Discourse as a Knowledge Resource for Sentence Disambiguation

Tetsuya Nasukawa Naohiko Uramoto

IBM Research, Tokyo Research Laboratory

1G23-14, Shimotsuruma, Yamato-shi Kanagawa-keu 242 Japan

Phone 81 462 73-4574 FAX 81 462 73-7413

{nasukawa, uramoto} Ctrl net Ibm com

Abstract

A consistent text contains rich resources of in formation such as collocation patterns that can be used to resolve ambiguities within its sentences For example attachment ambigui ties in a sentence can be resolved by selecting a candidate attachment that matches attach ments found in other sentences in the same dis course Thus, discourse can be regarded as a valuable knowledge resource for sentence anal ysis In this paper, we examine some features of discourse as a knowledge resource and pro- pose a framework for natural language process- ing that provides a simple algorithm for using information extracted from discourse together with information stored in knowledge bases The experimental results of using our frame- work to disambiguate sentences in technical documents offer good prospects for improving the at curacy of a broad coverage natural lan- guage processing system that handles various texts without constructing knowledge bases for each text in advance Some noteworthy fea tures of discourse information are also deduced from the results of our experiments

1 Introduction

In natural language processing knowledge resources are indispensable for disambiguating input sentences and a large quantity of such resources is necessary to improve the accuracy of sentence analysis However, construct- ing knowledge resources from hand-coded data requires enormous amounts of time and human resources and thus leads to a knowledge acquisition bottleneck Much research has been devoted to developing methods for re- solving ambiguities by using large quantities of knowl- edge resources that can be acquired semi-automatically from on the texts [Brown *et al* 1991 Gale et *al* 1992 Hindle and Rooth, 1993 Jensen and Binot, 1987 Na gao, 1990 Utsuro *et al* 1992, Uramoto 199I] Y*et* in the day to-day operation of a practical natural language processing system such as a machine translation sys tern it is inevitable that a new text will contain various ambiguities that cannot be resolved with the knowledge contained in the system Moreover a knowledge resource tends to be less effec tive for a text in a new domain

On the other hand a consistent text itself contains rich resources of information that can be used to re solve ambiguities within its sentences As shown by Gale et al [Gale *et al* 1992] and Nasukawa [Nasukawa, 1993], polysemous words within the same discourse tend to have the same word sense with a probability of over 95% and the accuracy of word sense disambiguation can be improved b\ applying discourse constraints in such a manner that each polysemous word in a discourse takes the same word sense Rinoshita et al [Rinoshita *et al* 1993] and Nasukawa [Nasukawa 1993] have reported that co-occurrence between words extracted from a text improves the accuracy with which attachment ambigui- ties in the text are resolved Furthermore information extrac ted from discourse by means of a simple algorithm makes it easier to select the correct antecedent of a pro noun from among candidate noun phrases as shown by Nasukawa [Nasukawa, 1994]

In this paper, we study the potential usefulness of in formation extracted from discourse in contrast to infor- mation stored in knowledgebases, and develop a method based on a simple algorithm that resolves ambiguities in sentences by using information extracted from discourse together with information stored in knowledge bases

In the next section we examine a couple of features of discourse information that allow more accurate sen- tence analysis than is possible with information stored in knowledge bases Then, in the third section, we pro- pose a framework of natural language processing that extracts information from a discourse and disambiguates each sentence in the discourse by using the discourse in formation together with information stored in knowledge bases Finally, in the fourth section we give the results of experiments on the use of our framework to resolve structural ambiguities

2 Information extracted from discourse

In this section we examine two features in discourse that are applicable to sentence disambiguation and compare the information extracted from discourse with the infor mation stored in knowledge bases, in terms of its effec tiveness for sentence disambiguation

2 1 Discourse constraint and discourse preference

The first feature we discuss in this section is the "discourse constraint" that polysemous words within a discourse have the same word sense According to Gale et al [Gale e*t* a/, 1992] and Nasukawa [Nasukawa, 1993] there is a high probability (98% according to Gale et al) that polvsemous. words within the *same* discourse actually carry the same word sense If this discourse constraint is assumed to hold a result of word sense disambiguation applied in one sentence can be shared with all the morphologic all} identical words within the discourse

Since a valid word sense disambiguation cannot alwavs be made for every word appearing in a text this procedure of sharing the results of word sense disambiguation improves the total accuracy of word-sense disambiguation For example if a polysemous word appears alone in a title, a heading, or a list item there is no way to disambiguate its word sense except bv referring to information in other sentences

