`, and a bridging anaphor is a mention whose information status is `mediated/bridging`.

[5]The existing OntoNotes coreference annotation was automatically carried over to the information status task by marking all mentions in a coreference chain (apart from the first mention in the chain) as `old`. Annotators were not allowed to override this annotation but could add other old and coreference annotations.

[6]We notice that the overall κ values for the fine-grained categories are higher than coarse-grained categories. The reason is that the hierarchy scheme motivated by the linguistic insights is organized in a way where a category lower down the tree is more often confused with a category higher up in the tree in a different branch of the

	A-B	A-C	B-C
Overall Percentage coarse	87.5	86.3	86.5
Overall κ coarse	77.3	75.2	74.7
Overall Percentage fine	86.6	85.3	85.7
Overall κ fine	80.1	77.7	77.3

Table 3.1: Agreement results.

show that the scheme is overall highly reliable, with not too many differences between the different annotator pairings[7].

Table 3.2 shows the individual category agreement for all nine fine-grained categories, where all categories but one are merged and then κ is computed as usual.

	A-B	A-C	B-C
κ Non-mention	81.5	78.9	86.0
κ Old	80.5	83.2	79.3
κ New	76.6	74.0	74.3
κ Mediated/Knowledge	82.1	78.4	74.1
κ Mediated/Syntactic	88.4	87.8	87.6
κ Mediated/Aggregate	87.0	85.4	86.0
κ Mediated/Function	6.0	83.2	6.9
κ Mediated/Comparative	81.8	78.3	81.2
κ Mediated/Bridging	70.8	60.6	62.3

Table 3.2: Agreement results for individual categories.

The results show that high reliability is achieved for most individual categories[8]. The agreement score for the category `bridging` is more annotator-dependent and relatively lower compared to other categories. This reflects the difficulty in recognizing `bridging`

---

tree than with its direct siblings in the tree, i.e., `mediated/bridging` mentions are often confused with `new` mentions whereas some mediated categories such as `mediated/syntactic` or `mediated/comparative` are very easy to recognize.

[7]Often, annotation is considered highly reliable when κ exceeds 0.80 and marginally reliable when between 0.67 and 0.80 (Carletta, 1996). However, the interpretation of κ is still under discussion (Artstein & Poesio, 2008).

[8]The low reliability of the very rare category `function`, when involving Annotator B, was solely explained by Annotator B forgetting about this category completely and only using it once. When two annotators remembered the category, it was actually easy to annotate reliably (κ 83.2 for the pairing A-C).

mentions. However, the reliability of the category `bridging` is still higher, sometimes considerably, than other previous attempts at bridging annotation (Fraurud, 1990; Poesio, 2003; Gardent & Manuélian, 2005; Riester et al., 2010). In particular, the bridging annotations of the pairing A-B was used to create a consistent gold standard. The agreement of selecting bridging antecedents is around 80% for all annotator pairings.

IS distribution. Table 3.3 shows the IS distribution in ISNotes among 50 texts which contain 1,726 sentences in total.

Texts	50	
Sentences	1,726	
Mentions	10,980	
old	3237	29.5%
coref	3,143	28.6%
generic_deictic_pr	94	0.9%
mediated	3,708	33.8%
syntactic	1,592	14.5%
world knowledge	924	8.4%
bridging	663	6.0%
comparative	253	2.3%
aggregate	211	1.9%
func	65	0.6%
new	4,035	36.7%

Table 3.3: IS distribution in ISNotes. The last column indicates the percentage of each IS category relative to the total number of mentions.

Two things are notable here. First, `new` mentions are the largest proportion (36.7%) of the three coarse categories. However, in the Switchboard corpus which contains 147 dialogues (see Section 2.1.2 in Chapter 2 for a detailed description), only 14.4% of NPs fall into the `new` type. This comparison reflects that the news genre tends to introduce more new information compared to the conversation genre. Second, `syntactic` mentions are the largest proportion among all subcategories of the type `mediated`. On the contrary, the percentages of the corresponding categories in the Switchboard corpus are only 1.0% (`med/bound`) and 2.7% (`med/poss`). This is because that in Switchboard, two types of bridging (i.e., *med/part* (1.7%) and *med/set* (20.2%)) include some non-anaphoric, syntactically linked part-of and set-member relations, which correspond to `mediated/syntactic` in ISNotes.

Bridging relations. In ISNotes, the semantic relations between anaphor and antecedent are extremely diverse. Table 3.4 shows that among 683 bridging pairs, only 2.3% of them have actions/verbs antecedents, only 6.6% of bridging pairs have a set/membership relation in which the bridging anaphor is a subset or a member of the antecedent, and only 13.5% of bridging pairs have a part-of/attribute-of relation in which the anaphor is a part or an attribute of the antecedent. 77.6% of bridging pairs fall under the category "other", without further distinction. This includes encyclopedic relations such as *restaurant* – **the waiter** as well as context-specific relations such as *palms* – **the thieves**.

Relation Type	Bridging Pairs	
Action	16	(2.3%)
Set/Membership	45	(6.6%)
part-of/attribute-of	92	(13.5%)
Other	530	(77.6%)
Total	683	(100.0%)

Table 3.4: Bridging pairs distribution w.r.t. relation types.

3.2 Corpus Analysis: Bridging Anaphora

In ISNotes, bridging anaphors are mentions which belong to the IS type `mediated/bridging`. This section analyzes the types of bridging anaphora in ISNotes from three aspects. First, we examine the bridging anaphora distribution with regard to the part of speech (POS) of a bridging anaphor's head. Second, we examine the modifications of bridging anaphora, and finally, we examine the distribution of bridging anaphora in terms of whether a bridging anaphor shares the same antecedent with another bridging anaphor or not.

Bridging anaphora types. We extracted the head word of each bridging anaphor. We first convert the Penn Treebank annotation to dependency parse trees using the Penn Converter tool (Johansson & Nugues, 2007). Then we select the head word of a bridging anaphor by examining the dependency parse tree of the NP: the word which is not dependent on any other word in the NP is chosen to be the head word of the NP. On the basis of the part of speech of the head word, we classify bridging anaphors into the following categories:

- **Common Noun:** the POS tag of the head word is *NN* or *NNS*.

- **Proper Name:** the POS tag of the head word is *NNP* or *NNPS*, or the NP is a named entity according to the named entity annotations.

- **Pronoun:** the POS tag of the head word begins with *PRP*, or the head word appears in a list of pronouns[9].

- **Other:** the head word does not fit into any of the above categories. This includes numbers (e.g., *$30,000*), adjectives (e.g., *the poor*) and gerunds (e.g., *his attending*).

Table 3.5 shows the bridging anaphora distribution with regard to the POS tag of the head word compared to other NPs. To test the significance of the difference between the distribution of bridging anaphora and other NPs, we carried out a χ^2 test using R (R Development Core Team, 2011) for consistency in a 4 x 2 table[10]. This gave $\chi^2 = 306.2353$, $df = 3$ and $p - value < 2.2e - 16$. The statistics suggest that the bridging anaphora distribution is significantly different to other NPs, with a much higher proportion of common nouns and much lower proportion of proper names and pronouns than other NPs.

NP Type	Bridging Anaphors		Other NPs	
Common Noun	605	(91.3%)	5,859	(56.8%)
Proper Name	42	(6.3%)	2,762	(26.8%)
Pronoun	14	(2.1%)	1,564	(15.1%)
Other	2	(0.3%)	132	(1.3%)
Total	663	(100.0%)	10,317	(100.0%)

Table 3.5: Bridging anaphora distribution w.r.t. the POS tag of the head word.

Bridging anaphora modifications. In ISNotes, bridging anaphora can be any noun phrase and are not limited to definite NPs as in Vieira & Poesio (2000), Poesio et al. (2004a), Gardent & Manuélian (2005), and Riester et al. (2010). Here we focus on a formal definition of definiteness instead of a semantic one. Following the definition of *Definite Description* in Vieira & Poesio (2000), we are interested in how many bridging anaphors are definite descriptions (i.e., NPs modified by the definite article *the*). Therefore we first examine bridging anaphors'

---

[9]Some pronouns have other POS tags such as *DT* (e.g., *this, that*) or *CD* (e.g., *one*). Therefore we compile an extra list of pronouns to improve the recall of pronoun detection. The whole list is: {*all, another, any, anybody, anyone, anything, both, each, either, everybody, everyone, everything, few, little, many, more, much, most, nobody, none, nothing, neither, one, other, others, some, somebody, something, someone, this, that, these, those*}.

[10]We use Pearson's χ^2 test with Yates's continuity correction.

modifications in terms of determiners. We again extract the head word for each bridging ana-
phor, and then look for words within the NP that depend upon the head word and precede it.
We classify bridging anaphors into one of four categories:

- **The:** the NP head is modified by the article *the* which normally indicates definite NPs.

- **A/An:** the NP head is modified by articles *a*, *an* which normally indicate indefinite NPs.

- **Other-determiner:** the NP head is modified by other determiners (e.g., demonstratives,
  possessives or quantifiers) which do not belong to the above two categories.

- **Non-determiner:** the NP head is not modified by any determiners, such as **residents** or
  relief efforts.

Table 3.6 shows the bridging anaphora distribution with regard to determiners in ISNotes:
only 38.5% of bridging anaphors are modified by *the*, most bridging anaphors (44.9%) are not
modified by any determiners. This calls into question the strategy of prior approaches to limit
themselves to definite bridging anaphora (i.e., bridging anaphora which are modified by *the*)
only. We find significant differences between the distribution of bridging anaphors and other
NPs regarding determiners ($\chi^2 = 175.4356$, $df = 3$, $p - value < 2.2e - 16$).

NP Type	Bridging Anaphors		Other NPs	
The	255	(38.5%)	1,877	(18.2%)
A/An	70	(10.6%)	928	(9.0%)
Other-determiner	40	(6.0%)	790	(7.7%)
Non-determiner	298	(44.9%)	6,722	(65.1%)
Total	663	(100.0%)	10,317	(100.0%)

Table 3.6: Bridging anaphora distribution w.r.t. determiners.

Although the largest proportion of bridging anaphora is "non-determiner" in Table 3.6,
these bridging anaphors could have other premodifers or postmodifiers. To gain a better view
of syntactic properties of bridging anaphora, e.g., how many bridging anaphors are "bare"
(NPs without any modifiers), we distinguish five main categories for bridging anaphora ac-
cording to the modifier's position and type:

- **Head**: the NP head has neither premodifications nor postmodifications.

- **Determiner + Head:** the NP head is only premodified by a determiner.

- **Other premodification:** the NP head only has premodifications, and at least one premodifier is not determiner. This includes NPs modified by both determiners and other premodifiers (e.g., **this dying and distorted system**), and NPs modified by premodifiers other than determiners (e.g., **relief efforts**).

- **Postmodification:** the NP head only has postmodifications, such as **viewers who have not been with us since the beginning**.

- **Both:** The NP head has both premodifications and postmodifications, such as **the contributions the LDP members received**.

The results in Table 3.7 show that only 20.5% of bridging anaphors are "bare" without any modifiers (e.g., **support** or **officials**).

NP Type	Bridging Anaphors		Other NPs	
Head	136	(20.5%)	2,863	(27.8%)
Determiner + Head	171	(25.8%)	1,075	(10.4%)
Premodification	258	(38.9%)	2,796	(27.1%)
Postmodification	20	(3.0%)	995	(9.6%)
Both	78	(11.8%)	2,589	(25.1%)
Total	663	(100.0%)	10,317	(100.0%)

Table 3.7: Bridging anaphora distribution w.r.t. modifications.

The distribution of bridging anaphora is significantly different from other NPs regarding modifications ($\chi^2 = 248.121$, $df = 4$, $p - value < 2.2e - 16$). We notice that bridging anaphora are likely to have a simple internal syntactic structure with regard to modification. As a result, 89.9% of bridging anaphors do not contain any other mentions.

Bridging anaphora: sibling and non-sibling. We define a bridging anaphor to be a *sibling anaphor* if it has the same antecedent with at least one other bridging anaphor. On the contrary, a *non-sibling* bridging anaphor does not share the antecedent with any other bridging anaphors.

Table 3.8 shows that most bridging anaphors (61.4%) are *sibling* anaphors. These *sibling* anaphors compose 118 "sibling clusters". A sibling cluster contains all sibling anaphors which share the same antecedent. We further examine these sibling clusters from two dimensions:

Type	Bridging Anaphors	
Sibling	407	(61.4%)
Non-Sibling	256	(38.6%)
Total	663	(100.0%)

Table 3.8: Distribution of *sibling* and *non-sibling* bridging anaphors.

(1) the number of sibling anaphors each sibling cluster contains; and (2) the average distance measured in words between sibling anaphors for each sibling cluster.

For a cluster c which contains n sibling anaphors, the average distance between sibling anaphors is calculated as below:

$$\text{distance}_c = \frac{\sum_{i,j \leq n} \text{distance}(i,j)}{n \cdot (n-1)} \tag{3.6}$$

The results in Table 3.9 show the distribution of sibling clusters in two dimensions: cluster size and average distance between cluster members. However, these two dimensions are not strongly correlated (Figure 3.1). The reason is that in a small sibling cluster consisting of only two sibling anaphors, these two sibling anaphors could be distant from each other but both could be close to the (different) instantiations of their common antecedent. In addition, we note that most sibling clusters are concentrated in the lower-left area in Figure 3.1.

Sibling Clusters	Min	Max	Mean	Median	Standard Deviation
Size	2.00	17.00	3.50	3.00	2.31
Distance	2.00	1160.60	154.89	57.63	224.86

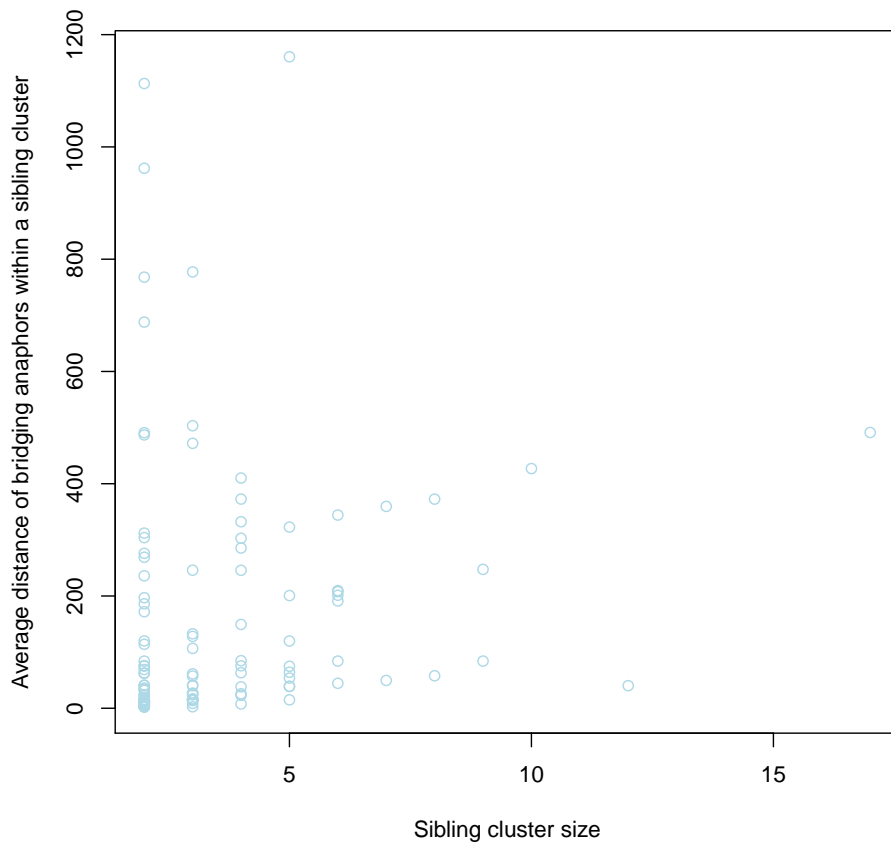Table 3.9: Distribution of sibling clusters.

Figure 3.1: Scatter plot of sibling cluster sizes versus sibling cluster average distances.

3.3 Corpus Analysis: Bridging Antecedents

In this section, we first analyze the distribution of bridging antecedents from the syntactic aspect. We then examine the salience of bridging antecedents.

Bridging antecedents: entities and events. Bridging antecedents in ISNotes can be entities or events. An entity antecedent could have several instantiations (represented by different mentions which are coreferent[11]) within the text. In very few cases, an entity antecedent is a non-mention NP, e.g., *parakeet* in Example 3.7, *food* in Example 3.8 and *East German* in Example 3.9. An event antecedent is represented by verbs or clauses/sentences (e.g., Example 3.10 and Example 3.11). Table 3.10 shows that most bridging antecedents (89.9%) are entities.

---

[11]Here we neglect verbs in event coreference.

(3.7)   For example, one *parakeet* owner returning home found that her apartment, like
            many others in the Marina, didn't have heat. [. . .]
            A warm foster home has been found.

(3.8)   On Aug. 1, the state tore up its controls, and *food* prices leaped.
            Without **buffer stocks**, **inflation** exploded.

(3.9)   *East German* leader Krenz said he was willing to hold talks with opposition groups
            pressing for internal changes.
            The **Communist Party** chief, facing what is viewed as the nation's worst unrest in
            nearly 40 years, also said he would allow East Germans to travel abroad freely.

(3.10)   . . . the drug still *lacks* federal approval for use in the youngest patients.
            As **a result**, many youngsters have been unable to obtain the drug . . .

(3.11)   *So what does George Bush really believe?*
            The answer is so murky that it is beginning to get this popular president in trouble
            . . .

Type	Bridging Antecedent	
Entity (mention NPs)	345	(89.1%)
Entity (non-mention NPs)	3	(0.8%)
Event (verbs and clauses)	39	(10.1%)
Total	387	(100.0%)

Table 3.10: Frequency of entity antecedents and event antecedents in ISNotes.

Bridging antecedents: globally salient and locally salient.   Poesio et al. (2004a) claimed
that bridging anaphora are sensitive to the *local* rather than the *global* focus (Grosz & Sid-
ner, 1986). Here we focus on examining the salience of entity antecedents represented by
mentions.

First, we assume that bridging antecedents are salient in the sense that they are omitted
later when the associative anaphors (bridging anaphors) are mentioned. We further assume
that the difference between salience, i.e., whether the antecedent is globally salient or locally
salient, may have an influence on the interpretation of bridging anaphora. However, it is
difficult to measure the salience of an entity directly. Here we use some heuristics to classify
entity antecedents into two categories: *globally salient* and *locally salient*.

An entity antecedent is globally salient if it: (1) appears in the headline, or (2) has the
document span ratio $\geq r$; otherwise it is locally salient. The document span ratio of an entity

is calculated via the span of text (measured in sentences) in which the entity is mentioned divided by the number of sentences in the whole document. Each entity antecedent could be linked to several bridging anaphors. We define a link between a bridging anaphor and an entity antecedent as a "*bridging pair*".

Table 3.11 shows the distribution of bridging antecedents regarding global or local salience as well as the corresponding bridging pairs under each value of $r \in \{0.6, 0.7, 0.8, 0.9\}$. We find that although the globally salient antecedents are much more infrequent than the locally salient antecedents for every value of r, they connect to more bridging anaphors with higher average number of pairs per antecedent. For instance, when r equals 0.8, we get 78 globally salient antecedents and 267 locally salient antecedents. These 78 globally salient antecedents connect to 257 bridging anaphors while the 267 locally salient antecedents connect to 383 bridging anaphors. As a result, on average, each globally salient antecedent participates in 3.3 bridging pairs, whereas each locally salient antecedent only participates in 1.4 bridging pairs.

Antecedent Type	Bridging Antecedent		Bridging Pair		Average Pairs per Ante.
$r = 0.6$					
Globally Salient	95	(27.5%)	290	(45.3%)	3.1
Locally Salient	250	(72.5%)	350	(54.7%)	1.4
Total	345	(100.0%)	640	(100.0%)	1.9
$r = 0.7$					
Globally Salient	89	(25.8%)	279	(43.6%)	3.1
Locally Salient	256	(74.2%)	361	(56.4%)	1.4
Total	345	(100.0%)	640	(100.0%)	1.9
$r = 0.8$					
Globally Salient	78	(22.6%)	257	(40.2%)	3.3
Locally Salient	267	(77.4%)	383	(59.8%)	1.4
Total	345	(100.0%)	640	(100.0%)	1.9
$r = 0.9$					
Globally Salient	71	(20.6%)	239	(37.3%)	3.4
Locally Salient	274	(79.4%)	401	(62.7%)	1.5
Total	345	(100.0%)	640	(100.0%)	1.9

Table 3.11: Globally and locally salient entity antecedents and their corresponding bridging pairs in ISNotes.

We further analyze the relations between the salience of bridging antecedents and the topology of bridging anaphors, i.e., how often a globally (locally) salient entity antecedent connects to sibling and non-sibling anaphors separately. The results are shown in Table 3.12, Table 3.13, Table 3.14 and Table 3.15 for r equaling 0.6, 0.7, 0.8 and 0.9 respectively.

We notice that in terms of the linkage to sibling and non-sibling anaphors, there are significant differences between globally salient entity antecedents and locally salient entity antecedents[12].

For instance, when r equals 0.8 (see Table 3.14), among 78 globally salient antecedents, 48 of them connect to 227 sibling anaphors and 30 of them connect to non-sibling anaphors. In contrast, among 267 locally salient antecedents, only 70 of them connect to 186 sibling anaphors while most of them (197) connect to non-sibling anaphors. As a result, on average, each globally salient antecedent connects to 4.7 sibling anaphors whereas each locally salient antecedent only connects to 2.7 sibling anaphors.

Figure 3.2 shows the difference more clearly. It plots sibling and non-sibling anaphors in terms of the linkage to globally or locally salient antecedents. The size of a circle or a square indicates the size of the corresponding sibling anaphor cluster (from 2 to 17), or the size of a non-sibling anaphor (1). When r equals 0.8, it is clear that **globally salient antecedents** (marked with red circles) **connect to a higher proportion of sibling anaphors and a lower proportion of non-sibling anaphors compared to locally salient antecedents** (marked with blue squares).

Antecedent Type	Bridging Antecedent		Bridging Pair		Average Pairs per Ante.
Globally Salient	95	(100.0%)	290	(100.0%)	3.1
– with Sibling Anaphors	57	(60.0%)	252	(86.9%)	4.4
– with Non-sibling Anaphors	38	(40.0%)	38	(13.1%)	1.0
Locally Salient	250	(100.0%)	350	(100.0%)	1.4
– with Sibling Anaphors	61	(24.4%)	161	(46.0%)	2.6
– with Non-sibling Anaphors	189	(75.6%)	189	(54.0%)	1.0

Table 3.12: Globally and locally salient entity antecedents and their corresponding bridging pairs by sibling and non-sibling anaphors, $r = 0.6$.

---

[12] $\chi^2 = 37.2024, df = 1, p-value = 1.065e-09$ for $r = 0.6$; $\chi^2 = 38.9484, df = 1, p-value = 4.352e-10$ for $r = 0.7$; $\chi^2 = 31.9137, df = 1, p-value = 1.612e-08$ for $r = 0.8$; $\chi^2 = 26.1483, df = 1, p-value = 3.162e-07$ for $r = 0.9$.

Antecedent Type	Bridging Antecedent		Bridging Pair		Average Pairs per Ante.
Globally Salient	89	(100.0%)	279	(100.0%)	3.1
– with Sibling Anaphors	55	(61.8%)	245	(87.8%)	4.5
– with Non-sibling Anaphors	34	(38.2%)	34	(12.2%)	1.0
Locally Salient	256	(100.0%)	361	(100.0%)	1.4
– with Sibling Anaphors	63	(24.6%)	168	(46.5%)	2.7
– with Non-sibling Anaphors	193	(75.4%)	193	(53.5%)	1.0

Table 3.13: Globally and locally salient entity antecedents and their corresponding bridging pairs by sibling and non-sibling anaphors, $r = 0.7$.

Antecedent Type	Bridging Antecedent		Bridging Pair		Average Pairs per Ante.
Globally Salient	78	(100.0%)	257	(100.0%)	3.3
– with Sibling Anaphors	48	(61.5%)	227	(88.3%)	4.7
– with Non-sibling Anaphors	30	(38.5%)	30	(11.7%)	1.0
Locally Salient	267	(100.0%)	383	(100.0%)	1.4
– with Sibling Anaphors	70	(26.2%)	186	(48.6%)	2.7
– with Non-sibling Anaphors	197	(73.8%)	197	(51.4%)	1.0

Table 3.14: Globally and locally salient entity antecedents and their corresponding bridging pairs by sibling and non-sibling anaphors, $r = 0.8$.

Antecedent Type	Bridging Antecedent		Bridging Pair		Average Pairs per Ante.
Globally Salient	71	(100.0%)	239	(100.0%)	3.4
– with Sibling Anaphors	43	(60.6%)	211	(88.3%)	4.9
– with Non-sibling Anaphors	28	(39.4%)	28	(11.7%)	1.0
Locally Salient	274	(100.0%)	401	(100.0%)	1.5
– with Sibling Anaphors	75	(27.4%)	202	(50.4%)	2.7
– with Non-sibling Anaphors	199	(72.6%)	199	(49.6%)	1.0

Table 3.15: Globally and locally salient entity antecedents and their corresponding bridging pairs by sibling and non-sibling anaphors, $r = 0.9$.
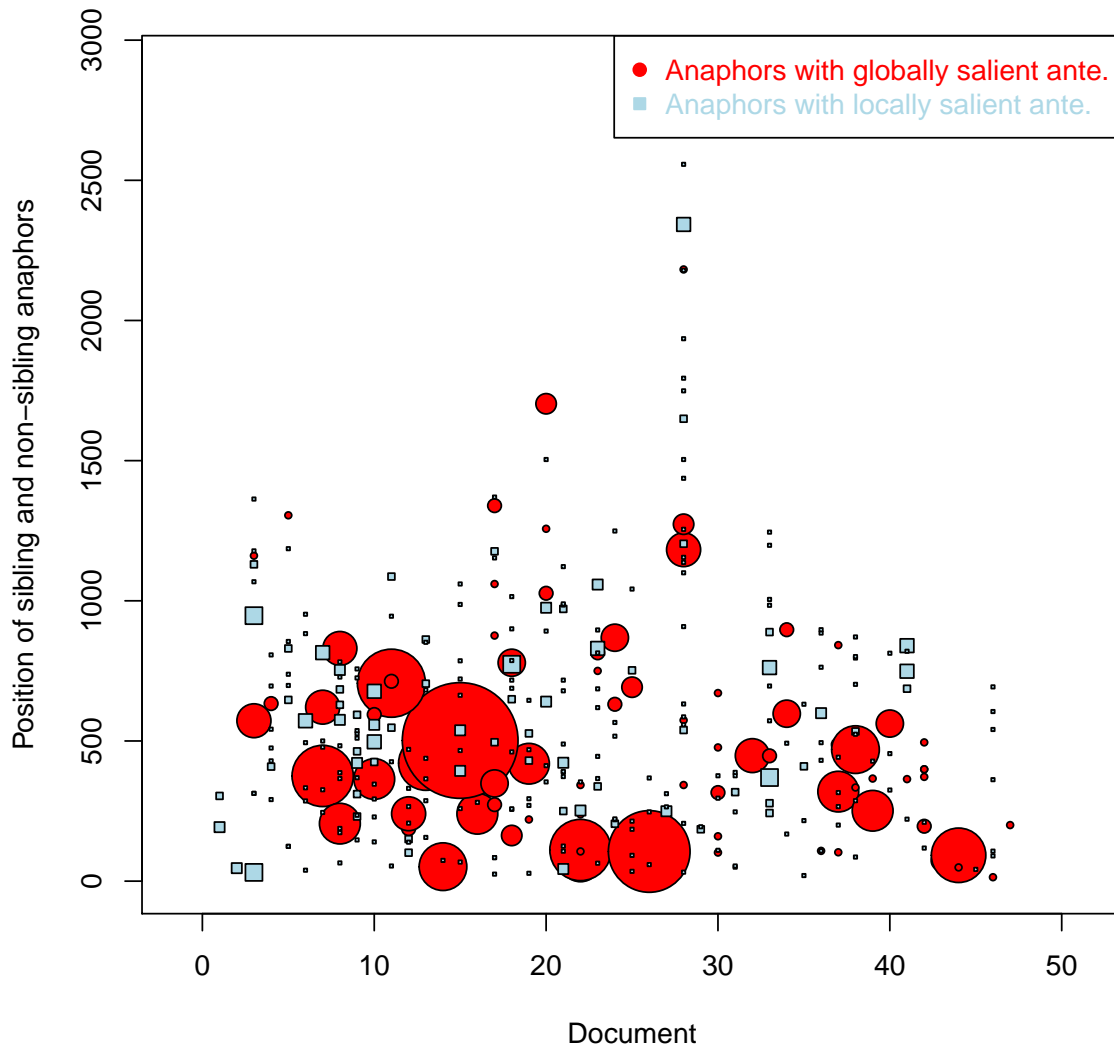
Figure 3.2: Scatter plot of sibling and non-sibling anaphors in terms of the linkage to glob-
ally or locally salient antecedents. The size of a circle or a square indicates the size of the
corresponding sibling anaphor cluster (from 2 to 17), or the size of a non-sibling anaphor (1).
The position of a non-sibling anaphor is the distance of its head word to the beginning of the
document measured in words. The position of a sibling anaphor cluster is the average dis-
tance of all anaphors within the cluster to the beginning of the document. The globally salient
antecedents are calculated under $r = 0.8$.

3.4 Corpus Analysis: Bridging Pair Distance

We define the distance between a bridging anaphor to its antecedent as the distance between the anaphor to its closest preceding antecedent instantiation. We measure the distances for 683 bridging pairs between 663 bridging anaphors and 387 antecedents[13]. Figure 3.3 shows the distribution of bridging pairs in terms of distance in sentences. Although Poesio et al. (2004a) claimed that bridging anaphora are sensitive to the *local* rather than the *global* focus (Grosz & Sidner, 1986), we find that bridging is a relatively local phenomenon in ISNotes: 76.92% of anaphors have antecedents occurring in the same or up to two sentences prior to the anaphor.
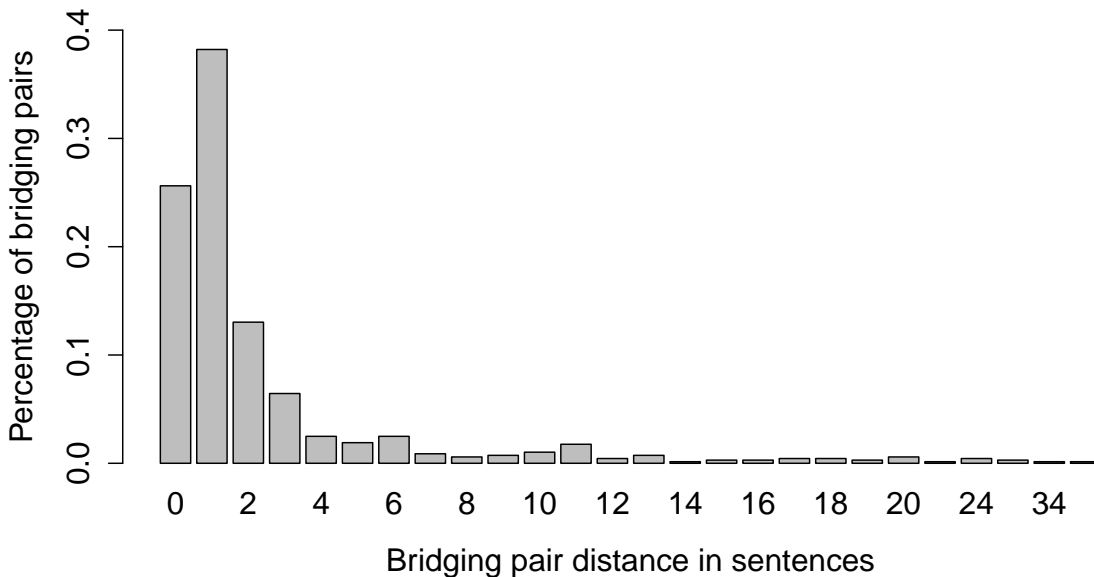

Figure 3.3: The distribution of bridging pairs w.r.t. distance in sentences.

Table 3.16 and Table 3.17 show the distribution of the distances of bridging pairs, measured in sentences and words, respectively. We notice that, in general, event antecedents are closer to their bridging anaphors compared to entity antecedents.

We then focus on bridging pairs with entity antecedents only. In the following, we examine the relations among the bridging pair distance, the topology of bridging anaphors (i.e., sibling anaphors and non-sibling anaphors), and the salience of bridging antecedents.

---

[13]As mentioned in Section 3.1, sometimes a bridging anaphor could have several antecedents when all of them are the missing mandatory roles of the bridging anaphor. Therefore the number of bridging pairs (683) is slighter more than the number of bridging anaphors (663).

Bridging Distance	Min	Max	Mean	Median	Standard Deviation
Bridging Pair with Entity Ante.	0.00	35.00	2.63	1.00	2.31
Bridging Pair with Event Ante.	0.00	5.00	1.28	1.00	4.55

Table 3.16: Distribution of the distances of bridging pairs measured by sentences.

Bridging Distance	Min	Max	Mean	Median	Standard Deviation
Bridging Pair with Entity Ante.	2.50	929.00	61.97	24.00	101.53
Bridging Pair with Event Ante.	5.00	95.50	27.31	20.00	22.57

Table 3.17: Distribution of the distances of bridging pairs measured by words, i.e., the number of words between the head word of the bridging anaphor and the head word (or the main verb) of its closest antecedent instantiation.

Bridging pair distance and the salience of bridging antecedents. We investigate the distribution of the distances of bridging pairs with regard to globally or locally salient antecedents. We use the same criteria with $r = 0.8$ from the previous section to define the salience of bridging antecedents, i.e., an entity antecedent is globally salient if it: (1) appears in the headline, or (2) has the document span ratio ≥ 0.8. Other entity antecedents are locally salient.

Bridging Distance	Min	Max	Mean	Median	Standard Deviation
Bridging Pairs with Globally Salient Ante.	0.00	34.00	3.26	1.00	5.10
Bridging Pairs with Locally Salient Ante.	0.00	35.00	2.22	1.00	4.13

Table 3.18: Distribution of the bridging pair distance (measured by sentences) between globally salient antecedents and locally salient antecedents.

Table 3.18 and Table 3.19 show that bridging pairs with globally salient antecedents tend to be more distant than those with locally salient antecedents. The average bridging pair distance with globally salient antecedents is 3.26 sentences or 75.42 words, whereas the average bridging pairs distance with locally salient antecedents is 2.22 sentences or 53.84 words.

Bridging Distance	Min	Max	Mean	Median	Standard Deviation
Bridging Pairs with Globally Salient Ante.	2.00	933.00	75.42	31.00	113.73
Bridging Pairs with Locally Salient Ante.	2.00	676.00	53.84	22.00	92.36

Table 3.19: Distribution of the bridging pair distance (measured by words) between globally salient antecedents and locally salient antecedents.

Bridging pair distance and the topology of bridging anaphors. We examine the distribution of the distances of bridging pairs in terms of the topology of the bridging anaphor, i.e., whether the bridging anaphor is a sibling anaphor or a non-sibling anaphor. The results in Table 3.20 and Table 3.21 show that bridging pairs with sibling anaphors are more distant than those with non-sibling anaphors. The average bridging pair distance with sibling anaphors is 3.19 sentences or 74.73 words, whereas the average bridging pairs distance with non-sibling anaphors is 1.53 sentences or 39.17 words.

Bridging Distance	Min	Max	Mean	Median	Standard Deviation
Bridging Pairs with Sibling Anaphors	0.00	35.00	3.19	1.00	5.07
Bridging Pairs with Non-sibling Anaphors	0.00	24.00	1.53	1.00	2.89

Table 3.20: Distribution of the bridging pair distance (measured by sentences) between sibling anaphors and non-sibling anaphors.

Bridging Distance	Min	Max	Mean	Median	Standard Deviation
Bridging Pairs with Sibling Anaphors	2.00	933.00	74.73	31.50	111.30
Bridging Pairs with Non-sibling Anaphors	2.00	676.00	39.17	19.00	76.10

Table 3.21: Distribution of the bridging pair distance (measured by words) between sibling anaphors and non-sibling anaphors.

Bridging pair distance, the salience of antecedents and the topology of bridging anaphors. In order to further understand the relations between bridging pair distance, the salience of antecedents and the topology of bridging anaphors, we fit a simple linear regression model

in R (R Development Core Team, 2011) on the data which contain the following three variables for each bridging pair[14]:

- *distance*: the distance of the bridging pair measured by words;

- *salience:* the salience of the antecedent (*global*[15] or *local*);

- *sibling:* the topology of the anaphor (*sibling* or *non-sibling*).

Table 3.22 shows the statistical results from R by applying a linear regression model for each pair of variables. We notice that both *salience* and *distance* are significantly correlated with *sibling* separately at the level of $p < 0.001$, *distance* and *salience* are significantly correlated at the level of $p < 0.01$.

variable pair	t-value	p-value
salience $\sim$ sibling	10.966	< 0.001
distance $\sim$ sibling	4.247	< 0.001
distance $\sim$ salience	2.636	0.00858

Table 3.22: The relations between bridging pair distance, the salience of antecedents and the topology of bridging anaphors.

The relations between these three variables are illustrated in a 3D scatter plot in Figure 3.4[16]. Figure 3.4 plots sibling and non-sibling anaphors in terms of: (1) the linkage to globally or locally salient antecedents; and (2) the distance between anaphor and antecedent measured in words. The distance between a sibling anaphor cluster and the shared antecedent is the average distance of all anaphors within the cluster to their corresponding antecedent instantiations. The size of a circle or a square indicates the size of the corresponding sibling anaphor cluster (from 2 to 17), or the size of a non-sibling anaphor (1). Figure 3.4 shows that **globally salient antecedents** (marked with red circles) **connect to a higher proportion of sibling anaphors and a lower proportion of non-sibling anaphors compared to locally salient antecedents** (marked with blue squares). Moreover, **bridging pairs with globally salient antecedents are more distant than those with locally salient antecedents, and bridging pairs with sibling anaphors are more distant than those with non-sibling anaphors**.

---

[14]We only consider bridging pairs with entity antecedents represented by mentions.

[15]The same criteria with $r = 0.8$ from the previous section is again used to define the salience of bridging antecedents.

[16]Figure 3.4 adds one more dimension (the distance between bridging anaphor and antecedent) on the basis of Figure 3.2 in Section 3.3.
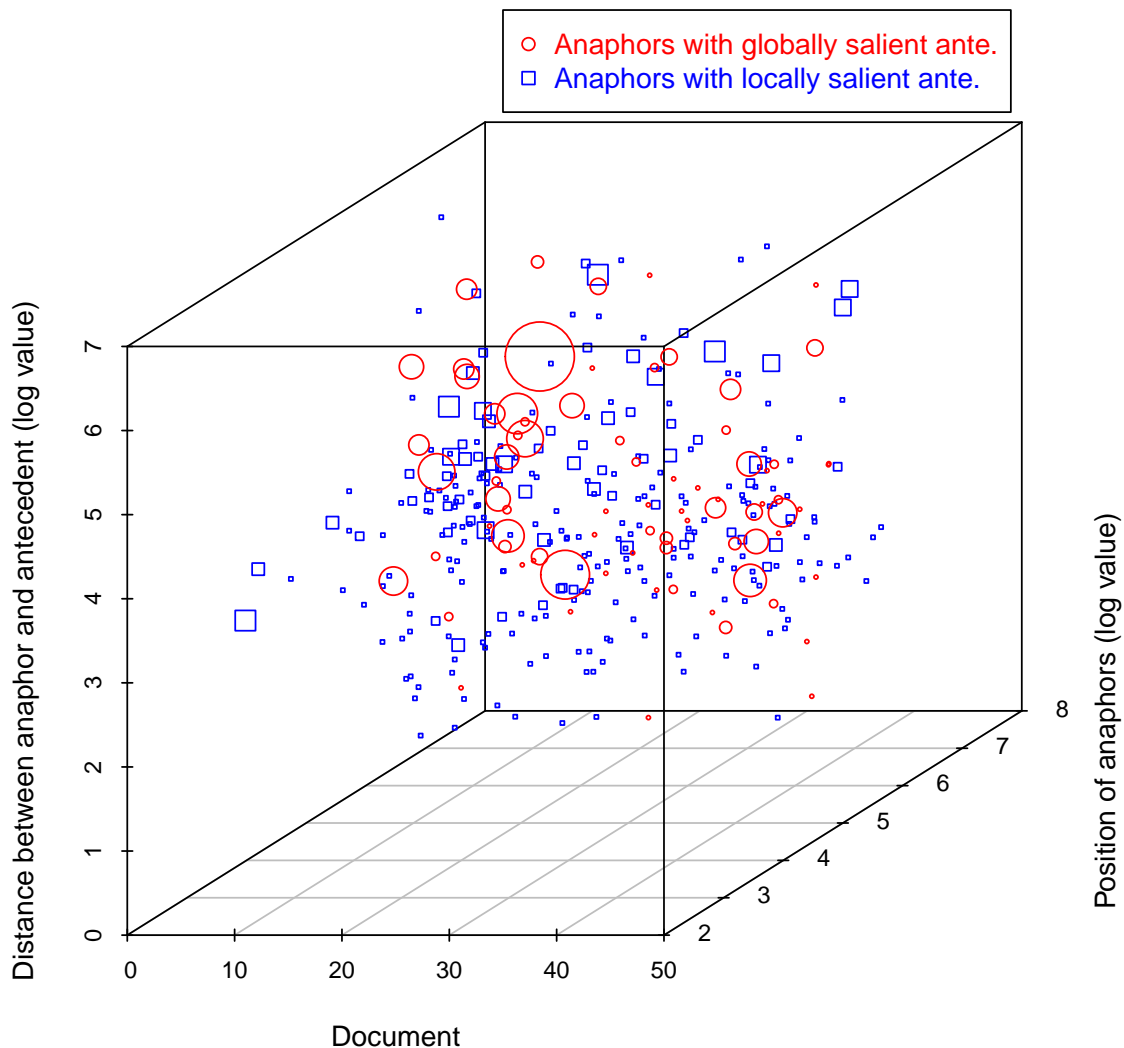
Figure 3.4: 3D scatter plot of sibling and non-sibling anaphors in terms of: (1) the linkage to globally or locally salient antecedents; and (2) the distance between anaphor and antecedent measured in words. The size of a circle or a square indicates the size of the corresponding sibling anaphor cluster (from 2 to 17), or the size of a non-sibling anaphor (1). The globally salient antecedents are calculated under $r = 0.8$ in Section 3.3. The size of a circle or a square indicates the size of the corresponding sibling anaphor cluster (from 2 to 17), or the size of a non-sibling anaphor (1). The position of a non-sibling anaphor in a document is the distance of its head word to the beginning of the document measured in words. The position of a sibling anaphor cluster is the average distance of all anaphors within the cluster to the beginning of the document. The distance between a sibling anaphor cluster and the shared antecedent is the average distance of all anaphors within the cluster to their corresponding antecedent instantiations.

3.5 Corpus Analysis: Bridging and Discourse Relations

The interaction between entity coherence and discourse relations is widely discussed in the literature (Hobbs, 1979; Hobbs et al., 1993; Asher & Lascarides, 1998; Knott et al., 2001; Cimiano, 2006; Kehler & Rohde, 2013). However, there is only little research on connections between bridging and discourse relations. Asher & Lascarides (1998) and Cimiano (2006) are exceptions to this. They focus on connections between bridging and discourse relations and model bridging by integrating discourse structure and semantics from a formal semantics viewpoint.

There are a variety of theories about discourse relations (Grimes, 1975; Longacre, 1976; Mann & Thompson, 1987; Webber & Joshi, 1998; Webber et al., 2003; Wolf & Gibson, 2005). Among all of them, the lexically-grounded approach (Webber & Joshi, 1998; Webber et al., 2003) is followed by the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), the biggest publicly available corpus annotated with discourse relations.

In this section, we first give a brief introduction to the PDTB in Section 3.5.1. We then present a corpus study of the interaction between bridging and discourse relations in Section 3.5.2. We focus on discourse relations annotated in the PDTB because it overlaps with the ISNotes corpus.

3.5.1 Penn Discourse Treebank

The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) is the largest annotated corpus of discourse relations. It contains 2,159 annotated texts from The Wall Street Journal, leading to a total of 40,600 annotated relations. Each discourse relation is assumed to hold between two arguments (Arg1 and Arg2).

In the PDTB, two types of relations are annotated. In *explicit relations*, a discourse connective such as "as a result", "because" or "and" is present (see Example 3.12). The arguments of explicit discourse relations can be anywhere in the text.

(3.12)   It continues to gain strength in the chamber <u>but</u> remains far short of the two-thirds
           majority required to prevail over Mr. Bush.

However, not all discourse relations are realized as explicit connectives. In Example 3.13, it is clear that the second sentence is the result of the first sentence. In the PDTB, such *implicit relations* are annotated only between adjacent sentences within a paragraph. Furthermore, three distinct labels are used for adjacent sentences where neither a discourse connective is present nor an implicit connective could be provided: *AltLex* for cases where relations are alternatively lexicalized by some non-connective expressions (Example 3.14); *EntRel* for cases only an entity-based coherence relations could be perceived between the adjacent sentences

(Example 3.15); finally, *NoRel* for cases where no discourse relations or entity-based relation could be perceived between the adjacent sentences.

(3.13)   By 1973, after their second child was born, it had become clear to Ms. Volokh and her husband Vladimir, a computer scientist, that they wanted to leave the U.S.S.R. *As a result*$_{implicit}$, Ms. Volokh quit her job, to remove herself from the public eye.

(3.14)   Other investors have lost millions in partnerships that bought thoroughbred race-horses or stallion breeding rights. *One big problem*$_{AltLex}$ has been the thoroughbred racehorse market.

(3.15)   The ambitious Warsaw project still awaits approval by city officials. Its developer is a Polish American, Sasha Muniak.

Apart from the argument structure of discourse relations, the PDTB provides sense annotations for *explicit*, *implicit* and *AltLex* relations. The sense tags are organized hierarchically with three levels: class, type and subtype. The top level has four classes representing four major semantic classes: `Temporal`, `Contingency`, `Comparison` and `Expansion`.

3.5.2   Interaction Between Bridging and Discourse Relations

Louis & Nenkova (2010) empirically assess the claims about the interaction between two types of local coherence, i.e., discourse relations and entity coherence based on coreference only. In a corpus consisting of 509 Wall Street Journal articles both annotated for coreference (from OntoNotes) and discourse relations (from the PDTB), they analyzed adjacent sentences within paragraphs and found that adjacent sentences connected by an `Expansion` relation are less likely to share entities compared to other discourse relations. They then speculate that bridging can be applied in some `Expansion` relations to create (local) entity coherence but cannot prove that due to the lack of reliable annotation for bridging. Since the texts annotated in ISNotes are also annotated in the PDTB, we combine annotations from these two corpora, leading to a overlap portion consisting of 48 texts[17].

In the following, we first calculate the distribution of discourse relations for this overlap portion, we then measure the co-occurrence between the bridging anaphors and the discourse relations from different perspectives. We define that a discourse relation co-occurs with a bridging anaphor if the second argument (Arg2) of the discourse relation contains the bridging anaphor[18]. We choose the second argument of a discourse relation because most discourse

---

[17]The whole ISNotes corpus consists of 50 texts. However, two short texts do not have bridging annotation.

[18]If a bridging anaphor co-occurs with two discourse relations where one's second argument is embedded in the other's second argument, only the one with the embedded second argument is considered to co-occur with the bridging anaphor.

relations appear in an Arg1-Arg2 order and such linear order is in line with "antecedent-bridging anaphor" structure.

Distribution of discourse relations.   The full PDTB contains 2,159 texts with 40,600 annotated relations. The portion of the PDTB that overlaps with the ISNotes corpus consists of 48 texts and 1,575 relations. The distribution of discourse relations in terms of relation types and class sense tags[19] for the overlap portion and the full PDTB are shown in Table 3.23 and Table 3.24 respectively. The distribution of discourse relations in the overlapping portion is broadly similar as in the full PTDB corpus: in Table 3.24, the distribution of discourse relations regarding class sense tags between the overlapping portion and the full PDTB is not significantly different at the level of $p < 0.01$ ($\chi^2 = 8.1039$, $df = 3$, $p - value = 0.04391$). However, we also notice that the distribution of discourse relations regarding the relation types in the overlapping portion (Table 3.23) is significantly different from in the full PDTB, with a slightly higher proportion of implicit relations ($\chi^2 = 21.9228$, $df = 4$, $p - value = 0.0002076$).

Relation Type	# in 48 bridging texts		# in full PDTB	
Explicit	667	(42.35%)	18,459	(45.47%)
Implicit	710	(45.08%)	16,053	(39.54%)
AltLex	25	(1.59%)	624	(1.5%)
EntRel	165	(10.48%)	5,210	(12.83%)
NoRel	8	(0.51%)	254	(0.63%)
Total	1,575	(100.00%)	40,600	(100.00%)

Table 3.23: Distribution of discourse relations w.r.t. discourse relation types in 48 texts from ISNotes.

Class Sense Tag	# in 48 bridging texts		# in full PDTB	
Temporal	180	(12.84%)	4,650	(12.71%)
Contingency	296	(21.11%)	8,042	(21.98%)
Comparison	285	(20.33%)	8,394	(22.94%)
Expansion	641	(45.72%)	1,5506	(42.38%)
Total	1,402	(100.00%)	3,6592	(100.00%)

Table 3.24: Distribution of discourse relations w.r.t. class sense tags in 48 texts from ISNotes.

---

[19]Sense tags are annotated for *explicit*, *implicit* and *AltLex* but not for *EntRel* and *NoRel*.

Bridging anaphora and discourse relations co-occurrence — discourse relations' perspective. In the overlap portion, 28.8% of discourse relations (454 out of 1,575) co-occur with bridging anaphors. For discourse relations whose class sense is annotated, 28.6% of them (401 out of 1402) co-occur with bridging anaphors. We examine the distribution of these discourse relations which co-occur with bridging anaphors with regard to discourse relation types and class sense tags respectively (see Table 3.25 and Table 3.26).

With regard to discourse relations types, we notice that the distribution of discourse relations which co-occur with bridging anaphors is significantly different from those which do not co-occur with bridging anaphors at the level of $p < 0.05$ (see Table 3.25). Among all relations types, *Implicit* and *EntRel* are likely to co-occur with bridging anaphora when compared to *Explicit*[20].

Relation Type	# Co-occurring with Bridging Anaphors		# Not Co-occurring with Bridging Anaphors	
Explicit	170	(37.44%)	497	(44.34%)
Implicit	219	(48.24%)	491	(43.80%)
AltLex	12	(2.64%)	13	(1.16%)
EntRel	51	(11.23%)	114	(10.17%)
NoRel	2	(0.44%)	6	(0.54%)
Total	454	(100.00%)	1,121	(100.00%)

Table 3.25: Distribution of discourse relations which co-occur with bridging anaphors w.r.t. discourse relation types.

In terms of class sense tags, the distribution of discourse relations which co-occur with bridging anaphors are significantly different from those which do not co-occur with bridging anaphors at the level of $p < 0.01$ (Table 3.26). We observe that discourse relations with `Expansion` sense tags are likely to co-occur with bridging anaphora compared to those with other sense tags.

In addition, we test Louis & Nenkova's assumption in the ISNotes corpus, i.e., bridging accounts for local coherence for some `Expansion` relations which do not share entities. Following Louis & Nenkova (2010), we collect pairs of adjacent sentences which are connected by one of the following discourse relations: `Temporal`, `Contingency`, `Comparison`, `Expansion` and `EntRel`. We also exclude a small number of cases (i.e., 15 cases) where a pair of adjacent sentences is connected by more than one discourse relation. Table 3.27 shows the total number of different discourse relations and the corresponding proportions of

---

[20]Although *AltLex* has the same trend, i.e., it seems likely to co-occur with bridging anaphora, it is unreliable to draw such conclusions given the number in this category in Table 3.25 is too small.

Class Sense Tag	# Co-occurring with Bridging Anaphors		# Not Co-occurring with Bridging Anaphors	
Temporal	38	(9.48%)	142	(14.19%)
Contingency	73	(18.20%)	223	(22.28%)
Comparison	77	(19.20%)	208	(20.78%)
Expansion	213	(53.12%)	428	(42.67%)
Total	401	(100.00%)	1,001	(100.00%)

Table 3.26: Distribution of discourse relations which co-occur with bridging anaphors w.r.t. class sense tags.

co-existing with "*coreference*" and "- *coreference + bridging*" respectively. "*coreference*" means adjacent sentences which share entities, whereas "- *coreference + bridging*" means adjacent sentences which do not share entities but contain a bridging relation across sentence boundaries (i.e., the antecedent is located in one sentence and the bridging anaphor is located in another sentence). Among all discourse relations, we find that adjacent sentences with Expansion relations are least likely to share entities and are most likely to co-exist with the type of "- *coreference + bridging*". It seems that the above observation supports Louis & Nenkova's assumption described previously. However, the occurrences of co-existence between some discourse relations and "- *coreference + bridging*" in Table 3.27 are rather few. Due to small sample size, it might not be possible to show statistical significance. A larger corpus containing annotations for different types of local coherence (e.g., discourse relations, coreference, bridging) may give us a better understanding of the phenomenon.

Class Sense Tag	# adjacent pairs	coreference		-coreference + bridging	
Core					
Temporal	50	35	(70.0%)	3	(6.0%)
Contingency	191	117	(61.3%)	11	(5.8%)
Comparison	155	101	(65.2%)	7	(4.5%)
Weak					
Expansion	422	214	(50.7%)	40	(9.5%)
EntRel	157	96	(61.1%)	4	(2.5%)

Table 3.27: Total number of different discourse relations and the corresponding proportions of co-existing with coreference and bridging.

Bridging anaphora and discourse relations co-occurrence — bridging anaphora's perspective. In total, 68.5% of bridging anaphors (454 out of 663) co-occur with discourse relations and 60.5% of bridging anaphors (401 out of 663) co-occur with discourse relations in which class senses are annotated.

To understand whether the co-occurrence of a bridging anaphor and a discourse relation has an effect on the distance between the bridging anaphor to its antecedent, we divide the bridging anaphors into two groups: *local* and *non-local*. We define a bridging anaphor as *local* if its antecedent appears in the same, or up to two sentences prior to the anaphor. We then examine the distribution of different groups of bridging anaphors (*local*, *non-local*) in terms of whether a bridging anaphor co-occurs with a discourse relation or not (*co-occurring with DR*, *not co-occurring with DR*). Table 3.28 shows that bridging anaphors co-occurring with discourse relations are likely to have close antecedents (i.e., *local* bridging anaphors) compared to those that do not co-occur with discourse relations. The distributions of *local* bridging anaphors is significantly different from the *non-local* ones at the level of $p < 0.01$ ($\chi^2 = 10.4186$, $df = 1$, $p - value = 0.001247$). The reason is that the discourse relation annotation in the PDTB mostly focuses on local coherence, and that most discourse relations are within one sentence or between the adjacent sentences[21].

Bridging Anaphora Type	# Co-occurring with DR		# Not co-occurring with DR	
Local Bridging Anaphora	366	(80.62%)	144	(68.90%)
Non-local Bridging Anaphora	88	(19.38%)	65	(31.10%)
Total	454	(100.00%)	209	(100.00%)

Table 3.28: Distribution of *local* and *non-local* bridging anaphors w.r.t. types of co-occurring with discourse relations.

The results in Table 3.28 indicate that *local* bridging anaphors are likely to co-occur with discourse relations. To further understand which class sense tags of discourse relations account more for this correlation, we divide *co-occurring with DR* in Table 3.28 into four different groups according to the class sense tag of a discourse relation. The results are shown in Table 3.29. We found that the distribution of local bridging anaphors concerning different (fine-grained) types of co-occurrence with discourse relations is not significantly different from those *non-local* ones ($\chi^2 = 2.0458$, $df = 3$, $p - value = 0.563$).

Although the existence of discourse relations does have an influence on the scope of bridging anaphora (i.e., *local* or *non-local*) (Table 3.28), it seems that there is no significant differ-

---

[21]In the PDTB, implicit relations are annotated between all successive pairs of sentences within paragraphs. Although explicit relations are not constrained with the locations of the arguments, 91% of them are within one sentence or between the adjacent sentences.

ence between different sense tags regarding the correlation with the scope of bridging anaphora (Table 3.29). However, given the relatively infrequent occurrence of some sense tags in Table 3.29, we think more data might give a better understanding of the problem.

Class Sense Tag	# Local Bridging Anaphors		# Non-local Bridging Anaphors	
`Temporal`	31	(9.69%)	7	(8.64%)
`Contingency`	60	(18.75%)	13	(16.05%)
`Comparison`	57	(17.81%)	20	(24.69%)
`Expansion`	172	(53.75%)	41	(50.62%)
Total	320	(100.00%)	81	(100.00%)

Table 3.29: Distribution of *local* and *non-local* bridging anaphors w.r.t. fine-grained types of co-occurring with discourse relations.

3.6   Summary

In this chapter, we have described the ISNotes corpus which is used throughout the thesis. The corpus contains around 11,000 NPs annotated for information status including 663 bridging anaphors and their antecedents in 50 texts taken from the WSJ portion of the OntoNotes corpus. ISNotes is reliably annotated for bridging: for bridging anaphor recognition, κ is over 60 for all three possible annotator pairings (κ is over 70 for two expert annotators); for selecting bridging antecedents, agreement is around 80% for all annotator pairings.

On the basis of this corpus, we have analyzed bridging phenomenon thoroughly from different perspectives. The results of these analyses as well as the linguistic knowledge of bridging from various theories will guide us to design computational models and features for identifying bridging anaphora and finding links to antecedents in the following chapters (i.e., Chapter 5, Chapter 6 and Chapter 7). Before diving into the problem of automatic bridging resolution, in the next chapter, we will describe computational methods as well as lexical semantic resources used for the problems that we address in this thesis.

Chapter 4

Methods and Resources

In this chapter we describe computational methods as well as lexical semantic resources used across the three problems that we address in this thesis (i.e., *bridging anaphora recognition*, *bridging anaphora resolution* and *bridging resolution*). These methods and resources are general, and have been used in a wide variety of problems in natural language processing. Since they appear frequently in the following three chapters (Chapter 5 – Chapter 7), we carve their descriptions out as subsections of this chapter. We first describe computational methods in Section 4.1: Section 4.1.1 describes the theory and algorithms of Markov logic networks; Section 4.1.2 explains the basics of support vector machines. We then describe lexical semantic resources in Section 4.2. Section 4.2.1 focuses on the principle and the calculation details of the Dunning root log-likelihood ratio. We also describe how we utilize this method to create a distributional semantic resource for bridging resolution. Finally, Section 4.2.2 and Section 4.2.3 briefly summarize information contained in two lexical semantic resources (WordNet and the General Inquirer lexicon) and explain how we use them for our problems.

4.1 Computational Methods

This section describes computational methods that we explore to address bridging resolution and its two subtasks (i.e., *bridging anaphora recognition* and *bridging anaphora resolution*). The linguistic knowledge about bridging from the empirical study in the previous chapter as well as from various theoretical studies should be brought to bear in designing model structures for bridging resolution. Therefore we utilize two machine learning methods (i.e., Markov logic networks and support vector machines) to model our target problems because they are able to "encode" our linguistic knowledge about the problems. We detail these two methods in Section 4.1.1 and Section 4.1.2 respectively.

4.1.1 Markov Logic Networks

Markov logic networks[1] (MLNs) (Domingos & Lowd, 2009) are a statistical relational learning framework that combines *first order logic* and *Markov networks*. They have been successfully applied in several NLP tasks such as semantic role labeling (Meza-Ruiz & Riedel, 2009), information extraction (Poon & Domingos, 2010) and coreference resolution (Poon & Domingos, 2008). Here we first compare Markov logic networks with other statistical relational learning approaches. We then focus on Markov logic networks and describe their representation, inference and learning algorithms, as well as implementations.

Markov logic networks and statistical relational learning. In contrast to traditional machine learning approaches (e.g., *decision trees*) which assume that the data is drawn independently and identically from some distribution (i.i.d.) and therefore mainly focus on "attribute-value" representation without considering relational aspects of the data, *statistical relational learning* (SRL) "attempts to represent, reason, and learn in domains with complex relational and rich probabilistic structure" (Getoor & Taskar, 2007, p.3). Intuitively, we would like to infer certain attributes of one data instance on the basis of its own other attributes and of other related data instances. This requires that all correlated data instances are inferred "collectively" rather than "separately".

In recent years, many approaches have been proposed in the field of statistical relational learning, e.g., probabilistic relational models (PRMs) (Getoor et al., 2001), relational Markov networks (RMNs) (Taskar et al., 2002), Markov logic networks (MLNs), and BLOG (Milch et al., 2007). Among many SRL approaches, we compare a few of them along several dimensions, i.e., representation, probabilistic semantics, learning, and inference (see Table 4.1). Table 4.1 shows that Markov logic networks are the most developed framework compared to other SRL approaches. First, Markov logic networks provide us with a simple yet flexible language to construct the appropriate models for bridging resolution. For instance, in Chapter 6, we use Markov logic networks to model that "sibling anaphors share the same antecedent". Moreover, our task specific models can benefit from the advances in inference and learning algorithms under this framework. In the remainder of this section, we turn our attention to Markov logic networks, describing their representation, inference and learning algorithms, as well as implementations.

Markov logic networks: representation. By combining first order logic and Markov networks, Markov logic networks can be seen as a template language which explores first order logic formulas to instantiate Markov networks. A *Markov network* or a *Markov random field*

---

[1]The terms "Markov logic" and "Markov logic networks" are both used in the literature to refer to the method. In this thesis, we use "Markov logic networks".

	Representation	Probabilistic semantics	Learning	Inference
PRMs	frame-based formalism	Bayesian networks	(1) parameter (2) structure	*belief propagation*
RMNs	logic formalism (SQL queries)	Markov networks	(1) parameter	*belief propagation*
MLNs	first order logic	Markov networks	(1) parameter (2) structure	sampling-based (e.g., *MC-SAT*, *lazySAT*) optimization-based (e.g., *MaxWalkSAT*, *CPI*)
BLOG	first order logic	Bayesian networks (allows unknown objects)	none to date	*MCMC*

Table 4.1: Comparison of different SRL approaches.

is an undirected probabilistic graphic model that represents the joint distribution over a set of variables $X = (X_1, X_2, ..., X_n) \in \mathcal{X}$. The graph has a node for each variable, and each clique in the graph is associated with a potential function ϕ. A potential function is a non-negative real-valued function of the state of the corresponding clique. The joint distribution represented by a Markov network is given by

$$P(X = x) = \frac{1}{Z} \prod_k \phi_k(x_{\{k\}}), \tag{4.1}$$

where $x_{\{k\}}$ is the state of the kth clique and Z is the partition function, which is given by $Z = \sum_{x \in \mathcal{X}} \prod_k \phi_k(x_{\{k\}})$. By replacing each clique potential as an exponentiated weighted sum of features of the state, we can represent Markov networks as *log-linear models*:

$$P(X = x) = \frac{1}{Z} \exp \left(\sum_i w_i f_i(x) \right), \tag{4.2}$$

where each feature $f_i(x)$ corresponds to each possible state $x_{\{k\}}$ of each clique in Equation 4.1, and the weight w_i of the feature $f_i(x)$ is $\log \phi_k(x_{\{k\}})$. Here we consider binary features, i.e., $f_i(x) \in \{0, 1\}$.

We now use a concrete example to explain how first order logic formulas are constructed and how the corresponding Markov network is instantiated in Markov logic networks. Suppose we need to predict information status for mentions, on the basis of the attributes of each mention (e.g., whether the mention has the same head with a previous mention or whether the mention is a pronoun) and of the relations between mentions (e.g., whether a syntactic parent-child relation is held between two mentions). For instance, in Example 4.3, we want to

predict the information statuses for the two mentions "*his*" and "*his uncle*" are `old` and `mediated/syntactic`, given the evidence that "*his*" is a pronoun, "*his uncle*" has the same head word with a previous mention, and "*his uncle*" is a syntactic parent of "*his*".

(4.3) $[[\text{his}]_{old} \text{ uncle}]_{mediated/syntactic}$

In first order logic, formulas are constructed using four types of symbols: *constants, variables, functions* and *predicates*. Constant symbols represent objects that we are interested in (e.g., mentions in our running example, such as *his uncle* or *his*), variable symbols range over objects in the domain (e.g., m), function symbols map objects to objects (e.g., SyntacticParentOf), predicate symbols represent relations among objects (e.g., SyntacticParentChild) or attributes of objects (e.g., IsHeadMatch or IsPronoun). A *term* is a variable or constant representing an object in the domain, such as *his uncle* or m. An *atomic formula* or *atom* is a predicate symbol applied to a tuple of terms, such as IsHeadMatch (m) or SyntacticParentChild (*his uncle*, *his*). Formulas are recursively constructed from atomic formulas using logic connectives (i.e., $\neg$, $\wedge$, $\vee$, $\Rightarrow$, $\Leftrightarrow$) and quantifiers (i.e., $\exists$, $\forall$). The first column in Table 4.2 shows three example formulas. $f1$ specifies that if a mention m has the same head as a previous mention, then its information status is `old`. $f2$ states that if a mention m is a pronoun, then its information status is `old`. $f3$ says that if two mentions m_1 and m_2 have a syntactic parent-child relation and m_2's information status is `old`, then the information status for m_1 is `mediated/syntactic`. A *ground term* is a term which contains only constants. A *ground atom* or *ground predicate* is an atom whose arguments are all ground terms. A *ground formula* is a formula that contains only ground atoms. A *possible world* assigns a truth value to each possible ground atom. We say that a possible world x satisfies a formula f if f is true in x. For instance, the possible world {IsPronoun (*his*), IsOld (*his*)} satisfies $f2$, the possible world {IsHeadMatch (*his uncle*), $\neg$ IsOld (*his uncle*)} does not satisfy $f1$.

A *first-order knowledge base* (KB) is a set of formulas in first order logic. It describes possible worlds that satisfy all formulas. Such hard constrains are softened in Markov logic networks: each possible world is mapped to a probability instead of checking whether it satisfies all formulas or not. Therefore Markov logic networks allows worlds which do not satisfy all formulas. To do so, each formula in the first-order KB is associated with a weight which reflects how strong the corresponding constraint is. For instance, the second column in Table 4.2 tells that a world that satisfies $f2$ is more probable than a world that satisfies $f1$.

A Markov logic network is defined as a set of pairs (f_i, w_i), where f_i is a formula in first-order logic and w_i is a real number (Domingos & Lowd, 2009). Together with a finite set of constants $C = \{c_1, c_2, \ldots, c_{|C|}\}$, it defines a Markov network $M_{L,C}$ by creating one binary variable for each possible ground predicate and adding one feature for each possible ground formula. The value of the binary variable is 1 if the corresponding ground predicate is true, and 0 otherwise. Similarly, the value of the feature is 1 if the corresponding ground formula

Formula	Weight
$f1$ $\forall m$ IsHeadMatch $(m) \Rightarrow$ IsOld (m)	7.2
$f2$ $\forall m$ IsPronoun $(m) \Rightarrow$ IsOld (m)	30.0
$f3$ $\forall m_1 \forall m_2$ (SyntacticParentChild $(m_1, m_2) \wedge$ IsOld $(m_2))$ $\Rightarrow$ IsMediatedSyntactic (m_1)	8.5

<div align="center">Table 4.2: An MLN example.</div>

is true, and 0 otherwise. The weight of a feature is the weight of the first-order formula that instantiates it. Such a network is called a *ground Markov network*. The probability of a world x specified by the ground Markov network $\mathbf{M}_{L,C}$ is defined as a *log-linear model*:

$$P(X = x) = \frac{1}{Z} \exp \left(\sum_i w_i n_i(x) \right), \qquad (4.4)$$

where $n_i(x)$ is the number of true groundings of f_i in x. The normalization factor Z is the partition function:

$$Z = \sum_x \sum_i w_i n_i(x). \qquad (4.5)$$

Figure 4.1 shows the graphical structure of a ground Markov network by applying the MLN defined in Table 4.2 to the constants {"*his uncle*", "*his*"}. Different colors indicate the cliques in the network originated from different formulas.

This ground Markov network can be used to infer the information statuses for the mentions (i.e., the values for the hidden predicates or the query atoms, such as IsMediatedSyntactic("*his uncle*") = ?), given the following observed values for other ground predicates (the evidence):

- SyntacticParentChild ("*his uncle*", "*his*") = true

- SyntacticParentChild ("*his*", "*his uncle*") = false

- IsHeadMatch ("*his*") = false

- IsHeadMatch ("*his uncle*") = true

- IsPronoun ("*his*") = true
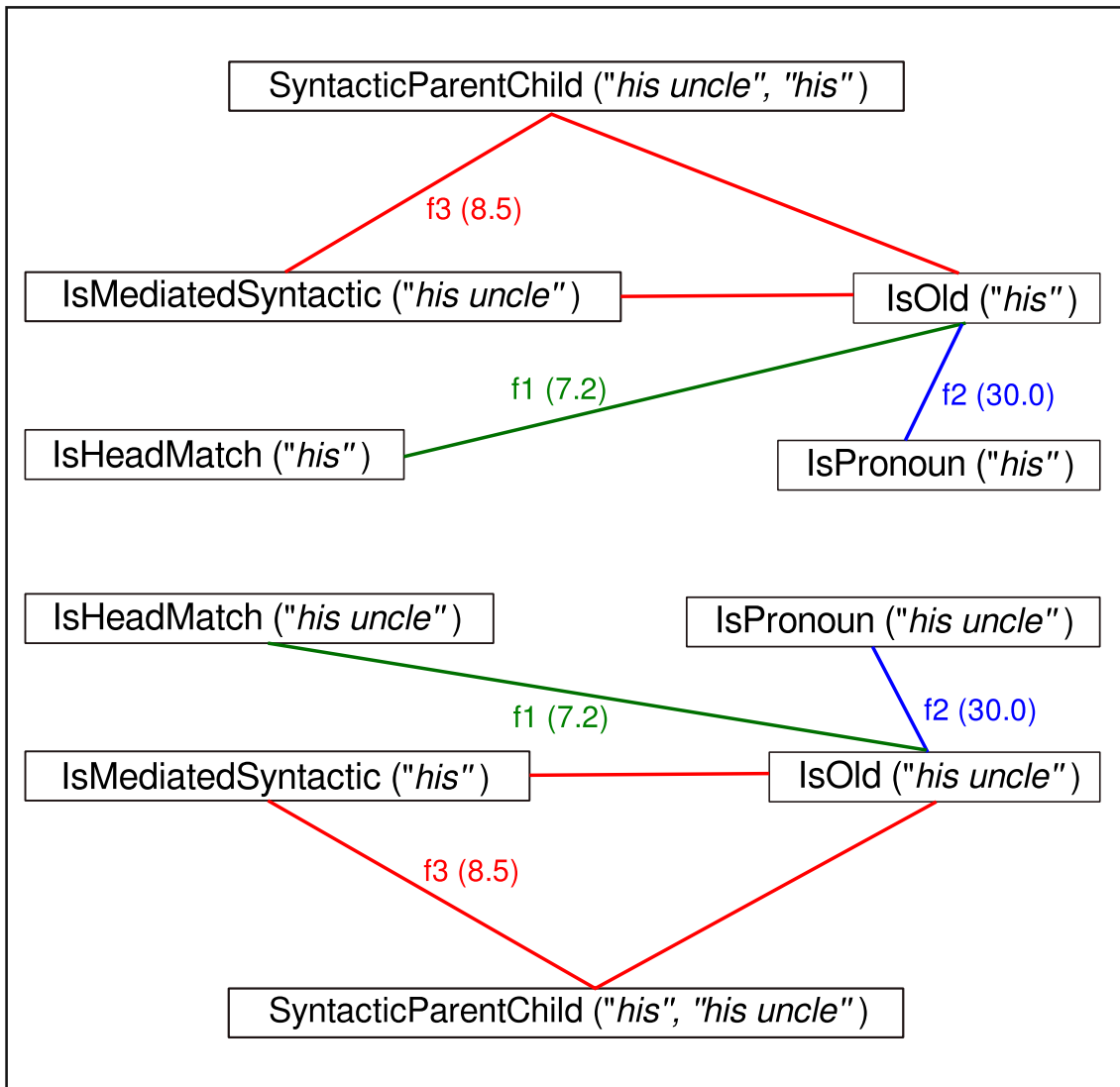
- IsPronoun ("*his uncle*") = false

Figure 4.1: The ground Markov network obtained by instantiating the three formulas in Table 4.2 to the constants {"*his uncle*", "*his*"}.

Figure 4.2 shows two states of the ground Markov networks (Figure 4.1) respectively. The clique with solid line indicates its potential is e^{w_i}, with w_i being the weight of the corresponding formula. The clique with dashed line indicates its potential is e^0. These two states only differ in the values of hidden predicates, i.e., the information statuses for "*his uncle*" and "*his*" are `mediated/syntactic` and `old` respectively in *state1*, whereas both mentions are `old` in *state2*. It is easy to see that *state1* is more probable than *state2* according to Equation 4.4.
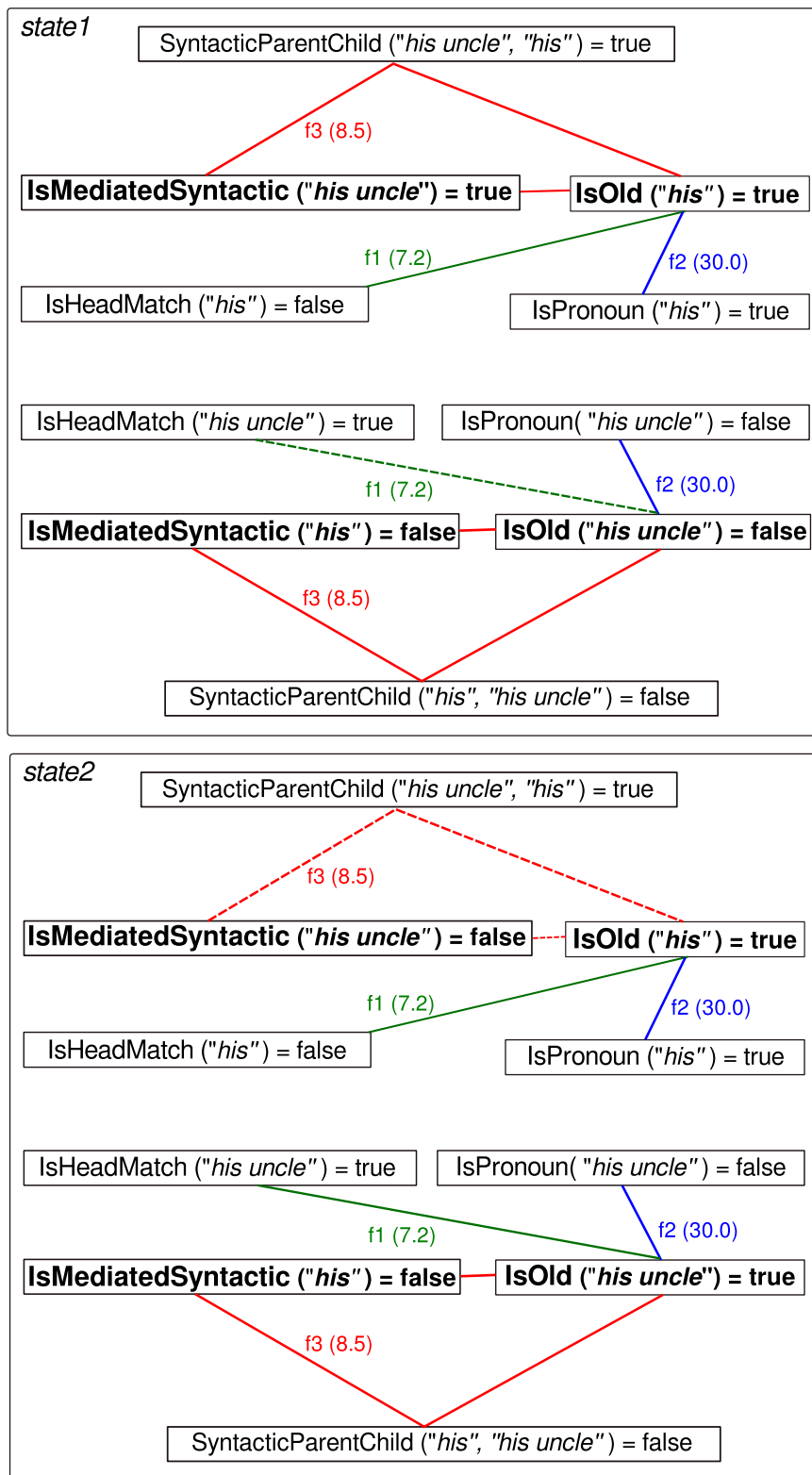
Figure 4.2: Two possible states of the Ground Markov network in Figure 4.1. Bold font indicates query atoms.

Markov logic networks: inference. There are two types of inference in Markov logic networks: marginal/conditional inference as well as *Maximum A Posteriori* (MAP) inference. The former gives the probabilities of the query atoms (hidden predicates) to be true over all possible events or the probabilities of the query atoms to be true given groundings of other formulas as evidence, while the latter finds the *most probable assignment* to the query atoms given the evidence[2]. As in Markov networks, inference in Markov logic networks is intractable. Therefore different approximate inference methods for MLNs are proposed, such as *lifted belief propagation* (Singla & Domingos, 2008), *MC-SAT* (Poon & Domingos, 2006), *MaxWalkSAT* (Kautz et al., 1996), and *cutting plane inference* (CPI) (Riedel, 2008). Here we briefly describe CPI that we use in this thesis.

CPI incrementally instantiates fractions of the complete ground Markov network. It performs as a meta method that uses another solver (base solver) to solve the actual partial problems. CPI proceeds as below:

1. Initially use a base solver to solve a partial grounding G^0. Normally G^0 consists of all groundings of formulas that only contain one hidden predicate. In this case finding a solution which maximizes G^0 is easy because the hidden predicates do not interact.

2. In each iteration i, we maintain the corresponding partial grounding G^i. First, find a solution y that maximizes the partial grounding G^i by exploring a base solver. Then find all ground formulas F that do not maximally contribute to the total probability given the current solution y and add them to the current partial grounding. If no more new ground formulas are found or a maximum number of iterations is reached, the algorithm is terminated and the solution with the highest score is returned.

3. Resolve the new partial grounding and return to step 2.

Riedel (2008) proves that when CPI returns the solution of iteration i, the error is bound by the sum of the error of the base solver on the partial problems and the sum of absolute weights of newly found ground formulas at this iteration. Therefore the performance of CPI is dependent on the performance of the base solver. Riedel (2008) also empirically shows that in

---

[2]Strictly, MAP inference in this context means *most probable explanation* (MPE) inference. MPE inference tries to find the most likely assignment to all of the non-evidence variables. That is, in a full joint probability distribution over Ω, given the evidence (i.e., a subset E of random variables in the network and an instantiation e to these variables) and the non-evidence variables $W = \Omega - E$, MPE inference finds the most likely assignment to the variables in W given the evidence $E = e$: $\arg\max_w P(w, e)$. MAP inference is more general than MPE inference. It tries to find the most likely assignment to the variables in Y given the evidence $E = e$. If we let $Z = \Omega - Y - E$, MAP inference is to compute: $\arg\max_y \sum_z P(y, z|e)$. It is clear that the "true" MAP inference contains both a conditional probability query and an MPE query. However, in NLP literature, the term MAP is often used to mean MPE.

two real world tasks, i.e., semantic role labeling and the joint entity resolution, CPI enables the approximate method (e.g., *MaxWalkSAT*) to be faster and more accurate. Moreover, it makes an exact method (e.g., *integer linear programming*) more efficient while remaining exact.

Markov logic networks: learning. Learning in Markov logic networks includes *weight learning* as well as *structure learning*. The former is learning weights for the formulas, while the latter is learning (additional) formulas and the corresponding weights all together from the data. Structure learning in MLNs often applies inductive logic programming (ILP) techniques to construct formulas by optimizing a likelihood-type measure (Kok & Domingos, 2005; Huynh & Mooney, 2008). In this thesis we only explore weight learning given that all formulas are designed manually on the basis of the knowledge about the problems. We now give a brief overview of different methods proposed previously to learn weights for MLNs.

In *supervised learning* where the non-evidence atoms are observed in the training set, MLNs weights can be learned *generatively* or *discriminatively*. In generative training, we try to maximize the likelihood of the training data (Equation 4.4). This requires us to calculate the gradient of the log-likelihood with respect to the weights:

$$\frac{\partial}{\partial w_i} P_w(X = x) = n_i(x) - \sum_{x'} P_w(X = x') n_i(x'), \qquad (4.6)$$

where the sum is over all possible data instances x', and $P_w(X = x')$ is $P(X = x')$ computed using the current weight vector w. Formula 4.6 can be understood as "that the ith component of the gradient is the difference between the number of true groundings of the ith formula in the data and its expectation according to the current model". However, computing these expectations requires inference over the model, which can be very expensive. Also most fast numeric optimization methods (e.g., conjugate gradient with line search) require computing the likelihood itself and hence the partition function Z, which is intractable. Instead of optimizing the log-likelihood, Richardson & Domingos (2006) propose to maximize the pseudo-likelihood:

$$P_w(X = x) = \prod_{l=1}^{n} P_w(X_l = x_l | MB_x(X_l)), \qquad (4.7)$$

where $MB_x(X_l)$ is the state of the Markov blanket[3] of X_l in the data. Computing the pseudo-likelihood and its gradient does not require inference, and therefore is efficient.

However, Singla & Domingos (2005) show that training MLNs generatively using pseudo-likelihood could lead to poor results when inference across non-neighboring variables is required. Instead, they propose *discriminative* training for MLNs by optimizing the *conditional*

---

[3]The Markov blanket of a node is the minimal set of nodes that renders it independent of the remaining network. In a Markov network, this is the node's neighbors in the graph.

likelihood of Y given X:

$$P(Y = y | X = x) = \frac{1}{Z_x} \exp \left(\sum_{i \in F_y} w_i n_i(x, y) \right), \tag{4.8}$$

where X is a set of evidence atoms, Y is a set of query atoms, F_y is the set of all MLN clauses[4] with at least one grounding involving a query atom, $n_i(x, y)$ is the number of true groundings of the ith clause involving query atoms, Z_x is the partition function by summing all x out.

Singla & Domingos (2005) then apply Collins's *voted perceptron* algorithm (Collins, 2002) to learn the weight vector w: weights are initialized to the corresponding clauses' log odds of being true in the data, then in each step t the weight vector w is updated as below:

$$w_{i,t} = w_{i,t-1} + \eta g, \tag{4.9}$$

where η is a learning rate, g is the gradient. Similar as before, the derivative of the conditional log-likelihood with respect to a weight is the difference between the number of the true groundings of the corresponding clause in the data, and its expected counts according to the current model:

$$\frac{\partial}{\partial w_i} P_w(y|x) = n_i(x, y) - \sum_{y'} P_w(y'|x) n_i(x, y')$$
$$= n_i(x, y) - E_w[n_i(x, y)] \tag{4.10}$$

Computing the expected counts $E_w[n_i(x, y)]$ is intractable. However, they can be approximated by the counts in the most probable explanation (MPE) state. This is often called as *Maximum A Posteriori* (MAP) inference. *MaxWalkSat* is explored by Singla & Domingos (2005) to approximate the MAP inference.

Since *MaxWalkSat* is not guaranteed to find the MPE state, Lowd & Domingos (2007) explore several alternatives, including contrastive divergence, per-weight learning rates, diagonal Newton, and scaled conjugate gradient. They find that scaled conjugate gradient is the best-performing method in experiments on standard statistical relational learning datasets.

In *unsupervised learning* or *semi-supervised learning* where the non-evidence atoms are not observed or partially observed in the training set, MLNs weights can be learned by applying the *expectation maximization* (EM) algorithm (Dempster et al., 1977) or using gradient descent (Poon & Domingos, 2008). However, unlike in the supervised learning scenario, the

---

[4]For automated inference, it is often convenient to convert formulas to a clausal form (also known as *conjunctive normal form* (CNF)). A clause is a disjunction of atomic formulas. For instance, the clausal form of $f2$ $\forall m$ IsPronoun $(m) \Rightarrow$ IsOld (m) in Table 4.2 is: $\neg$ IsPronoun $(m) \lor$ IsOld (m).

conditional likelihood objective function in the unsupervised or semi-supervised learning scenarios is not convex anymore, so the above mentioned algorithms can only guarantee to find a local optimum. Therefore initialization is an important step, so that the local optimum found by the algorithm is an improvement over an already reasonable starting point.

Markov logic networks: implementations.   There are several open-source software packages that implement Markov logic networks. We summarize some of them from several dimensions, i.e., weight learning (supervised and unsupervised/semi-supervised), structure learning, MAP inference, and probability inference. Table 4.3 shows that Alchemy is the most complete software implementation for existing algorithms of Markov logic networks. However, we use *thebeast* in this thesis for our experiments because it implements *cutting plane inference* which makes an exact inference method (i.e., *integer linear programming*) more efficient while remaining exact.

	Learning			Inference	
	Weight		**Structure**	**MAP**	**Probability**
	Supervised	**Unsupervised** **Semi-supervised**			
Alchemy	generative training: *pseudo-likelihood* discriminative training: (1) *voted perceptron* (2) *diagonal Newton* (3) *scaled conjugate gradient*	*EM*	(1) *hypergraph lifting* (2) *inductive logic programming* (3) *structural motifs*	(1) *MaxWalkSat* (2) *LazySAT*	(2) *lifted belief propagation* (2) *MC-SAT* (3) *Gibbs sampling*
thebeast	online discriminative training: (1) *MIRA*, (2) *perceptron* (3) *passive aggressive*	–	–	*CPI*	–
RockIt	online discriminative training: *voted perceptron*	–	–	(1) *CPI* (2) *CPA*	*Gibbs sampling*
Tuffy	discriminative training: *diagonal Newton*	–	–	*WalkSAT*	*MC-SAT*

Table 4.3: Comparison of different software packages that implement MLNs.

4.1.2 Support Vector Machines

Support vector machines (SVMs) are a theoretically well-founded machine learning method introduced by Vapnik (1995; 1998) for classification and regression. It is successfully applied in many applications across different domains, e.g., isolated handwritten digit recognition (Cortes & Vapnik, 1995), text classification (Joachims, 1998), and face detection in images (Osuna et al., 1997). In Chapter 5 of this thesis, we explore SVMs to identify minority categories (e.g., *bridging*) of information status in a cascading collective classification model because SVMs have the advantage of dealing well with high-dimensional lexical features. Moreover, weighted SVMs can handle the problem of learning from imbalanced datasets in which negative instances heavily outnumber the positive instances. For example, in bridging anaphora recognition, bridging is a relatively rare category compared to many other information status classes.