The second feature is "discourse preference namelv a tendency for each word to modify or be modified by similar words within a discourse During our analysis of computer manuals, we noticed that the same collocation patterns appeared frequently For example a collocation pattern in which the noun cursor occurred as the suit ject of the verb *move* appeared more than the limes on one page of a computer manual By assuming discourse preference, we obtain another chit on how improve *the* accuracy of sentence analysis For example consider the following sentences extracted from a computer manual

(2 1) *Type your user name on the user line*
(2 2) *You probably typed it on the line correctly but*

Sentence (2 2) actually comes immediately after the sentence (2 1) in the manual In sentence (2 1) the modifiee of the prepositional phrase *on the user line* can be either *Type* or your *user name* but in sentence (2 2) *on th(line* uniquely modifies *typed, t*herefore, by applying the discourse preference that the noun *line* tends to modify the verb *type* with the preposition *on* we can resolve the structural ambiguity of sentence (2 1) Furthermore the candidates for the antecedent of the pronoun *it* in sentence (2 2) are your user *name* and *the user line* m sentence (2 1), since both of them satisfy syntactic constraints To select the actual antecedent we can apply discourse preference and thus your *user* name is preferred as an object of the verb *type* since a collocation pattern in which your *user name* is the object of the verb *typp* appears in sentence (2 1)

2 2 Comparison with information stored in knowledge bases

In this paper we assume that

- A general natural language processing system holds one or more knowledge bases that contain information for sentence disambiguation
- Data are initially stored in the knowledge bases and updated through day-to-day operations
- Multiple knowledge bases ma\ be provided, to improve the accuracy of disambiguation by suppying appropriate information for various domains

With the aim of relieving the knowledge acquisition bottleneck we consider an example based approach to sentence disambiguation in which disambiguation is based on information extracted from on line texts For example, a sentence analyzer called SENA [Uramoto 199l] analyzes a sentence and extracts word to word relationships that basically consist of both a modifier word and a modifiee word along with a marker of the relation between them (such as preposition case marker and so on) and prefers an attachment candidate that matches word-to word relationships in its knowledge base (in this case an example-base) Word toword relationships can be extrac ted from on line texts seni automatically with a little human intervention to eliminate ambiguities in the data [Nagao 1990 Uramoto 199lJ Within the framework of the example-based approach it is easy to correct a specific disambiguation error l)\ adding hand coded word toword relationships and an additional set of hand coded word-to-word relationships may easily improve (the accuracy with which sentences in a specific text can be disambignaled However a knowledge base that contains only word to word relationships has a disadvantage in its coverage [Uramoto 1994] Since the vocabulary as well as the usage of specific words vanes from text to text depending on the writer and the topic improving the cove rage of word to-word relationships in a practical domain is an important issue in the area of example based s\sterns However It may reduce the particularity of some word to weird relationships sue h as those extracted from an idiomatic expression in which the modifier modifiee relationship may not be general

On the other *hand* information extracted from a discourse contains some ideal features that can be applied to each sentence in the discourse A sentence to be disambiguated and the discourse share the same wnter and the same topic thus information extracted from a discourse is more reliable for disimbiguating a sentence inside the discourse than information in a knowledge base However the discourse cannot be the only resource for sentence disambiguation because

- Not all ambiguities can be covered bv information extracted from 1 the discourse
- Information extrac ted from the discourse mav itself contain ambiguities
- t Correct information may not always be extracted from the discourse because of ellipses and analysis failures and TOT Various other reasons

Examples of these reasons are provided in Section 4 along with the results of experime nts on computer manuals To improve the accuracy of sentence disambiguation we propose to use information extracted from discourse together with information stored in knowledge bases The next section describes the framework of our method

3 Framework for using discourse information

In this paper we focus on structural ambiguities rather than ail} of the other ambiguities in sentences such as word sense ambiguities since information e\tra.c ted from discourse can resolve structural ambiguities with out being supplemented bv other information and the effects of discourse information r<in be evaluated separatelv from those of infoniiatioii stored m knowledge bases Hesolution of other ambiguities such as selection of a word sense for a polvsemous word, requires liifonnation initially stored in knowledge bases and discourse information is applied in such a manner that mformatiou extracted from knowledge bases is circulated within the discourse In this section therefore we bneflj de stnhc a framework for using information extracted from discourse to disambiguate a sentence hj focusing on its structural ambiguities