In this section, we first explain the principle of SVMs from two perspectives, i.e., the regularization framework of *loss + regularization* and the geometric approach. We then briefly describe the kernel functions in SVMs, the traditional ways of conducting multi-class classification with binary SVMs, and weighted SVMs for imbalanced classification,

Linear binary SVMs: a regularization view. Given n training examples $\{x_i, y_i\}$, where $i = 1, ..., n$, $x_i \in \mathbb{R}^d$, $y_i \in \{1, -1\}$, we want to learn a classifier $f(x) = \text{sign}(w \cdot x + b)$, so that it has the best performance on the training data as well as on other unseen data. SVMs achieve this by optimizing the following function:

$$\min_{w,b} (\overbrace{C \underbrace{\sum_{i=1}^{n} \max\{0, 1 - y_i f(x_i)\}}_{\text{training error (empirical risk)}} + \underbrace{\frac{1}{2}||w||^2}_{\text{complexity term}})}^{\text{structural risk minimization}}, \tag{4.11}$$

where the first term is the hinge loss function (see Figure 4.3) which makes the model aware of the error (and how severe the error is), and the second term is the L_2 regularization term which lets the model avoid large magnitudes in w. C is a constant ($C > 0$) which controls the trade-off between these two terms, i.e., achieving a low error on the training data and minimizing the norm of the weights. Higher C corresponds to letting the model put more effort on classifying all the training data correctly. However, this could cause overfitting as the model has higher complexity. On the contrary, lower C results in a more simple model with lower complexity. In this case, the training error may increase as the model allows some of the data to be misclassified.

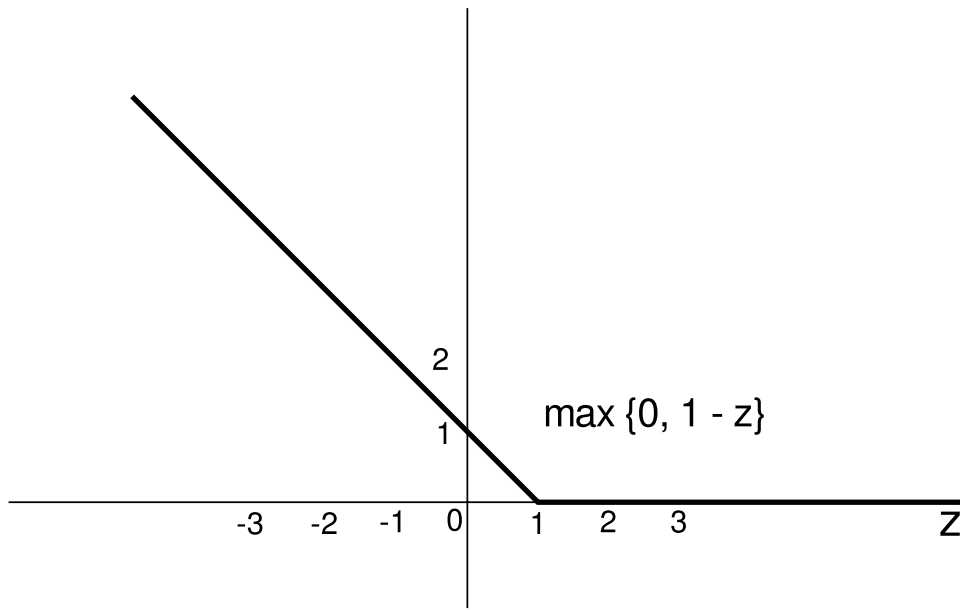The optimization formulation in Formula 4.11 is also known as the primal problem of

Figure 4.3: The hinge function decreases linearly for $z < 1$ and but remains 0 for $z \geq 1$.

SVMs. This problem can be converted into an equivalent dual problem[5] using Lagrange multipliers:

$$\max_{\alpha} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i \cdot x_j \rangle \right)$$
$$s.t. \ \ 0 \leq \alpha_i \leq C, i = 1, 2, ...n \qquad (4.12)$$
$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

where α is the vector of n non-negative Lagrange multipliers to be determined. This maximization problem is known as a *quadratic programming* (QP) problem. Once the solution of problem 4.12 is obtained, w can be calculated as follows:

$$w = \sum_{i} \alpha_i y_i x_i \qquad (4.13)$$

The above formula (Formula 4.13) means the vector w is just a linear combination of the training examples. Notice that there is a Lagrange multiplier α_i for every training instance. In

---

[5]In mathematical optimization theory, optimization problems may be viewed from either of two perspectives: the primal problem or the dual problem. When the problem is convex and satisfies a constraint qualification, the value of an optimal solution of the primal problem is given by the dual problem.

the optimal solution, only the training instances for which $\alpha_i > 0$ are called "support vectors", all other training instances have $\alpha_i = 0$ and do not influence the decisions for the new test instances. Intuitively, the support vectors are the "borderline cases" in the decision function that we try to learn. The value of α_i can be thought of a measure which reflects how important the example x_i is in determining the classification boundary.

In order to compute b, we collect the set of support vectors S by finding all x_i such that $0 < \alpha_i \leq C$, we then calculate b as follows:

$$b = \frac{1}{|S|} \sum_{s \in S} (y_s - w \cdot x_s) \qquad (4.14)$$

Many algorithms have been developed for training SVMs, such as *gradient descent* and *quadratic programming*. In the test stage, in order to predict the class tag for the new instance x', we substitute w in Formula 4.13 back to the decision function $f(x) = \text{sign}(w \cdot x + b)$, accordingly we get:

$$f(x') = \text{sign}\left(\sum_i \alpha_i y_i \langle x_i \cdot x' \rangle + b\right) \qquad (4.15)$$

Formula 4.15 tells us that the value of the decision function depends only on the support vectors and their Lagrange multipliers, and the optimum value for b. So we do not need to compute w explicitly in order to predict the class tag for the new instance.

Linear binary SVMs: a geometric view. Again suppose we have n training examples $\{x_i, y_i\}$, where $i = 1, ..., n$, $x_i \in \mathbb{R}^d$, $y_i \in \{1, -1\}$. Assume we have some hyperplanes which separate the positive examples from the negative ones (i.e., *separating hyperplanes*). The points x which lie on the hyperplane satisfy $w \cdot x + b = 0$, where $|b|/||w||$ is the perpendicular distance from the hyperplane to the origin, and $||w||$ is the Euclidean norm of w. Let d_+ (d_-) be the shortest distance from the separating hyperplane to the closest positive (negative) example and define the "margin" of a separating hyperplane to be $d_+ + d_-$, the goal of an SVM is to select a separating hyperplane with the largest margin.

Intuitively, choosing the separating hyperplane with the largest margin will give us better chance to correctly predict data points which are not in the training set but are very close to the hyperplane. On the contrary, in a hyperplane that separates the training points correctly but allows some points to be very close to the hyperplane itself, these is a fair chance that a new point which is close to the hyperplane would be misclassified (see Figure 4.4 for the illustration).

To derive the optimization function, suppose all the training examples satisfy the following constraints:

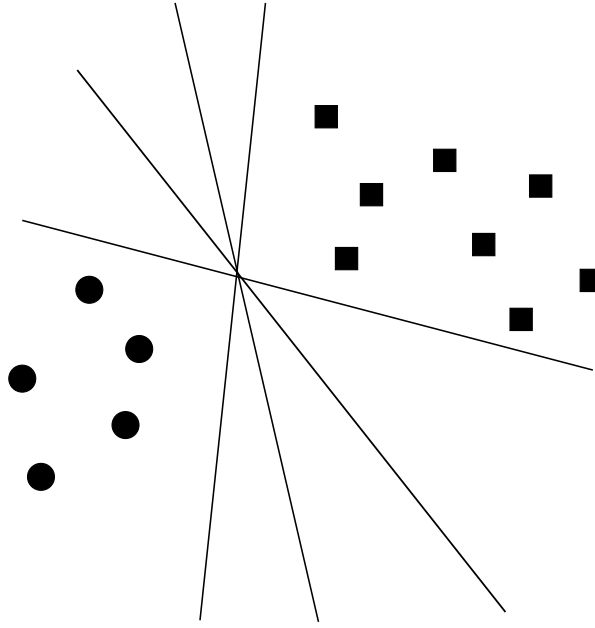$$y_i(x_i \cdot w + b) \geq 1 \ \ \forall i \qquad (4.16)$$

Figure 4.4: Hyperplanes which separate the classes.

So the points for which the equality in Formula 4.16 holds lie on the hyperplanes $H1 : x_i \cdot w + b = 1$ or $H2 : x_i \cdot w + b = -1$. The perpendicular distance of $H1$ and $H2$ to the origin are $|b - 1|/||w||$ and $|b + 1|/||w||$ respectively. Hence $d_+ = d_- = 1/||w||$ and the margin is $2/||w||$. Therefore we can find the pair of hyperplanes which gives the maximum margin by minimizing $||w||^2$, subject to constraints 4.16 (see Figure 4.5 for the illustration).

The above algorithm could be extended to the non-separable cases by relaxing the constraints 4.16 when necessary (Figure 4.6). Specifically, we introduce positive slack variables ξ_i, $i = 1, \ldots, n$ in the constraints to allow "outliers", i.e., ξ_i is a penalty for misclassified data. Hence we get the new constraints as below:

$$y_i(x_i \cdot w + b) \geq 1 - \xi_i \ \ \forall i$$
$$\xi_i > 0 \ \ \forall i \tag{4.17}$$

With the new constraints 4.17, we change the optimization function as:

$$\min \ \left(\frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \xi_i \right) \tag{4.18}$$

Intuitively, the first term in Formula 4.18 corresponds to finding a separating hyperplane with a margin as large as possible, so that the hyperplane can guarantee good prediction performance. The second term of Formula 4.18 tries to minimize the number of classification errors.
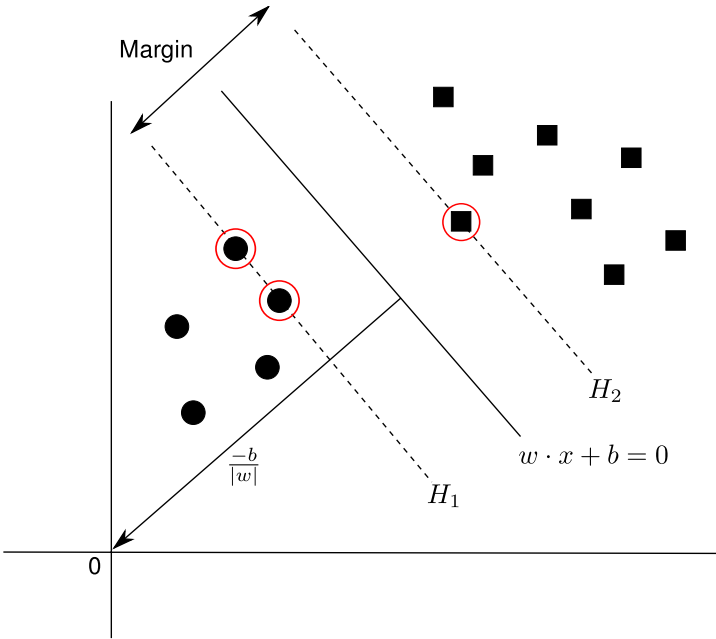
Figure 4.5: An SVM selects the hyperplane with the largest possible margin. Support vectors are marked with red circles.
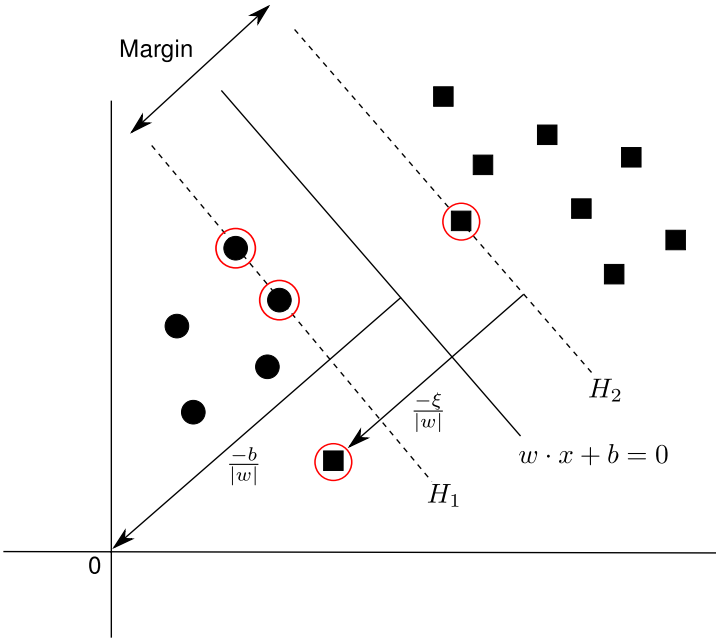

Figure 4.6: Linear hyperplane through two non-linearly separable classes.

These two terms are combined with a parameter C. A larger C assigns a higher penalty to classification errors, a small C maximizes the margin so that the optimum separating hyperplane is less sensitive to the errors from the training set. It is easy to see that the above problem derived from the geometric view (i.e., optimizing Equation 4.18 with the constraints 4.17) is equal to the primal problem (Formula 4.11) derived from the *loss* + *regularization* framework.

Non-linear binary SVMs: kernel functions. In the above Lagrangian formulation of the problem (Formula 4.12 for training and Formula 4.15 for testing), the training examples only appear in the form of dot products between vectors. This is a crucial property of SVMs. It allows us to generalize SVMs to the nonlinear case. The idea is mapping the input variables into a feature space of a higher dimension and then performing a linear classification in that higher dimensional space (Figure 4.7).
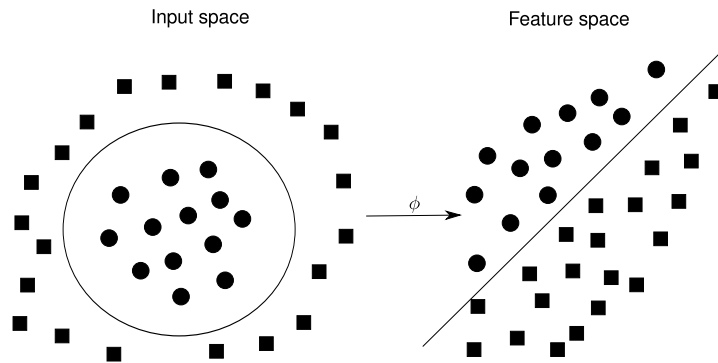

Figure 4.7: Linear separation of two classes in the high dimensional feature space.

For n training examples $\{x_i, y_i\}$, where $i = 1, ..., n$, $x_i \in \mathbb{R}^d$, $y_i \in \{1, -1\}$, we could define a function ϕ, which maps the d dimensional input vector x_i into a higher d' dimensional vector z, so that the new training data $\{\phi(x_i), y_i\}$ is linearly separable by a hyperplane in the new feature space. However, choosing ϕ directly is difficult. Moreover, if the new feature space has very high dimensionality ($d' \gg d$), then computing the dot products $\langle \phi(x_i) \cdot \phi(x_j) \rangle$ could be computationally intractable. Therefore kernel functions are proposed so that $K(x_i, x_j) = \langle \phi(x_i) \cdot \phi(x_j) \rangle$. The idea is to enable computations to be performed in the original input space rather than in the high-dimensional feature space.

There are several widely used kernel functions for SVMs, such as linear kernel, polynomial kernel, and Gaussian radical basis functions (RBF). Among all of them, *tree kernels* (Collins & Duffy, 2002; Moschitti, 2006) are developed to measure the similarity of tree structures. They can be used to capture useful patterns for identifying the target object. Such patterns are often implicitly encoded in the tree representation and are hard to design manually. In NLP, tree kernels are successfully applied in various tasks, such as relation extraction (Zelenko

et al., 2003), syntactic parsing (Collins & Duffy, 2002), and question answer classification (Moschitti et al., 2007).

Multi-class classification with SVMs.   SVMs are originally defined for two class problems. The common methods to apply SVMs in multi-class classification tasks are the *one-versus-one* approach or the *one-versus-all* approach.

In a k-class classification task, the one-versus-one approach trains a two-class SVM model for any pair of two classes from the training set. This yields $k(k-1)/2$ SVM models. In the testing stage, a voting procedure assigns the instance to the class with the maximum number of votes.

In contrast to the one-versus-one approach, the one-versus-all procedure requires a small number of models. In the previous example, only k SVM classifiers are needed using the one-versus-all approach. The ith classifier is trained with all instances from the ith class labeled as $+1$, and all other instances labeled as -1. In the testing stage, the instance is assigned to the class in which the corresponding classifier has the largest margin among all classifiers. Although less classifiers are needed in the one-versus-all approach, the training data set may be imbalanced due to the large number of instances with label -1.

Weighted SVMs for imbalanced classification.   In an imbalanced data set in which negative instances heavily outnumber the positive instances, the SVM classifier will favor the negative class because the simplest hypothesis is the one that classifies almost all instances as negative. Furthermore, the positive instances can be treated as noise and ignored completely by the classifier. In order to deal with imbalanced classification problems, weighted SVMs use different penalties (C^+ and C^-) for the two classes. The most unfavorable type of error has a higher penalty. For instance, the classification error for a positive instance should be more expensive than for a negative instance. Now the dual problem is as below:

$$\max_{\alpha} \ \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i \cdot x_j \rangle \right)$$
$$s.t. \ \ 0 \le \alpha_i \le C^+, i = 1, 2, \ldots n \ \ for \ y_i = +1$$
$$0 \le \alpha_i \le C^-, i = 1, 2, \ldots n \ \ for \ y_i = -1 \tag{4.19}$$
$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

In this thesis, we incorporate weighted SVMs into a cascading collective classification model for bridging anaphora recognition (Chapter 5).

4.2 Lexical Semantic Resources

4.2.1 A Distributional Semantic Resource for Bridging Resolution

In this section, we first explain the principle of the Dunning root log-likelihood ratio. We then detail how we utilize this method to create a distributional semantic resource for bridging resolution.

Dunning root log-likelihood ratio. As a measure of strength of association, the log-likelihood ratio statistic[6] was first introduced by Dunning (1993) to the NLP community. Dunning (1993) shows that the likelihood ratio test is more appropriate for text analysis tasks than the χ^2 test. The reason is that the normality assumption which χ^2 is based on breaks down for language usage due to the Zipf's law effect (Powers, 1998). That is, given a corpus, most of the distinct words will occur only a small number of times. For instance, in the Brown corpus (Kucera & Francis, 1967) which contains over one million words, the word "the" is the most frequently occurring word and accounts for nearly 7% of all word occurrences. Indeed, among 50,000 word types in the Brown corpus, only 135 are needed to account for half of the corpus, and 80% appear five or fewer times. To deal with the dominance of "rare" events in text analysis tasks, Dunning (1993) proposes likelihood ratio tests since they "do not depend so critically on assumptions of normality" and "allow comparison to be made between the significance of the occurrences of both rare and common phenomenon" (Dunning, 1993, p.65-66). Although Moore (2004) shows that Fisher's exact test would produce more accurate p-values than the likelihood ratio test, it is difficult to compute Fisher's exact test when the sample size is greater than 10^{11} due to floating point overflow on current ordinary computers (64-bit operations). Moreover, Moore (2004) confirms that an LLR score[7] above 10 is a reliable indicator of a significant association by comparing the noise estimates between LLR scores and the gold standard (i.e., Fisher's exact test) in a bilingual word association (English - French) experiment on a corpus consisting of 500,000 aligned sentence pairs.

We now explain the calculation details of computing the association strength between two events a and b (LLR_{ab}) in a data sample using the log-likelihood ratio measure. Table 4.4 shows the necessary counts we need to compute LLR_{ab}: C(a, b) is the number of times that a and b occurs together in the data sample, C(-a, b) and C(a, -b) is the number of times that b and a occurs without each other in the data sample, C(-a, -b) is the number of times that something has been observed that is neither a nor b, C(x) (x = a, -a, b, -b) is the number of times that x occurs in the data sample.

The log-likelihood ratio score LLR_{ab} based on binomial distribution is calculated as below:

---

[6]Its preferred name among statisticians is G^2.

[7]Moore (2004) defines LLR to be $G^2/2$.

	Event a	Everything except a	Total
Event b	C(a, b)	C(-a, b)	C(b)
Everything except b	C(a, -b)	C(-a, -b)	C(-b)
Total	C(a)	C(-a)	N

Table 4.4: The contingency table containing the counts for calculating the association between the event a and b.

$$LLR_{ab} = 2 \sum_{a? \in \{a, -a\}} \sum_{b? \in \{b, -b\}} C(a?, b?) \log \frac{C(a?, b?)}{E(a?, b?)}, \qquad (4.20)$$

where C(a?, b?) is the observed frequency, E(a?, b?) is the expected frequency under the null hypothesis. Let N denote the total sample size, then

$$E(a?, b?) = \frac{C(a?)C(b?)}{N} \qquad (4.21)$$

Replacing E(a?, b?) in Formula 4.20 with Formula 4.21, we get:

$$LLR_{ab} = 2 \sum_{a? \in \{a, -a\}} \sum_{b? \in \{b, -b\}} C(a?, b?) \log \frac{p(a?|b?)}{p(a?)} \qquad (4.22)$$

Note that $C(a?, b?) = p(a?, b?)N$ and $p(a?, b?) = p(a?|b?)p(b?)$. We therefore arrive at the following formula:

$$LLR_{ab} = 2N \underbrace{\sum_{a? \in \{a, -a\}} \sum_{b? \in \{b, -b\}} p(a?, b?) \log \frac{p(a?, b?)}{p(a?)p(b?)}}_{MI(a,b)} \qquad (4.23)$$

Formula 4.23 reveals the relation between log-likelihood ratio and mutual information: the log-likelihood ratio value LLR_{ab} is as 2N times as the mutual information value MI (a, b).

However, using the raw log-likelihood ratio score defined above, we will get counterintuitive results for the following examples[8]. Suppose we would like to compare the association strength between different terms and a specific cluster which consists of many documents. Table 4.5 shows the necessary counts for *term1* and *term2* respectively. For instance, *term1* occurs in 904 documents in the cluster and in 1,144 documents outside the cluster, while *term2* occurs only in 36 documents in the cluster but in 60,280 documents outside the cluster.

---

[8]The example as well as the formula of the signed root log-likelihood ratio are compiled from the forum discussion: http://mail-archives.apache.org/mod_mbox/mahout-user/201001.mbox.

	C(a, b)	C(a, -b)	C(-a, b)	C(-a, -b)	LLR$_{ab}$	Root LLR$_{ab}$
term1	904	21,060	1144	68,3012	3569.5	59.7
term2	36	21,928	60,280	62,3876	3622.0	- 60.2

Table 4.5: The log-likelihood ratio values and the signed root log-likelihood ratio values for the example of measuring term-cluster association.

Intuitively, the association strength between *term1* and the cluster should be higher than the association strength between *term2* and the cluster. Unfortunately this is not the case when comparing the raw log-likelihood ratio score LLR$_{ab}$ in Table 4.5. The reason lies in that the raw log-likelihood ratio can have a large value whenever there is an anomaly. In this example, *term2* is rare in the cluster and common outside the cluster, therefore it is an anomaly. Such cases can be fixed by applying the signed root log-likelihood ratio measure as below:

$$rootLLR_{ab} = \text{sgn}\left(\frac{C(a,b)}{C(b)} - \frac{C(a,-b)}{C(-b)}\right) \cdot \sqrt{LLR_{ab}} \qquad (4.24)$$

The signed root log-likelihood ratio (also called Dunning root log-likelihood ratio in this thesis) measure defined in Formula 4.24 has two advantages over the raw log-likelihood ratio measure:

> (1) It is positive where C(a, b) is bigger than expected, negative where it is lower.
> (2) If there is no difference, it is asymptotically normally distributed. This allows people to talk about "number of standard deviations" which is a more common frame of reference than the χ^2 distribution.
> (from Dunning's comments on $rootLLR_{ab}$ in the forum discussion)

Using the signed root log-likelihood ratio measure, it is easier to see that *term2* is an anomaly and *term1* has a stronger association with the cluster compared to *term2* (see the last column in Table 4.5).

Create a distributional semantic resource for bridging resolution by applying Dunning root log-likelihood ratio.   In this thesis, we apply the signed root log-likelihood ratio measure in big corpora to calculate the association strength between the head word of a (potential) bridging anaphor and the head word of an antecedent candidate given certain prepositions. This yields a distributional semantic resource for bridging resolution. Table 4.6 shows a fragment of this resource which contains the association strength (i.e., Root LLR$_{ab}$) between the head word of the bridging anaphor "**reasonable changes**" and the head words of

antecedent candidates under the preposition "*of*" using the Tipster corpus (Harman & Liberman, 1993). For instance, in the corpus we observe 61 times that "*changes/change*" is modified by "*structures/structure*" via the preposition "*of*"[9], 3,703 times that "*changes/change*" is modified by an NP other than "*structures/structure*" via the preposition "*of*", 8,497 times that "*structures/structure*" modifies an NP other than "*changes/change*" via the preposition "*of*", 8,010,544 times that the "*of*" preposition structure "*np1 of np2*" occurs where *np1* is not "*changes/change*" and *np2* is not "*structures/structure*". The results (ranked according to the Root LLR_{ab} score) in Table 4.6 show that Root LLR_{ab} could be used to filter out unlikely antecedent candidates and highlight the semantically possible antecedent candidates.

Root LLR_{ab}	C(a, b)	C(a, -b)	C(-a, b)	C(-a, -b)	anaphor	antecedent candidate
14.81	61	3,703	8,497	8,010,544	changes	the structure
7.83	61	17,199	8,497	7,997,048	changes	the state
0.13	4	3,508	8,554	8,010,739	changes	design of the structure
- 2.18	2	6,407	8,556	8,007,840	changes	A sales tax increase
- 2.33	1	5,051	8,557	8,009,196	changes	The claims
- 3.68	1	9,437	8,557	8,004,810	changes	Last week
- 3.9	0	7,120	8,558	8,007,127	changes	court
- 3.97	0	7,383	8,558	8,006,884	changes	The men

Table 4.6: The signed root log-likelihood ratio values between the bridging anaphor **reasonable changes** and antecedent candidates.

4.2.2 WordNet

WordNet (Fellbaum, 1998) is a lexical database of English, in which nouns, verbs, adjectives, and adverbs are grouped into synsets which represent distinctive concepts. Synsets are interlinked by conceptual-semantic and lexical relations, such as *synonym*, *antonym*, *hypernym*, *hyponym*, and *meronymy*. The newest version of WordNet (WordNet 3.0) contains 155,278 unique strings and 117,659 synsets.

In this thesis, we use WordNet in two ways. In order to predict meronymy bridging, we use WordNet to decide whether a part-of relation exists between two nouns. We also use WordNet to assign words to certain semantic classes, on the basis of whether these words are hyponyms of some specific abstract nouns. For instance, words that are hyponyms of

---

[9]It is worth noting that these three tokens do not always appear successively. We are interested in how many times that "*structures/structure*" modifies "*changes/change*" via the preposition "*of*". Therefore "*changes of the building's structures*" is also counted as one occurrence. We explore the automatic POS tag and chunk information to approximate such non-successive cases.

person/people/native/male/female/inhabitant are assigned to the semantic class *person*. The detected part-of relations and the semantic class assignments are used as features in Chapter 5, Chapter 6 and Chapter 7.

4.2.3   General Inquirer Lexicon

The General Inquirer (Stone et al., 1966) is an IBM 7090 program developed by Philip J. Stone and his colleagues in the 1960s at the Harvard Laboratory of Social Relations. The program was originally designed for content analysis research problems in the behavioral sciences, e.g., distinguishing real suicide notes from simulated ones. According to Stone et al. (1966), the General Inquirer processes natural language texts by (a) identifying words and phrases that belong to categories specified by the investigator; (b) performing graphical and statistical analysis based on occurrence or specified co-occurrence counting of these categories; (c) producing a report containing sentences with significant content words or phrases.

The web page of "How the General Inquirer is used" describes the General Inquirer as below[10]:

> The General Inquirer is basically a mapping tool. It maps each text file with counts on dictionary-supplied categories. The currently distributed version combines the "Harvard IV-4" dictionary content-analysis categories, the "Lasswell" dictionary content-analysis categories, and five categories based on the social cognition work of Semin and Fiedler, making for 182 categories in all. Each category is a list of words and word senses. A category such as "self reference" may contain only a dozen entries, mostly pronouns. Currently, the category "negativ" is our largest with 2291 entries.

The General Inquirer lexicon contains 11,788 entries and 182 categories. Each entry is assigned to one or several categories. Table 4.7 shows a fragment of the General Inquirer lexicon. The first column lists the entries, the second column indicates the source of an entry (e.g., *H4Lvd* or *H4*), the last column briefly explains the meanings for some entries, and other columns are different categories which are used to classify entries from different aspects. In this thesis, the General Inquirer lexicon is used as a lexical resource. We use it for the sole purpose of deciding whether words belong to specific categories, e.g., *role person, increase, decrease, space*. These categories assignments are used as features in Chapter 5, Chapter 6 and Chapter 7.

---

[10]http://www.wjh.harvard.edu/~inquirer/3JMoreInfo.html

Entry	Source	Negative	Fail	Decrease	Space	...	Defined
A	H4Lvd					...	indefinite singular article
Abandon	H4Lvd	Negative	Fail			...	
Abate	H4Lvd	Negative		Decrease		...	
Aboard	H4Lvd				Space	...	
Above#1	H4Lvd					...	59% prep-adv: Higher than
Above#2	H4Lvd				Space	...	21% noun-adj-adv: Previously stated
Adjacent	H4Lvd				Space	...	
Adultery	H4	Negative				...	
...							

Table 4.7: A fragment of the General Inquirer lexicon. "#n" differentiates senses. 5,395 words have definitions.

Chapter 5

Bridging Anaphora Recognition

Recognizing that a bridging anaphor is present is an integral part of the bridging resolution process. In this chapter, we cast *bridging anaphora recognition* as a subtask of learning fine-grained information status (IS). Each mention in a text gets assigned one IS class, bridging being one possible class. Some novel aspects of our current approach include a joint inference model that collectively predicts IS classes of mentions together, as well as a cascading collective classification method that addresses the data sparseness problem due to bridging occurring less frequently than many other IS classes. The research described in this chapter is an extension of our previous work (Markert et al., 2012; Hou et al., 2013a). The chapter is organized as follows. Section 5.1 defines the task. It also discusses some related empirical work and the motivation for the task. Section 5.2 proposes a cascading collective classification model for this task after discussing linguistic constraints among several IS classes and the wide variation of bridging anaphora. We describe the features used in this task in detail in Section 5.3. Section 5.4 presents the experimental setup and reports the results achieved for bridging anaphora recognition in the ISNotes corpus, with comparisons with several baselines. Finally, Section 5.5 summarizes this chapter.

5.1 Task

5.1.1 Task Definition

Bridging anaphora recognition is the problem of deciding which mentions are bridging anaphors. A mention is a noun phrase (NP) which refers to a discourse entity and carries information status. According to Prince (1981), a discourse entity is a discourse-model object, e.g., an individual (existent in the real world or not), a class of individuals, an exemplar, a substance, or a concept. All discourse entities in a discourse model are represented by noun phrases in a text. However, not all noun phrases in a text represent discourse entities. Such

NPs are annotated as *non-mentions* in ISNotes. Non-mentions include expletive or pleonastic *it/there*, reflexive pronouns that are used as emphasisers, parts of idioms, or parts of proper names. A more detailed discussion of non-mentions can be found in the ISNotes annotation scheme (Markert, 2013).

In this thesis, we handle bridging anaphora recognition as part of information status classification. Each mention in a text gets assigned one of the eight IS classes, bridging being one possible class. The eight classes are: `old`, `mediated/syntactic`, `mediated/world-Knowledge`, `mediated/bridging`, `mediated/comparative`, `mediated/aggregate`, `mediated/function`, and `new`. Section 3.1 of Chapter 3 provides more detailed information about these eight IS classes.

5.1.2   Background

Most previous empirical research on bridging concentrates on antecedent selection only (Poesio & Vieira, 1998; Poesio et al., 2004a; Markert et al., 2003; Lassalle & Denis, 2011), assuming that bridging anaphora recognition has already been performed. Recent work on bridging anaphora recognition models it as a subtask of learning fine-grained information status (Rahman & Ng, 2012; Cahill & Riester, 2012). Each mention in a text gets assigned one IS class that describes its accessibility to the reader at a given point in a text, bridging being one possible class. Under this framework, we reported moderate results for bridging in written news text (ISNotes) (Hou et al., 2013a) on the basis of our previous work (Markert et al., 2012), whereas Rahman & Ng (2012) reported high results for the four subcategories of bridging which are annotated in the Switchboard dialogue corpus by Nissim et al. (2004). This discrepancy is due to differences in corpus size and genre as well as in bridging definition. Bridging in Switchboard includes non-anaphoric, syntactically linked part-of and set-member relationships (such as *the building's lobby*)[1], as well as comparative anaphora which being marked by surface indicators such as *other, another, etc*[2]. Both types are much easier to identify than anaphoric bridging cases which we address in this thesis[3]. In addition, many non-anaphoric lexical cohesion cases have been annotated as bridging in Switchboard as well[4]. Another work on bridging anaphora recognition was carried out by Cahill & Riester (2012). Although their definition for bridging is similar to ours[5], they did not report the result for the *bridging* subcategory.

---

[1]This corresponds to `mediated/syntactic` in ISNotes.

[2]This corresponds to `mediated/comparative` in ISNotes.

[3]See also the high results for our specific categories (i.e., `mediated/syntactic` and `mediated/bridging`) in Section 5.4.5.

[4]See Section 2.1.2 for the detailed description of bridging annotation in the Switchboard corpus.

[5]See Section 2.1.2 for the definition of bridging in Cahill & Riester (2012).

5.1.3   Motivation for the Task

First, bridging anaphora recognition on its own can be valuable for applications. For example, prosody is influenced by information status without needing antecedent knowledge (Baumann & Riester, 2013).

Second, we argue that it is possible to recognize bridging anaphora without knowing the antecedent information. It seems that a joint approach (i.e., predicting bridging anaphors and their antecedents together) is more attractive since some antecedents could trigger subsequent bridging anaphors. For instance, in Example 5.1, the antecedent *the Polish center* triggers the anaphor **walls**. However, bridging anaphora can be indicated by referential patterns without world knowledge about the anaphor/antecedent NPs, as the nonsense Example 5.2 shows: **the wug** is clearly a bridging anaphor although we do not know the antecedent[6].

(5.1)   If Mr. McDonough's plans get executed, as much as possible of *the Polish center* will be made from aluminum, steel and glass recycled from Warsaw's abundant rubble. **The windows** will open. **The carpets** won't be glued down and **walls** will be coated with non-toxic finishes.

(5.2)   The blicket couldn't be connected to the dax. **The wug** failed.

Similarly, Clark (1975) distinguishes between bridging via necessary, probable and inducible parts/roles[7]. He argues that only in the first case the antecedent triggers the bridging anaphor in the sense that we already spontaneously think of the anaphor when we read/hear the antecedent. For instances, **walls** in Example 5.1 are necessary parts of the antecedent *the Polish center* according to common sense knowledge. However, there is no guarantee that a building must have windows. Instead, a building probably has windows. So **The windows** (Example 5.1) are probable parts of the antecedent (*the Polish center*). Furthermore, carpets are certainly not necessary parts of a building. Therefore **The carpets** (Example 5.1) are inducible parts of the antecedent (*the Polish center*). In the probable/inducible cases, the bridging anaphor accommodates itself into the context and is induced by the need for an antecedent.

5.2   Model

In this section, we propose a cascading collective classification approach to model bridging anaphora recognition. We show how linguistic knowledge can be used to derive the constraints used in our collective classification (Section 5.2.1), and how this collective classifier

---

[6]We thank an anonymous reviewer for pointing this out.
[7]See Table 1.1 in Chapter 1 for examples of different types given by Clark (1975).

is integrated into a cascading collective classification system to address the minority class (`bridging`) identification problem in a multi-class setting (Section 5.2.2).

5.2.1 Collective Classification

Background.   Statistical Relational Learning (SRL) (Getoor & Taskar, 2007) addresses the problem of performing probabilistic inference on data instances that are correlated. Collective classification is one of the important tasks in SRL research (Jensen et al., 2004; Macskassy & Provost, 2007). In collective classification, related data instances are classified simultaneously rather than independently which is done in traditional classification. In traditional machine learning, the data is considered to be drawn independently and identically from some distribution (i.i.d.). In contrast to i.i.d., *autocorrelation* is a widely observed characteristic of relational data in which the value of a variable for one instance is highly correlated with the value of the same variable on another instance. By exploiting relational autocorrelation, some studies (Taskar et al., 2001; Neville & Jensen, 2003; Domingos & Lowd, 2009) show that collective classification can significantly outperform independent supervised classification on various relational datasets. Jensen et al. (2004) claim that such improvements attributed to collective classification result from "a clever factoring of the space of statistical dependencies in relational data" (Jensen et al., 2004, p.598). In datasets with strong autocorrelation, this factoring (i.e., modeling dependencies between: (1) the class label of an object and attributes on this object, and (2) the class label of an object and the class labels of adjoining objects) provides the collective classification with informative relational information which then greatly reduce the model's bias at a minimum cost.

Collective classification has been widely applied in various NLP tasks, such as part-of-speech tagging (Lafferty et al., 2001), webpage categorization (Taskar et al., 2002), opinion mining (Somasundaran et al., 2009; Burfoot et al., 2011), and entity linking (Fahrni & Strube, 2012).

Motivation for the model: linguistic relations among IS categories.   Among all eight IS categories, two `mediated` subcategories account for accessibility via syntactic links to another `old` or `mediated` mention. `Mediated/syntactic` is used when at least one child of a mention is `mediated` or `old`, with child relations restricted to the following:

- Possessive pronouns or Saxon genitives (e.g., [[his]$_{old}$ father]$_{mediated/syntactic}$)

- Of-genitives (e.g., [The alcoholism of [his]$_{old}$ father]$_{mediated/syntactic}$)

- Proper name premodifiers
  (e.g., [The [Federal Reserve]$_{mediated/worldKnowledge}$ boss]$_{mediated/syntactic}$)

- Other Prepositional phrases
  (e.g., [professors at [Cambridge]$_{mediated/worldKnowledge}$]$_{mediated/syntactic}$)

The other `mediated` subcategory `mediated/aggregate` is for coordinations in which at least one of the children is `old` or `mediated`, e.g., *Not only George Bush but also Barack Obama* is `mediated/syntactic` as *Barack Obama* is `mediated/worldKnowledge`.

In the above two cases, a mention's IS depends directly on the IS of its children. Therefore, we explore collective classification to capture such dependencies.

Detailed model. We use M to denote the set of n mentions in a document D, and S to denote the set of eight IS classes. Let s_m be the IS class associated with a mention $m \in M$, S_M be the IS class assignments for all mentions in M, S_M^n be the set of all possible IS class assignments for M. The collective IS classification task can be represented as a log-linear model:

$$P(S_M|M; w) = \frac{\exp(w \cdot \Phi(M, S_M))}{\sum_{S_{M'} \in S_M^n} \exp(w \cdot \Phi(M, S_M'))} \tag{5.3}$$

where w is the model's weight vector, $\Phi(M, S_M)$ is a global feature vector which takes the entire IS class assignments for all mentions in M into account. We define $\Phi(M, S_M)$ as:

$$\Phi(M, S_M) = \sum_{l \in F_l} \sum_{m \in M} \Phi_l(m, s_m) + \sum_{g \in F_g} \sum_{m_i, m_j \in M} \Phi_g(s_{m_i}, s_{m_j}) \tag{5.4}$$

where $\Phi_l(m, s_m)$ is a local feature function that looks at the mention m and the target IS class s_m, $\Phi_g(s_{m_i}, s_{m_j})$ is a global feature function that looks at the target IS class assignments for m_i and m_j at once.

This log-linear model can be represented using Markov logic networks (MLNs) which were described in Section 4.1.1. In a ground Markov network for this task, the probability distribution over the possible world S_M is given by

$$P(S_M) = \frac{1}{Z} \exp\left(\sum_i w_i n_i(S_M)\right) \tag{5.5}$$

where $n_i(S_M)$ is the number of true groundings of a local or a global feature function F_i in S_M. Table 5.1 shows the formula templates that we design to model this problem in MLNs. $p1$ is the hidden predicate that we want to predict, i.e., the information status s of a mention m. $f1$ models that each mention can only belong to one IS class, f_g and f_l are the templates of joint inference formulas and non-joint inference formulas respectively. The details of specific formulas (features) instantiating f_g and f_l are described in Section 5.3.1 and Section 5.3.2.

Hidden predicates	
$p1$ $hasIS(m, s)$	

Formulas

Hard constraints

$f1$ $\forall m \in M : |s \in S : hasIS(m, s)| = 1$

Joint inference formula template

f_g (w) $\forall m_i, m_j \in M \; \forall s_{m_i}, s_{m_j} \in S :$ jointInferenceFormula_Constraint (m_i, m_j)
$\rightarrow hasIS(m_i, s_{m_i}) \; \wedge \; hasIS(m_j, s_{m_j})$

Non-joint inference formula template

f_l (w) $\forall m \in M \; \forall s \in S :$ non-jointInferenceFormula_Constraint (m, s)
$\rightarrow hasIS(m, s)$

Table 5.1: Hidden predicates and formulas used for bridging anaphora recognition. m represents a mention, M the set of mentions in the whole document, s an IS class, S the set of eight IS classes, and w the weight learned from the data for the specific formula.

We use *thebeast*[8] to learn weights for the formulas and to perform inference. *thebeast* employs cutting plane inference (Riedel, 2008) to improve the accuracy and efficiency of MAP inference for MLNs.

5.2.2   Cascading Collective Classification

Motivation for the model: the wide variation of bridging anaphora and their relative rarity. Bridging anaphors are rarely marked by surface features. Indeed, even the common practice (Vieira & Poesio, 2000; Lassalle & Denis, 2011; Cahill & Riester, 2012) to limit bridging anaphora to definite NPs does not seem to be correct: in Section 3.2, we found that less than 40% of the bridging anaphors in ISNotes are modified by *the*, and most bridging anaphors (44.9%) are not modified by any determiners. Furthermore, bridging anaphora are diverse with regard to syntactic form and function: bridging anaphora can be definite NPs (Examples 5.7 and 5.9), indefinite NPs (Example 5.8), or bare NPs (Examples 5.6, 5.11 and 5.12). The only frequent syntactic property shared is that bridging anaphora tend to have a simple internal structure with regard to modification[9]. Bridging anaphora are also easily confused with generics: *friends* is used as a bridging anaphor in Example 5.12 but generically in Example 5.13.

---

[8]http://code.google.com/p/thebeast

[9]In Chapter 3, we found that around 90% of bridging anaphors in the ISNotes corpus do not contain any other mentions.

(5.6)   In June, farmers held onto *meat, milk and grain*, waiting for July's usual state-directed price rises. The Communists froze **prices** instead.

(5.7)   To reduce it at *the fund's building*, workers rubbed beeswax instead of polyurethane on the floors in the executive's office. [...] **The budget** was only $400,000.

(5.8)   Still, *employees* do occasionally try to smuggle out a gem to two. [...] **A food caterer** stashed stones in the false bottom of a milk pail.

(5.9)   *His truck* is parked across the field, in a row of grain sellers. [...] The farmer at **the next truck** shouts, "Wheat!"

(5.10)   The survey found that over a three-year period 22% of *the firms* said employees or owners had been robbed on their way to or from work or while on the job. Crime was the reason that **26%** reported difficulty recruiting personnel and that **19%** said they were considering moving.

(5.11)   Mr. Leavitt, 37, was elected chairman earlier this year by *the company's* new board [...]. His father, David S. Leavitt, was **chairman** and **chief executive** until his death in an accident five years ago, ...

(5.12)   *She* made money, but spent more. **Friends** pitched in.

(5.13)   Friends are part of the glue that holds life and faith together.

In the initial experiment, we find that our collective classification model presented in the previous section (Section 5.2.1) works well to identify most IS categories except `mediated/bridging` (see Table 5.12 in Section 5.4.5). In fact, the collective classification model misclassifies many `mediated/bridging` mentions as `new`. This is due to the wide variation within the phenomenon, the resulting lack of easily identifiable surface markers and the relative rarity of bridging compared to many other IS classes. Such multi-class imbalance problems (i.e., learning from imbalanced data within a multi-class setting) are still an open research topic (Abe et al., 2004; Zhou & Liu, 2010; Wang & Yao, 2012). The classification accuracy is not a fair measure to be optimized when facing imbalanced classes (Fawcett, 2006). Accuracy may be artificially high in case of extremely imbalanced data: majority classes are favored, while minority classes are not recognized. Prediction is biased toward the classes with the highest priors. Such a bias gets stronger within the multi-class setting. To address this problem while still keeping the strength of collective inference within a multi-class setting, we integrate our collective classification model (Section 5.2.1) into a cascading collective classification system inspired by Omuya et al. (2013).

Detailed Model. Figure 5.1 shows the framework of the cascading collective classification system. Unlike in the multi-class setting, learning from imbalanced data in the binary setting has been well-studied over the past years (He & Garcia, 2009). Our cascading collective classification system combines the binary classifiers for minority categories and a collective classifier for all categories in a cascaded way. Specifically, for the five minority classes (i.e., `mediated/function`, `mediated/aggregate`, `mediated/compar-ative`, `mediated/bridging`, and `mediated/worldKnowledge`) that each makes up less than the expected $\frac{1}{8}$ of the data set, we develop five binary classifiers with SVM^{light} (Joachims, 1999)[10] using all non-joint features from the collective classification model described in the previous section (i.e., features instantiating f_l in Table 5.1)[11] and apply them in order from rarest to more frequent category. Whenever a minority classifier predicts true, this class is assigned. When all minority classifiers say false, we back off to the multi-class classification system exploring collective inference described in Section 5.2.1. We show in Section 5.4 that such a framework substantially improves bridging anaphora recognition without jeopardizing performance on other IS classes.
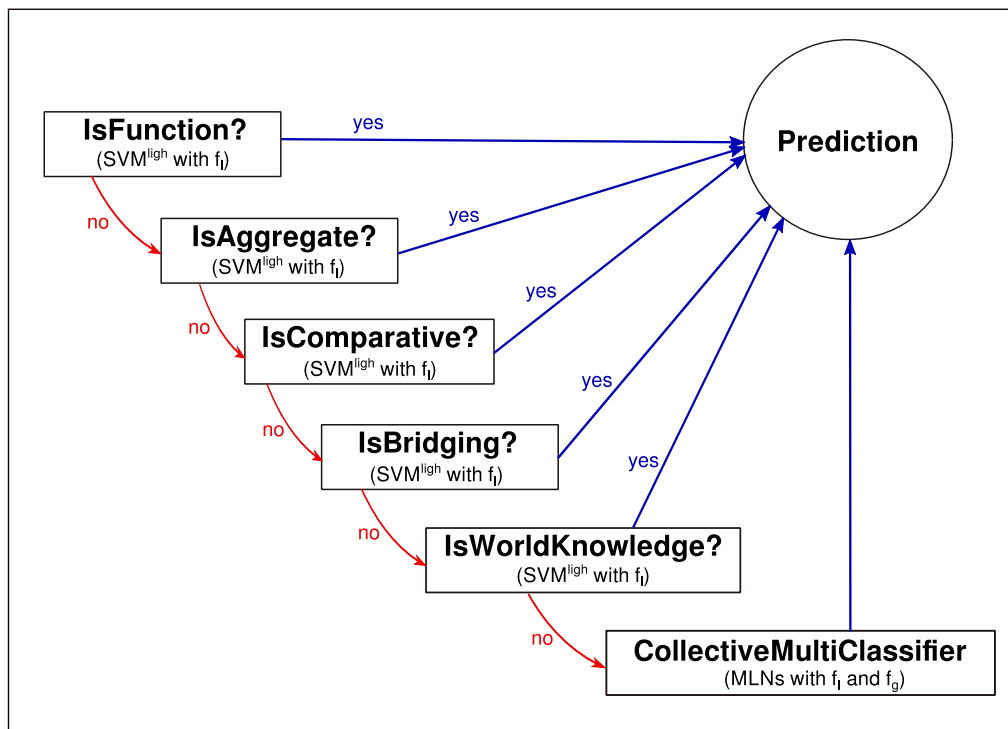
Figure 5.1: The cascading collective classification system.

---

[11]The non-joint features are equal to non-joint inference formulas in Section 5.2.1 and will be detailed in Section 5.3.2.

5.3 Feature Design

In this section, we describe all features used in our model presented in Section 5.2. Section 5.3.1 details the relational features which instantiate the joint inference formula template f_g in Table 5.1, whereas Section 5.3.2 details non-relational features which instantiate the non-joint inference formula template f_l in Table 5.1.

5.3.1 Relational Features

Syntactic hasChild relations. We link a mention m_1 to a mention m_2 via a *hasChild* relation if (i) m_2 is a possessive or prepositional modification of m_1; or (ii) m_2 is a proper name premodifier of m_1. For instance, the mention *[professors at Cambridge]* is linked to the mention *[Cambridge]* via a *hasChild* relation.

Syntactic hasChildCoordination relations. We link a mention m_1 to a mention m_2 via a *hasChildCoordination* relation if m_1 is a coordination and m_2 is one of its children. For example, the mention *[Not only George Bush but also Barack Obama]* is linked to the mention *[Barack Obama]* via a *hasChildCoordination* relation.

Syntactic conjunctionOf relations. We hypothesize that there are certain IS patterns for mentions which are syntactically parallel. For instance, syntactically parallel mentions may have the same IS class. Therefore we link a mention m_1 to a mention m_2 via a *conjunctionOf* relation if both m_1 and m_2 are the children of a coordination structure. For example, the mention *[George Bush]* is linked to the mention *[Barack Obama]* via a *conjunctionOf* relation as both are the children of the coordination *[Not only George Bush but also Barack Obama]*.

5.3.2 Non-relational Features

In this section, we first describe features used in previous work (Nissim, 2006; Rahman & Ng, 2011) on IS classification (Section 5.3.2.1). We then detail the new additional features that we design for recognizing several IS categories (`old`, `mediated/worldKnowledge`, `mediated/comparative`, and `mediated/function`) and for recognizing bridging anaphora in Section 5.3.2.2 and Section 5.3.2.3 respectively. Finally, a full table of all non-relational features is shown in Section 5.3.2.4 where all features are organized into different groups according to different criteria.

5.3.2.1   Features From Previous Work

Table 5.2 summarizes the features that we adapt from previous work (Nissim, 2006; Rahman & Ng, 2011) for IS classification. They have shown positive effects on IS classification: a mention which has the same string as a previous mention is likely to be `old` ($f1$). As claimed by Prince (1992), subjects are more likely to be `old` ($f6$). Also pronouns tend to be `old` ($f7$) and indefinite NPs often are `new` (Hawkins, 1978) ($f5$). $f4$ NPlength is motivated by linguistic studies which show that "items that are new to the discourse tend to be complex and items that are given tend to be simple" (Arnold et al., 2000, p.34). Therefore `new` mentions are likely to be long while `old` or `mediated` mentions are likely to be short. Although Nissim intends to apply "head match" ($f3$) to capture cases of *mediated/set* annotated in a dialogue corpus, such as *my children – your children*[12], we think this feature is more useful to predict `old` mentions since "head match" is widely used in coreference resolution (Vieira & Poesio, 2000; Soon et al., 2001)

Building upon Nissim's work (2006), Rahman & Ng (2011) explore lexical features to capture the correlations between certain lexical forms and IS categories ($f8$), e.g., mentions which include the lexical unit *his* are not likely to be `new`.

Feature	Value
Features from Nissim (2006)	
$f1$ FullPrevMention (n)	*{yes, no, NA}*[13]
$f2$ FullMentionTime (n)	*{first, second, more, NA }*
$f3$ PartialPreMention (n)	*{yes, no, NA}*
$f4$ NPlength (int)	numeric, e.g., 5
$f5$ Determiner (n)	*{def, indef, dem, poss, bare, NA}*
$f6$ GrammaticalRole (n)	*{subject, subjpass, object, predicate, pp, other}*
$f7$ NPType (n)	*{common noun, proper noun, pronoun, other}*
Feature from Rahman & Ng (2011)	
$f8$ Unigrams (l)	e.g., *his*, *the*, *China*

Table 5.2: Non-relational features from previous work. "n" indicates nominal features, "l" lexical features, "int" integer. A nominal feature draws the feature value from a restricted set. A lexical feature indicates the presence or absence of a lexicon unit in a mention.

In the following, we detail how we extract these features for mentions. The Penn Treebank annotation as well as the named entity annotation in ISNotes are explored for feature

---

[12]The example is from Nissim (2006). Without context, it is not clear why the relation between two mentions in this example is *set/membership* instead of *identity*.

[13]We changed the value of "$f1$ FullPrevMention" from "numeric" to *{yes, no, NA}*.

extraction. We convert the Penn Treebank annotation to dependency parse trees using the Penn Converter tool (Johansson & Nugues, 2007). Following Nissim (2006), the value "NA" stands for "not applicable" and is used for pronouns.

$f1$ **FullPrevMention.** $f1$ considers whether a mention has the same string as a previous mention in the whole document.

$f2$ **FullMentionTime.** As a categorical version of $f1$, this feature categorizes the "oldness" of a mention into three categories (i.e., *first, second, more*) according to how many times it has the same string as previous mentions in the whole document.

$f3$ **PartialPreMention.** $f3$ decides whether a mention has the same head word as a previous mention in the whole document. The head word of a mention is detected by examining the dependency parse tree of the NP: the word which is not dependent on any other word within the NP is chosen to be the head word of the NP.

$f4$ **NPlength.** $f4$ calculates how many words a mention contains.

$f5$ **Determiner.** $f5$ classifies mentions into the following categories according to their determiner modifications:

- **Definite (def):** the head word of the mention is modified by *the*.

- **Indefinite (indef):** the head word of the mention is modified by determiners such as *a*, *an*, or *one* which normally indicate indefinite NPs[14].

- **Demonstrative (dem):** the head word of the mention is modified by demonstrative determiners, i.e., *this, that, these*, and *those*.

- **Possessive (poss):** the head word of the mention is modified by possessive pronouns or Saxon genitives, such as *his staff* or *the city's black leaders*.

- **Bare:** the head word of the mention is not modified by any determiners, such as *relief efforts*.

---

[14]The whole list of determiners we used to detect indefinite NPs is: {*a, an, one, some, any, either, neither, no, all, each, every, another, whichever, which, what, whatever*}.

ƒ6 **GrammaticalRole.** Nissim (2006) designs four categories for this feature, i.e., *subject, subjpass, pp*, and *other*. We include two more categories on the basis of this. Thus ƒ6 classifies mentions into the following categories according to their grammatical roles in the sentences by checking the dependency relation (DEPREL) tags of the mentions' head words in the dependency parse trees:

- **Subject:** the DEPREL tag of the head word is *SBJ* (subject).

- **Passive Subject (subjpass):** the DEPREL tag of the head word is *OBJ* (object) and the head word appears before its dependent verb.

- **Object:** the DEPREL tag of the head word is *OBJ* (object) and the head word appears after its dependent verb.

- **Predicate:** the DEPREL tag of the head word is *OPRD* (predicative complement of raising/control verb) or *PRD* (predicative complement).

- **Preposition (pp):** the DEPREL tag of the head word is *PMOD* (modifier of preposition).

- **Other:** the DEPREL tag of the head word does not fit into any of the above categories.

ƒ7 **NP Type.** ƒ7 classifies mentions into the following categories on the basis of the part of speech (POS) of their head words:

- **Common Noun:** the POS tag of the head word is *NN* or *NNS*.

- **Proper Name:** the POS tag of the head word is *NNP* or *NNPS*, or the mention is a named entity according to the named entity annotations.

- **Pronoun:** the POS tag of the head word begins with *PRP*, or the head word appears in a list of pronouns[15].

- **Other:** the head word does not fit into any of the above categories. This includes numbers (e.g., *$30,000*), adjectives (e.g., *the poor*), and gerunds (e.g., *his attending*).

---

[15]Some pronouns have other POS tags such as *DT* (e.g., *this, that*) or *CD* (e.g., *one*). Therefore we compile an extra list of pronouns to improve the recall of pronoun detection. The whole list is: {*all, another, any, anybody, anyone, anything, both, each, either, everybody, everyone, everything, few, little, many, more, much, most, nobody, none, nothing, neither, one, other, others, some, somebody, something, someone, this, that, these, those*}.

$f8$ **Unigrams.** We first create the unigram lexicon by collecting all words of all mentions from the training data. Then for each unit (a word) of the lexicon, we consider whether it appears in the mention under consideration.

5.3.2.2 New Features for Recognizing Some IS Categories

Table 5.3 lists several new features we designed for the `old` category and for the three `mediated` categories, i.e., `mediated/worldKnowledge`, `mediated/comparative`, and `mediated/function`.

Feature	Value
Features for identifying `old`	
$f1$ PartialPreMentionTime (n)	*{first, second, more, NA}*
$f2$ ContentWordPreMention (b)	*{yes, no, NA}*
Feature for identifying `mediated/worldKnowledge`	
$f3$ IsFrequentProperName (b)	*{yes, no}*
Feature for identifying `mediated/comparative`	
$f4$ PreModByCompMarker (b)	*{yes, no}*
Feature for identifying `mediated/function`	
$f5$ DependOnChangeVerb (b)	*{yes, no}*

Table 5.3: Non-relational features for recognizing some IS categories. "b" indicates binary features, "n" nominal features. A nominal feature draws the feature value from a restricted set. The value "NA" stands for "not applicable" and is used for pronouns.

We describe the details of these features as well as the motivation for devising these features in the following:

$f1$ **PartialPreMentionTime.** This feature is a categorial version of $f3$ *PartialPreMention* (head match) in Table 5.2. We calculate how many times the mention under consideration has the same head word as previous mentions in the whole document and categorize the value into three types, i.e., *first, second, more*.

$f2$ **ContentWordPreMention.** We detect whether a content word of the mention under consideration appears in previous mentions in the document. We define the content word as a word within the mention which has one of the following POS tags: *NN, NNS, NNP, NNPS, JJ*. This feature is designed to capture a partial string match between an `old` or a `mediated/bridging` mention and its previously mentioned entity (antecedent), e.g., *a hostile*

takeover attempt – the takeover threat (coreference), *the state –* **state gasoline taxes** (bridging)[16].

$f3$ **IsFrequentProperName.**   Previously unmentioned proper names, if they appear frequently in many documents, are more likely to be hearer-old and hence of the IS class `mediated/worldKnowledge`. To approximate this, we extract a list of proper names that occur in at least 100 documents in the Tipster corpus (Harman & Liberman, 1993). We then decide whether the mention under consideration appears in this list or not.

$f4$ **PreModByCompMarker.**   `Mediated/comparative` mentions are often indicated by surface clues such as premodifiers (e.g., *other* or *another*) which make clear that the entity is compared to another one. We use a small list of ten such markers[17] as well as the presence of adjectives or adverbs in the comparative form to detect `mediated/comparative` mentions.

$f5$ **DependOnChangeVerb.** $f5$ determines whether a number mention is the object of an increase/decrease verb (using a list extracted from Inquirer[18]) and therefore is likely to be the IS class `mediated/function`.

5.3.2.3   New Features for Recognizing Bridging Anaphora

The features presented in the last two sections (Section 5.3.2.1 and Section 5.3.2.2) are effective for most IS categories except bridging. Unlike other IS categories, bridging anaphors are rarely marked by surface features and can have almost limitless variations (see the motivation part in Section 5.2.2 for a detailed discussion).

However, we observe that bridging anaphors are often licensed because of discourse structure and/or lexical or world knowledge. With regard to discourse structure, Grosz et al. (1995) observe that bridging is often needed to establish entity coherence between two adjacent sentences (Examples 5.1, 5.2, 5.7, 5.8, 5.9, 5.10 and 5.12). With regard to lexical and world knowledge, relational noun phrases (Examples 5.6, 5.7, 5.11 and 5.12), building parts (Example 5.1), set membership elements (Example 5.10), or, more rarely, temporal/spatial modification (Example 5.9) may favor a bridging reading. Motivated by these observations, we develop discourse structure and lexico-semantic features indicating bridging anaphora. We

---

[16]In the initial experiment, we find this feature has more effect on the `old` category compared to the `mediated/bridging` category.

[17]The full list is: *{other, another, such, different, similar, additional, comparable, same, further, extra}.*

[18]The full increase/decrease verb list is *{increase, raise, rise, climb, swell, ascend, jump, leap, scale, stretch, become, double, extend, grow, improve, strengthen, fall, drop, cut, slow, ease, reduce, descend, lower, slip}.*

also design features to separate genericity from bridging anaphora. Table 5.4 lists all features for recognizing bridging anaphors. We detail these features in the following.

Feature	Value
Discourse structure	
$f1$ IsCoherenceGap (b)	*{yes, no}*
$f2$ IsSentFirstMention (b)	*{yes, no}*
$f3$ IsDocFirstMention (b)	*{yes, no}*
Lexico-semantics	
$f4$ IsArgumentTakingNP (b)	*{yes, no}*
$f5$ IsWordNetRelationalNoun (b)	*{yes, no}*
$f6$ IsInquirerRoleNoun (b)	*{yes, no}*
$f7$ SemanticClass (n)	a list of 16 classes, e.g. *location, organization*
$f8$ IsBuildingPart (b)	*{yes, no}*
$f9$ IsSetElement (b)	*{yes, no}*
$f10$ PreModSpatialTemporal (b)	*{yes, no}*
$f11$ IsYear (b)	*{yes, no}*
$f12$ PreModifiedByCountry (b)	*{yes, no}*
Identifying generic NPs	
$f13$ AppearInIfClause (b)	*{yes, no}*
$f14$ NPNumber (n)	*{singular, plural, unknown}*
$f15$ VerbPosTag (l)	e.g., *VBG*
$f16$ IsFrequentGenericNP (b)	*{yes, no}*
$f17$ GeneralWorldKnowledge(l)	e.g., *the sun*, *the wind*
$f18$ PreModByGeneralQuantifier (b)	*{yes, no}*
Mention syntactic structure	
$f19$ HasChildMention (b)	*{yes, no}*

Table 5.4: Non-relational features for recognizing bridging anaphora. "b" indicates binary features, "n" nominal features, "l" lexical features. A nominal feature draws the feature value from a restricted set. A lexical feature indicates the presence or absence of a lexicon unit in a mention.

Discourse structure features (Table 5.4, $f1$ - $f3$). Bridging occurs frequently in sentences where otherwise there would be no entity coherence to previous sentences/clauses (see Grosz et al. (1995) and Poesio et al. (2004b) for discussions about bridging, entity coherence and

centering transitions in the *centering* framework). This is especially true for *topic* NPs[19] (Halliday & Hasan, 1976) in such sentences.

We follow these insights by identifying coherence gap sentences (see Examples 5.1, 5.7, 5.8, 5.9, 5.12 and also 5.2): a sentence has a *coherence gap* ($f1$ *IsCoherenceGap*) if it has none of the following three coherence elements: (1) entity coreference to previous sentences, as approximated via string match or the presence of pronouns, (2) comparative anaphora approximated by mentions modified via a small set of comparative markers, or by the presence of adjectives or adverbs in the comparative form (see also Table 5.3, $f4$ *PreModByCompMarker*), or (3) proper names. We approximate the topic of a sentence via the first mention ($f2$ *IsSentFirstMention*).

$f3$ *IsDocFirstMention* models that bridging anaphors do not appear at the beginning of a text.

Semantic features (Table 5.4, $f4$ - $f12$).   In contrast to generic patterns, our semantic features capture lexical properties of nouns that make them more likely to be the head of a bridging NP. Drawing on theories of noun types by Löbner (1985) as well as often-discussed bridging sub-classes in prior work (Lassalle & Denis, 2011; Clark, 1975; Poesio & Vieira, 1998), we create features to capture several typical kinds of bridging anaphora using various resources.

Lexico-semantic features: general relational nouns.   Löbner (1985) distinguishes between relational nouns that take on at least one obligatory semantic role (such as *friend*) and sortal nouns (such as *table* or *flower*). He points out that relational nouns are more frequently used as bridging than sortal nouns (see Examples 5.6, 5.7, 5.11 and 5.12). We design $f4$ *IsArgumentTakingNP* and $f5$ *IsWordNetRelationalNoun* to capture general relational nouns.

$f4$ decides whether the argument taking ratio of a mention's head is bigger than some threshold k. We calculate the argument taking ratio α for a mention using NomBank (Meyers et al., 2004). For each mention, α is calculated via its head frequency in the NomBank annotation divided by the head's total frequency in the WSJ corpus in which the NomBank annotation is conducted. The value of α reflects how likely an NP is to take arguments. For instance, the value of α is 0.90 for *husband* but 0.31 for *children*.

We also extract a list containing around 4,000 relational nouns from WordNet, then determine whether the mention head appears in the list or not ($f5$ *IsWordNetRelationalNoun*).

However, the obligatory semantic role for a relational noun can of course also be filled NP-internally instead of anaphorically. We use the features $f12$ *PreModifiedByCountry* (such as *the Egyptian president*) and $f19$ *HasChildMention* (for complex NPs that are likely to fill needed roles NP-internally) to address this.

---

[19]A *topic* NP is what the sentence is about.

Lexico-semantic features: role terms and kinship terms. Role terms (e.g. *chairman*) and kinship terms (e.g. *husband*) are two specific types of relational nouns which are likely to be used anaphorically (Löbner, 1985). *f6 IsInquirerRoleNoun* determines whether the mention head appears in a list containing around 500 nouns that specify professional roles. The list consists of nouns from the General Inquirer lexicon (Stone et al., 1966) under the "*role*" category.

However, this list also includes kinship terms such as *wife* and *husband*. Therefore in *f7 SemanticClass*, we explore WordNet to divide person mentions into three categories: *rolePerson*, *relativePerson*, and *person⋆*. We integrate these three categories into a list containing 16 coarse-grained semantic classes. The detailed information about the semantic classes is summarized in Table 5.5. We describe the process details of *f7 SemanticClass* below.

Semantic Class Type	Definition
RolePerson	Professional roles, based on a list extracted from WordNet containing around 100 nouns which specify professional roles (e.g., *chairman*, *president* or *professor*)
RelativePerson	Relatives, based on a list extracted from WordNet containing 110 relative nouns (e.g., *husband*, *daughter* or *friend*)
Person⋆	People that are neither rolePerson nor relativePerson
Organization	Companies, agencies, institutions, etc.
GPE	Countries, cities, states
Location	Non-GPE locations, mountain ranges, bodies of water
NORP	Nationalities or religious or political groups
Event	Named hurricanes, battles, wars, sports events, etc.
Product	Vehicles, weapons, food, etc. (Not services)
Date	Absolute or relative dates or periods
Time	Times smaller than a day
Percent	Percentage (including "%")
Money	Monetary values, including unit
Ordinal	"first", "second"
Cardinal	Numerals that do not fall under another type (ordinal)
Other	NPs that do not fall under the above 15 types

Table 5.5: The detailed information for 16 semantic classes. The definitions of these semantic classes (except the first three categories) are from the OntoNotes guidelines for named entity annotation (Weischedel et al., 2011).

To assign a semantic class to a mention, we first consult the OntoNotes named entity an-

notation. When such information is not available, we use WordNet to decide whether the head lemma of the common noun mention is a hyponym of *{person, people, native, male, female, inhabitant}, {location}* or *{organization}*. All personal pronouns are assigned to the coarse-grained *person* category. Finally, we further divide the *person* category (including mentions annotated as *person* in OntoNotes, mentions being hyponyms of *{person, people, native, male, female, inhabitant}* in WordNet as well as personal pronouns) into three semantic classes using two lists extracted from WordNet. One list contains around 100 nouns which specify professional roles (e.g., *mayor*, *director*, or *president*), the other contains 110 relative nouns (e.g., *husband*, *daughter*, or *friend*). We assign the semantic class *rolePerson* or *relativePerson* to a person mention according to whether its head is present in the respective list. Person mentions with heads not present in these two lists belong to the semantic class *person⋆*.

Lexcio-semantic features: part terms. Because part-of relations are typical bridging relations (see Example 5.1 and Clark (1975)), we use $f8$ *IsBuildingPart* to determine whether the mention head is a part of the building or not, using a list containing 45 nouns extracted from Inquirer[20].

Lexico-semantic features: set bridging. $f9$ *IsSetElement* is used to identify set/membership bridging cases (see Example 5.10), by checking whether the mention head is a number or indefinite pronoun (i.e., *one, some, none, many, most*) or modified by *each*, *one*. However, not all numbers are bridging cases (such as *1976*) and we use $f11$ *IsYear* to exclude such cases.

Lexico-semantic features: spatial/temporal bridging. Lassalle & Denis (2011) note that some bridging anaphors are indicated by spatial or temporal modifications (see Example 5.9). We use $f10$ *PreModSpatialTemporal* to detect such cases by compiling around 20 such adjectives/adverbs from Inquirer[21].

Features to detect generic NPs (Table 5.4, $f13$ - $f18$). Generic NPs (Example 5.13) are easily confused with bridging anaphora. Inspired by Reiter & Frank (2010) who build on a wide variety of previous linguistic research on genericity, we develop features ($f13$-$f18$) to exclude generics.

First, hypothetical entities are likely to refer to generic entities (Mitchell et al., 2002), such as *a person* in Example 5.14[22]. We approximate this by determining whether the NP appears in

---

[20]The list contains lexical items from Inquirer under the "BldgPt" category, such as *window* or *room*.

[21]The whole list is: *{final, first, last, next, prior, succeed, second, nearby, previous, close, above, adjacent, behind, below, bottom, early, formal, future, before, after, earlier, later}*.

[22]The example is from the ACE-2 annotation guidelines for genericity.

an if-clause ($f13$ *AppearInIfClause*). Also the NP's number (e.g., *singular* or *plural*) and the clause tense/mood may play a role to decide genericity (Reiter & Frank, 2010). The former is detected on the basis of the POS tag of the mention's head word ($f14$ *NPNumber*). The latter is often reflected by the verb form of a clause (where the mention is present), such as "*VBG*" or "*MD VB VBG*". So we extract the POS tags of the clause verbs (from the training data) and use them as lexical features ($f15$ *VerbPosTag*).

(5.14)   If *a person* steps over the line, they must be punished.

Some NPs are commonly used generically, such as *children*, *men*, or *the dollar*. The ACE-2 corpus (Mitchell et al., 2002) (distinct from our corpus) contains annotations for genericity. We collect all NPs from ACE-2 that are always used generically ($f16$ *IsFrequentGenericNP*). We also try to learn NPs that are uniquely identifiable without further description or anaphoric links such as *the sun* or *the pope*. We do this by extracting common nouns which are annotated as `mediated/worldKnowledge` from the training set and use these as lexical features ($f17$ *GeneralWorldKnowledge*).

Finally, motivated by the ACE-2 annotation guidelines for genericity, we identify six quantifiers that may indicate genericity, i.e., *all*, *no*, *neither*, *every*, *any*, and *most* ($f18$ *PreModBy-GeneralQuantifier*).

Mention syntactic structure feature (Table 5.4, $f19$).   Feature $f19$ *HasChildMention* models that bridging anaphors most often have a simple internal structure and usually do not contain any other mentions[23].

5.3.2.4   The Full List of Non-relational Features

Table 5.6 shows all non-relational features we use for recognizing bridging anaphora under a fine-gained IS classification setting. We structure the features along two dimensions. In one dimension, we divide features into three different groups according to the extent of the discourse which the feature utilizes, i.e., *discourse level*, *sentence level*, and *NP level*. In another dimension, we associate each feature with one or more types (i.e., *surface, syntactic, semantic*) on the basis of which types of information are mainly explored to extract the feature value. *Surface* features only consider the surface forms of the tokens/strings of mentions. *Syntactic* features are extracted by exploring (dependency) parse tree information. *Semantic* features include lexical semantic features as well as features indicating certain abstract semantic aspects of NPs, such as $f33$ *PreModByCompMarker* and $f30$ *PreModByGeneralQuantifier*.

---

[23]In Chapter 3, we find that around 90% of bridging anaphors in the ISNotes corpus do not contains any other mentions.

Feature	Type	Value
Discourse level		
$f1$ FullPrevMention (n)	surface	*{yes, no, NA}*
$f2$ FullMentionTime (n)	surface	*{first, second, more, NA }*
$f3$ PartialPreMention (n)	surface	*{yes, no, NA}*
$f4$ PartialPreMentionTime (n)	surface	*{first, second, more, NA}*
$f5$ ContentWordPreMention (b)	surface	*{yes, no, NA}*
$f6$ IsCoherenceGap (b)	surface, syntactic semantic	*{yes, no}*
$f7$ IsDocFirstMention (b)	surface	*{yes, no}*
Sentence level		
$f8$ AppearInIfClause (b)	syntactic	*{yes, no}*
$f9$ VerbPosTag (l)	syntactic	e.g., *VBG*
$f10$ IsSentFirstMention (b)	surface	*{yes, no}*
$f11$ DependOnChangeVerb (b)	syntactic, semantic	*{yes, no}*
NP level		
$f12$ NPlength (int)	surface	numeric, e.g., *5*
$f13$ Unigrams (l)	surface	e.g., *his*, *the*, *China*
$f14$ Determiner (n)	syntactic	*{def, indef, dem, poss, bare, NA}*
$f15$ GrammaticalRole (n)	syntactic	*{subject, subjpass, object, predicate, pp, other}*
$f16$ NPType (n)	syntactic	*{common noun, proper noun, pronoun, other}*
$f17$ NPNumber (n)	syntactic	*{singular, plural, unknown}*
$f18$ HasChildMention (b)	syntactic	*{yes, no}*
$f19$ IsFrequentProperName (b)	semantic	*{yes, no}*
$f20$ IsArgumentTakingNP (b)	semantic	*{yes, no}*
$f22$ IsWordNetRelationalNoun (b)	semantic	*{yes, no}*
$f22$ IsInquirerRoleNoun (b)	semantic	*{yes, no}*
$f23$ SemanticClass (n)	semantic	a list of 16 classes, e.g. *time*
$f24$ IsBuildingPart (b)	semantic	*{yes, no}*
$f25$ IsSetElement (b)	semantic	*{yes, no}*
$f26$ PreModSpatialTemporal (b)	semantic	*{yes, no}*
$f27$ IsYear (b)	semantic	*{yes, no}*
$f28$ PreModifiedByCountry (b)	semantic	*{yes, no}*
$f29$ PreModByGeneralQuantifier (b)	semantic	*{yes, no}*
$f30$ IsFrequentGenericNP (b)	semantic	*{yes, no}*
$f31$ GeneralWorldKnowledge(l)	semantic	e.g., *the sun*, *the wind*
$f32$ PreModByCompMarker (b)	semantic	*{yes, no}*

Table 5.6: Non-relational feature set, "b" indicates binary features, "n" nominal features, "l" lexical features. A nominal feature draws the feature value from a restricted set. A lexical feature indicates the presence or absence of a lexicon unit in a mention.

5.4 Experiments and Results

5.4.1 Experimental Setup

We conduct experiments on the ISNotes corpus. All experiments are performed via 10-fold cross-validation on documents. We use the OntoNotes named entity and syntactic annotation for feature extraction. The value of the parameter k in the feature $f4$ *IsArgumentTakingNP* (Section 5.3.2.3) is estimated for each fold separately: we first choose ten documents randomly from the training set for each fold as the development set to estimate the value of the parameter k[24], then the whole training set is trained again using the optimized parameter. For our experiments, statistical significance is measured using McNemar's χ^2 test (McNemar, 1947). In Section 5.4.3, Section 5.4.4 and Section 5.4.5, the word *significantly* means a statistically significant difference in performance between two models at the level of $p < 0.01$.

5.4.2 Evaluation Metrics

To evaluate the performance of our model on bridging anaphora recognition, we employ several standard measures (i.e., *recall, precision, F-score, accuracy*) used in the NLP field. We use *recall*, *precision* and *F-score* to measure the performance of the model on each IS category, *accuracy* is used to measure the overall performance of the model on all IS categories. The calculations of the four measures are briefly described below:

$$\text{recall}_i = \frac{\mid \text{correctly predicted mentions for the IS class i} \mid}{\mid \text{gold mentions for the IS class i} \mid} \tag{5.15}$$

$$\text{precision}_i = \frac{\mid \text{correctly predicted mentions for the IS class i} \mid}{\mid \text{predicted mentions for the IS class i} \mid} \tag{5.16}$$

$$\text{F-score}_i = 2 \cdot \frac{\text{recall}_i \cdot \text{precision}_i}{\text{recall}_i + \text{precision}_i} \tag{5.17}$$

$$\text{accuracy} = \frac{\mid \text{correctly predicted mentions} \mid}{\mid \text{all mentions} \mid} \tag{5.18}$$

It is worth noting that there is not a best way to evaluate a model. For instance, on a highly skewed data set consisting of ten bridging anaphors (positive instances) and 90 non-bridging mentions (negative instances), a classifier which solely predicts the "negative category" would score an accuracy of 90%. This high accuracy does not reflect the performance of the classifier for identifying bridging anaphora. Instead, different metrics give us different insights into how a classification model performs. By combining all of these metrics we hope to give a representative picture of a model's performance.

---

[24]The parameter is estimated using a grid search over $k \in \{0.5, 0.6, 0.7.0.8, 0.9\}$.

5.4.3 Evaluation of New Non-relational Features

To test the effectiveness of our new non-relational features presented in Section 5.3.2.2 and Section 5.3.2.3, we reimplemented two local classifiers presented in Nissim (2006) and Rahman & Ng (2011) for IS classification as comparison baselines (hence *Nissim* and *RahmanNg*), using their feature and algorithm choices. We then add our new features (features to detect several IS classes from Table 5.3 in Section 5.3.2.2 and features to recognize bridging anaphora from Table 5.4 in Section 5.3.2.3) to the two baselines respectively. We also examine the performance of our collective classification model with only non-relational features.

5.4.3.1   Comparison With *Nissim*

The configurations of the baseline (*Nissim*) as well as our local models (built upon the baseline with more features) are described below:

Nissim. Algorithm *Nissim* is a decision tree J48[25] with standard settings in WEKA (Witten & Frank, 2005). Features $f1$ - $f7$ from Table 5.2 (Section 5.3.2.1) are used.

Nissim + newLocal1. On the basis of *Nissim*, features from Table 5.3 (Section 5.3.2.2) are added. These new features are designed for the `old` category and for the three `mediated` categories, i.e., `mediated/worldKnowledge`, `mediated/comparative`, and `mediated/function`.

Nissim + newLocal1 + newLocal2. On the basis of *Nissim + newLocal1*, features designed for the `mediated/bridging` category from Table 5.4 (Section 5.3.2.3) except two lexical features (i.e., $f15$ *VerbPosTag* and $f17$ *GeneralWorldKnowledge*) [26] are added.

Table 5.7 shows the results of the original local classifier (*Nissim*) and of the new local classifiers built on *Nissim*. The statistically significant improvement in *Nissim + newLocal1* over the original local classifier (*Nissim*) comes from the additional non-relational features from Table 5.3. In particular, comparative anaphora are recognized reliably via a small set of comparative markers (with an F-score of 83.3 in *Nissim + newLocal1*), and the inclusion of the features $f3$ *IsFrequentProperName* and $f5$ *DependOnChangeVerb* considerably improves the results for `mediated/worldKnowledge` and `mediated/function` respectively.

---

[25]The J48 algorithm is a decision tree based algorithm which is an adaptation of the popular C4.5 classifier developed by Quinlan (1993).

[26]We exclude lexical features because J48 cannot handle arbitrary lexical features well. Nissim (2006) also reported that the unigram feature turned out to negatively affect the performance of IS classification.

	Nissim			*Nissim + newLocal1*			*Nissim + newLocal1 + newLocal2*		
	R	P	F	R	P	F	R	P	F
old	85.1	82.7	83.9	85.6	82.5	84.0	85.6	85.4	85.5
med/worldKnowledge	62.3	64.4	63.3	64.2	72.0	67.8	63.3	76.3	69.2
med/syntactic	41.6	59.7	49.0	44.8	61.8	52.0	59.2	63.9	**61.4**
med/aggregate	28.4	36.8	32.1	31.8	44.7	37.1	34.6	44.5	38.9
med/function	0.0	NA	NA	38.5	89.3	53.8	58.5	76.0	**69.2**
med/comparative	0.4	7.7	0.7	84.6	82.0	83.3	83.0	78.1	80.5
med/bridging	4.4	23.0	7.4	5.3	24.5	8.9	20.7	41.5	**27.6**
new	82.7	62.3	71.1	82.0	65.4	72.8	79.7	68.7	73.8
acc		67.6			70.4			**72.6**	

Table 5.7: Experimental results: compared to the baseline *Nissim*. Bolded scores indicate significant improvements relative to all other models ($p < 0.01$).

However, we notice that the inclusion of the additional local features from Table 5.3 in *Nissim + newLocal1* fails to recognize bridging (with an F-score of only 8.9). The features that we designed to capture some common properties of bridging anaphora from Table 5.4 help *Nissim + newLocal1 + newLocal2* to improve the results for bridging and for several other IS classes substantially over *Nissim + newLocal1*. Still, the performance for recognizing bridging anaphors remains quite low compared to other IS classes.

5.4.3.2 Comparison With *RahmanNg*

The configurations of the baseline (*RahmanNg*) as well as our local models (built upon the baseline with more features) are described below:

RahmanNg. Algorithm *RahmanNg* (Rahman & Ng, 2011) utilizes a binary SVM with a composite kernel, i.e., *SVM-LIGHT-TK* (Joachims, 1999; Moschitti, 2006). It adapts the one-versus-all strategy for classifying coarse-grained IS classes (i.e., old, mediated, and new) in the Switchboard corpus[27]. Features $f1$ - $f8$ from Table 5.2 plus a tree kernel feature are

---

[27]Rahman & Ng (2011) classify coarse-grained IS classes on the Switchboard corpus. However, the same authors explore an almost totally different feature set and a new algorithm for fine-grained IS classification on the same corpus (Rahman & Ng, 2012). The feature set includes unigrams, markables and two other features heavily depending on the complex rules, which the authors manually designed on the basis of Nissim's IS annotation guidelines (Nissim et al., 2004). Since the rule-based features are data specific, we do not reimplement the whole approach in Rahman & Ng (2012) as a new baseline.

included in *RahmanNg*.

The tree kernel feature generalizes the syntactic context of a mention by extracting the mention's parent and sibling nodes without lexical leaves. Specifically, given a mention m and its corresponding syntactic constituent tree, the generalized syntactical subtree for the mention m is extracted as follows: first, find the root node $root(m)$ which spans all and only the words in m; second, find the immediate parent node of $root(m)$, i.e., $parent(root(m))$; finally, replace each leaf node in $parent(root(m))$ with a node labeled X and replace the whole subtree rooted at $root(m)$ with a leaf node labeled Y.

Figure 5.2 and Figure 5.3 show the syntactic tree for the sentence where the mention "robbers with guns" is present and the generalized syntactical tree for this mention respectively. Although this feature captures the syntactic context of a mention, it does not capture the underlying structure of the mention itself. For instances, the underlying structure of the mention "robbers with guns" is covered by the single node "Y".
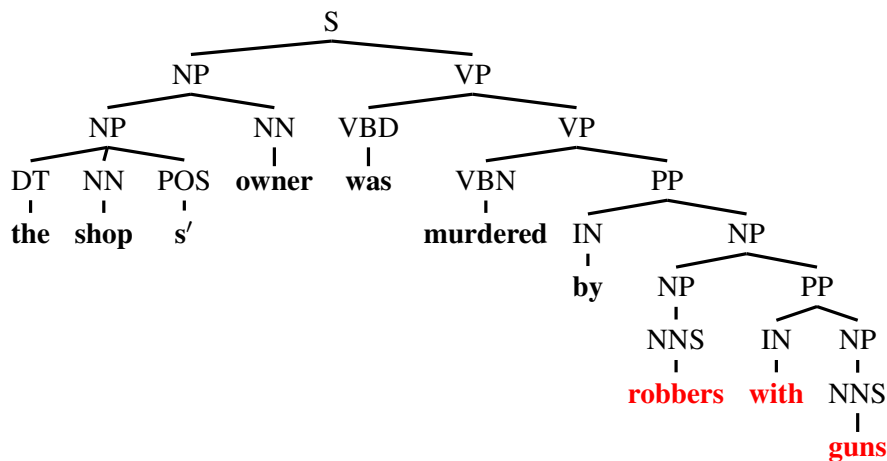
Figure 5.2: The whole syntactic tree for the sentence where the mention "robbers with guns" is present.

Figure 5.3: The generalized syntactic tree for the mention "robbers with guns".

	RahmanNg			RahmanNg + newLocal1			RahmanNg + newLocal1 + newLocal2		
	R	P	F	R	P	F	R	P	F
old	85.3	87.1	86.2	85.7	86.9	86.3	86.8	86.7	86.8
med/worldKnowledge	66.6	69.6	68.0	67.1	73.5	70.1	64.9	81.2	**72.2**
med/syntactic	57.3	72.2	63.9	55.8	72.8	63.2	66.3	71.7	**68.9**
med/aggregate	26.5	75.7	39.3	25.1	73.6	37.5	29.4	78.5	**42.8**
med/function	24.6	51.6	33.3	56.9	84.1	67.9	44.6	85.3	58.6
med/comparative	26.5	85.9	40.5	79.4	81.7	80.6	79.1	81.0	80.0
med/bridging	11.6	45.6	18.5	8.9	44.7	14.8	12.4	61.2	20.6
new	87.8	66.7	75.8	87.6	67.6	76.3	87.4	70.0	**77.8**
acc		73.3			74.4			**76.2**	

Table 5.8: Experimental results: compared to the baseline *RahmanNg*. Bolded scores indicate significant improvements relative to all other models ($p < 0.01$).

***RahmanNg + newLocal1*.** On the basis of *RahmanNg*, features from Table 5.3 (Section 5.3.2.2) are added. These new features are designed for the `old` category and for the three `mediated` categories, i.e., `mediated/worldKnowledge`, `mediated/comparative`, and `mediated/function`.

***RahmanNg + newLocal1 + newLocal2*.** On the basis of *RahmanNg + newLocal1*, features designed for the `mediated/bridging` category from Table 5.4 (Section 5.3.2.3) are added.

Table 5.8 shows the comparison of the new local classifiers built on Rahman & Ng's framework (2011) to the original local classifier (*RahmanNg*). The continuous improvements for the overall IS classification performance using the additional non-relational features in *RahmanNg + newLocal1* and *RahmanNg + newLocal1 + newLocal2* have a similar pattern as in *Nissim + newLocal1* and *Nissim + newLocal1 + newLocal2*.

However, it seems that the new features designed for bridging only have a limited effect in *RahmanNg + newLocal1 + newLocal2* compared to *RahmanNg*. This is because unigrams explored in *RahmanNg* potentially encapsulate some of the lexical knowledge for bridging anaphora recognition. We also observe that although the overall IS classification performance of *RahmanNg + newLocal1 + newLocal2* is significantly better than *Nissim + newLocal1 +*

newLocal2, the former is worse than the latter with regard to bridging anaphora recognition. We assume that the standard one-versus-all strategy for a multi-class setting explored by Rahman & Ng (2011) is not suitable for identifying a minority class which lacks strong indicators.

5.4.3.3   Collective Classification With Different Non-relational Features

We further evaluate our collective classification model based on Markov logic networks with only non-relational features. The purpose of this experiment is to assess the effects of non-relational features under our collective classification framework described in Section 5.2.1. We therefore exclude all relational features described in Section 5.3.1. In fact, the collective classification models in this section are equal to their non-relational counterparts when relational features are not provided. The configurations of the collective classification models with different non-relational features are described below:

Collective_local_I.   The model *Collective_local_I* is based on Markov logic networks, with only non-relational features (from previous work) from Table 5.2 (Section 5.3.2.1). We use *thebeast* to learn weights for the formulas and to perform inference.

Collective_local_II.   On the basis of *Collective_local_I*, new non-relational features from Table 5.3 (Section 5.3.2.2) are added. These new features are designed for the `old` category and for the three `mediated` categories, i.e., `mediated/worldKnowledge`, `mediated/comparative`, and `mediated/function`.

Collective_local_III.   On the basis of *Collective_local_II*, new non-relational features designed for the `mediated/bridging` category from Table 5.4 (Section 5.3.2.3) are added. This model is equal to the collective classification model described in Section 5.2.1 without relational features.

Table 5.9 shows the results of our collective classification algorithm with different non-relational feature sets. Using the additional non-relational features in *Collective_local_II* and *Collective_local_III* improves the overall IS classification performance significantly compared to the models with less features. We also notice that *Collective_local_III* performs best with regard to bridging anaphora recognition among all local classifiers.

	Collective_local_I			Collective_local_II			Collective_local_III		
	R	P	F	R	P	F	R	P	F
old	84.8	83.4	84.1	85.0	83.7	84.3	85.2	84.5	84.8
med/worldKnowledge	57.9	62.0	59.9	60.2	75.5	67.0	64.7	82.3	**72.4**
med/syntactic	63.1	65.3	64.2	63.5	66.7	65.1	70.9	67.7	**69.3**
med/aggregate	57.3	59.6	58.5	56.9	56.6	56.7	61.6	63.4	**62.5**
med/function	3.1	25.0	5.5	52.3	82.9	64.2	56.9	86.0	**68.5**
med/comparative	57.7	67.3	62.1	79.8	83.5	81.6	82.6	83.3	82.9
med/bridging	24.1	35.1	28.6	21.7	40.6	28.3	25.5	49.0	**33.5**
new	76.9	70.4	73.5	80.0	70.2	74.8	80.8	72.9	**76.6**
acc		71.2			73.3			**75.5**	

Table 5.9: Experimental results: collective classification with only non-relational features. Bolded scores indicate significant improvements relative to all other models ($p < 0.01$).

5.4.4 Evaluation of Collective Classification

In the last section we showed the effectiveness of our non-relational features for recognizing bridging and several other IS categories under different algorithms. However, all algorithms from the last section are "local" classifiers in the sense that they predict the IS class for each instance (mention) separately. In this section, we compare the two best local classifiers (i.e., *RhamanNg + newLocal1 + newLocal2* and *Collective_local_III*) to our collective classification model (*Collective*). *Collective* is the model based on Markov logic networks, with all relational features from Section 5.3.1 and all non-relational features from Section 5.3.2 (i.e., Table 5.2, Table 5.3, and Table 5.4). The details of this model are described in Section 5.2.1. We use *thebeast* to learn weights for the formulas and to perform joint inference.

Table 5.10 shows the results of the collective classification model and of the two best local classifiers. The relational features we explored in *Collective* lead to a significant improvement in accuracy over all local classifiers. The improvement is centered on the categories of `mediated/syntactic` and `mediated/aggregate` as well as their distinctions from `new`. Such improvement is in accordance with the linguistic relations among IS categories we analyzed in Section 5.2.1.

	RahmanNg + newLocal1 + newLocal2			Collective_local_III			Collective		
	R	P	F	R	P	F	R	P	F
old	86.8	86.7	86.8	85.2	84.5	84.8	85.7	84.7	85.2
med/worldKnowledge	64.9	81.2	72.2	64.7	82.3	72.4	64.2	80.5	71.4
med/syntactic	66.3	71.7	68.9	70.9	67.7	69.3	82.7	80.1	**81.4**
med/aggregate	29.4	78.5	42.8	61.6	63.4	62.5	71.6	78.2	**74.8**
med/function	44.6	85.3	58.6	56.9	86.0	68.5	56.9	90.2	69.8
med/comparative	79.1	81.0	80.0	82.6	83.3	82.9	81.4	84.8	83.1
med/bridging	12.4	61.2	20.6	25.5	49.0	33.5	25.9	49.9	34.1
new	87.4	70.0	77.8	80.8	72.9	76.6	84.5	75.7	**79.9**
acc		76.2			75.5			**78.9**	

Table 5.10: Experimental results: comparing the collective classifier to the local classifiers. Bolded scores indicate significant improvements relative to all other models ($p < 0.01$).

5.4.5 Evaluation of Cascading Collective Classification

In the last section, we show that our collective classification model (*Collective*) improves the result over the local classifiers significantly. However, the improvement comes from other IS categories than bridging. Although the new non-relational features we proposed to identify bridging (Section 5.3.2.3) show positive effects on our target IS class (i.e., `mediated/bridging`) in all local classifiers (i.e., *Nissim + newLocal1 + newLocal2*, *RahmanNg + newLocal1 + newLocal2* and *Collective_local_III* in Section 5.4.3), the result of `mediated/bridging` is still low. One reason is the relative rarity of bridging compared to many other IS classes. In a multi-class setting, prediction is biased toward the classes with the highest priors. Our cascading collective classification model (*CascadedCollective*) aims to address this problem. *CascadedCollective* is the system described in Section 5.2.2 with all non-relational features (Section 5.3.2) for the minority binary classifiers (based on SVMs) and all features (Section 5.3.1 and Section 5.3.2) for the collective classifier[28] (based on MLNs). We compare it with *Collective* as well as with *CascadedCollective - bridgingFeat*. The latter is based on *CascadedCollective* but removes the new non-relational features for bridging anaphora recognition both from the minority binary classifiers and from the collective classifier. We summarize the configurations for these three models in Table 5.11.

---

[28]The collective classifier in *CascadedCollective* is the same as *Collective*.

Model	Algorithm	Features
Collective	MLNs	relational features (Section 5.3.1)
		non-relational features from previous work (Table 5.2)
		non-relational features for some IS classes (Table 5.3)
		non-relational features for bridging anaphora (Table 5.4)
CascadedCollective	SVMs	non-relational features from previous work (Table 5.2)
		non-relational features for some IS classes (Table 5.3)
		non-relational features for bridging anaphora (Table 5.4)
	MLNs	relational features (Section 5.3.1)
		non-relational features from previous work (Table 5.2)
		non-relational features for some IS classes (Table 5.3)
		non-relational features for bridging anaphora (Table 5.4)
CascadedCollective - bridgingFeat	SVMs	non-relational features from previous work (Table 5.2)
		non-relational features for some IS classes (Table 5.3)
	MLNs	relational features (Section 5.3.1)
		non-relational features from previous work (Table 5.2)
		non-relational features for some IS classes (Table 5.3)

Table 5.11: Configurations for different models for IS classification.

Table 5.12 shows the results of our cascading collective classification system and the other two models described above. Compared to the collective classification model *Collective*, using the cascading collective classification system *CascadedCollective* improves bridging results substantially while the performance on other categories does not worsen. The algorithm needs both the features for bridging anaphora recognition (Section 5.3.2.3) as well as the cascaded modeling (Section 5.2.2) to achieve this improvement as the comparison to *CascadedCollective - bridgingFeat* shows: the latter lowers overall accuracy as it tends to overgenerate rare classes (including bridging) with low precision if the features are not strong enough. Our novel features (addressing linguistic properties of bridging) and the cascaded algorithm (addressing data sparseness) appear to be complementary.

	collective			CascadedCollective			CascadedCollective - bridgingFeat		
	R	P	F	R	P	F	R	P	F
old	85.7	84.7	85.2	83.1	85.7	84.4	79.8	85.3	82.5
med/worldKnowledge	64.2	80.5	71.4	65.6	79.5	71.9	64.3	73.3	68.5
med/syntactic	82.7	80.1	81.4	82.2	80.5	81.3	77.4	78.6	78.0
med/aggregate	71.6	78.2	74.8	71.1	77.7	74.3	66.8	75.4	70.9
med/function	56.9	90.2	69.8	61.5	83.3	70.8	61.5	80.0	69.6
med/comparative	81.4	84.8	83.1	83.4	82.7	83.1	81.0	82.3	81.7
med/bridging	25.9	49.9	34.1	48.7	43.8	**46.1**	35.7	24.7	29.2
new	84.5	75.7	79.9	81.3	77.7	79.5	77.6	75.9	76.8
acc		78.9			78.4			74.4	

Table 5.12: Experimental results for bridging anaphora recognition: comparing the cascading collective classifier to the collective classifier. The bolded score indicates a significant improvement relative to all other models ($p < 0.01$).

5.4.6   Feature Analysis for Bridging Anaphora Recognition

To assess the impact of features for bridging anaphora recognition in our best system (*CascadedCollective*), we perform a feature ablation study in which each main feature group in Section 5.3.2.3 (i.e., *Lexico-semantics*, *Discourse structure* and *Identifying generic NPs* in Table 5.4) is removed from *CascadedCollective* in turn. The results in Table 5.13 show that semantic features have the most impact. Discourse structure and genericity information features have less of an impact. We believe the latter to be due to noise involved in extracting these features (such as approximating coreference for the coherence gap feature) as well as genericity recognition still being in its infancy (Reiter & Frank, 2010).

	F-score for bridging
CascadedCollective	46.1
- Discourse structure	43.3
- Lexico-semantics	**33.7**
- Identifying generic NPs	42.2

Table 5.13: Results of feature ablation experiments for bridging anaphora recognition. The bolded score indicates a significant difference compared to *CascadedCollective* ($p < 0.01$)

5.4.7 Error Analysis for Bridging Anaphora Recognition

To gain a better understanding of the performance for bridging anaphora recognition in our best system (*CascadedCollective*) , we analyze the results of recognizing bridging anaphors from different perspectives.

First, we examine the confusion matrix (Table 5.14) of our best model *CascadedCollective*. Table 5.14 only shows the numbers which are related to bridging.

$C \rightarrow$ $G \downarrow$	old	new	brid	syn	comp	aggr	func	know
old	-	-	175	-	-	-	-	-
new	-	-	193	-	-	-	-	-
brid	66	251	323	10	2	1	0	10
synt	-	-	10	-	-	-	-	-
comp	-	-	2	-	-	-	-	-
aggr	-	-	0	-	-	-	-	-
func	-	-	0	-	-	-	-	-
know	-	-	35	-	-	-	-	-

Table 5.14: Confusion matrix of *CascadedCollective* for bridging anaphora recognition. "C" indicates classifier tags, "G" gold tags. "brid" stands for *mediated/bridging*, "syn" *mediated/syntactic*, "comp" *mediated/comparative*, "aggr" *mediated/aggregate*, "func" *mediated/function*, "know" *mediated/worldKnowledge*.

We observe that the highest proportion of recall errors is due to the fact that 251 bridging anaphors are misclassified as `new`. Also the precision errors mostly stem from `new` and `old` mentions which are misclassified as `mediated/bridging`. Among 175 `old` mentions which are misclassified as `mediated/bridging`, most of them are definite NPs with simple syntactic structures (i.e., *the + head*, such as *the president* or *the economy*) and do not have the same string as a previous mention. Also among 193 `new` mentions which are misclassified as `mediated/bridging`, 96 cases are "bare" NPs (NPs whose head is not modified by any determiners) with simple syntactic structure, such as *control* or *property owners*, 60 cases are definite NPs with simple structure, such as *the food supply* or *the back*. These results reveal that the lexical semantic knowledge we explored is not adequate to capture bridging. Indeed, such knowledge only indicates that some NPs are more likely to be used as bridging anaphora than others. As the previous examples have shown ("friends" in Example 5.12 and 5.13), the NP's IS also depends on how it is embedded into the discourse, which is only partially modeled in our approach.

Next, we investigate the performance of our model (*CascadedCollective*) on recognizing

bridging anaphors modified by *the* and other bridging anaphors (bridging anaphors not modified by *the*) respectively. Accordingly, we evaluate the model's performance on bridging anaphora recognition on NPs modified by *the* and other NPs (NPs not modified by *the*) separately. The results in Table 5.15 show that **recognizing definite bridging anaphors modified by *the* is harder than recognizing bridging anaphors which are not modified by *the***.

Bridging Anaphora Type	R	P	F
All	48.7	43.8	46.1
Anaphors modified by *the*	53.0	35.1	42.2
Anaphors not modified by *the*	45.8	53.7	49.4

Table 5.15: Results of bridging anaphora recognition in *CascadedCollective* with regard to determiners.

To understand the reasons for this, we create a set of mentions (*filtered mentions*) which likely contain bridging anaphors while excluding obvious non-bridging anaphors. A mention is added to *filtered mentions* if it: (1) is not modified by demonstrative determiners; and (2) is a common noun; and (3) does not contain any other mentions; and (4) is not modified by comparative markers. We then analyze the main IS distributions for definite filtered mentions (*filtered mentions* modified by *the*) and other filtered mentions (*filtered mentions* not modified by *the*) separately. Table 5.16 shows that in the group of *definite filtered mentions*, `old` and `bridging` are the two main categories, whereas in the group of *other filtered mentions*, `new` and `bridging` are the two main groups.

IS class	Definite filtered mentions		Other filtered mentions	
`old`	502	(51.6%)	48	(2.1%)
`mediated/bridging`	213	(21.9%)	319	(14.1%)
`new`	158	(16.2%)	1,825	(80.8%)
`mediated/worldKnowledge`	97	(10.0%)	31	(1.4%)

Table 5.16: IS distribution for different groups of *filtered mentions*. *filtered mentions* are mentions that: (1) are not modified by demonstrative determiners; and (2) are common nouns; and (3) do not contain any other NPs; and (4) are not modified by comparative markers.

We further examine how some "surface" features behave for recognizing bridging anaphors in these two groups. Table 5.17 shows that two surface features (*argumentTakingRatio*[29] *> 0.5* and *semanticClass = rolePerson*) can cover most of bridging anaphors in the group of

---

[29]The argument taking ratio of a mention is calculated by exploring NomBank. See *f4 IsArgumentTakingNP* in Section 5.3.2.3 for the detailed information about how to calculate a mention's argument taking ratio.

Surface Feature	Bridging anaphora in *definite filtered mentions*		Bridging anaphora in *other filtered mentions*	
argumentTakingRatio > 0.5	86	(40.4%)	229	(71.8%)
semanticClass = rolePerson	9	(4.2%)	48	(15.0%)
Total	213	(100.0%)	319	(100.0%)

Table 5.17: Distribution of bridging anaphora in different groups of *filtered mentions* w.r.t. different conditions.

other filtered mentions. It seems that most bridging anaphors in the group of *definite filtered mentions* are hard to recognize by just using surface features. Table 5.18 shows some examples of definite bridging anaphors that are not covered by these two surface features. Actually, these definite bridging anaphors need to be put into context to understand their "associative anaphor" usages.

Examples of definite bridging anaphors from the *definite filtered mentions* group that are not rolePersons and whose argumentTakingRatios < 0.5

the scars, the woman, the cliche, the local council, the paralegal, the guys, the cafeteria, the hill, the desks, the sand, the diamonds, the south, the board, the characters, the screening plants, the ride, the few exceptions, the large and lucrative market, the market, the opening show, the takeover, the building, the state, the firms, the final package, the expensive unknowns, the city, the wait, the best yearlings, the right focus, the next truck, the phrases, the first time, the justices, the trees, …

Table 5.18: Definite filtered bridging anaphora examples.

5.5   Summary

In this chapter, we have presented a cascading collective model for bridging anaphora recognition within a multi-class classification setting. The model is motivated by the linguistic properties of the task, i.e., linguistic relations among several IS categories, the wide variation of bridging and its relative rarity compared to many other IS categories. The model combines the binary classifiers for minority categories and a collective classifier for all IS categories in a cascaded way. The system addresses the multi-class imbalance problem (for rare categories without strong indicators) within a multi-class setting while still keeping the strength of the collective classifier by exploring relational autocorrelation (for several IS cat-

egories in which such relational autocorrelation exists). Our system achieves substantial improvements both for the overall IS classification accuracy as well as for bridging anaphora recognition over the reimplementations of the two previous systems (Nissim, 2006; Rahman & Ng, 2011). We show that the improvements come from both the new features we designed to recognize bridging (Section 5.3.2.3) as well as our cascaded modeling (Section 5.2.2).

Although we are interested in recognizing bridging anaphora, our approach produces fine-grained IS classification as an additional outcome. IS has been claimed to be beneficial for a number of NLP tasks, though the results have been mixed. Nenkova et al. (2007) used IS as a feature for generating pitch accent in conversational speech. Since IS is restricted to noun phrases, pitch accent, however, can be assigned to any word in an utterance, the experiments were not conclusive. For determining constituent order of German sentences, Cahill & Riester (2009) achieve significant improvement over the baseline by incorporating features modeling IS. Rahman & Ng (2011) showed that IS is a useful feature for coreference resolution.

In the next chapter, we turn our attention to the second subtask of bridging resolution: antecedent selection for bridging anaphora.

Chapter 6

Antecedent Selection for Bridging Anaphora

As a subtask of bridging resolution, *antecedent selection for bridging anaphora* or *bridging anaphora resolution* is the main focus of most previous empirical research on bridging resolution (Poesio & Vieira, 1998; Poesio et al., 2004a; Markert et al., 2003; Lassalle & Denis, 2011). However, previous work on this subtask is restricted. It makes untested assumptions about bridging anaphora types or bridging relations, such as limiting it to definite NPs (Poesio & Vieira, 1998; Poesio et al., 2004a; Lassalle & Denis, 2011) or to part-of bridging relations (Poesio et al., 2004a; Markert et al., 2003; Lassalle & Denis, 2011).

In this chapter, we present a joint inference model to select antecedents for bridging anaphora. We break new ground for this subtask by considering all bridging relations and anaphora types. Our main contributions lie in the following aspects: (1) we model the *sibling anaphors clustering* (i.e., syntactically or semantically related anaphors are likely to be *sibling anaphors*, and hence share the same antecedent) and *bridging anaphora resolution* jointly; (2) we develop novel semantic, syntactic and salience features based on linguistic intuition to capture various bridging relations; and (3) we propose a new method to select antecedent candidates for bridging anaphors by exploring discourse relation `Expansion` and by modeling salience from different perspectives. The method reflects the different interpretive preferences (*local* or *global* focus) of bridging anaphors.

The work described in this chapter is an extension of our previous work (Hou et al., 2013b). The chapter is organized as follows. Section 6.1 discusses the related empirical work and defines the task. Section 6.2 details our joint inference model for this problem, followed by two sections that focus on linguistically motivated features we designed for this task (Section 6.3) and the new method to select antecedent candidates for bridging anaphors (Section 6.4) respectively. Section 6.5 reports the results achieved under different settings in ISNotes and analyzes the results from different aspects. Finally, Section 6.6 summarizes this chapter.

6.1   Task

Background.   Previous work on automatic bridging anaphora resolution suffers from focusing on subproblems, e.g., only part-of bridging (Poesio et al., 2004a; Markert et al., 2003) or definite NP anaphora (Poesio et al., 2004a; Markert et al., 2003; Lassalle & Denis, 2011). Also the evaluation setup is sometimes not clear: the high results in Poesio et al. (2004a) cannot be used for comparison as the evaluation is unrealistic, i.e., the authors distinguish only between the correct antecedent and *one* or *three* false candidates (baseline of 50% for the former)[1]. They also restrict the phenomenon to part-of relations. Lassalle & Denis (2011) use a similar method to Poesio et al. (2004a) to resolve bridging anaphora in a French corpus where the anaphors are restricted to definite descriptions and bridging relations are limited to meronymic relations. In contrast to Poesio et al. (2004a), they evaluate the model in a realistic setting and report an accuracy of 23%.

There is a partial overlap between bridging resolution and implicit semantic role labeling (Ruppenhofer et al., 2010)[2].   However, work on implicit semantic role labeling is mostly focused on few predicates and/or assumes that the gold standard local semantic argument structure (i.e., target words, frames, locally realized semantic roles and their fillers) is provided (e.g., Silberer & Frank (2012), Laparra & Rigau (2012), Gerber & Chai (2012), Laparra & Rigau (2013)). We consider all bridging anaphors in running text.

Task definition and evaluation metric.   *Antecedent selection for bridging anaphora* or *bridging anaphora resolution* is the task of choosing antecedents among candidates for the given bridging anaphors. In ISNotes, bridging is annotated mostly between a mention (bridging anaphor) and an entity (antecedent)[3], so that a bridging anaphor could have multiple links to different instantiations of the same entity (entity information is based on the OntoNotes coreference annotation). In this thesis, we only consider resolving bridging anaphors with entity antecedents, leaving resolving bridging anaphors with non-entity antecedents for future work.

For this task, we measure *accuracy* on the basis of the number of bridging anaphors, instead of on the basis of all links between bridging anaphors and their antecedent instantiations. We calculate how many bridging anaphors are correctly resolved by a model among all bridging anaphors. In the mention-entity setting (bridging anaphor – entity antecedent) where the

---

[1]They also report relatively high results in the "hard" test in which they try to find the correct antecedents for six bridging anaphors among all possible candidates. We think cross-validations would be more appropriate considering that a test set containing only six cases is rather small.

[2]A detailed description of implicit semantic role labeling and the analysis of differences between bridging resolution and implicit semantic role labeling can be found in Section 2.2 in Chapter 2.

[3]There are a few cases where bridging is annotated between an NP and a non-entity antecedent (e.g., *verbs* or *clauses*).

gold entity information is given, a bridging anaphor is counted as correctly resolved by a model if the model links the anaphor to its entity antecedent. In the mention-mention setting (bridging anaphor – mention antecedent) where the gold entity information is not given, a bridging anaphor is counted as correctly resolved by a model if the model links the anaphor to one of its preceding antecedent instantiations.

6.2   Model

Motivation for the model: the sibling anaphors phenomenon.   In Chapter 3, we analyzed the sibling anaphors phenomenon in detail. Sibling anaphors are bridging anaphors that share the same antecedent, while "non-siblings" are anaphors that do not share an antecedent with any other anaphor. For instance, in Example 6.1, **The windows**, **The carpets** and **walls** are sibling anaphors. In ISNotes, we found that most bridging anaphors (61.4%) are sibling anaphors, and globally salient antecedents are likely to connect to sibling anaphors compared to the locally salient antecedents. Such sibling anaphors phenomenon (*sibling anaphors clustering*) can help our target task (*bridging anaphora resolution*), i.e., sibling anaphors should be resolved to the same antecedent. Therefore we explore a joint inference approach to model *sibling anaphors clustering* and *bridging anaphora resolution* together.

(6.1)   If Mr. McDonough's plans get executed, as much as possible of *the Polish center* will be made from aluminum, steel and glass recycled from Warsaw's abundant rubble. **The windows** will open. **The carpets** won't be glued down and **walls** will be coated with non-toxic finishes.

Detailed model.   We use A to denote the set of n bridging anaphors in a document D, and E to denote the set of antecedent candidate entities in the whole document. Let c_{a_i/a_j} be a sibling anaphors clustering assignment for bridging anaphors $a_i, a_j \in A$, C_A be a sibling anaphors clustering result for all bridging anaphors in A, C_A^n be the set of all possible sibling anaphors clustering results for A; e_a be an antecedent assignment for a bridging anaphor $a \in A$, E_A be an antecedent assignment result for all bridging anaphors in A, E_A^n be the set of all possible antecedent assignment results for A. The joint inference model for *sibling anaphors clustering* and *bridging anaphora resolution* can be represented as a log-linear model:

$$P(C_A, E_A | A; w) = \frac{\exp(w \cdot \Phi(A, C_A, E_A))}{\sum_{E_A' \in E_A^n, C_A' \in C_A^n} \exp(w \cdot \Phi(A, C_A', E_A'))} \tag{6.2}$$

where w is the model's weight vector, $\Phi(A, C_A, E_A)$ is a "global" feature vector which takes the entire clustering and antecedent assignments for all bridging anaphors in A into account. We define $\Phi(A, C_A, E_A)$ as:

$$\Phi(A, C_A, E_A) = \sum_{l \in F_c} \sum_{a_i, a_j \in A} \Phi_l(a_i, a_j, c_{a_i/a_j}) + \sum_{k \in F_r} \sum_{a \in A} \Phi_k(a, e_a)$$
$$+ \sum_{g \in F_g} \sum_{a_i, a_j \in A} \Phi_g(c_{a_i/a_j}, e_{a_i}, e_{a_j}) \tag{6.3}$$

where $\Phi_l(a_i, a_j, c_{a_i/a_j})$ and $\Phi_k(a, e_a)$ are two local feature functions for *sibling anaphors clustering* and *bridging anaphora resolution* respectively. The former looks at two bridging anaphors a_i and a_j whereas the latter looks at the bridging anaphor a and the antecedent candidate e_a. $\Phi_g(c_{a_i/a_j}, e_{a_i}, e_{a_j})$ is a global feature function that looks at the antecedent assignments for a_i and a_j at the same time.

Like our collective classification model presented in the previous chapter (Section 5.2.1), this log-linear model can also be represented using Markov logic networks (MLNs) which were described in Section 4.1.1. In a ground Markov network for this task, the probability distribution over the possible world C_A, E_A is given by

$$P(C_A, E_A) = \frac{1}{Z} \exp \left(\sum_i w_i n_i(C_A, E_A) \right) \tag{6.4}$$

where $n_i(C_A, E_A)$ is the number of true groundings of a local or a global feature function F_i in C_A, E_A. We use *thebeast*[4] to learn weights for the formulas and to perform inference. *thebeast* employs cutting plane inference (Riedel, 2008) to improve the accuracy and efficiency of MAP inference for MLNs.

Table 6.1 shows the formulas and the formula templates that we design to model this problem in MLNs. $p1$ and $p2$ are two hidden predicates that we try to predict, i.e., choosing the antecedent for the bridging anaphor $a1$ and deciding whether bridging anaphors a_1 and a_2 are sibling anaphors. $f1$ and $f2$ model that each bridging anaphor can have at most one antecedent[5] and the bridging anaphor should not appear before its antecedent[6] respectively. $f3$ and $f4$ model the reflectivity and transitivity of *sibling anaphors clustering*. $f5$ and $f6$ model the mutual relations between the two hidden predicates, i.e., sibling anaphors share the same antecedent.

f_c is the formula template for *sibling anaphors clustering*, f_{r1} and f_{r2} are the formula templates for *bridging anaphora resolution*. The details of specific formulas instantiating f_c and f_{r1}/f_{r2} are described in Section 6.3.1 and Section 6.3.2 accordingly. In formulas which

---

[4] http://code.google.com/p/thebeast

[5] Since we only consider entity antecedents, bridging anaphors with non-entity antecedents do not have antecedents in our model. We also do not model that some bridging anaphors have multiple entity antecedents (see Example 3.5 in Chapter 3), because this happens rarely.

[6] Specifically, a bridging anaphor should not appear before the first instantiation of its entity antecedent.

Hidden predicates	
$p1$	$isBridging(a_1, e)$
$p2$	$hasSameAntecedent(a_1, a_2)$
Formulas	
Hard constraints	
$f1$	$\forall a \in A : \; |e \in E : isBridging(a, e)| \leq 1$
$f2$	$\forall a \in A \; \forall e \in E : \; hasPairDistance(e, a, d) \wedge d < 0 \rightarrow \neg isBridging(a, e)$
$f3$	$\forall a_1, a_2 \in A : \; a_1 \neq a_2 \wedge hasSameAntecedent(a_1, a_2) \rightarrow hasSameAntecedent(a_2, a_1)$
$f4$	$\forall a_1, a_2, a_3 \in A : \; a_1 \neq a_2 \wedge a_1 \neq a_3 \wedge a_2 \neq a_3 \wedge hasSameAntecedent(a_1, a_2)$ $\wedge hasSameAntecedent(a_2, a_3) \rightarrow hasSameAntecedent(a_1, a_3)$
$f5$	$\forall a_1, a_2 \in A \; \forall e \in E : \; a_1 \neq a_2 \wedge hasSameAntecedent(a_1, a_2) \wedge isBridging(a_1, e)$ $\rightarrow isBridging(a_2, e)$
$f6$	$\forall a_1, a_2 \in A \; \forall e \in E : \; a_1 \neq a_2 \wedge isBridging(a_1, e) \wedge isBridging(a_2, e)$ $\rightarrow hasSameAntecedent(a_1, a_2)$
Formula template for sibling anaphors clustering	
f_c	$\forall a_1, a_2 \in A : \; \text{siblingAnaphorsClusteringFormula\_Template} (a_1, a_2)$ $\rightarrow hasSameAntecedent(a_1, a_2)$
Formula template for bridging anaphora resolution	
f_{r1}	$\forall a \in A \; \forall e \in E : \; \text{bridgingAnaResolutionFormula\_Template1} (a, e) \rightarrow isBridging(a, e)$
f_{r2}	$\forall a \in A \; \forall e \in E_a : \; \text{bridgingAnaResolutionFormula\_Template2} (a, e) \rightarrow isBridging(a, e)$

Table 6.1: Hidden predicates and formulas used for bridging anaphora resolution. a_1, a_2, a_3 represent bridging anaphors, A the set of bridging anaphors in the whole document, e the antecedent candidate entity, E_a the set of the antecedent candidate entities for a according to a's discourse scope, and E the set of antecedent candidate entities in the whole document.

instantiate f_{r2}, the set of the antecedent candidates (E_a) for each bridging anaphor a is constructed on the basis of the anaphor's discourse scope (i.e., *local* or *non-local*). We describe this method in detail in Section 6.4.

6.3   Feature Design

In this section, we detail all features that we design for our joint inference model described in the previous section. Section 6.3.1 describes the features instantiating the formula template f_c for *sibling anaphors clustering* in Table 6.1, whereas Section 6.3.2 explains the features instantiating the formula templates f_{r1} and f_{r2} for *bridging anaphora resolution* in Table 6.1.

6.3.1   Features for Sibling Anaphors Clustering

Table 6.2 shows the formulas used to predict sibling anaphors. Each formula is associated with a weight w learned from the training data. The polarity of the weights is indicated by the leading $+$ or $-$. Formulas $f1$-$f3$ explore two different ways (syntactic and semantic) to predict sibling anaphors.

Formulas for *sibling anaphors clustering*			
f1	$+$	(w)	$\forall a_1, a_2 \in A \ syntacticParallelStructure(a_1, a_2)$
			$\rightarrow hasSameAntecedent(a_1, a_2)$
f2	$+$	(w)	$\forall a_1, a_2 \in A \ sameHead(a_1, a_2) \rightarrow hasSameAntecedent(a_1, a_2)$
f3	$+$	(w)	$\forall a_1, a_2 \in A \ relatedTo(a_1, a_2) \rightarrow hasSameAntecedent(a_1, a_2)$

Table 6.2: Formulas used for sibling anaphors clustering. a_1, a_2 represent bridging anaphors, A the set of bridging anaphors in the whole document, and w the weight learned from the data for the specific formula.

$f1$ captures that syntactically parallel bridging anaphors are likely to be sibling anaphors. We define bridging anaphors a_1 and a_2 to be syntactically parallel if (i) a_1 and a_2 are the children of a coordination structure (e.g., **business manager**, **bookkeeper** and **publicist** in Example 6.5); or (ii) $a1$ and $a2$ are both subjects/objects of the verbs in the conjoined clauses (e.g., **Three** and **two** in Example 6.6).

(6.5)   When her husband and son founded *their computer company, Vesoft*, she worked as **business manager**, **bookkeeper** and **publicist**.

(6.6)   Back in 1964, the FBI had *five black agents*. **Three** were chauffeurs for J. Edgar Hoover, and **two** cleaned his house.

Semantically related bridging anaphors are likely to be sibling anaphors. $f2$ models this via surface forms, i.e., bridging anaphors sharing the same head word are sibling anaphors (e.g., two occurrences of **residents** in Example 6.7).

In $f3$, we explore SVM^{light} to predict semantically related bridging anaphors which do not share the same head word (such as **limited access** and **one last entry** in Example 6.7), on the basis of the two types of features. In the first type of features, we use WordNet-based similarity measures implemented by Pedersen et al. (2004) to calculate similarity scores between the head words of bridging anaphors. In the second type of features, we measure the distance between the head words of bridging anaphors.

(6.7)   After being inspected, *buildings with substantial damage* were color-coded. Green allowed **residents** to re-enter; yellow allowed **limited access**; red allowed **residents one last entry** to gather everything they could within 15 minutes.

6.3.2 Features for Bridging Anaphora Resolution

Table 6.3 shows the formulas used for bridging anaphora resolution. Each formula is associated with a weight w learned from the training data. The polarity of the weights is indicated by the leading $+$ or $-$. For some formulas the final weight consists of a learned weight w multiplied by a score d (e.g., the inverse distance between antecedent and anaphor). In these cases, the final weight for a ground formula in a ground Markov network does not just depend on the respective formula, but also on the specific constants. We indicate such combined weights by the term $w \cdot d$.

All numeric features (i.e., $f5$, $f7$, and $f11$) in Table 6.3 are normalized among all antecedent candidates of one anaphor. Given a bridging anaphor a_i, its antecedent candidate set E_{a_i} ($e_{ij} \in E_{a_i}$) and the numeric score S_{ik} for the pair $\{a_i, e_{ik}\}$, the normalized value of S_{ik} (i.e., $NormS_{ik}$) is calculated (set to values between 0 and 1) as below:

$$NormS_{ik} = \frac{S_{ik} - \min_j S_{ij}}{\max_j S_{ij} - \min_j S_{ij}} \tag{6.8}$$

The other variants of numeric features (i.e., $f6$, $f8$, and $f12$) tell whether the score of an anaphor-antecedent candidate pair is the highest among all pairs for this anaphor.

We now describe the features as well as their intuitions in the following.

6.3.2.1 Frequent Bridging Relations

Four common bridging relations could be captured by the semantic classes of anaphor and antecedent (Table 6.3: $f1$-$f4$). It is worth noting that in formulas $f1$ and $f2$ (modeling that a bridging anaphor with the semantic class *role person* prefers GPE or organization antecedents), we do not penalize the antecedent candidates that are far away from the anaphor. This is because in news articles, it is common that a globally salient GPE or organization entity is introduced in the beginning, then later an NP denoting the related roles (such as *president* or *chairman*) is used directly without referring to its antecedent explicitly. However, in formula $f3$ (modeling that a bridging anaphor with the semantic class *relativePerson*, such as *mother* or *husband*, prefers close person antecedents) and $f4$ (modeling temporal relations, such as *September* – **a year** in Example 6.9), we prefer close antecedents by including the distance between antecedent and anaphor into the weights since these two bridging relations are local (coherence) phenomena.

(6.9)   Production of cars rose to 801,835 units in *September*. [. . .] Total truck production fell 22% from **a year** to 315,546 units.

Formulas for *bridging anaphora resolution*			
Semantic class features			
f1	+	(w)	$\forall a \in A\ \forall e \in E:\ hasSemanticClass(a, \text{``}gpeRolePerson\text{''}) \land$
			$hasSemanticClass(e, \text{``}gpe\text{''}) \land hasPairDistance(e, a, d) \land d > 0$
			$\rightarrow isBridging(a, e)$
f2	+	(w)	$\forall a \in A\ \forall e \in E:\ hasSemanticClass(a, \text{``}otherRolePerson\text{''}) \land$
			$hasSemanticClass(e, \text{``}org\text{''}) \land hasPairDistance(e, a, d) \land d > 0$
			$\rightarrow isBridging(a, e)$
f3	+	$(w \cdot d)$	$\forall a \in A\ \forall e \in E:\ hasSemanticClass(a, \text{``}relativePerson\text{''})$
			$\land hasSemanticClass(e, \text{``}person \star\text{''}) \land hasPairDistanceInverse(e, a, d)$
			$\rightarrow isBridging(a, e)$
f4	+	$(w \cdot d)$	$\forall a \in A\ \forall e \in E:\ hasSemanticClass(a, \text{``}date|time\text{''})$
			$\land hasSemanticClass(e, \text{``}date|time\text{''}) \land hasPairDistanceInverse(e, a, d)$
			$\rightarrow isBridging(a, e)$
Semantic features			
f5	+	$(w \cdot d)$	$\forall a \in A\ \forall e \in E_a:\ relativeRankPrepPattern(a, e, d) \rightarrow isBridging(a, e)$
f6	+	(w)	$\forall a \in A\ \forall e \in E_a:\ isTopRelativeRankPrepPattern(a, e) \rightarrow isBridging(a, e)$
f7	+	$(w \cdot d)$	$\forall a \in A\ \forall e \in E_a:\ relativeRankVerbPattern(a, e, d) \rightarrow isBridging(a, e)$
f8	+	(w)	$\forall a \in A\ \forall e \in E_a:\ isTopRelativeRankVerbPattern(a, e) \rightarrow isBridging(a, e)$
f9	+	$(w \cdot d)$	$\forall a \in A\ \forall e \in E_a:\ isPartOf(a, e) \land hasPairDistanceInverse(e, a, d)$
			$\rightarrow isBridging(a, e)$
Salience features			
f10	+	(w)	$\forall a \in A\ \forall e \in E_a:\ predictedGlobalAnte(e) \land hasPairDistance(e, a, d)$
			$\land d > 0 \rightarrow isBridging(a, e)$
f11	+	$(w \cdot d)$	$\forall a \in A\ \forall e \in E_a:\ relativeRankDocSpan(a, e, d) \rightarrow isBridging(a, e)$
f12	+	(w)	$\forall a \in A\ \forall e \in E_a:\ isTopRelativeRankDocSpan(a, e) \rightarrow isBridging(a, e)$
Lexical features			
f13	−	(w)	$\forall a \in A\ \forall e \in E_a:\ isSameHead(a, e) \rightarrow isBridging(a, e)$
f14	+	(w)	$\forall a \in A\ \forall e \in E_a:\ isPremodOverlap(a, e) \rightarrow isBridging(a, e)$
Syntactic features			
f15	−	(w)	$\forall a \in A\ \forall e \in E_a:\ isCoArgument(a, e) \rightarrow isBridging(a, e)$
f16	+	(w)	$\forall a \in A\ \forall e \in E_a:\ synParallelStructure(a, e) \rightarrow isBridging(a, e)$
f17	+	(w)	$\forall a \in A\ \forall e \in E_a:\ isClosestNominalModifer(a, e) \rightarrow isBridging(a, e)$
f18	+	(w)	$\forall a \in A\ \forall e \in E_a:\ isPredictSetBridging(a, e) \rightarrow isBridging(a, e)$

Table 6.3: Formulas used for bridging anaphora resolution. a represents a bridging anaphor, A the set of bridging anaphors in the whole document, e the antecedent candidate entity, E_a the set of the antecedent candidate entities for a according to a's discourse scope, and E the set of antecedent candidate entities in the whole document.

6.3.2.2 Semantic Features

Preposition pattern (Table 6.3: $f5$ and $f6$**).** We explore the preposition pattern to capture the semantic connectivity between a bridging anaphor and its antecedent. The *NP of NP* pattern proposed by Poesio et al. (2004a) is useful for part-of and attribute-of relations (e.g., *entry of buildings*) but cannot cover all bridging relations (such as *sanctions against a country*). Therefore we extend this pattern to a generalized *preposition pattern* to capture the diverse semantic relations between anaphor and antecedent. This is done by utilizing the Dunning root log-likelihood ratio and big corpora (i.e., Gigaword (Parker et al., 2011) and Tipster (Harman & Liberman, 1993)) to create a distributional semantic resource for bridging resolution. The principle of the Dunning root log-likelihood ratio and an example of how we explore this method to create the distributional semantic resource are detailed in Section 4.2.1 of Chapter 4. Here we focus on the process of the feature extraction.

First, we extract the three most highly associated prepositions for each anaphor using big corpora (e.g., *{against, on, in}* for the bridging anaphor **sanctions**). Then for each anaphor-antecedent candidate pair, we use their head words to create the query *"anaphor preposition antecedent"* (e.g., "sanction against/on/in country") which is executed against big corpora. To improve recall, we take lowercase, uppercase, singular and plural forms of the head word into account. We also replace proper names with fine-grained named entity types (using a gazetteer).

All raw hit counts of the queries are converted into the Dunning root log-likelihood ratio scores, then normalized using Formula 6.8 within all antecedent candidates of one anaphor. Table 6.4 shows an excerpt of the raw hit counts of the preposition pattern queries, the corresponding Dunning root log-likelihood ratio scores, and the normalized scores for the bridging anaphor **sanctions** and its antecedent candidates.

Anaphor	Antecedent Candidate	RawCount	RootLLR	NormalizedScore
sanctions	*the country*	6817	81.44	1.00
sanctions	*apartheid*	26	4.8	0.32
sanctions	*further punishment*	9	-1.88	0.26
sanctions	...	...	...	...

Table 6.4: An example of preposition pattern feature.

Verb pattern (Table 6.3: $f7$ and $f8$**).** A set/membership relation between anaphor and antecedent is often hard to capture by the *preposition pattern* because the anaphor often has no common noun head (see Example 6.10). However, in such a bridging relation, the antecedent should be semantically compatible with the verb that the anaphor depends on. For instance, in

Example 6.10, *"employees poked"* is reasonable whereas *"gem poked"* is illogical. Therefore, we explore the verb pattern to measure the compatibility between the antecedent candidates and the dependent verb of the anaphor.

(6.10)   Still *employees* do occasionally try to smuggle out a gem or two.

            . . .

            Another poked a hole in the heel of his shoe.

            . . .

            None made it past the body searches.

First, we hypothesize that anaphors whose lexical head is an indefinite pronoun or a number are potential set bridging cases. We then extract the dependent verbs for these potential set bridging anaphors. For instance, in Example 6.10, *poked* is the dependent verb for the set anaphor **Another**. Finally, for each antecedent candidate, subject-verb, verb-object or preposition-object[7] queries are executed against the Web 1T 5-gram corpus (Brants & Franz, 2006). In this example, *employees poked* and *gem poked* are two possible queries. The raw hit counts of the queries are transformed into the Dunning root log-likelihood ratio scores, then normalized as described in Formula 6.8 among all pairs for one anaphor.

WordNet Part-of relation (Table 6.3: $f9$).   To capture part-of bridging, We use WordNet to decide whether a part-of relation is held between an anaphor and its antecedent candidates. To improve recall, we use hyponym information of the antecedent, i.e., if an antecedent e is a hypernym of x and an anaphor a is a meronym of x, then a is also a meronym of e.

6.3.2.3   Salience Features

Salient entities are preferred as bridging antecedents. Although Poesio et al. (2004a) claimed that bridging anaphora are sensitive to the *local* rather than the *global* focus (Grosz & Sidner, 1986), we find that bridging anaphors with distant antecedents are also common when the antecedent is the global focus of a document (see Section 3.3 in Chapter 3). We capture salience from two different perspectives (Table 6.3: $f10$-$f12$).

$f10$ models that the globally salient entity is preferred to be the antecedent for bridging anaphors[8]. We now describe how we predict the globally salient entity for each document.

For each bridging anaphor $a \in A$ and each entity $e \in E$, let $score(a, e)$ be the preposition pattern score (Dunning root log-likelihood ratio score described in the preposition pattern

---

[7]The query form (i.e., subject-verb, verb-object or preposition-object) are decided by the syntactic relation between the anaphor and its dependent verb/preposition.

[8]Strictly, the meaning of $f10$ should be stated as: the globally salient entity e is preferred to be the antecedent for a bridging anaphor a if e is in the set of the antecedent candidate entities for a according to a's discourse scope.

features in Section 6.3.2.2) for the pair (a, e), we calculate the global semantic connectivity score e_{sal} for each $e \in E$ as below:

$$e_{sal} = \sum_{a \in A} score(a, e) \tag{6.11}$$

If an entity appears in the title and also has the highest global semantic connectivity score among all entities in E, then this entity is predicted as the globally salient entity for this document. Note that the globally salient entity here is based on semantic connectivity to all anaphors in the document and that not every document has a globally salient entity.

$f11$ and $f12$ capture salience by computing the *"antecedent document span"* of an antecedent candidate. We compute the span of text (measured in sentences) in which the antecedent candidate entity is mentioned. This is divided by the number of sentences in the whole document. This score is normalized using Formula 6.8 for all antecedent candidates of one anaphor.

6.3.2.4 Surface Features

The *isSameHead* feature (Table 6.3: $f13$) checks whether antecedent candidates have the same head as the anaphor: this is rarely the case in bridging (except in some cases of set bridging and spatial/temporal sequence, see Example 6.12) and can therefore be used to exclude antecedent candidates. The *isPremodOverlap* feature (Table 6.3: $f14$) determines the antecedent for compound noun anaphors whose head is prenominally modified by the antecedent head (see Example 6.13).

(6.12)   *His truck* is parked across the field, [...] The farmer at **the next truck** shouts ...

(6.13)   ...it doesn't make *the equipment needed to produce those chips*. And IBM worries that the Japanese will take over **that equipment market**, ...

6.3.2.5 Syntactic Features

CoArgument (Table 6.3: $f15$). The *isCoArgument* feature is based on the intuition that the subject cannot be the bridging antecedent of the object in the same clause. This feature excludes (some) close antecedent candidates. In Example 6.13, the antecedent candidate *the Japanese* isCoArgument with the anaphor **that equipment market** therefore should not be its antecedent.

Intra-sentential syntactic parallelism (Table 6.3: $f16$). If a noun phrase precedes a bridging anaphor in a different clause within the same sentence and both occupy the same syntactic role (*syntactic parallelism*), it is likely that this noun phrase is the antecedent of the bridging

anaphor. In Example 6.14, the anaphor and the antecedent are both objects of the verbs in the conjoined clauses. In Example 6.15, the anaphor's governing verb modifies the antecedent's governing verb. In Example 6.16, the anaphor and the antecedent both occupy the subject positions of the two conjoined clauses.

(6.14)   Poland must privatize *industry* and eliminate **subsidies** to stabilize its currency.

(6.15)   Many Japanese think it only natural that the organization or their members would donate to *politicians*, the way many Japanese do, to win **favor or support**.

(6.16)   *One building* was upgraded to red status while people were taking things out, and **a resident who wasn't allowed to go back inside** called up the stairs to his girlfriend . . .

Inter-sentential syntactic modification (Table 6.3: $f17$). Recent work on implicit semantic role labeling assumes that different occurrences of the same predicate (nominal or verbal predicate) in a document likely maintain the same argument fillers (Laparra & Rigau, 2013). Here we follow this assumption but apply it to nominal predicates only: different occurrences of the same nominal predicate are likely to have the same argument fillers indicated by nominal modifiers. Therefore we can identify the antecedent of a bridging anaphor by analyzing the nominal modifiers of other nouns which have the same head word as the anaphor. While Laparra & Rigau's work (2013) on implicit semantic role labeling is restricted to ten predicates annotated by Gerber & Chai (2012), we consider all bridging anaphors in ISNotes. In $f17$, we predict antecedents for bridging anaphors by performing the following two steps:

1. For each bridging anaphor a, we take its head lemma form a_h and collect all syntactic modifications of a_h in the document. We consider prenominal modification, possessive modification as well as prepositional postmodification. All realizations of these modifications which precede a form the antecedent candidates set $Ante_a$.

2. We choose the most recent mentioned entity from $Ante_a$ as the predicted antecedent for the bridging anaphor a.

In Example 6.17, to resolve the bridging anaphor **heavy damage**, we first check the other occurrences of the lemma "damage" and analyze their nominal modifiers, i.e., one modifier is "area" (supported by damage in the six - county San Francisco Bay area) and the other modifier is "quake" (supported by quake damage). We then collect all mentions whose syntactic head is "area" or "quake" in $Ante_a$ (i.e., *the six-county San Francisco Bay area* and *the quake, which registered 6.9 on the Richter scale*). Finally, the most recent entity in $Ante_a$ is predicted to be the antecedent (i.e., *the quake, which registered 6.9 on the Richter scale*).

(6.17)   Estimates of [damage in [the six - county San Francisco Bay area]] neared $5 billion, excluding the cost of repairing the region's transportation system.

. . .

Part of the bridge collapsed in *[the quake, which registered 6.9 on the Richter scale]*.

. . .

While many of these buildings sustained **heavy damage**, little of that involved major structural damage.

. . .

On Friday, during a visit to California to survey [quake damage], President Bush promised to "meet the federal government's obligation" to assist relief effort.

Hypertheme antecedent prediction for set sibling anaphors (Table 6.3: $f18$). The VerbPattern features (Table 6.3: $f7$ and $f8$) only apply to a small part of typical set bridging cases, such as **Another** and **None** in Example 6.18. However, the other set bridging anaphors (i.e. **One man** and **A food caterer**) are not covered by the VerbPattern features since it is hard to tell that they are set bridging anaphors from the surface form.

(6.18)   Still, *[employees]*$_{hypertheme}$ do occasionally try to smuggle out a gem or two.
         [One man]$_{theme}$ wrapped several diamonds in the knot of his tie.
         [Another]$_{theme}$ poked a hole in the heel of his shoe.
         [A food caterer]$_{theme}$ stashed stones in the false bottom of a milk pail.
         [None]$_{theme}$ made it past the body searches and X-rays of mine security.

We observe that set bridging anaphors frequently occur in clusters, where multiple anaphors refer to the same antecedent, e.g., **One man**, **Another**[9], **A food caterer**, and **None** are all elements of the set provided by *employees* in Example 6.18. We call this phenomenon *set sibling anaphors*. We also observe that all four bridging anaphors occupy the same syntactic position in their respective sentences. The information structure pattern we observe here is called the *Hypertheme-theme* structure by Daneš (1974).

We therefore explore a heuristic method to predict the "themes" (set sibling anaphors) and their "Hypertheme" (antecedent). Given that set sibling anaphors likely appear in a parallel structure, i.e., first subject positions in adjacent sentences, we first predict set sibling anaphors by expanding "typical" set bridging anaphors (e.g., **Another** and **None** in Example 6.18) to their syntactically parallel neighbors (e.g., **One man** and **A food caterer** in Example 6.18). We then predict the closest mention among all plural, subject mentions from the sentence immediately preceding the first anaphor as the antecedent for all (predicted) set sibling anaphors. If such a mention does not exist, the closest mention among all plural, object mentions

---

[9]Here "Another" is both a comparative anaphor and a bridging anaphor.

from the sentence immediately preceding the first anaphor is predicted to be the antecedent. In Example 6.18, *employees* is predicted to be the antecedent for all (predicted) set sibling anaphors.

6.4   Antecedent Candidate Selection Based on the Anaphor's Discourse Scope

In this section, we propose a new method (i.e., *d-scope-salience*) to select antecedent candidates for an anaphor on the basis of its discourse scope. *d-scope-salience* is applied in the formulas $f5$-$f18$ in Table 6.3 (Section 6.3.2) to form the antecedent candidates for bridging anaphors[10]. Before we detail the method, we first explain its motivation and define anaphors' discourse scopes.

Motivation.   Bridging anaphora resolution needs to tackle two interacting problems: (1) first, we create a list of antecedent candidates, so that this list includes the true antecedent while it contains as few false candidates as possible; (2) then, among all candidates, we should be able to choose the right antecedent according to different criteria (e.g., semantic information, world knowledge, or contextual clues). Once the competitive false candidates are removed from the candidate list in (1), we can have better access to the true antecedent in (2). For this task, most previous work (Markert et al., 2003; Poesio et al., 2004a; Lassalle & Denis, 2011) simply uses a static sentence window to construct the list of antecedent candidates. Although Poesio et al. (2004a) claimed that bridging anaphora are sensitive to the *local* rather than the *global* focus (Grosz & Sidner, 1986), we find that bridging anaphors with distant antecedents are also common when the antecedent is the global focus of a document. Indeed, around 24% of anaphors in ISNotes have antecedents that are three or more sentences away. So the method based on a static sentence window to choose the antecedent candidates has problems: if the window is too small, we miss too many correct antecedents; if it is too large, we include too much noise in learning.

We address this problem by proposing the *discourse scope* for an anaphor. Discourse entities have different scopes: some contribute to the main topic and can interact with other distant entities (globally salient entities), while others only focus on subtopics and can only interact with nearby entities (locally salient entities). As an example shown in Figure 6.1, a globally salient entity (e.g., *Marina* in s1) has a long (forward) lifespan, so that it can be accessed by both close and distant (non-local) anaphors (e.g., **a resident** in s2 and **residents** in

---

[10]Semantic class constraints ($f1$-$f4$ in Table 6.3) are strong indications for bridging, therefore the antecedent access scope of an anaphor in these constraints is not strongly connected to the anaphor's discourse scope. The pilot experiments also support this assumption.

No DiscourseRel	non-local	s1: In the hard - hit *Marina* neighborhood, life after the earthquake is often all too real, but sometimes surreal. s2: Some scenes: -- Saturday morning, **a resident** was given 15 minutes to scurry into a sagging buildinig and reclaim what she could of her life's possesions. ...
Expansion.Restatement	local	s24: After being inspected, *buildings with substantial damage* were color - coded. s25: Green allowed **residents** to re-enter; yellow allowed **limited access**; red allowed **residents one last entry** to gather everything they could within 15 minutes. ...
Expansion.Conjunction	local	s34: *One building* was upgraded to red status while people were taking things out, and **a resident who wasn't allowed to go back inside** called up **the stairs** to his girl friend, telling her keep sending things down to **the lobby.** ...
No DiscourseRel	non-local	s36: Enforcement of restricted - entry rules was sporadic, **residents** said.
Discourse Relation	anaLifeSpan	Bridging: →

Figure 6.1: Global and local salience in bridging.

s36). In contrast, a locally salient entity (e.g., *buildings with substantial damage* in s24) has a short (forward) lifespan, therefore it can only be accessed by nearby subsequent anaphors (e.g., **residents** and **limited access** in s25). Accordingly, anaphors which have non-local discourse scopes can access both locally and distant globally salient entities, whereas anaphors which have local discourse scopes can only access nearby locally salient entities. In consequence, we can add globally or locally salient entities to antecedent candidate lists for bridging anaphors according to their discourse scopes. The challenge is how to decide the discourse scopes for bridging anaphors automatically and how to model salience properly.

We hypothesize that there is a connection between bridging and discourse relations, so that some discourse relations can indicate the discourse scope of an anaphor. There are various theories of discourse relations. In this thesis we follow the one used in the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) since the texts annotated in ISNotes are also annotated in the PDTB.

The PDTB defines a hierarchy of relations. As one of the top level relations, the `Expansion` relation has a micro nucleus-satellite structure. In an `Expansion` relation, the second argument elaborates on the first argument. We assume that in this relation, most entities in the second argument should contribute to local entity coherence instead of global entity coherence, therefore bridging anaphors contained in the second argument of an `Expansion` relation are

likely to have local discourse scopes. In Section 3.5, we find that the existence of discourse relations has an influence on the scope of bridging anaphora. Although it is not clear which type of discourse relations has a dominant influence due to the lack of sufficient statistics for some relation types, intuitively the semantic nature of the `Expansion` relation as explained above should account more for local discourse scopes of bridging anaphora. Therefore we explore the `Expansion` relation to prevent anaphors with local discourse scopes to access distant globally salient entities.

Anaphors' discourse scopes.   Two types of discourse scopes are defined for bridging anaphora: *local* and *non-local*. If a bridging anaphor appears in the argument 2 of an `Expansion` relation, it has a *local* discourse scope; otherwise, it has a *non-local* discourse scope.

Antecedent candidate selection for an anaphor based on its discourse scope (*d-scope-salience*).   We first model the global and local salience from different perspectives. For each bridging anaphor $a \in A$, we define three antecedent candidate sets according to different salience measures: $E_A^{\text{globalSal1}}$, $E_A^{\text{globalSal2}}$ and E_a^{localSal}. The details of each set are described in the following:

- $E_A^{\text{globalSal1}}$ includes the top p percent salient entities in the text measured through the numbers of mentions in coreference chains.

- $E_A^{\text{globalSal2}}$ is the set of the globally salient entities measured by the global semantic connectivity score (described in $f10$ in Section 6.3.2.3). For each document, we create a list by ranking all entities according to their semantic connectivity to all anaphors. An entity is added to $E_A^{\text{globalSal2}}$ if it ranks among the top k in this list and appears in the headline.

- The set E_a^{localSal} consists of the locally salient entities. We approximate the entity's local salience according to the head position of the mention (which represents the entity) in the parse tree of the sentence. A mention preceding a from the same sentence as well as from the previous two sentences is added to E_a^{localSal} if the distance from its head to the root of the sentence's dependency parse tree is less than a threshold t.

We then select the antecedent candidates for an anaphor on the basis of its discourse scope: as shown in Figure 6.2, for a *local* anaphor, only locally salient entities from the local window (E_a^{localSal}) are allowed; for a *non-local* anaphor, apart from E_a^{localSal}, globally salient entities ($E_A^{\text{globalSal1}}$ and $E_A^{\text{globalSal2}}$) are also allowed.
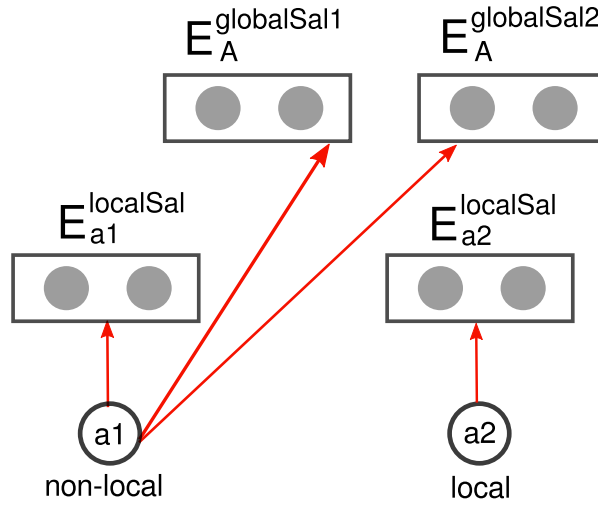
Figure 6.2: Antecedent candidate selection strategy based on anaphors' discourse scopes.

6.5 Experiments and Results

6.5.1 Experimental Setup

We conduct experiments on the ISNotes corpus. All experiments are performed via 10-fold cross-validation on documents. We use the OntoNotes named entity and syntactic annotation as well as the PDTB annotation for feature extraction. In each fold, we first choose ten documents randomly from the training set as the development set to estimate the values of the parameters p, k and t in $E_A^{\text{globalSal1}}$, $E_A^{\text{globalSal2}}$ and E_a^{localSal} respectively (Section 6.4)[11], then the whole training set is trained again using the optimized parameters. For our experiments, statistical significance is measured using McNemar's χ^2 test (McNemar, 1947). In Section 6.5.3, Section 6.5.4 and Section 6.5.5, the word *significantly* means a statistically significant difference in performance between two models at the level of $p < 0.01$.

6.5.2 Mention-Entity Setting and Mention-Mention Setting

We consider two experimental settings for bridging anaphora resolution: the mention-entity setting and the mention-mention setting.

In the mention-entity setting, the entity information is based on the OntoNotes coreference annotation. We resolve bridging anaphors (mentions) to the entity antecedents. In this

---

[11]The parameter is estimated using a grid search over $p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$, $k \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, and $t \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

controlled experiment setting, features are extracted by exploring the entity information. For instance, the semantic class of an entity is the majority semantic class of its all mention instantiations. The raw hit counts of the preposition pattern query for a bridging anaphor a and its antecedent candidate e ($f5$ and $f6$ in Section 6.3.2.2) is the maximum count among all instantiations of e. The distance between a bridging anaphor a and its antecedent candidate e is the distance between a and the closest mention instantiation of e preceding a.

In the mention-mention setting, we resolve bridging anaphors (mentions) to the mention antecedents (i.e., the instantiations of entity antecedents). In this setting, we use "string match" for $f11/f12$ in Section 6.3.2.2 and $E_A^{\text{globalSal1}}$ in Section 6.4 to measure the salience of the mention antecedent candidates.

6.5.3   Evaluation of Our New Features for Bridging Anaphora Resolution and of the Method to Select Antecedent Candidates (*d-scope-salience*)

To evaluate the effectiveness of our features for bridging anaphora resolution (Table 6.3 in Section 6.3.2) and the new method to select antecedent candidates (*d-scope-salience*, Section 6.4), we compare a local pairwise model to a reimplementation of a previous system (Poesio et al., 2004a) on the mention-entity setting.

Improved baseline.   We reimplement the algorithm from Poesio et al. (2004a) as a baseline. Since they did not explain whether they conducted the experiments under the mention-mention setting or the mention-entity setting, we assume they treated antecedents as entities and use a two and five sentence window for antecedent candidate selection[12].

Table 6.5 shows the feature set proposed by Poesio et al. (2004a) for part-of bridging. *Google distance* is the inverse value of Google hit counts for the *NP of NP* pattern query (e.g., *the windows of the center*). *WordNet distance* is the inverse value of the shortest path length between an anaphor and an antecedent candidate among all synset combinations. These features are supposed to capture the meronymy relation between anaphor and antecedent. The other features measure the salience of an antecedent candidate under the assumption that a salient entity is more likely to be the bridging antecedent. For instance, the feature *local first mention* checks whether an antecedent candidate is realized in the first position of a sentence within the previous five sentences of the anaphor, and the feature *global first mention* checks whether an antecedent candidate is realized in the first position of a sentence anywhere.

Since GoogleAPI is not available any more, we use the Web 1T 5-gram corpus (Brants &

---

[12]Poesio et al. (2004a) used a five sentence window for antecedent candidate selection, because all antecedents in their corpus are within the previous five sentences of the anaphors.

Franz, 2006) to extract the *Google distance* feature. We improve it by taking all information about entities via coreference into account as well as by replacing proper names with fine-grained named entity types (using a gazetteer). All other features in Table 6.5 are extracted as described in Poesio et al. (2004a). A Naive Bayes classifier with standard settings in WEKA (Witten & Frank, 2005) is used. In order to evaluate their model in the more realistic setting, i.e., the ability to choose the antecedent among all possible candidates for a bridging anaphor, we apply the *best first* strategy (Ng & Cardie, 2002) to predict the antecedent for each anaphor.

Group	Feature	Value
lexical	Google distance	numeric
	WordNet distance	numeric
salience	utterance distance	numeric
	local first mention	boolean
	global first mention	boolean

Table 6.5: Poesio et al.'s feature set.

Pairwise models.   The pairwise model is widely used in coreference resolution (Soon et al., 2001). We adapt it for bridging anaphora resolution[13]: given an anaphor mention a and the set of antecedent candidate entities E_a which appear before a, we create a pairwise instance (a, e) for every $e \in E_a$. A binary decision whether a is bridged to e is made for each instance (a, e) separately. Finally, we explore the *best first* strategy to choose one antecedent for each bridging anaphor. The different configurations of our pairwise models are described as below:

Pairwise model I.   In *pairwise model I*, we use the *preposition pattern* features ($f5$ and $f6$ in Section 6.3.2.2) plus Poesio et al.'s salience features from Table 6.5. We use a two sentence window as it performed on a par with the five sentence window in the baseline. We replace Naive Bayes with SVM^{light} because it can deal better with imbalanced data[14].

Pairwise model II.   On the basis of *pairwise model I*, other features from Table 6.3 (i.e., $f1$-$f4$, $f7$-$f18$) are added. It is worth noting that *Pairwise model II* uses the same strategy as *Pairwise model I* (i.e., two sentence window) to form the antecedent candidate sets for anaphors.

---

[13]Unlike in coreference resolution, we treat an anaphor as a mention and an antecedent as an entity. The anaphor is the first mention of the corresponding entity in the document.

[14]The SVM^{light} parameter which handles data imbalance is set according to the ratio between positive and negative instances in the training set.

		acc
improved baseline	*2 sent. + NB*	18.9
	5 sent. + NB	18.4
pairwise model	*pairwise model I*	29.1
	pairwise model II	39.3
	pairwise model III	**46.0**

Table 6.6: Results for bridging anaphora resolution: comparing the pairwise models to baselines. The bolded score indicates a significant improvement over all other models ($p < 0.01$).

Pairwise model III. On the basis of *pairwise model II*, we apply our new method (*d-scope-salience*, Section 6.4) to select the antecedent candidates for bridging anaphors.

Table 6.6 shows the performances of our pairwise models as well as the baselines. We observe that *pairwise model I* already outperforms two improved baselines by about 10%. This is attributed to the normalization of the *preposition pattern* feature (Formula 6.8 in Section 6.3.2), as well as its generalization (from the preposition "*of*" to appropriate prepositions for each anaphor) to capture more diverse semantic relations. The continuous significant improvements shown in *pairwise model II* and *pairwise model III* verify the contributions of our other features for bridging anaphora resolution (Section 6.3.2) and the advanced antecedent candidate selection strategy (Section 6.4).

To assess the impact of our features for bridging anaphora resolution, we perform a feature ablation study in which each main feature group in Table 6.3 (Section 6.3.2) is removed from *Pairwise model III* in turn. The results in Table 6.7 show that semantic features and syntactic features have the most impact.

	acc
pairwise model III	46.0
- Semantic class features	42.3
- Semantic features	**35.7**
- Salience features	42.7
- Lexical features	44.6
- Syntactic features	**41.0**

Table 6.7: Results of feature ablation experiments for bridging anaphora resolution. The bolded scores indicate a significant difference compared to *pairwise model III* ($p < 0.01$).

6.5.4 Evaluation of the Joint Inference Model on the Mention-Entity Setting

In the last section we show the effectiveness of our new features as well as the method to form the antecedent candidate sets for bridging anaphora resolution. However, the pairwise models we explored in the last section are "local" models in the sense that they choose the antecedent for each bridging anaphor separately. In this section we evaluate our joint inference model (*joint*), which models *bridging anaphora resolution* and *sibling anaphors clustering* jointly, with comparison with the best local model from the last section, i.e., *pairwise model III*. The model *joint* is the system described in Section 6.2 with all features for *sibling anaphors clustering* (Section 6.3.1) and all features for *bridging anaphora resolution* (Section 6.3.2). We use *thebeast*[15] to learn weights for the formulas and to perform inference. *thebeast* employs cutting plane inference (Riedel, 2008) to improve the accuracy and efficiency of MAP inference for MLNs. The model *joint - sibling* is based on *joint* but removes formulas for *sibling anaphors clustering* (f_c in Table 6.1) as well as formulas for joint inference ($f3$-$f6$ in Table 6.1). In fact, the model *joint - sibling* is equal to its non-joint counterpart (i.e., the local model *pairwise model III*) when joint inference features are not provided.

Table 6.8 shows the performances of our joint inference model and the two local models. Our joint inference model (*joint*) performs significantly better than the local models. The gain comes from that the joint inference model captures the mutually supportive relations between *sibling anaphors clustering* and *bridging anaphora resolution*. It confirms our assumption that the additional informative information from *sibling anaphors clustering* can help us to resolve bridging anaphora.

		acc
local	*pairwise model III*	46.0
	joint - sibling	46.4
joint inference	*joint*	**50.7**

Table 6.8: Results for bridging anaphora resolution: comparing the joint inference model to the local models. The bolded score indicates a significant improvement over all other models ($p < 0.01$).

6.5.5 Evaluation of Different Settings

In this section, we compare our best model from the last section (*joint*) on different settings (i.e., the mention-entity setting and the mention-mention setting): the system $joint_{me}$

---

[15]http://code.google.com/p/thebeast

	acc
$joint_{me}$	**50.7**
$joint_{mm}$	39.8
$joint_{me\_mm}$	44.2

Table 6.9: Experimental results for bridging anaphora resolution on different settings. The bolded score indicates a significant improvement over all other models ($p < 0.01$).

is trained and tested on the mention-entity setting, the system $joint_{mm}$ is trained and tested on the mention-mention setting, the system $joint_{me\_mm}$ is trained on the mention-entity setting but tested on the mention-mention setting.

Table 6.9 shows the results for bridging anaphora resolution on different settings. We notice that the performance of $joint_{mm}$ drops dramatically compared to $joint_{me}$. This is due to the noise involved in representing *sibling anaphors clustering* (e.g., two sibling anaphors may not share the same antecedent anymore in the mention-mention setting) as well as some features becoming weak on the mention-mention setting. For instance, in Example 6.19 where two sibling anaphors **Employees** and **workers** share the same entity antecedent represented by three coreferent mentions (*Mobil Corp.*, *the company's*, and *Mobil*), these two sibling anaphors do not always share the same mention antecedent on the mention-mention setting, e.g., *Mobil* is not the antecedent of the bridging anaphor **Employees**. Furthermore, knowing that *Mobil* is a company by exploring the entity information (in the mention-entity setting) can help resolve the bridging anaphor **workers** whereas this information is not available in the mention-mention setting. In fact, in Example 6.19, the mention *the company's* appears only once and is distant from the bridging anaphor **workers**, therefore it is not included as an antecedent candidate for the anaphor **workers** on the mention-mention setting.

(6.19) $\boxed{Mobil\ Corp.}$ is preparing to slash the size of its work force in the U.S., possibly as soon as next month, says individuals familiar with $\boxed{the\ company's}$ strategy. [...] **Employees** haven't yet been notified.

. . .

Some $\boxed{Mobil}$ executives were dismayed that a reference to the cutbacks was included in the earning report before **workers** were notified.

However, we find that $joint_{me\_mm}$ performs significantly better compared to $joint_{mm}$. This indicates that the model trained on the mention-entity setting represents the phenomenon better (with more reasonable weights learned for each formula) compared to the model trained on a noisy setting (mention-mention setting). Furthermore, $joint_{me\_mm}$ is the more realistic setting

for bridging anaphora resolution, and the accuracy score of 44.2 is the upper bound in the current model for bridging resolution, i.e., recognizing bridging anaphors and finding links to antecedents.

6.5.6 Error Analysis

The results in Table 6.9 in the last section show that even the best model (i.e., $joint_{me}$) on the mention-entity setting can only resolve around half of the bridging anaphors correctly. To understand the problems better, we analyze the joint inference model ($joint_{me}$) from different perspectives.

First, we notice that anaphors with distant antecedents are harder to resolve. Table 6.10 shows the comparison of correctly resolved anaphors with regard to anaphor-antecedent distance. Anaphors with antecedents that are three or more sentences away are harder to resolve compared to those with close antecedents (i.e., sentence distance equals to 0, 1 or 2). This tendency is reflected more clearly in Table 6.11 in which we only consider bridging pairs with entity antecedents[16].

	# pairs	**acc** in $joint_{me}$
sentence distance		
0	175	59.4
1	260	47.0
2	90	50.0
≥ 3	158	44.3

Table 6.10: Comparison of the percentage of correctly resolved anaphors with regard to anaphor-antecedent distance. Percentage is calculated relative to all bridging pairs.

	# pairs with entity Ante.	**acc** in $joint_{me}$
sentence distance		
0	171	60.8
1	231	52.8
2	84	53.6
≥ 3	154	45.5

Table 6.11: Comparison of the percentage of correctly resolved anaphors regarding anaphor-antecedent distance. Percentage is calculated relative to bridging pairs with entity antecedents.

---

[16]Most of bridging pairs with non-entity antecedents have a sentence distance of 1.

Second, we observe that non-sibling anaphors are harder to resolve compared to sibling anaphors. Table 6.12 shows that only 30.9% of non-sibling anaphors are resolved correctly whereas 63.1% of sibling anaphors are resolved correctly.

	# pairs	**acc** in $joint_{me}$
sibling anaphors	407	63.1
non-siblings	256	30.9

Table 6.12: Comparison of the percentage of correctly resolved anaphors with regard to sibling and non-sibling anaphors.

The relatively higher accuracy of resolving sibling anaphors is attributed to the modeling of global salience as well as the modeling of *sibling anaphors clustering* in our approach. However, this also leads to the majority errors of resolving non-sibling anaphors: non-sibling anaphors are wrongly resolved to the globally salient entities. This is due to the noise from our approximation of anaphors' discourse scopes as well as the context-specific semantic knowledge we employ being still insufficient. For instance, in one document two laws are discussed (one is a globally salient entity, the other is a locally salient one) and a later anaphor **the veto** is wrongly resolved to the globally salient one. In another text about the stealing of Sago Palms in California, we found **the thieves** as a bridging anaphor with the antecedent *his prized miniature palms*, which is a context-specific bridging relation. In addition, around 13.7% of non-sibling anaphors have a non-entity antecedent which are not handled by the current model.

6.6   Summary

In this chapter, we have provided a joint inference model for bridging anaphora resolution. The approach models two mutually supportive tasks (i.e., *bridging anaphora resolution* and *sibling anaphors clustering*) jointly, on the basis of the observation that semantically/syntactically related anaphors are likely to be sibling anaphors, and hence share the same antecedent (Section 6.2). In contrast to previous work (Poesio & Vieira, 1998; Poesio et al., 2004a; Markert et al., 2003; Lassalle & Denis, 2011), we do not limit ourselves to definite bridging anaphors or to specific relations (e.g., part-of relation), instead we consider all anaphor types and all relation types. We then propose various features to capture the unrestricted phenomenon (Section 6.3). Our approach also models the interpretive preferences (*local* or *global* focus) of bridging anaphors by selecting antecedent candidates for bridging anaphors on the basis of their discourse scopes (i.e., *local* or *non-local*) derived from discourse relation `Expansion` (Section 6.4). Our system achieves a considerable improvement for bridging anaphora resolution over the reimplementation of a previous system (Poesio et al., 2004a). We show that the improvement

comes from (1) linguistically motivated features (Section 6.5.3), (2) our new method to form the antecedent candidate sets for bridging anaphors (Section 6.5.3), and (3) the joint inference (Section 6.5.4).

So far, we have presented two models for two subtasks of bridging resolution (i.e., *bridging anaphora recognition* and *bridging anaphora resolution*) in the previous chapter (Chapter 5) and this chapter respectively. In the next chapter, we combine these two models in a pipeline way for full bridging resolution, i.e., recognizing bridging anaphors and finding links to antecedents.

Chapter 7

Unrestricted Bridging Resolution

Bridging resolution recovers the various non-identity relations between anaphors and antecedents. It plays an important role in establishing entity coherence in a text. Bridging resolution includes two subtasks: (1) recognizing bridging anaphors and (2) finding the correct antecedent among candidates. In recent empirical work, these two subtasks have been tackled separately: (Markert et al., 2012; Cahill & Riester, 2012; Rahman & Ng, 2012; Hou et al., 2013a) handle bridging anaphora recognition as part of information status (IS) classification, while (Poesio et al., 1997; 2004a; Markert et al., 2003; Lassalle & Denis, 2011; Hou et al., 2013b) concentrate on antecedent selection only, assuming that bridging anaphora recognition has already been performed. One exception is Vieira & Poesio (2000). They propose a rule-based system for processing definite NPs. However, their definition of bridging includes cases where anaphors and antecedents are coreferent but do not share the same head (*different-head coreference*). In this thesis, we restrict *bridging* to non-coreferential cases. We also exclude *comparative anaphora* (Modjeska et al., 2003)[1]. In addition, Vieira & Poesio (2000) report results for the whole anaphora resolution but do not report results for bridging resolution only. Another exception is Rösiger & Teufel (2014). They apply a coreference resolution system with several additional semantic features to find bridging links in scientific text where bridging anaphors are limited to definite NPs. They report preliminary results using the CoNLL scorer. However, we argue that the coreference resolution system and the evaluation metric for coreference resolution are not suitable for bridging resolution since bridging is not a set problem[2].

In this chapter, we propose a two-stage statistical model and a rule-based system for the challenging task of unrestricted bridging resolution (i.e., recognizing bridging anaphora and

---

[1]See Section 1.1 in Chapter 1 for a detailed discussion of the working definition for `bridging` in this thesis.

[2]In coreference resolution, the results are represented as sets of mentions where mentions in one set refer to the same entity. A coreference relation satisfies reflexivity, symmetry, and transitivity. However, the results in bridging resolution are represented as pairs of mentions in which two mentions in a pair often refer to different entities. A bridging relation is non-transitive and asymmetric.

finding links to antecedents), where unlike previous work, bridging anaphors are not limited to definite NPs and semantic relations between anaphors and their antecedents are not restricted to meronymic relations. Part of the work described in this chapter has been presented in our conference paper (Hou et al., 2014). The chapter is organized as follows. Section 7.1 defines the task and describes the evaluation metric. Section 7.2 and Section 7.3 details the two-stage statistical model and the rule-based system we proposed for this task respectively. In Section 7.4, we report the results of our systems for unrestricted bridging resolution, with comparison with a reimplementation of a previous system (Vieira & Poesio, 2000) as well as a learning-based pairwise model. Finally, Section 7.5 summarizes the chapter.

7.1   Task Definition and Evaluation Metrics

Bridging resolution is the task of recognizing bridging anaphora and finding links to antecedents. This thesis considers resolving "unrestricted bridging" in the sense that bridging anaphors are not limited to definite NPs and semantic relations between anaphors and their antecedents are not restricted to meronymic relations. However, we only consider entity antecedents[3], leaving resolving bridging with non-entity antecedents for future work.

For this task, we use an evaluation metric based on the number of bridging anaphors. All systems should predict one unique antecedent for each predicted bridging anaphor. A link predicted by a system is counted as correct if it recognizes the bridging anaphor correctly and links the anaphor to any instantiation of its antecedent preceding the anaphor[4].

We use *recall*, *precision* and *F-score* to measure the performance of a system for bridging resolution. The calculation of each measure is briefly described below:

$$\text{recall} = \frac{\mid \text{correct links predicted by the system} \mid}{\mid \text{gold bridging anaphors} \mid} \tag{7.1}$$

$$\text{precision} = \frac{\mid \text{correct links predicted by the system} \mid}{\mid \text{total links predicted by the system} \mid} \tag{7.2}$$

$$\text{F-score} = 2 \cdot \frac{\text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}} \tag{7.3}$$

---

[3]Entity antecedents are mentions in ISNotes, see Section 3.3 in Chapter 3 for the discussion of entity antecedents and event antecedents.

[4]In ISNotes, bridging is annotated mostly between an mention (bridging anaphor) and an entity (antecedent), so that a bridging anaphor could have multiple links to different mention instantiations of the same entity.

7.2 A Two-stage Model for Unrestricted Bridging Resolution

The two-stage model for bridging resolution combines the two models described in Chapter 5 and Chapter 6 (i.e., the *cascading collective classification model* for bridging anaphora recognition and the *joint inference model* for bridging anaphora resolution) in a pipeline way.

Figure 7.1 shows the framework of our two-stage model. Given mentions extracted from the documents, the system first predicts bridging anaphors by applying the cascading collective classification model (detailed in Chapter 5). Then it predicts antecedents for these predicted bridging anaphors (in the mention-mention setting) by applying the joint inference model trained on the mention-entity setting (detailed in Chapter 6). We summarize the configurations of our two-stage model in Table 7.1.

Model	Algorithm	Features
Stage 1: Cascading Collective Classification for Bridging Anaphora Recognition (Section 5.2.2)		
CascadedCollective (*Section 5.4.5*)	SVMs	non-relational features from previous work (Table 5.2)
		non-relational features for some IS classes (Table 5.3)
		non-relational features for bridging anaphora (Table 5.4)
	MLNs	relational features (Section 5.3.1)
		non-relational features from previous work (Table 5.2)
		non-relational features for some IS classes (Table 5.3)
		non-relational features for bridging anaphora (Table 5.4)
Stage 2: Joint Inference Model for Bridging Anaphora Resolution (Section 6.2)		
training: $joint_{me}$ testing: $joint_{mm}$ (*Section 6.5.5*)	MLNs	features for sibling anaphors clustering (Table 6.2)
		features for bridging anaphora resolution (Table 6.3)

Table 7.1: Configurations of the two-stage model for unrestricted bridging resolution.

In the experiment, we find that this two-stage model suffers from low precision. We observe that diverse bridging relations and relatively small-scale data for each type of relation make generalization difficult for the learning-based approach. Therefore we propose a rule-based system in which we aim to resolve bridging with higher precision. We describe this system in the next section.
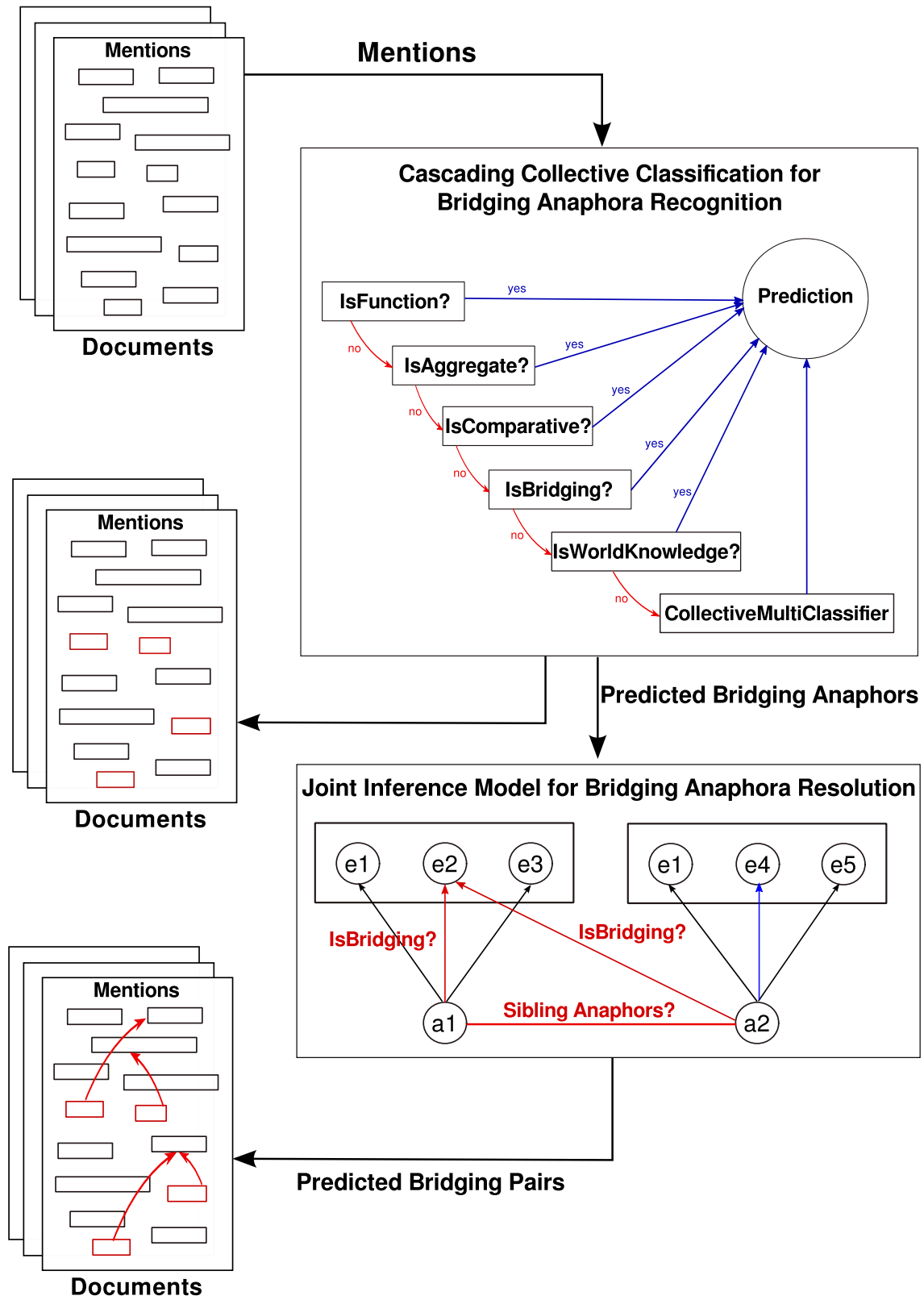
Figure 7.1: The two-stage model for unrestricted bridging resolution.

7.3 A Rule-based System for Unrestricted Bridging Resolution

In this section, we propose a rule-based system for unrestricted bridging resolution in the ISNotes corpus. The system consists of eight rules which we carefully design on the basis of linguistic intuitions, i.e., how the nature of bridging is reflected by various lexical, syntactic and semantic features.

We choose ten documents randomly from the corpus as the development set. Then we design rules for finding "bridging links" among all mentions in a document on the basis of the generalizations of bridging in the linguistic literature as well as our inspections of bridging annotations in the development set. The system consists of two components: bridging link prediction and post processing.

7.3.1 Bridging Link Prediction

The bridging link prediction component consists of eight rules. Löbner (1985; 1998) interprets bridging anaphora as a particular kind of functional concept, which in a given situation assign a necessarily unique correlate to a (implicit) possessor argument. He distinguishes between relational nouns (e.g., part terms, kinship terms, and role terms) and sortal nouns, and points out that relational nouns are more frequently used as bridging anaphora than sortal nouns. Rule1 to Rule4 in our system aim to resolve such relational nouns. We design Rule5 and Rule6 to capture set bridging. Finally, Rule7 and Rule8 are motivated by previous work on implicit semantic role labeling (Laparra & Rigau, 2013) which focuses on few predicates.

For all mentions in a document, each rule r is applied separately to predict a set of potential bridging links. Every rule has its own constraints on bridging anaphora and antecedents respectively. Bridging anaphors are diverse with regard to syntactic form and function: they can be modified by definite or indefinite determiners, furthermore they can take the subject or other positions in sentences (see Section 5.2.2 in Chapter 5 for examples to illustrate the variety of bridging anaphora). The only frequent syntactic property shared is that bridging anaphors most often have a simple internal structure concerning modifications. Therefore we first create an initial list of potential bridging anaphors A which excludes mentions that have a complex syntactic structure. A mention is added to A if it does not contain any other mentions and do not have modifications strongly indicating comparative NPs (such as *other symptoms*)[5]. Since head match is a strong indicator of coreference anaphora for definite NPs (Vieira & Poesio, 2000; Soon et al., 2001), we further exclude definite mentions from A if they

---

[5]A small list of ten markers such as *such, another …* and the presence of adjectives or adverbs in the comparative form are used to predict comparative NPs. See also *f4 PreModByCompMarker* in Section 5.3.2.2 for the full list of the markers.

are modified by *the* and have the same head as a previous mention. Then a set of potential bridging anaphors A_r is chosen from A on the basis of r's constraints on bridging anaphora. Finally, for each potential bridging anaphor $ana \in A_r$, a single best antecedent $ante$ from a list of candidate mentions (C_{ana}) is chosen if the rule's constraint on antecedent selection is applied successfully.

Every rule has its own scope to form the antecedent candidate set C_{ana}. Instead of using a static sentence window to construct the list of antecedent candidates like most previous work for resolving bridging anaphora (Poesio et al., 1997; Markert et al., 2003; Poesio et al., 2004a; Lassalle & Denis, 2011), we use the development set to estimate the proper scope for each rule. The scope is influenced by the following factors: (1) the nature of the target bridging link (e.g., set bridging is a local coherence phenomenon where the antecedent often occurs in the same or up to two sentences prior to the anaphor); and (2) the strength of the rule's constraint to select the antecedent among candidates (e.g., in Rule8, the ability to select the antecedent decreases with increasing the scope to contain more antecedent candidates). In the following, we describe the motivation for each rule and its constraints in detail.

Rule1: building part NPs.   To capture typical part-of bridging (Example 7.4), we extract a list of 45 nouns which specify building parts (e.g., *room* or *roof*) from the General Inquirer lexicon (Stone et al., 1966)[6]. A common noun phrase from A is added to A_{r1} if: (1) its head appears in the building part list; and (2) it does not contain any nominal premodifications. Then for each potential bridging anaphor $ana \in A_{r1}$, the NP with the strongest semantic connectivity to the potential anaphor ana among all mentions preceding ana from the same sentence as well as from the previous two sentences is predicted to be the antecedent.

The semantic connectivity of an NP to a potential anaphor is measured via the hit counts of the preposition pattern query (*anaphor preposition NP*) in big corpora. An initial effort to extract part-of relations using WordNet yields low recall on the development set. Therefore we use semantic connectivity expressed by prepositional patterns (e.g., *the basement of the house*) to capture underlying semantic relations. Such syntactic patterns are also explored in Poesio et al. (2004a) to resolve meronymy bridging. The "preposition pattern" part in Section 6.3.2.2 (Chapter 6) provides a detailed description for the calculation of semantic connectivity.

(7.4)   At age eight, Josephine Baker was sent by her mother to *a white woman's house* to do chores in exchange for meals and a place to sleep – a place in **the basement** with coal.

---

[6]The list contains lexical items from Inquirer under the "BldgPt" category. This list is also used in $f8$ *IsBuildingPart* in Section 5.3.2.3 (Chapter 5) to recognize bridging anaphors.

Rule2: relative person NPs. This rule is used to capture the bridging relation between a relative (e.g., *The husband*) and its antecedent (e.g., *She*). A list of 110 such relative nouns is extracted from WordNet. However, some relative nouns are frequently used generically instead of being bridging, such as *children*. To exclude such cases, we compute the argument taking ratio α for an NP using NomBank (Meyers et al., 2004). For each NP, α is calculated via its head frequency in the NomBank annotation divided by the head's total frequency in the WSJ corpus in which the NomBank annotation is conducted. The value of α reflects how likely an NP is to take arguments. For instance, the value of α is 0.90 for *husband* but 0.31 for *children*. To predict bridging anaphors more accurately, a conservative constraint is used. A mention from A is added to A_{r2} if: (1) its head appears in the relative person list; (2) The argument taking ratio α of its head is bigger than 0.5[7]; and (3) it does not contain any nominal or adjective premodifications. Then for each potential bridging anaphor $ana \in A_{r2}$, the closest non-relative person NP among all mentions preceding ana from the same sentence as well as from the previous two sentences is chosen as its antecedent.

Rule3: GPE job title NPs. In news articles, it is common that a globally salient geopolitical entity (hence GPE, e.g., *Japan* or *U.S.*) is introduced in the beginning, then later a related job title NP (e.g., *officials* or *the prime minister*) is used directly without referring to this GPE explicitly. To resolve such bridging cases accurately, we compile a list of twelve job titles which are related to GPEs[8]. An NP from A is added to A_{r3} if its head appears in this list and does not have a country premodification (e.g., *the Egyptian president*). Then for each potential bridging anaphor $ana \in A_{r3}$, the most salient GPE NP among all mentions preceding ana is predicted as its antecedent. We use the NP's frequency in the whole document to measure its salience. In case of a tie, the closest one is chosen to be the predicted antecedent.

Rule4: role NPs. Compared to Rule3, Rule4 is designed to resolve more general role NPs to their implicit possessor arguments. We extract a list containing around 100 nouns which specify professional roles from WordNet (e.g., *chairman*, *president* or *professor*)[9]. An NP from A is added to A_{r4} if its head appears in this list. Then for each potential bridging anaphor $ana \in A_{r4}$, the most salient proper name NP which stands for an organization among all mentions preceding ana from the same sentence as well as from the previous four sentences is chosen as its antecedent (if such an NP exists). Recency is again used to break ties.

---

[7]The same idea is explored in $f4$ *IsArgumentTakingNP* in Section 5.3.2.3 (Chapter 5) to recognize bridging anaphors.

[8]The full list is: *{president, governor, senator, minister, official, mayor, ambassador, autocrat, chancellor, premier, commissioner, dictator}*.

[9]This list is also used in $f7$ *SemanticClass* in Section 5.3.2.3 (Chapter 5) to assign the semantic class for a mention.

Rule5: percentage NPs.   In set bridging as shown in Example 7.5, the anaphor (**Seventeen percent**) is indicated by a percentage expression from A, which is often in the subject position. The antecedent (*the firms*) is predicted to be the closest NP which modifies another percentage NP via the preposition "of" among all mentions occurring in the same or up to two sentences prior to the potential anaphor.

(7.5)   22% of *the firms* said employees or owners had been robbed on their way to or from work. **Seventeen percent** reported their customers being robbed.

Rule6: other set member NPs.   In set bridging, apart from percentage expressions, numbers or indefinite pronouns are also good indicators for bridging anaphora. For such cases, the anaphor is predicted if it is: (1) a number expression (e.g., **One** in Example 7.6) or an indefinite pronoun (e.g., **some**, as shown in Example 7.7)[10] from A; and (2) a subject NP. The antecedent is predicted to be the closest NP among all plural, subject mentions preceding the potential anaphor from the same sentence as well as from the previous two sentences (e.g., *Reds and yellows* in Example 7.7). If such an NP does not exist, the closest NP among all plural, object mentions preceding the potential anaphor from the same sentence as well as from the previous two sentences is chosen to be the predicted antecedent (e.g., *several problems* in Example 7.6).

(7.6)   This creates *several problems*. **One** is that there are not enough police to satisfy small businesses.

(7.7)   *Reds and yellows* went about their business with a kind of measured grimness. **Some** frantically dumped belongings into pillowcases.

Rule7: argument-taking NPs I.   Laparra & Rigau (2013) found that different instances of the same predicate in a document likely maintain the same argument fillers. Here we follow this assumption but apply it to nominal predicates and their nominal modifiers only: different instances of the same nominal predicate likely maintain the same argument fillers indicated by nominal modifiers[11]. First, a common noun phrase from A is added to A_{r7} if: (1) its argument taking ratio α is bigger than 0.5; (2) it does not contain any nominal or adjective premodifications; and (3) it is not modified by determiners (e.g., *a*, *an* or *one*)[12] which usually indicate indefinite NPs and introduce new discourse referents (Hawkins, 1978). Then for each potential bridging anaphor $ana \in A_{r7}$, we choose the antecedent by performing the following steps:

---

[10]The same information is used in $f9$ *IsSetElement* in Section 5.3.2.3 (Chapter 5) to recognize set bridging anaphora.

[11]The same idea is used to design the feature $f17$ *inter-sentential syntactic modification* in Section 6.3.2.5 (Chapter 6) for bridging anaphora resolution.

[12]See $f5$ *determiner* in Section 5.3.2.1 (Chapter 5) for the full list of such determiners.

1. We take ana's head lemma form ana_h and collect all its syntactic modifications in the document. We consider nominal premodification, possessive modification as well as prepositional postmodification. All mention realizations of these modifications which precede ana form the antecedent candidates set C_{ana}.

2. We choose the most recent NP from C_{ana} as the predicted antecedent for the potential bridging anaphor ana.

In Example 7.8, we first predict the two occurrences of **residents** as bridging anaphors. Since in the text, other occurrences of the lemma "resident" are modified by "Marina" (supported by <u>Marina residents</u>) and "buildings" (supported by <u>some residents of badly damaged buildings</u>), we collect all mentions whose syntactic head is "Marina" or "buildings" in C_{ana} (i.e., *Marina*, *badly damaged buildings* and *buildings with substantial damage*). Then among all NPs in C_{ana}, the most recent NP is chosen to be the antecedent (i.e., *buildings with substantial damage*).

(7.8) She finds the response of <u>Marina residents</u> to the devastation of their homes "incredible".

...

Out on the streets, <u>some residents of badly damaged buildings</u> were allowed a 15 minute scavenger hunt through their possessions.

...

After being inspected, *buildings with substantial damage* were color - coded.

Green allowed **residents** to re-enter; red allowed **residents** one last entry to gather everything they could within 15 minutes.

Rule8: argument-taking NPs II. Prince (1992) found that discourse-old entities are more likely to be represented by NPs in subject position. Although she could not draw a similar conclusion when collapsing *Inferrables* (= *bridging* in this thesis) with *Discourse-old Non-pronominal*, we find that in the development set, an argument-taking NP in the subject position is a good indicator for bridging anaphora (e.g., **participants** in Example 7.9). A common noun phrase from A is collected in A_{r8} if: (1) its argument taking ratio α is bigger than 0.5; (2) it does not contain any nominal or adjective premodifications; and (3) it is in the subject position. Semantic connectivity again is used as the criteria to choose the antecedent: for each potential bridging anaphor $ana \in A_{r8}$, the NP with the strongest semantic connectivity to ana among all mentions preceding ana from the same sentence as well as from the previous two sentences is predicted to be the antecedent.

(7.9)   Initial steps were taken at *Poland's first international environmental conference,*
          which I attended last month. [...] While Polish data have been freely available
          since 1980, it was no accident that **participants** urged the free flow of information.

7.3.2   Post-processing

In the bridging link prediction component, each rule is applied separately. To resolve the
conflicts between different rules (e.g., two rules predict different antecedents for the same
potential anaphor), a post processing step is applied. We first order the rules according to their
precision for predicting bridging pairs (i.e., recognizing bridging anaphors and finding links to
antecedents) in the development set. When a conflict happens, the rule with the highest order
has the priority to decide the antecedent. Table 7.2 summarizes the rules described in Section
7.3.1, the numbers in square brackets in the first column indicate the order of the rules. Table
7.3 shows the precisions of bridging anaphora recognition and bridging pairs prediction for
each rule in the development set. Firing rate is the proportion of bridging links predicted by
rule r among all predicted links.

7.4   Experiments and Results

7.4.1   Experimental Setup

We conduct all experiments on the ISNotes corpus. We use the OntoNotes named entity and
syntactic annotations to extract features. Ten documents containing 113 bridging anaphors
from the ISNotes corpus are set as the development set to estimate parameters for the rule-
based system. The remaining 40 documents are used as the test set. In order to compare the
results of different systems directly, we evaluate all systems on the test set.

7.4.2   Evaluation and Discussion

We compare our two-stage model (*pipeline model*) and our rule-based system (*rule system*)
with a reimplementation of a previous rule-based system from Vieira & Poesio (2000) (*base-
line*) as well as a learning-based pairwise model (*pairwise model*). We describe each system
in the following:

Baseline.   We reimplement the rule-based system from Vieira & Poesio (2000) as a baseline.
The original algorithm focuses on processing definite NPs. It classifies four categories for the
definite NPs: *discourse new*, *direct anaphora* (same-head coreference), *lenient bridging*, and

rule	anaphor	antecedent	scope
rule1 [2]	building part NPs	the NP with the strongest semantic connectivity to the potential anaphor	two
rule2 [5]	relative person NPs	the closest non-relative person NP	two
rule3 [6]	GPE job title NPs	the most salient GPE NP	all
rule4 [7]	role NPs	the most salient organization NP	four
rule5 [1]	percentage NPs	the closest NP which modifies another percentage NP via the preposition "of"	two
rule6 [3]	other set member NPs	the closest subject, plural NP; otherwise the closest object, plural NP	two
rule7 [4]	argument-taking NPs I	the closest NP whose head is an unfilled role of the potential anaphor (such a role is predicted via syntactic modifications of NPs which have the same head as the potential anaphor)	all
rule8 [8]	argument-taking NPs II	the NP with the strongest semantic connectivity to the potential anaphor	two

Table 7.2: Rules for unrestricted bridging resolution. The scope of antecedent candidates are verified in the development set: "all" represents all mentions preceding the potential anaphor from the whole document, "four" mentions occurring in the same or up to four sentences prior to the potential anaphor, "two" mentions occurring in the same or up to two sentences prior to the potential anaphor.

rule	anaphora	anaphora recognition precision	bridging pairs prediction precision	firing rate
rule1 [2]	building part NPs	75.0%	50.0%	6.1%
rule2 [5]	relative person NPs	69.2%	46.2%	6.1%
rule3 [6]	GPE job title NPs	52.6%	44.7%	19.4%
rule4 [7]	role NPs	61.7%	32.1%	28.6%
rule5 [1]	percentage NPs	100.0%	100.0%	2.6%
rule6 [3]	other set member NPs	66.7%	46.7%	7.8%
rule7 [4]	argument-taking NPs I	53.8%	46.4%	6.1%
rule8 [8]	argument-taking NPs II	64.5%	25.0%	25.5%

Table 7.3: Precision of bridging anaphora recognition and bridging pairs prediction for each rule in the development set. The numbers in square brackets in the first column indicate the order of the rules.

Unknown. This algorithm also finds antecedents for NPs which belong to *direct anaphora* or *lenient bridging*.

Since Vieira & Poesio (2000) include *different-head coreference* into their *lenient bridging* category[13], we further divide their *lenient bridging* category into two subcategories: *different-head coreference* and *bridging*. Figure 7.2 shows the details of the division after failing to classify an NP as *discourse new* or *direct anaphora*. The details of the whole system are provided in Appendix A.
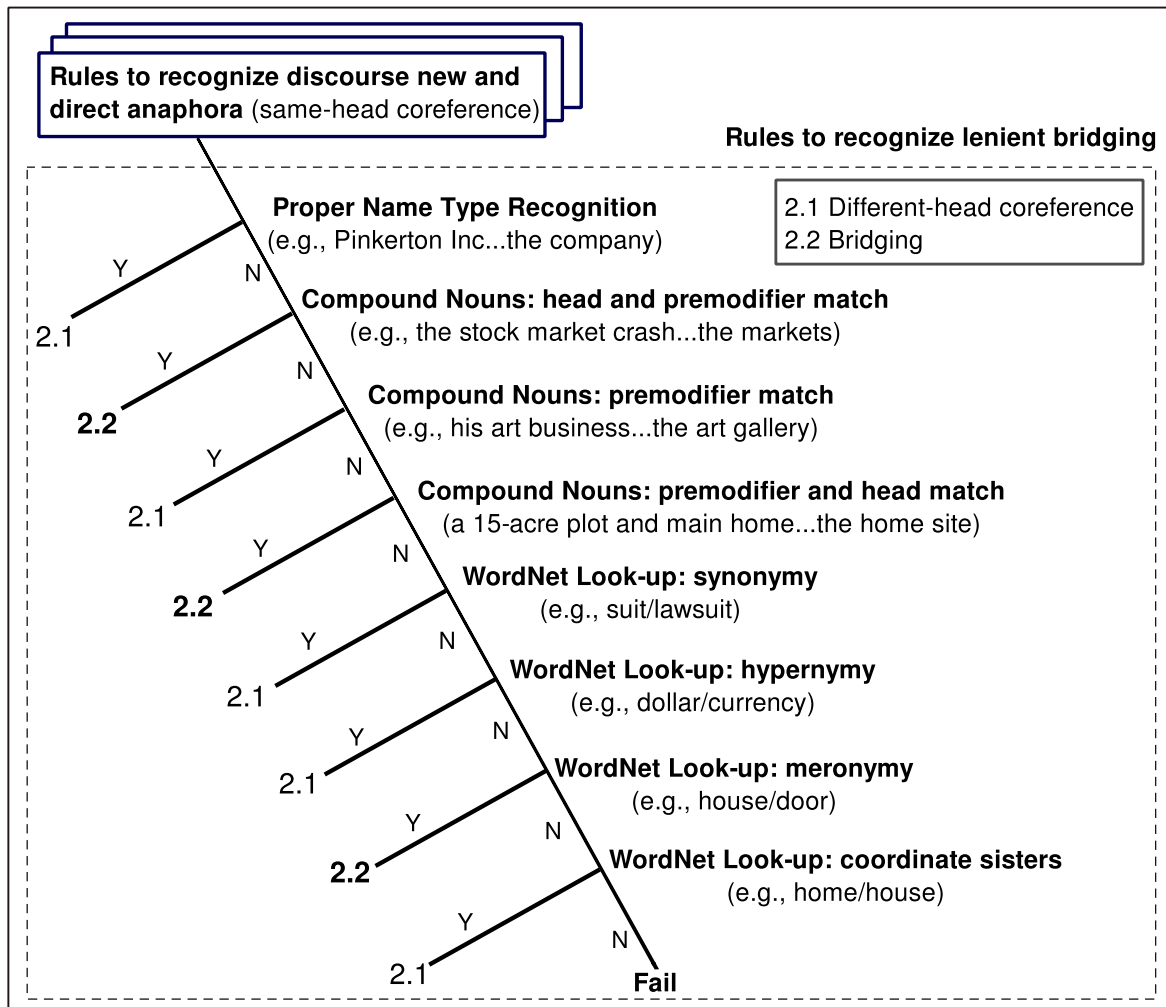
Figure 7.2: Vieira & Poesio's (2000) original algorithm for processing definite NPs. We further divide their lenient bridging category into two subcategories: *2.1 Different-head coreference* and *2.2 Bridging*.

---

[13]See Section 2.1.2 for the detailed description of *lenient bridging* annotation in Vieira & Poesio (2000).

Pairwise model. We adapt the pairwise model which is widely used in coreference resolution (Soon et al., 2001) for unrestricted bridging resolution. Like in the rule-based system, we first create an initial list of possible bridging anaphors A_{ml} with one more constraint. The purpose is to exclude as many obvious non-bridging anaphoric mentions from the list as possible. A mention is added to A_{ml} if: (1) it does not contain any other mentions; (2) it is not modified by premodifications which strongly indicate comparative NPs; and (3) it is not a pronoun or a proper name. Then for each NP $a \in A_{ml}$, a list of antecedent candidates C_a is created by including all mentions preceding a from the same sentence as well as from the previous two sentences[14]. We create a pairwise instance (a, c) for every $c \in C_a$. In the decoding stage, the *best first* strategy (Ng & Cardie, 2002) is used to predict the bridging links. Specifically, for each $a \in A_{ml}$, we predict the bridging link to be the most confident pair (a, c_{ante}) among all instances with the positive prediction. We provide this pairwise model with the same non-relational features as our two-stage model described in Section 7.2, i.e., features from Table 5.6 described in Section 5.3.2.4 (Chapter 5) and features from Table 6.3 described in Section 6.3.2 (Chapter 6). We use SVM^{light} to conduct the experiments[15]. The experiment is performed via 10-fold cross-validation on the whole corpus. However, to compare the learning-based approach to the rule-based system described in Section 7.3 directly, we report the results of *pairwise model* on the same test set as the rule-based system.

Rule system. The system described in Section 7.3.

Pipeline model. The model described in Section 7.2. The experiment is conducted via 10-fold cross-validation on the whole corpus. Like in *pairwise model*, we report the result of *pipeline model* on the same test set as our rule-based system.

Results. Table 7.4 shows the performances of our rule-based system (*rule system*), our two-stage model (*pipeline model*), and the two baselines (*baseline* and *pairwise model*) for bridging resolution and bridging anaphora recognition. The F-score for bridging anaphora recognition of our two-stage model (*pipeline model*) represents significant improvements over all other systems at $p < 0.01$ (randomization test[16]). For bridging resolution, both *pipeline model* and *rule system* significantly outperform *baseline* and *pairwise model* regarding F-score at $p < 0.01$ (randomization test). However, the difference between *pipeline model* and *rule system* with regard to bridging resolution is not significant.

---

[14]In ISNotes, 76% of antecedents occur in the same or up to two sentences prior to the anaphor. Initial experiments show that increasing the window size more than two sentences decreases the performance.

[15]To deal with data imbalance, the SVM^{light} parameter is set according to the ratio between positive and negative instances in the training set.

[16]We use the package from `https://github.com/smartschat/art` to perform two-sided paired approximate randomization tests.

	Bridging resolution			Bridging anaphora recognition		
	R	P	F	R	P	F
baseline	2.9	13.3	4.8	4.5	21.0	7.4
pairwise model	20.5	10.1	13.5	62.0	30.5	40.9
rule system	11.9	42.3	**18.6**	18.3	61.7	28.2
pipeline model	22.0	20.1	**21.0**	48.7	43.8	**46.1**

Table 7.4: Experimental results for unrestricted bridging resolution and bridging anaphora recognition. Bolded scores indicate significant improvements relative to the comparison models ($p < 0.01$).

Discussion: learning-based approach vs. rule-based approach. The low recall both for bridging anaphora recognition and bridging resolution in *baseline* is predictable, since it only considers meronymy bridging and compound noun anaphors whose head is prenominally modified by the head word of the antecedent (e.g., *the state* – **state gasoline taxes**). We notice that all learning-based approaches (i.e., *pairwise model* and *pipeline model*) perform significantly better than the rule-based approaches (i.e., *baseline* and *rule system*) with regard to bridging anaphora recognition. We also notice that our rule-based system (*rule system*) achieves higher precision but suffers from lower recall in the both tasks (i.e., bridging anaphora recognition and bridging resolution) compared to the learning-based approaches. However, our two-stage system (*pipeline model*) is not significantly better than our rule-based system (*rule system*) regarding bridging resolution. Although ISNotes is a reasonably sized corpus for bridging compared to previous work, diverse bridging relations, especially lots of context specific relations such as *pachinko* – **devotees** or *palms* – **the thieves**, lead to relatively small-scale training data for each type of relation. Therefore, it is difficult for the learning-based approach to learn effective rules to predict bridging links. In the future, we assume that our learning-based system could benefit from more training data.

Discussion: bridging anaphors modified by *the* vs. bridging anaphors not modified by *the*. We compare the performances of all systems for bridging resolution and bridging anaphora recognition with regard to different types of bridging anaphora, i.e., bridging anaphors modified by *the* and bridging anaphors which are not modified by *the*. Accordingly, we evaluate the performances of all systems for bridging resolution and bridging anaphora recognition on NPs modified by *the* and on NPs which are not modified by *the* separately.

The results in Table 7.5 and Table 7.6 show that *baseline* performs similar on the both types of bridging anaphora for the two tasks (i.e., bridging resolution and bridging anaphora recog-

	Bridging resolution			Bridging anaphora recognition		
	R	P	F	R	P	F
baseline	2.7	24.1	4.9	3.9	34.5	7.0
pairwise model	20.4	10.0	13.4	70.6	34.5	**46.4**
rule system	5.9	45.2	10.4	9.0	74.2	16.1
pipeline model	18.8	12.5	15.0	53.0	35.1	42.2

Table 7.5: Experimental results for unrestricted bridging resolution and bridging anaphora recognition, where bridging anaphors are modified by "*the*". Bolded scores indicate significant improvements relative to all other models ($p < 0.01$).

	Bridging resolution			Bridging anaphora recognition		
	R	P	F	R	P	F
baseline	2.9	10.5	4.5	4.9	17.5	7.7
pairwise model	20.6	10.2	13.6	56.6	27.9	37.4
rule system	16.9	41.8	24.1	24.0	60.9	34.2
pipeline model	24.0	28.6	26.1	45.8	53.7	**49.4**

Table 7.6: Experimental results for unrestricted bridging resolution and bridging anaphora recognition, where bridging anaphors are not modified by "*the*". Bolded scores indicate significant improvements relative to all other models ($p < 0.01$).

nition). We notice that *rule system* and *pipeline model* achieve better results on recognizing bridging anaphors which are not modified by *the* compared to recognizing bridging anaphors modified by *the*, whereas *pairwise model* behaves in an opposite manner. Here we compare the two learning-based systems on bridging anaphora recognition regarding these two types of bridging anaphora.

There are two main differences between *pairwise model* and *pipeline model*. First, *pairwise model* includes antecedent information and features between pairs, e.g., $f6$ *isTopRelativeRankPrepPattern*, $f14$ *isPremodOverlap* in Table 6.3. These features are not used in *pipeline model* for bridging anaphora recognition. Second, *pairwise model* only uses a static sentence window (two sentences) to construct pair instances. This has the effect that not all bridging anaphors in the corpus can be used as positive training instances for bridging anaphora recognition. For those bridging anaphors with non-entity antecedents or distant antecedents, the corresponding pair instances are marked as false during training. As a result, these bridging anaphors are not explored by *pairwise model* to recognize bridging anaphors,

whereas *pipeline model* uses all bridging anaphors to train the classifier to recognize bridging anaphors.

These two differences explain the different performances of the two learning-based systems on bridging anaphora recognition. First, we observe that *pairwise model* performs better than *pipeline model* on recognizing bridging anaphors modified by *the* (see Table 7.5). In Section 5.4.7 (Chapter 5), we already found that in our cascading collective classification model (which is used as the first component in *pipeline model* to recognize bridging anaphors), recognizing bridging anaphors modified by *the* is harder than recognizing bridging anaphors which are not modified by *the*. We therefore assumed that definite bridging anaphors (bridging anaphors modified by *the*) need to be put into the context to understand their "associative anaphor" usages. Consequently, *pairwise model* recognizes definite bridging anaphors better than *pipeline model* by exploring part of the context, i.e., the additional antecedent information and the corresponding pair features.

Second, it seems confusing that *pairwise model* performs worse than *pipeline model* on recognizing bridging anaphors not modified by *the*, given both models have the same non-relational features for bridging anaphora recognition[17]. The reason lies in the fact that *pairwise model* has less positive training instances compared to *pipeline model* for the task (see the second difference between these two learning-based systems explained before). In fact, around 30% of bridging anaphors are "missing" in *pairwise model* due to non-accessible antecedents. Among those "missing" bridging anaphors, 61% are not modified by *the*.

The above analysis suggests that in the future, it might be possible to improve the performance of *pipeline model* for bridging resolution by dealing with different types of bridging anaphora (i.e., bridging anaphors modified by *the* and bridging anaphors not modified by *the*) separately in the first stage.

7.5   Summary

In this chapter, we have proposed two approaches for the challenging task of full bridging resolution, i.e., recognizing bridging anaphora and finding links to antecedents. One is a learning-based system combining the two joint inference models for *bridging anaphora recognition* (described in Chapter 5) and *bridging anaphora resolution* (described in Chapter 6) in a two-stage framework (Section 7.2). The other is a rule-based system consisting of eight rules which target different bridging relations based on linguistic insights (Section 7.3). Both the two systems considerably outperform a reimplementation of a previous rule-based system

---

[17]Although *pipeline model* explores additional relational features for bridging anaphora recognition, we found that the relational features have more influences on other IS categories other than `bridging` (see Section 5.4.4 in Chapter 5).

(Vieira & Poesio, 2000) and a learning-based pairwise model on bridging resolution.

Our two-stage model performs significantly better than our rule-based system with regard to bridging anaphora recognition. Our rule-based system achieves higher precision but suffers from lower recall compared to the two-stage approach. However, the difference of these two systems on bridging resolution is not significant. We observe that diverse bridging relations and relatively small-scale data for each type of relation make generalization difficult for the learning-based approach. In the future, we speculate that the learning-based system could benefit from more training data.

This work is – to our knowledge – the first bridging resolution system that handles the unrestricted phenomenon (i.e., bridging anaphors are not limited to definite NPs and semantic relations between anaphors and their antecedents are not restricted to meronymic relations) in a realistic setting.

Chapter 8

Conclusions

In this chapter, we summarize the contributions of this thesis and the insights gained. We then discuss the limitations of the current work and various potential future directions.

8.1   Contributions

This thesis focuses on the problem of *bridging resolution*, the most challenging task of anaphora resolution. At the beginning of the thesis (Section 1.3), we raised a series of research questions centered on bridging characteristics and automatic bridging resolution. Here we revisit these questions, summarizing how they have been solved and highlighting our core contributions.

Characterizing bridging on the basis of a corpus analysis.   Previous corpus-linguistic studies on bridging are beset by several problems. First, the definition of bridging is extended too broadly to include coreferential NPs with lexical variety (Vieira, 1998) or other non-anaphoric NPs (Nissim et al., 2004). Second, reliability is not measured or low (Poesio & Vieira, 1998; Gardent & Manuélian, 2005; Nedoluzhko et al., 2009; Riester et al., 2010). Third, annotated corpora are small (Poesio, 2004; Caselli & Prodanof, 2006). Fourth, they are often based on strong untested assumptions about bridging anaphora types, antecedent types or bridging relations, such as limiting it to definite NP anaphora (Poesio & Vieira, 1998; Gardent & Manuélian, 2005; Caselli & Prodanof, 2006; Riester et al., 2010), to NP antecedents (all prior work) or to part-of relations between anaphor and antecedent (Poesio, 2004). On the contrary, the corpus used in this thesis (ISNotes) circumvents these problems, i.e., bridging is strictly used for anaphoric NPs which bear non-identity relations with antecedents, human bridging recognition is reliable, it also contains a medium number of **unrestricted** bridging cases in the sense that bridging anaphora/antecedents/relations are not limited to certain types.

To better understand the nature of the unrestricted phenomenon, we carried out a thorough

statistical analysis for bridging from different perspectives on the basis of the ISNotes corpus (Chapter 3). We summarize some of the important results as follows:

- **Syntactic property**: only 38.5% of bridging anaphors are definite NPs (i.e., NPs modified by "*the*"). This calls into question the strategy of prior approaches to limit themselves to this type of bridging. Moreover, bridging anaphora are diverse with regard to syntactic forms and functions.

- **Topological property**: most bridging anaphors (61.4%) are *sibling anaphors* which share the same antecedent with other bridging anaphors. We found that there is a difference between sibling anaphors and non-sibling anaphors in terms of the salience of antecedents, i.e., globally salient antecedents connect to a higher proportion of sibling anaphors and a lower proportion of non-sibling anaphors compared to locally salient antecedents. Also we found that bridging is a relatively local phenomenon, with 76.92% of anaphors having antecedents occurring in the same or up to two sentences prior to the anaphor. In addition, bridging pairs with sibling anaphors are more distant than those with non-sibling anaphors, and bridging pairs with globally salient antecedents tend to be more distant than those with locally salient antecedents.

- **Bridging and discourse relations**: we empirically assessed the interaction between bridging and discourse relations. We observed that *local* bridging anaphors are likely to co-occur with discourse relations, and that a larger proportion of *local* bridging anaphors co-occur with `Expansion` relations compared to *non-local* bridging anaphors. Moreover, we found that adjacent sentences with `Expansion` relations are least likely to share entities and are most likely to co-exist with the type of "- *coreference + bridging*" (i.e., adjacent sentences which do not share entities but contain a bridging relation across sentence boundaries).

The above analysis and results are important as they provide us with prior knowledge of linguistic structure when we design computational models to resolve bridging automatically. In other words, we should not expect statistical methods to provide the whole solution. Instead, linguistic knowledge should guide us to design the model structure.

Resolving bridging automatically.   We solve the problem of bridging resolution using a two-stage statistical model which targets *bridging anaphora recognition* and *bridging anaphora resolution* in a pipeline way. Given all mentions in a document, the first stage predicts bridging anaphors by exploring a cascading collective classification model; the second stage then finds the antecedents for all predicted bridging anaphors simultaneously by exploring a joint inference model. We summarize our contributions in the following:

1. We cast *bridging anaphora recognition* as a subtask of learning fine-grained information status (IS) (Chapter 5). Each mention in a text gets assigned one IS class, bridging being one possible class. Motivated by the linguistic properties of the task, i.e., linguistic relations among several IS categories, the wide variation of bridging anaphora as well as their relative rarity compared to many other IS categories, we design a **cascading collective classification** model for this task. The model combines the binary classifiers for minority categories and a collective classifier for all categories in a cascaded way. It addresses the multi-class imbalance problem (for rare categories without strong indicators) within a multi-class setting while still keeping the strength of the collective classifier by exploring relational autocorrelation (for several IS classes in which such relational autocorrelation exists). Our system achieves substantial improvements both for the overall IS classification accuracy as well as for bridging anaphora recognition over the reimplementations of two previous systems. Moreover, via a detailed comparison and analysis, we found that recognizing bridging anaphors modified by *the* is harder than recognizing bridging anaphors which are not modified by *the*.

2. We propose a **joint inference** model for bridging anaphora resolution (Chapter 6). The approach models two mutually supportive tasks (i.e., *bridging anaphora resolution* and *sibling anaphors clustering*) jointly, on the basis of the observation that semantically/syntactically related anaphors are likely to be *sibling anaphors*, and hence share the same antecedent. In addition, we augment this model by integrating an advanced candidate selection strategy which accounts for the interpretive preference (*local* or *global* focus) of bridging anaphors. Our model results in considerable improvements over the reimplementation of a previous system as well as a local pairwise model.

3. Finally, we resolve the task of unrestricted bridging resolution (i.e., recognizing bridging anaphors and finding links to antecedents) by combining the two joint inference models for *bridging anaphora recognition* and *bridging anaphora resolution* in a pipeline way. This two-stage model significantly outperforms a pairwise model and a reimplementation of a previous rule-based system. It also beats another advanced rule-based system (Hou et al., 2014) by a small margin. The work is – to our knowledge – the first bridging resolution system that handles the unrestricted phenomenon in a realistic setting.

8.2   Limitations

Although this thesis breaks new ground for bridging resolution from a computational model perspective, there are still many open problems in this field. Here we discuss some of the limitations of this work.

Reliance on gold annotations. All experiments in this thesis rely on gold mentions as well as the named entity and syntactic annotation. Nevertheless, given the difficulty of the task, it is essential to carry out the study under controlled conditions first. This helps us to concentrate on the task (bridging resolution) better. In the future, when the overall performance for bridging resolution reaches a reasonable level, it is possible to develop an end-to-end bridging resolution system by removing the current constraints.

Lack of large-scale datasets. Although the corpus used in the thesis contains a medium number of bridging cases compared to previous work, diverse bridging relations lead to relatively small-scale training data for each type of relation. As a result, it makes generalization difficult for the learning-based approach. Indeed, as we have already seen in Section 2.1.2 and Section 2.1.3, the lack of a large-scale, standard corpus hinders the research in this area: it is difficult to compare different models directly since they are based on different corpora and follow different bridging definitions. In addition, research in the field of NLP has shown that the performance of statistical models can benefit significantly from much larger training sets (Banko & Brill, 2001). However, annotating a large-scale dataset for bridging will take a great deal of time and effort. A possible option is to harvest bridging pairs by exploring unsupervised or semi-supervised learning approaches, as we will discuss in the next section.

Focus on entity antecedents only. In this thesis, although we design a model which resolves unrestricted bridging in the sense that bridging anaphors are not limited to definite NPs and bridging relations are not limited to part-of relations, we do not consider non-entity antecedents (e.g., verbs or clauses). The reason mainly lies in the rarity of the phenomenon, i.e., only 6% of the anaphors in ISNotes have a non-entity antecedent. However, it could be interesting to investigate the properties of bridging anaphors with non-entity antecedents in a larger scale corpus and to develop algorithms to resolve them.

8.3   Future Work

In this section, we discuss a few promising directions of future research that can be extensions of the work presented in this dissertation.

Interactions between bridging resolution, textual entailment and implicit semantic role labeling. *Bridging resolution, textual entailment* and *implicit semantic role labeling* are three standard tasks in NLP and have been studied respectively. They share some common properties. Recently, there are a few efforts that try to "bridge" boundaries between these tasks: Mirkin et al. (2010) show that textual entailment (TE) recognition can benefit from bridging resolution; Stern & Dagan (2014) extract potential instances of implicit predicate-argument relations from a RTE (Recognizing Textual Entailment) dataset using a semi-automatic method. They further improve the performance of textual entailment recognition by exploring implicit semantic role labeling. It would be interesting to further explore interactions between these three tasks.

For instance, although FrameNet itself does not cover text cohesion or anaphora information, the annotation of *null instantiation* (NI) in single sentences offers lexico-semantic/syntactic resources for bridging anaphora recognition. According to the frame semantics paradigm, the non-locally realized core semantic roles (also called *Core Frame Elements*) are considered as *null instantiations*. NIs are divided into three categories: *definite null instantiations* (DNI) are NIs whose fillers are accessible from the context. In contrast, the fillers for *indefinite null instantiations* (INI) and *constructional null instantiations* (CNI) are inaccessible from the context because such omissions are licensed by particular lexical items or grammatical constructions. The following examples are from FrameNet (Baker et al., 1998). They are all related to the *statement* frame which contains three main core frame elements, i.e., *speaker*, *message*, and *topic*. In Example 8.1, the filler of *message* should be accessible from the context (although it was not annotated), while in Example 8.2 and Example 8.3, the fillers of *speaker* are allowed to be unspecified from the context.

(8.1)   These **claims**$_{statement}$ have yet to be fully investigated , much less verified. (DNI$_{message}$)

(8.2)   The media have a right to publish defamatory **remarks**$_{statement}$ at the risk of paying heavy damages if they can not subsequently be justified. (INI$_{speaker}$)

(8.3)   It is no longer possible to make **claims**$_{statement}$ to understand a culture simply by classifying it in terms of its relations to a present western culture. (CNI$_{speaker}$)

In our model for bridging anaphora recognition, we notice that generic expressions are easily confused with bridging anaphora (Chapter 5). The theory of *null instantiation* along with its annotation provides a new perspective for bridging anaphora identification, i.e., certain lexical/grammatical patterns can indicate DNI and INI/CNI, which are related to bridging anaphora and generic expressions respectively. Therefore one might improve the performance for bridging anaphora recognition via learning such patterns from the FrameNet DNI/INI/CNI annotation.

Building a large-scale corpus for bridging semi-automatically. Large-scale annotated corpora such as Treebank (Marcus et al., 1993), PropBank (Palmer et al., 2005) and FrameNet (Baker et al., 1998) play an important role in NLP research. As benchmarks, they provide a basis for comparing results obtained by independent researchers and encourage the development of novel ideas and algorithms. However, the construction of such corpora by linguists is expensive in both time and money. Therefore, various researchers have studied how to build large-scale corpora efficiently, such as collecting annotations from cheap non-expert annotations via controlling labeler bias (Snow et al., 2008).

Given the difficulty of the task itself, we cannot expect that a large-scale corpus for bridging, which is reliably annotated by linguists, will occur any time soon. We also cannot expect that non-expert annotations can perform well for this task since the results of previous efforts to construct bridging corpora by experts are not always satisfactory (see discussion in Section 2.1.2). An option is to combine harvesting (potential) bridging pairs by exploring semi-supervised or unsupervised learning approaches with expert/non-expert annotations.

For instance, the work of Roth & Frank (2012) can be adapted to obtain potential bridging pairs automatically. Roth & Frank (2012) propose a method to identify implicit arguments by aligning and comparing predicate-argument structures across pairs of comparable texts under an unsupervised framework. They then find discourse antecedents of implicit arguments heuristically. Among all induced instances, the nominal predicates and the discourse antecedents (of the core implicit arguments of these predicates) can be seen as potential bridging instances.

Joint bridging resolution and coreference resolution. Identity coreference is a relatively well understood and well-studied instance of entity coherence. However, entity coherence can rely on more complex relations than identity, e.g., various bridging relations. Currently, these two tasks (i.e., bridging resolution and coreference resolution) are studied separately. In principle, they are both subtasks of *anaphora resolution* and should benefit from each other. For instance, in Example 8.4, knowing that "**The opening show**" is associated to "*Mancuso FBI*" and "the show" is coreferent with "*Mancuso FBI*", one can infer that "**The opening show**" and "the show" do not refer to the same entity.

(8.4)   Over the first few weeks, *Mancuso FBI* has sprung straight from the headlines.
           The opening show featured a secretary of defense designate accused of womanizing
           (a la John Tower).
           . . .
           Most of all though, the show is redeemed by the character of Mancuso.

Therefore an interesting direction for further research is to model these two tasks (i.e., *bridging resolution* and *coreferent resolution*) jointly. This joint framework can enhance the

entity-based coherence model (Barzilay & Lapata, 2008) as we have discussed in Section 1.2, and has the potential to benefit other related applications such as *readability assessment* or *sentence ordering*. Furthermore, as we will discuss in the next part, this joint framework can also be applied to relation extraction tasks to infer relations between entities which do not occur in the same sentence.

Relation extraction across sentence boundaries. Recently, there is a considerable interest in applying unsupervised or distantly supervised approaches for open-domain relation extraction (e.g., Hasegawa et al. (2004), Pantel & Pennacchiotti (2006), Yao et al. (2011), Surdeanu et al. (2012), Riedel et al. (2013)). The majority of research in this area only considers extracting relations between entities occurring in the same sentence. However, Swampillai & Stevenson (2010) found that 28.5% of the relations occur between entities in different sentence in an analysis of the MUC-6 corpus. Ji & Grishman (2011) reported that only 60.4% of all relations occur between entities in the same sentence in the training dataset of TAC-2010 Knowledge Base Population (KBP) track. Indeed, one could apply bridging resolution to extract relations across sentence boundaries. For instance, in Example 8.5, traditional relation extraction systems can extract the *person-position-company* relations encoded in s1: {*Roberb S. Ehrlich-chairman-Delmed, Roberb S. Ehrlich-president-Delmed, Roberb S. Ehrlich-chief executive-Delmed*}. By applying coreference resolution and bridging resolution, we know that "Mr Ehrlich" is coreferent with "Robert S.Ehrlich", "**a director**" and "**a consultant**" is bridged to "*Delmed*". We then can extract new relations across sentences: {*Roberb S. Ehrlich-director-Delmed, Roberb S. Ehrlich-consultant-Delmed*}.

(8.5)   s1: *Delmed* said Robert S.Ehrlich resigned as **chairman**, **president** and **chief executive**.
         s2: Mr Ehrlich will continue as **a director** and **a consultant**.

Information status and applications. In this thesis we also introduced an efficient collective algorithm for information status classification. The model achieves a reasonable performance on the ISNotes corpus, with an accuracy of 78.9% for fine-grained information status classification. In the future, it will be interesting to evaluate the utility of our information status classification system in different applications, such as pitch accent generation and coreference resolution.

Appendix A

The Baseline in Chapter 7

We reimplement the rule-based system for processing definite NPs proposed by Vieira & Poesio (2000). We adapt this algorithm to resolve bridging in ISNotes. This appendix provides the details of our implementation.

A.1   Preprocessing

All lexical units annotated as "mention" in ISNotes are treated as NPs. We use the syntactic tree annotation from OntoNotes to extract the head of an NP as well as its various modifications (if such a modification exists). In the preprocessing step, the following NPs which are rarely used as bridging anaphora are filtered out:

- pronouns such as *he* or *she*

- genitive NPs and NPs that are modified by possessive pronouns, such as *Mr. Bush's problem* and *his mother*

- demonstrative NPs that are modified by *this*, *that*, *these* and *those*

- comparative NPs that are modified by comparison markers[1] or modified by adjectives or adverbs in the comparative form[2]

---

[1] The full list of such markers is: *{other, another, such, different, similar, additional, comparable, same, further, extra}*.

[2] Later the "special predicate" node in the heuristic decision tree (Section A.2) also handles comparative NPs in certain forms.

A.2   A Heuristic Decision Tree

After preprocessing, each NP is passed to a heuristic decision tree to decide which category it belongs to. Vieira & Poesio (2000) distinguish four categories: *discourse new*, *direct anaphora* (same-head coreference), *lenient bridging*, and *Unknown*. However, their *lenient bridging* include anaphors which are coreferent with antecedents but do not share the same head noun, as well as *associative anaphora* (= *bridging anaphora* in this thesis) which are not coreferent with the antecedents. Therefore we split their *lenient bridging* into two categories: *different-head coreference* and *bridging*. The latter corresponds to our bridging definition. The decision tree of the algorithm is presented in Figure A.1 and Figure A.2.
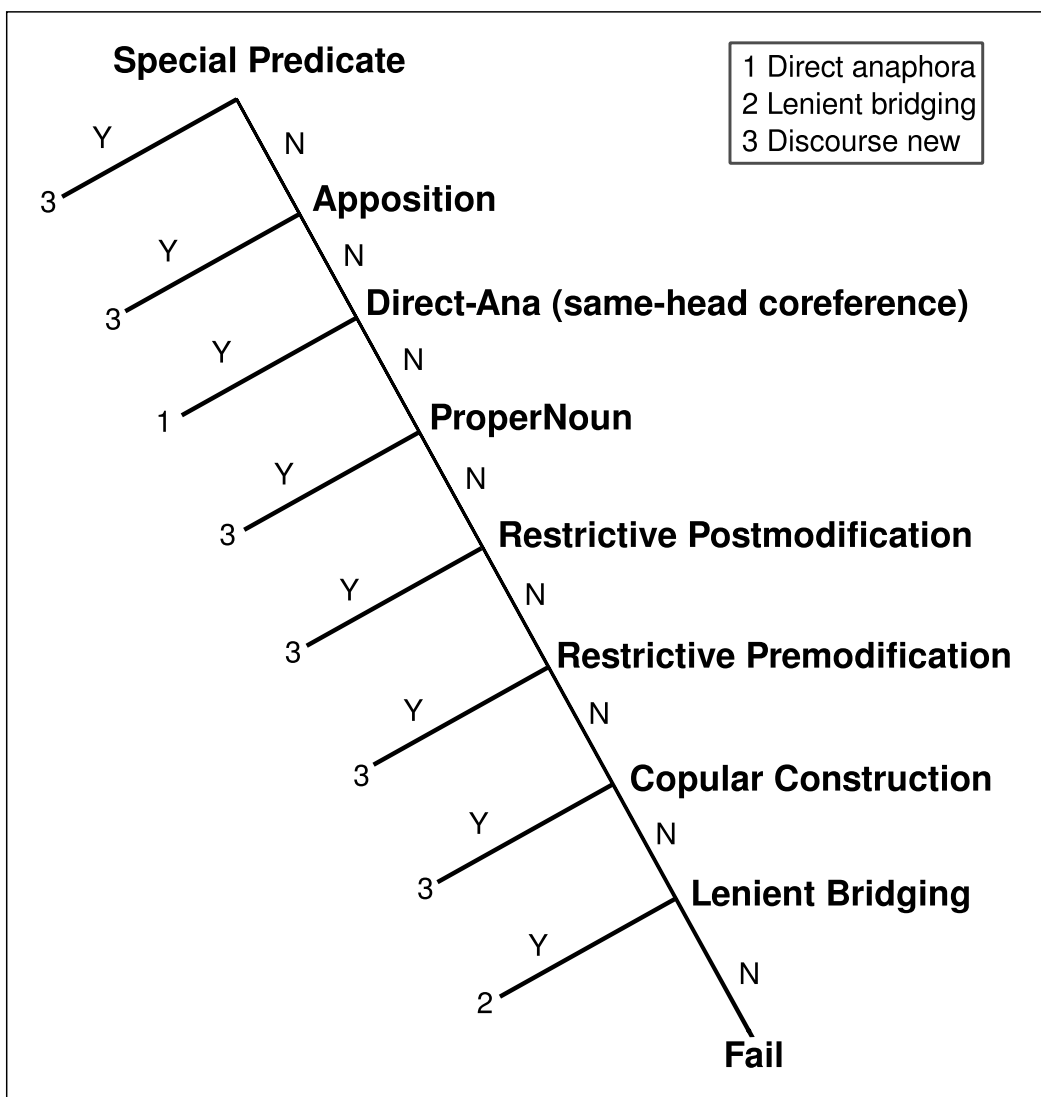
Figure A.1: Vieira & Poesio (2000)'s original algorithm for processing definite NPs.
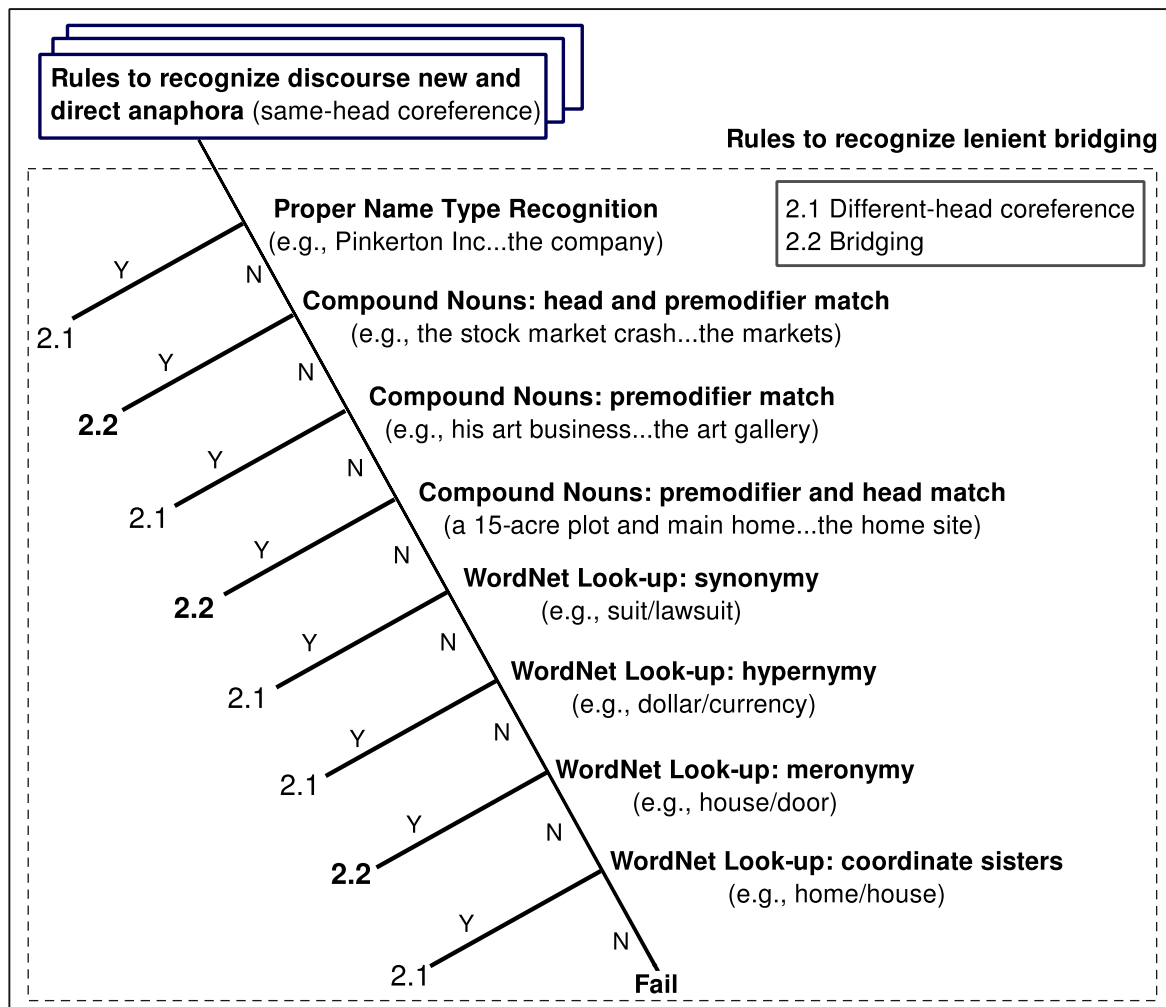
Figure A.2: Vieira & Poesio (2000)'s original algorithm for processing definite bridging NPs. We further divide their lenient bridging category into two subcategories: *2.1 Different-head coreference* and *2.2 Bridging*.

We describe the details of each processing node in the following.

Special predicate. The system checks whether an NP belongs to one of the following cases:

- NPs which has a complement[3] and its head lemma appears in the list: {*fact, result, conclusion, idea, belief, saying, remark*}.

- NPs whose modifier is in the list: {*first, last, best, most, maximum, minimum, only*} or NPs whose modifier is the comparative/superlative form of adjectives or adverbs. For

---

[3]We take an NP's post modification as its complement.

NPs whose modifier is *first* or *last* or comparatives, the presence of the complement is checked.

- NPs whose head lemma appears in the list: {*year, day, week, month, hour, time, morning, afternoon, night, period, quarter*}.

- NPs appears in the list: {*the noon, the sky, the pope, the weather*}.

An NP belonging to one of the above cases is predicted as *discourse new*.

Apposition. The system checks whether an NP is an appositive construction by exploring the syntactic tree. If so, the NP is predicted as *discourse new*.

Direct-Ana. The system checks whether an NP is a direct anaphor and predicts its antecedent. For an NP n, we construct an antecedent candidates set A by exploring a loose segmentation heuristic: indefinite NPs, possessive NPs and definite NPs from the same sentence preceding n as well as from the previous four sentences are added to A. Additionally, an NP a whose distance to n is more than four sentences is added to A when either:

- a is a direct anaphor predicted by the system before; or

- a is identical to n.

We predict n is a direct anaphor when an NP a_{ante} from A has the same head as n and either:

- the premodification[4] of n is a subset of the premodification of a_{ante}; or

- a_{ante} has no premodifiers.

If more than one such NP are present, the closest one to n is predicted as its antecedent.

ProperNoun. The system checks whether an NP is a proper noun. If so, the NP is predicted as *discourse new*. Vieira & Poesio (2000) decide whether an NP is a proper noun by checking if it is capitalized. We instead use the OntoNotes named entity annotation which provides us with more accurate proper name information.

---

[4]Here we exclude determiners from premodification.

Restrictive postmodification. The system checks whether an NP is (post) modified by a relative clause (Example A.1) or a prepositional phrase (Example A.2) or other non-finite post-modifiers (Example A.3). The non-finite post-modifiers include *ing*, *ed (participle)* and infinitive clauses. An NP with such restrictive postmodifiers is classified as *discourse new*.

(A.1)   *The girl who I met …*

(A.2)   *The political importance of California …*

(A.3)   It doesn't make *the equipment needed to produce those chips*.

Restrictive premodification. The system checks whether an NP has a proper noun premodifier. If so, the NP is considered as *discourse new*.

Copular construction. The system checks whether an NP occurs in a copular construction. If so, the NP is classified as *discourse new*. The NP is in a copular construction if either:

- it is a subject and the main verb of the clause is *to be*, *to seem* or *to become* and the complement of the verb is not an adjective phrase; or

- it is an object of the verb *to be*.

Proper name type recognition. The system tries to identify the *different-head coreference* relation between a common noun and its proper name antecedent, such as *IBM…the company*. Vieira & Poesio (2000) explore appositive construction, abbreviations like *Mr., Inc.* as well as WordNet to decide the type of a proper name. They query WordNet to confirm whether a proper name is an instance of one of elements from the list: {*country, city, state, continent, language, person*}. Apart from this, we further look up gazetteers to get type information for named entities. If an NP's head matches the type of a preceding proper name, the system considers the NP as *different-head coreference* and predicts the closest such proper name as its antecedent.

Compound nouns: head and premodifier match. The system checks whether an NP's head lemma is the same as the premodifier of a previous NP (Example A.4). If so, the NP is classified as *bridging*, and the closest such preceding NP is predicted as its antecedent.

(A.4)   *the stock market crash …* the markets

Compound nouns: premodifier match. The system checks whether an NP's premodifier is the same as the premodifier of a previous NP (Example A.5). If so, the NP is classified as *different-head coreference* and the closest such preceding NP is chosen as its antecedent.

(A.5) *his art business* . . . the art gallery

Compound nouns: premodifier and head match. The system checks whether an NP's premodifier is the same as the head of a previous NP (Example A.6). If so, the NP is classified as *bridging* and the closest such preceding NP is considered as its antecedent.

(A.6) *a 15-acre plot and main home* . . . the home site

WordNet look-up: synonymy. The system queries WordNet to decide whether an NP's head and a previous NP's head is in the same synset (e.g., *suit – lawsuit*). If so, the NP is considered as *different-head coreference* and the closest such preceding NP is predicted as its antecedent.

WordNet look-up: hypernymy. The system queries WordNet to decide whether a hypernymy or hyponymy relation is held between an NP's head and a previous NP's head (e.g., *dollar – currency*). If so, the NP is considered as *different-head coreference* and the closest such preceding NP is predicted as its antecedent.

WordNet look-up: meronymy. The system queries WordNet to decide whether a meronymy or holonymy relation is held between an NP's head and a previous NP's head (e.g., *house – door*). If so, the NP is considered as *bridging* and the closest such preceding NP is predicted as its antecedent.

WordNet look-up: coordinate sisters. The system queries WordNet to decide whether an NP's head and a previous NP's head are hyponyms of the same hypernym (e.g., *house* and *home* are hyponyms of *housing*). If so, the NP is considered as *different-head coreference* and the closest such preceding NP is chosen as its antecedent.

List of Figures

List of Tables

Bibliography

Abe, Naoki, Binanca Zadrozny & John Langford (2004). An iterative method for multi-class cost-sensitive learning. In *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* Seattle, Wash., 22–25 August 2004, pp. 3–11.

Agichtein, Eugene & Luis Gravano (2000). Extracting relations from large plain text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pp. 85–94.

Arnold, Jennifer E., , Anthony Losongco, Thomas Wasow & Ryan Ginstrom (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.

Artstein, Ron & Massimo Poesio (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Asher, Nicholas & Alex Lascarides (1998). Bridging. *Journal of Semantics*, 15:83–113.

Asher, Nicholas & Alex Lascarides (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Baker, Collin F., Charles J. Fillmore & John B. Lowe (1998). The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* Montréal, Québec, Canada, 10–14 August 1998, pp. 86–90.

Banko, Michele & Eric Brill (2001). caling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics,* Toulouse, pp. 26–33.

Banko, Michele, Michael J. Cafarella, Stephen Soderland, Matt Broadhead & Oren Etzioni (2007). Open information extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* Hyderabad, India, 6–12 January 2007, pp. 2670–2676.

Barzilay, Regina & Mirella Lapata (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Baumann, Stefan & Arndt Riester (2013). Coreference, lexical givenness and prosody in German. *Lingua*. Accepted.

Bengtson, Eric & Dan Roth (2008). Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pp. 294–303.

Björkelund, Anders, Kerstin Eckart, Arndt Riester, Nadja Schauffler & Katrin Schweitzer (2014). The extended DIRNDL corpus as a resource for coreference and bridging resolution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation,* Reykjavik, Iceland, 26–31 May 2014, pp. 3222–3228.

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge & Jamie Taylor (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 1247–1250.

Bos, Johan, Paul Buitelaar & Anne Marie Mineur (1995). Bridging as coercive accomodation. In E. Klein, S. Manandhar, W. Nutt & J. Siekmann (Eds.), *Working Notes of the Edinburgh Conference on Computational Logic and Natural Language Processing (CLNLP-95), Human Communications Research Centre, University of Edinburgh, Edinburgh, U.K.*

Brants, Thorsten & Alex Franz (2006). *Web 1T 5-gram Version 1*. LDC2006T13, Philadelphia, Penn.: Linguistic Data Consortium.

Brin, Sergey (1998). Extracting patterns and relations from the World Wide Web. In *Proceedings of the Workshop on the Web and Databases at the 6th International Conference on Extending Database Technology*, pp. 172–183.

Bunescu, Razvan (2003). Associative anaphora resolution: A Web-based approach. In *Proceedings of the EACL 2003 Workshop on The Computational Treatment of Anaphora,* Budapest, Hungary, 14 April, 2003, pp. 47–52.

Burfoot, Clinton, Steven Bird & Timothy Baldwin (2011). Collective classification of congressional floor-debate transcipts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* Portland, Oreg., 19–24 June 2011, pp. 1506–1515.

Cahill, Aoife & Arndt Riester (2009). Incorporating information status into generation ranking. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association*

for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, Singapore, 2–7 August 2009, pp. 817–825.

Cahill, Aoife & Arndt Riester (2012). Automatically acquiring fine-grained information status distinctions in German. In *Proceedings of the SIGdial 2012 Conference: The 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue,* Seoul, Korea, 5–6 July 2012, pp. 232–236.

Carletta, Jean (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carreras, Xavier & Lluís Màrquez (2004). Introduction to the CoNLL-2004 shared task: Semantic Role Labeling. In *Proceedings of the 8th Conference on Computational Natural Language Learning,* Boston, Mass., USA, 6–7 May 2004, pp. 89–97.

Carreras, Xavier & Lluís Màrquez (2005). Introduction to the CoNLL-2005 shared task: Semantic Role Labeling. In *Proceedings of the 9th Conference on Computational Natural Language Learning,* Ann Arbor, Mich., USA, 29–30 June 2005, pp. 152–164.

Caselli, Tommaso & Irina Prodanof (2006). Annotating bridging anaphors in Italian: In search of reliability. In *Proceedings of the 5th International Conference on Language Resources and Evaluation,* Genoa, Italy, 22–28 May 2006.

Cederberg, Scott & Dominic Widdows (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the 7th Conference on Computational Natural Language Learning,* Edmonton, Alberta, Canada, 31 May – 1 June 2003, pp. 111–118.

Chen, Desai, Nathan Schneider, Dipanjan Das & Noah A. Smith (2010). SEMAFOR: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2),* Uppsala, Sweden, 15–16 July 2010, pp. 264–267.

Cimiano, Philipp (2006). Ingredients of a first-order account of bridging. In *Proceedings of the 5th International Workshop on Inference in Computational Semantics,* Buxton, U.K., 20–21 April 2006, pp. 139–144.

Clark, Herbert H. (1975). Bridging. In *Proceedings of the Conference on Theoretical Issues in Natural Language Processing,* Cambridge, Mass., June 1975, pp. 169–174.

Clark, Herbert H. & Susan E. Haviland (1977). Comprehension and the given-new contract. In Roy Freedle (Ed.), *Discourse processes: Advances in research and theory*, Vol. 1, pp. 1–40. Norwood, NJ: Ablex.

Collins, Michael (2002). Discriminative training methods for hidden Markov models. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing,* Philadelphia, Penn., 6–7 July 2002, pp. 1–8.

Collins, Michael & Nigel Duffy (2002). New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* Philadelphia, Penn., 7–12 July 2002, pp. 1–8.

Cortes, Corinna & Vladimir Vapnik (1995). Support vector networks. *Machine Learning*, 20(3):273–297.

Culotta, Aron & Jeffrey Sorensen (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pp. 423–429.

Dagan, Ido, Oren Glickman & Bernardo Magnini (2006). The PASCAL recognising textual entailment challenge. In J. Quiñonero-Candela, I. Dagan & B. Magnini (Eds.), *Machine Learning Challenges*, pp. 177–190. Heidelberg, Germany: Springer.

Daneš, František (1974). Functional sentence perspective and the organization of the text. In F. Daneš (Ed.), *Papers on Functional Sentence Perspective*, pp. 106–128. Prague: Academia.

Das, Dipanjan, Nathan Schneider, Desai Chen & Noah A. Smith (2010). Probabilistic frame-semantic parsing. In *Proceedings of Human Language Technologies 2010: The Conference of the North American Chapter of the Association for Computational Linguistics,* Los Angeles, Cal., 2–4 June 2010, pp. 948–956.

Dempster, A. P., N. M. Laird & D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Domingos, Pedro & Daniel Lowd (2009). *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan Claypool Publishers.

Dunning, Ted (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Eckart, Kerstin, Arndt Riester & Katrin Schweitzer (2012). A discourse information radio news database for linguistic analysis. In Christian Chiarcos, Sebastian Nordhoff & Sebastian Hellmann (Eds.), *Linked Data in Linguistics*, pp. 65–76. Springer Berlin Heidelberg.

Eckert, Miriam & Michael Strube (2000). Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Erkü, Feride & Jeanette K. Gundel (1987). The pragmatics of indirect anaphors. In Jef Verschueren & Marcella Bertuccelli-Papi (Eds.), *The pragmatic perspective: Selected papers from the 1985 International Pragmatics Conference*, pp. 533–545. Amsterdam: John Benjamins.

Fahrni, Angela & Michael Strube (2012). Jointly disambiguating and clustering concepts and entities with Markov logic. In *Proceedings of the 24th International Conference on Computational Linguistics,* Mumbai, India, 8–15 December 2012, pp. 815–832.

Fan, James, Ken Barker & Bruce Porter (2005). Indirect anaphora resolution as semantic path search. In *K-CAP '05: Proceedings of the 3rd International Conference on Knowledge Capture*, pp. 153–160.

Fawcett, Tom (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874. Article No.10.

Fellbaum, Christiane (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Fraurud, Kari (1990). Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7:395–433.

Gardent, Claire & Hélène Manuélian (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues*, 46(1):115–140.

Garrod, Simon C. & Anthony J. Sanford (1982). The mental representation of discourse in a focussed memory system: Implications for the interpretation of anaphoric noun phrases. *Journal of Semantics*, (1):21–41.

Gerber, Matthew & Joyce Chai (2012). Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*, 38(4):756–798.

Getoor, Lise, Nir Friedman, Daphne Koller & Avi Pfeffer (2001). Learning probabilistic relational models. In S. Dzeroski & N. Lavrac (Eds.), *Relational Data Mining*, pp. 307–335. Springer-Verlag.

Getoor, Lise & Ben Taskar (Eds.) (2007). *Introduction to Statistical Relational Learning*. Cambridge, Mass.: MIT Press.

Gildea, Daniel & Daniel Jurafsky (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Grice, H. Paul (1975). Logic and conversation. In P. Cole & J.L. Morgan (Eds.), *Syntax and Semantics 3: Speech Acts*, pp. 41–58. New York, N.Y.: Academic Press.

Grimes, Joseph E. (1975). *The Thread of Discourse*. The Hague, Netherlands: Mouton.

Grosz, Barbara J., Aravind K. Joshi & Scott Weinstein (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Grosz, Barbara J. & Candace L. Sidner (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Gundel, Jeanette K., Nancy Hedberg & Ron Zacharski (2000). *Cognitive status and the form of indirect anaphors*.

Hahn, Udo, Michael Strube & Katja Markert (1996). Bridging textual ellipses. In *Proceedings of the 16th International Conference on Computational Linguistics,* Copenhagen, Denmark, 5–9 August 1996, Vol. 1, pp. 496–501.

Halliday, M. A. K. & Ruqaiya Hasan (1976). *Cohesion in English*. London, U.K.: Longman.

Harabagiu, Sanda, Dan Moldovan, Marius Paşca, Mihai Surdeanu, Rada Mihalcea, Roxana Gîrju, Vasile Rus, Finley Lăcătuşu, Paul Morărescu & Răzvan Bunescu (2001). Answering complex list and context questions with LCC's Question-Answering Server. In *Proceedings of the Tenth Text REtrieval Conference,* Gaithersburg, Md., 13–16 November 2001.

Harman, Donna & Mark Liberman (1993). *TIPSTER Complete*. LDC93T3A, Philadelphia, Penn.: Linguistic Data Consortium.

Harris, Zellig S. (1954). Distributional structure. *Word*, 10:146–162.

Hasegawa, Takaaki, Satoshi Sekine & Ralph Grishman (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pp. 415–422.

Hawkins, John A. (1978). *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. Atlantic Highlands, N.J.: Humanities Press.

He, Haibo & Edwardo A. Garcia (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Hearst, Marti A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 15th International Conference on Computational Linguistics,* Nantes, France, 23-28 August 1992, pp. 539–545.

Hirschman, Lynette & Nancy Chinchor (1997). *MUC-7 Coreference Task Definition,* `http://www.muc.saic.com/proceedings/`.

Hobbs, Jerry R. (1978). Resolving pronominal references. *Lingua*, 44:311–338.

Hobbs, Jerry R. (1979). Coherence and coreference. *Cognitive Science*, 3:67–90.

Hobbs, Jerry R., Mark E. Stickel, Douglas E. Appelt & Paul Martin (1993). Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

Hoffmann, Raphael, Congle Zhang, Xiao Ling, Luke Zettlemoyer & Daniel S Weld (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* Portland, Oreg., 19–24 June 2011, pp. 541–550.

Hou, Yufang, Katja Markert & Michael Strube (2013a). Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* Seattle, Wash., 18–21 October 2013, pp. 814–820.

Hou, Yufang, Katja Markert & Michael Strube (2013b). Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Atlanta, Georgia, 9–14 June 2013, pp. 907–917.

Hou, Yufang, Katja Markert & Michael Strube (2014). A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* Doha, Qatar, 25–29 October 2014, pp. 2082–2093.

Huynh, Tuyen N. & Raymond J. Mooney (2008). Discriminative structure and parameter learning for Markov logic networks. In *Proceedings of the 25th International Conference on Machine Learning,* Helsinki, Finland, 5–9 July 2008, pp. 416–423.

Irmer, Matthias (2009). *Bridging Inferences in Discourse Interpretation*, (Ph.D. thesis). Leipzig University.

Jensen, David, Jennifer Neville & Brian Gallagher (2004). Why collective inference improves relational classification. In *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* Seattle, Wash., 22–25 August 2004, pp. 593–598.

Ji, Heng & Ralph Grishman (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* Portland, Oreg., 19–24 June 2011, pp. 1148–1158.

Joachims, Thorsten (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning,* Chemnitz, Germany, 21–23 April 1998, pp. 137–142.

Joachims, Thorsten (1999). Making large-scale support vector machine learning practical. In Bernhard Schölkopf, Christopher J. C. Burges & Alexander J. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning*, pp. 169–184. Cambridge, MA, USA: MIT Press.

Johansson, Richard & Pierre Nugues (2007). Extended constituen-to-denpendency conversion for English. In *Proceedings of the 16th Nordic Conference of Computational Linguistics,* Tartu, Estonia, 25–26 May 2007, pp. 105–112.

Kambhatla, Nanda (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pp. 136–143.

Kamp, Hans & Uwe Reyle (1993). *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Dordrecht, The Netherlands: Kluwer.

Kautz, Henry, Bart Selman & Yueyen Jiang (1996). A general stochastic approach to solving problems with hard and soft constraints. In *The Satisfiability Problem: Theory and Applications*, pp. 573–586. American Mathematical Society.

Kehler, Andrew & Hannah Rohde (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1/2):1–38.

Knott, Alistair, Jon Oberlander, Michael O'Donnell & Chris Mellish (2001). Beyond elaboration: The interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord & W. Spooren (Eds.), *Text Representation: Linguistic and Psycholinguistic Aspects*, pp. 181–196. Amsterdam, The Netherlands: John Benjamins.

Kobayashi, Nozomi, Kentaro Inui & Yuji Matsumoto (2007). Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning,* Prague, Czech Republic, 28–30 June 2007, pp. 1065–1074.

Kok, Stanley & Pedro Domingos (2005). Learning the structure of Markov logic networks. In *Proceedings of the International Conference on Machine Learning,* Bonn, Germany, 7–11 August 2005, pp. 441–448.

Korzen, Iorn & Matthias Buch-Kromann (2011). Anaphoric relations in the Copenhagen dependency treebanks. In S. Dipper & H. Zinsmeister (Eds.), *Corpus-based Investigations of Pragmatic and Discourse Phenomena*, Vol. 3, Bochumer Linguistische Arbeitsberichte, pp. 83–98. University of Bochum, Bochum, Germany.

Kucera, W. Nelson & Henry Francis (1967). *Computational Analysis of Present Day English.* Providence, Rhode Island: Brown University Press.

Lafferty, John, Andrew McCallum & Fernando Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning,* Williamstown, Mass., 28 June – 1 July 2001, pp. 282–289.

Laparra, Egoitz & German Rigau (2012). Exploiting explicit annotations and semantic types for implicit argument resolution. In *Proceedings of the 6th IEEE International Conference on Semantic Computing (ICSC 2012), Palermo, Italy, 19-21 September 2012*, pp. 75–78.

Laparra, Egoitz & German Rigau (2013). ImpAr: A deterministic algorithm for implicit semantic role labelling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,* Sofia, Bulgaria, 4–9 August 2013, pp. 1180–1189.

Lassalle, Emmanuel & Pascal Denis (2011). Leveraging different meronym discovery methods for bridging resolution in French. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011),* Faro, Algarve, Portugal, 6–7 October 2011, pp. 35–46.

LDC (1993). *Switchboard.* Linguistic Data Consortium. University of Pennsylvania, Philadelphia, Penn.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu & Dan Jurafsky (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Lin, Dekang & Patrick Pantel (2001). DIRT-Discovery of inference rules from text. In *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining*, pp. 323–328.

Löbner, Sebastian (1985). Definites. *Journal of Semantics*, 4:279–326.

Löbner, Sebastian (1998). *Definite associative anaphora*. Unpublished Manuscript, Heinrich-Heine-Universität Düsseldorf.

Longacre, Robert (1976). *An Anatomy of Speech Notions*. Ghent, Belgium: The Peter de Ridder Press.

Louis, Annie & Ani Nenkova (2010). Creating local coherence: An empirical assessment. In *Proceedings of Human Language Technologies 2010: The Conference of the North American Chapter of the Association for Computational Linguistics,* Los Angeles, Cal., 2–4 June 2010, pp. 313–316.

Lowd, Daniel & Pedro Domingos (2007). Efficient weight learning for Markov logic networks. In *Proceedings of the 11th European Conference on Principles and Practices of Knowledge Discovery in Databases,* Warsaw, Poland, 17–21 September 2007, pp. 200–211.

Macskassy, Sofus A. & Foster Provost (2007). Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983.

Mann, William C. & Sandra A. Thompson (1987). *A theory of text organisation*. Technical Report ISI/RS-87-190: Information Sciences Institute at the University of Southern California, Marina del Rey, Cal. Available at: http://www.sfu.ca/rst/pdfs/Mann_Thompson_1987.pdf.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

Markert, Katja (2013). *Annotation scheme for informations status for ISNotes 1.0*. Available at: http://www.h-its.org/english/research/nlp/download/isnotes.php.

Markert, Katja, Yufang Hou & Michael Strube (2012). Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics,* Jeju Island, Korea, 8–14 July 2012, pp. 795–804.

Markert, Katja, Malvina Nissim & Natalia N. Modjeska (2003). Using the web for nominal anaphora resolution. In *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora.* Budapest, Hungary, 14 April 2003, pp. 39–46.

Matsui, Tomoko (2000). *Bridging and Relevance*. John Benjamins Publishing.

McKinlay, Andrew (2013). *Modelling Entity Instantiations*, (Ph.D. thesis). University of Leeds.

McKinlay, Andrew & Katja Markert (2013). Recognising sets and their elements: Tree kernels for entity instantiation identification. In *Proceedings of the 10th International Conference on Computational Semantics,* Potsdam, Germany, 19–22 March 2013, pp. 167–178.

McNemar, Quinn (1947). Note on the sampling errors of the difference between correlated proportions or percentages. *Phychometrika*, 12(2):153–157.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young & Ralph Grishaman (2004). Annotating noun argument structure for NomBank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* Lisbon, Portugal, 26–28 May 2004, pp. 803–806.

Meza-Ruiz, Ivan & Sebastian Riedel (2009). Jointly identifying predicates, arguments and senses using Markov logic. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics,* Boulder, Col., 31 May – 5 June 2009, pp. 155–163.

Milch, Brian, Bhaskara Marthi, Stuart Russell, David Sontag, Daniel L. Ong & Andrey Kolobov (2007). Blog: Probabilistic models with unknown objects. In L. Getoor & B. Taskar (Eds.), *Introduction to Statistical Relational Learning*, pp. 373–394. Cambridge, Mass.: MIT Press.

Mintz, Mike, Steven Bills, Rion Snow & Dan Jurafsky (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing,* Singapore, 2–7 August 2009, pp. 1003–1011.

Mirkin, Shachar, Ido Dagan & Sebastian Padó (2010). Assessing the role of discourse references in entailment inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pp. 1209–1219.

Mitchell, Alexis, Stephanie Strassel, Mark Przybocki, JK Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstain, Lisa Ferro & Beth Sundheim (2002). *ACE-2 Version 1.0*. LDC2003T11, Philadelphia, Penn.: Linguistic Data Consortium.

Modjeska, Natalia M., Katja Markert & Malvina Nissim (2003). Using the web in machine learning for other-anaphora resolution. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing,* Sapporo, Japan, 11–12 July 2003, pp. 176–183.

Moore, Robert C. (2004). On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* Barcelona, Spain, 25–26 July 2004, pp. 333–340.

Moschitti, Alessandro (2006). Making tree kernels practical for natural language learning. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* Trento, Italy, 3–7 April 2006, pp. 113–120.

Moschitti, Alessandro, Silvia Quarteroni, Roberto Basili & Suresh Manandhar (2007). Exploiting syntactic and shallow semantic kernels for question answer classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June 2007, pp. 776–783.

Nedoluzhko, Anna, Jiří Mírovskỳ & Petr Pajas (2009). The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague dependency treebank. In *Proceedings of the Third Linguistic Annotation Workshop at ACL-IJCNLP 2009,* Suntec, Singapore, 6–7 August 2009, pp. 108–111.

Nenkova, Ani, Jason Brenier, Anubha Kothari, Sasha Calhoun, Laura Whitton, David Beaver & Dan Jurafsky (2007). To memorize or to predict: Prominence labeling in conversational speech. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, N.Y., 22–27 April 2007, pp. 9–16.

Neville, Jennifer & David Jensen (2003). Collective classification with relational dependency networks. In *Proceedings of the 2nd International Workshop on Multi-Relational Data Mining (MRDM-2003) at KDD-03,* Waschington DC, USA, 27 August 2003, pp. 77–91.

Ng, Vincent (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pp. 1396–1411.

Ng, Vincent & Claire Cardie (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* Philadelphia, Penn., 7–12 July 2002, pp. 104–111.

Nissim, Malvina (2006). Learning information status of discourse entities. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* Sydney, Australia, 22–23 July 2006, pp. 94–012.

Nissim, Malvina, Shipara Dingare, Jean Carletta & Mark Steedman (2004). An annotation scheme for information status in dialogue. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* Lisbon, Portugal, 26–28 May 2004, pp. 1023–1026.

Omuya, Adinoyi, Vinodkumar Prabhakaran & Owen Rambow (2013). Improving the quality of minority class identification in dialog act tagging. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Atlanta, Georgia, 9–14 June 2013, pp. 802–807.

Osuna, Edgar, Robert Freund & Federico Girosi (1997). Training support vector machines: An application to face detection. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition,* San Juan, Puerto Rico, 17–19 June, 2004, pp. 130–136.

Palmer, Martha, Daniel Gildea & Paul Kingsbury (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Pang, Bo & Lillian Lee (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pantel, Patrick & Marco Pennacchiotti (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. pp. 113–120.

Parker, Robert, David Graff, Junbo Kong, Ke Chen & Kazuaki Maeda (2011). *English Gigaword Fifth Edition*. LDC2011T07.

Pedersen, Ted, Siddharth Patwardhan & Jason Michelizzi (2004). WordNet::Similarity – Measuring the relatedness of concepts. In *Companion Volume to the Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* Boston, Mass., 2–7 May 2004, pp. 267–270.

Poesio, Massimo (2003). Associate descriptions and salience: A preliminary investigation. In *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora.* Budapest, Hungary, 14 April 2003, pp. 31–38.

Poesio, Massimo (2004). The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue,* Cambridge, Mass., 30 April – 1 May 2004, pp. 154–162.

Poesio, Massimo, Tomonori Ishikawa, Sabine Schulte im Walde & Renata Vieira (2002). Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation,* Las Palmas, Canary Islands, Spain, 29–31 May 2002, pp. 1220–1225.

Poesio, Massimo, Rahul Mehta, Axel Maroudas & Janet Hitzeman (2004a). Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pp. 143–150.

Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio & Janet Hitzeman (2004b). Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3). 309-363.

Poesio, Massimo & Renata Vieira (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Poesio, Massimo, Renata Vieira & Simone Teufel (1997). Resolving bridging references in unrestricted text. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Text, Madrid, Spain, July 1997*, pp. 1–6.

Poon, Hoifung & Pedro Domingos (2006). Sound and efficient inference with probabilistic and deterministic dependencies. In *Proceedings of the 21st National Conference on Artificial Intelligence,* Boston, Mass., 16–20 July 2006, pp. 458–463.

Poon, Hoifung & Pedro Domingos (2008). Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pp. 650–659.

Poon, Hoifung & Pedro Domingos (2010). Unsupervised ontology induction from text. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pp. 296–305.

Powers, David MW (1998). Applications and explanations of Zipf's law. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing,* Granada, Spain, June 1998, pp. 151–160.

Prasad, Rashmi, Nikhil Dineh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi & Bonnie Webber (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation,* Marrakech, Morocco, 26 May – 1 June 2008.

Prince, Ellen F. (1981). Towards a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics*, pp. 223–255. New York, N.Y.: Academic Press.

Prince, Ellen F. (1992). The ZPG letter: Subjects, definiteness, and information-status. In W.C. Mann & S.A. Thompson (Eds.), *Discourse Description. Diverse Linguistic Analyses of a Fund-Raising Text*, pp. 295–325. Amsterdam: John Benjamins.

Pustejovsky, James (1995). *The Generative Lexicon.* Cambridge, Mass.: MIT Press.

Quinlan, J. Ross (1993). *C4.5: Programs for Machine Learning.* San Mateo, Cal.: Morgan Kaufman.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* Vienna, Austria. ISBN 3-900051-07-0.

Rahman, Altaf & Vincent Ng (2011). Learning the information status of noun phrases in spoken dialogues. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* Edinburgh, Scotland, U.K., 27–29 July 2011, pp. 1069–1080.

Rahman, Altaf & Vincent Ng (2012). Learning the fine-grained information status of discourse entities. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics,* Avignon, France, 23–27 April 2012, pp. 798–807.

Reiter, Nils & Anette Frank (2010). Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pp. 40–49.

Richardson, Matthew & Pedro Domingos (2006). Markov logic networks. *Machine Learning,* 62(1-2):107–136.

Riedel, Sebastian (2008). Improving the accuracy and efficiency of MAP inference for Markov logic. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence,* Helsinki, Finland, 9–12 July 2008, pp. 468–475.

Riedel, Sebastian, Limin Yao & Andrew McCallum (2010). Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Practice of Knowledge Discovery in Databases,* Barcelona, Spain, 20–24 September 2010, pp. 148–163.

Riedel, Sebastian, Limin Yao, Andrew McCallum & Benjamin M Marlin (2013). Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Atlanta, Georgia, 9–14 June 2013, pp. 74–84.

Riester, Arndt, David Lorenz & Nina Seemann (2010). A recursive annotation scheme for referential information status. In *Proceedings of the 7th International Conference on Language Resources and Evaluation,* La Valetta, Malta, 17–23 May 2010, pp. 717–722.

Rösiger, Ina & Simone Teufel (2014). Resolving coreference and associative noun phrases in scientific text. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics,* Gothenburg, Sweden, 26–30 April 2014, pp. 44–55.

Roth, Dan & Wen-tau Yih (2004). A linear programming formulation for global inference in natural language tasks. In *Proceedings of the 8th Conference on Computational Natural Language Learning,* Boston, Mass., USA, 6–7 May 2004, pp. 1–8.

Roth, Michael & Anette Frank (2012). Aligning predicates across monolingual comparable texts using graph-based clustering. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pp. 171–182.

Ruppenhofer, Josef, Caroline Sporleder, Roser Morante, Collin Baker & Martha Palmer (2010). SemEval-2010 Task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2),* Uppsala, Sweden, 15–16 July 2010, pp. 45–50.

Sasano, Ryohei & Sadao Kurohashi (2009). A probabilistic model for associative anaphora resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* Singapore, 6–7 August 2009, pp. 1455–1464.

Schulte im Walde, Sabine (1998). *Resolving Bridging Descriptions in High-Dimensional Space*, (Master's thesis). University of Edinburgh, Centre for Cognitive Science.

Schwarz, Monika (2000). *Indirekte Anaphern in Texten. Studien zur domänengebundenen Referenz und Kohärenz im Deutschen*. Tübingen, Germany: Niemeyer.

Sidner, Candace L. (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. Technical Report AI-Memo 537, Cambridge, Mass.: Massachusetts Institute of Technology, AI Lab.

Silberer, Carina & Anette Frank (2012). Casting implicit role linking as an anaphora resolution task. In *Proceedings of STARSEM 2012: The First Joint Conference on Lexical and Computational Semantics,* Montréal, Québec, Canada, 7–8 June 2012, pp. 1–10.

Singla, Parag & Pedro Domingos (2005). Discriminative training of Markov logic networks. In *Proceedings of the 20th National Conference on Artificial Intelligence,* Pittsburgh, Penn., 9–13 July 2005, pp. 868–873.

Singla, Parag & Pedro Domingos (2008). Lifted first-order belief propagation. In *Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence,* Chicago, Ill., 13–17 July 2008, pp. 1094–1099.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky & Andrew Ng (2008). Cheap and fast – But is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pp. 254–263.

Somasundaran, Swapna, Galileo Namata, Janyce Wiebe & Lise Getoor (2009). Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* Singapore, 6–7 August 2009.

Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Sperber, Dan & Deirdre Wilson (1986). *Relevance: Communication and Cognition*. Cambridge, MA, USA: Harvard University Press.

Stern, Asher & Ido Dagan (2014). Recognizing implied predicate-argument relationships in textual inference. In *Proceedings of the 52st Annual Meeting of the Association for Computational Linguistics,* Baltimore, USA, 22–27 June 2014.

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie & Cambridge Computer Associates (1966). *General Inquirer: A Computer Approach to Content Analysis*. Cambridge, Mass.: MIT Press.

Surdeanu, Mihai, Richard Johansson, Adam Meyers, Lluís Màrquez & Joakim Nivre (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the 12th Conference on Computational Natural Language Learning,* Manchester, UK, 16–17 August 2008, pp. 159–177.

Surdeanu, Mihai, Julie Tibshirani, Ramesh Nallapati & Christopher D Manning (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pp. 455–465.

Swampillai, Kumutha & Mark Stevenson (2010). Inter-sentential relations in information extraction corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation,* La Valetta, Malta, 17–23 May 2010, pp. 2637–2641.

Swampillai, Kumutha & Mark Stevenson (2011). Extracting relations within and across sentences. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing,* Hissar, Bulgaria, 12–14 September 2011, pp. 25–32.

Taskar, Ben, Pieter Abbeel & Daphne Koller (2002). Discriminative probabilistic models for relational data. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence,* Edmonton, Alberta, Canada, 1-4 August 2002, pp. 485–492.

Taskar, Ben, Eran Segal & Daphne Koller (2001). Probabilistic classification and clustering in relational data. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence,* Seattle, Wash., 4–10 August, 2001, pp. 870–876.

Tonelli, Sara & Rodolfo Delmonte (2010). VENSES++: Adapting a deep semantic processing system to the identification of Null instantiations. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2),* Uppsala, Sweden, 15–16 July 2010, pp. 296–299.

Tonelli, Sara & Rodolfo Delmonte (2011). Desperately seeking implicit arguments in text. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics,* Portland, Oregon, USA, 23 June 2011, pp. 54–62.

Van der Sandt, Rob A (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, 9(4):333–377.

Vapnik, Vladimir (1995). *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer-Verlag.

Vapnik, Vladimir N (1998). *Statistical Learning Theory*. John Wiley and Sons, Inc., New York.

Versley, Yannick (2011). *Resolving Coreferent Bridging In German Newspaper Text*, (Ph.D. thesis). University of Tübingen.

Vieira, Renata (1998). *Definite Description Processing in Unrestricted Text*, (Ph.D. thesis). University of Edinburgh.

Vieira, Renata & Massimo Poesio (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Vieira, Renata & Simone Teufel (1997). Towards resolution of bridging descriptions. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics,* Madrid, Spain, 7–12 July 1997, pp. 522–524.

Voorhees, Ellen M. (2001). Overview of the TREC 2001 question answering track. In *Proceedings of the Tenth Text REtrieval Conference,* Gaithersburg, Md., 13–16 November 2001.

Wang, Shuo & Xin Yao (2012). Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 42(4):1119–1130.

Webber, Bonnie L. & Aravind K. Joshi (1998). Anchoring a lexicalized Tree-Adjoining Grammar for discourse. In *COLING-ACL '98 Workshop on Discourse Relations and Discourse Markers, Montréal, Québec, Canada, 15 August, 1998*, pp. 86–92.

Webber, Bonnie L., Matthew Stone, Aravind Joshi & Alistair Knott (2003). Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–588.

Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin & Ann Houston (2011). *OntoNotes Release 4.0.* LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium.

Wilson, Deirdre & Dan Sperber (2002). Relevance theory. *Linguistics*, 14:249–287.

Witten, Ian H. & Eibe Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco, Cal.: Morgan Kaufmann.

Wolf, Florian & Edward Gibson (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

Yao, Limin, Aria Haghighi, Sebastian Riedel & Andrew McCallum (2011). Structured relation discovery using generative models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* Edinburgh, Scotland, U.K., 27–29 July 2011, pp. 1456–1466.

Yao, Limin, Sebastian Riedel & Andrew McCallum (2012). Unsupervised relation discovery with sense disambiguation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics,* Jeju Island, Korea, 8–14 July 2012, pp. 712–720.

Yates, Alexander & Oren Etzioni (2007). Unsupervised resolution of objects and relations on the Web. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, N.Y., 22–27 April 2007, pp. 121–130.

Zelenko, Dmitry, Chinatsu Aone & Anthony Richardella (2003). Kernel methods for relation extraction. *Journal of Machine Learning Research*, pp. 1083–1106.

Zhao, Shubin & Ralph Grishman (2005). Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Mich., 25–30 June 2005, pp. 419–426.

Zhou, Zhihua & Xuying Liu (2010). On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257.