3 1 Extraction of discourse information

The first step of our framework is to extrae t information from a discourse while the second step is to resolve the ambiguities in each sentence of the discourse

In the first step each sentence in the whole text given as a discourse is processed bv a sMitactu parwr and the position of each instance of everv lemma and its modifi.ee modifier relationships with other content words in the parser output are stored as discourse mforiiia tion In accumulation of discourse information prefer ence scores are given for each definite modifier modifier relationship, such as the modification of type t>\ on *the line* in sentence (2 2) and lower scores are gnen for eath ambiguous inodifice-modifier relationship such as the possible modification of *Type* b\ *on iht user line-* in seutence (2 1)

Table 1 shows an example of such information e\ traded from the text of a computer manual consisting of sentences such as

> *A server shares resource* with other workstations on the LAN*

In this table, scores of 10 for each definite modihee modifier relationship and 3 for each ambiguous inoditiee-modiher relationship art assigned Thus the sentence above yielded preference scores of 10for sm>ermodifying *share* as a subject, 3 for *workstation posslb.lv* modifying *share* with the preposition *with* and so on

3 2 Structural disambiguation

To resolve structural ambiguities, the discourse mforiiia tion is referred to for each modifiee c andidaU If the same modifiee modifier relationship that is the same collocation pattern, is found in the discourse information, the score of the collocation is extracted and added to the candidate s preference Value In order to improve the coverage of the discourse information, an on hue svnonvm dictionary [Collins 1984] is consulted so that a collocation pattern can be checked with a svnonjm in one side of the collocation (either the modifiee or the modifier) The? score after being multiplied bv a weighting value is then added to its preference value of the candidate For example to select the verb see as the modifiet of the propositional phrase, *for information,* in the sentence

> *Set Where to Look for Additional Task, Infor matton in "Up and Running" for innformation on which books to use*

a collocation pattern in which *for instruction* modifies see extracted from a later sentence in the same discourse (11 sentences after the sentence above

> *For instructions see "Up and Running*

is referred to since *instruction* is defined LS a svnonvm of *information* in the on line synonym dictionary [Collins, 1984]

In addition to the dictionary information coordinate structures in discourse information are referred to, and the coordinated words are also treated as synonyms in examining collocation pat terns For example in the sentence

> *In the MVS environment tht product runs* un *der TSO/E CICS and IMS*

TSO/E CICS and *IMS* art coordmated and thus treated as synoyms in the same discourse

Information on word to-word relationships stored in knowledge bases also provides preference stores, in addi tion to those given by the discourse information and a store is added to the preference value of cat h candidate modifier After evaluation of the preference value for each candidate modifice, the candidate with the highest preference is selected

4 Results and discussion

After implementing our framework on an English-to-Japanese machine translation system calletl Shalt2 [Takeda e*t al*, 1992] we processed a large number of technical documents such as patents technical letters and in particular computer manuals The results were encouraging and we found some interesting phenomena that should be taken into account when information ex traded from discourse is used along with information stored in knowledge bases

4 1 Coverage

The coverage of information used for structural disam bignation is shown in Figure 1 The figure reflects the results of disambiguating all the sententces in a computer manual consisting of 791 consecutiv e sentences by chang ing the size of the discourse from 10 to 791 sentences Except for the result of referring to the whole 791 sentences as a discourse all the results indicate the averages of the results obtained by referring to eat h of several sample areas as a discourse For example, to obtain data for the case in which the size of a discourse is 20 sentences, we examined 32 areas each consisting of 20 sentences, such as the 1st sentence to the 20th, the 51st to the 70th, and the 701st to the 720th

In this experiment, we disambiguated the modifiees of structurally ambiguous phrases such as prepositional phrases, present and past participle clauses, relative

Table 1 Example of discourse information on the verb "share'

	POS	Relation	Word (preference value)
Modifiers	Noun	from	list (3)
		with	user (30) workstation (3)
		OBJ	printer (50) directory (50) resource (20)
		on	network (3) LAN (3)
		SUBJ	server (10)
	Verb	when	complete (10)
		if	exist (10)
	Adverb	DIRECT	for example (10) also (10)
Modifiees	POS	Relation	Word (preference value)
	Noun	PASTPART	network printer (3) list (3)
		for	procedure (3)
		preinf	subdirectory (3) procedure (3)
	Verb	preinf	want (10) install (10) create (3) follow (3)
		PASTPART	identify (3) type (3)
		for	go (3)

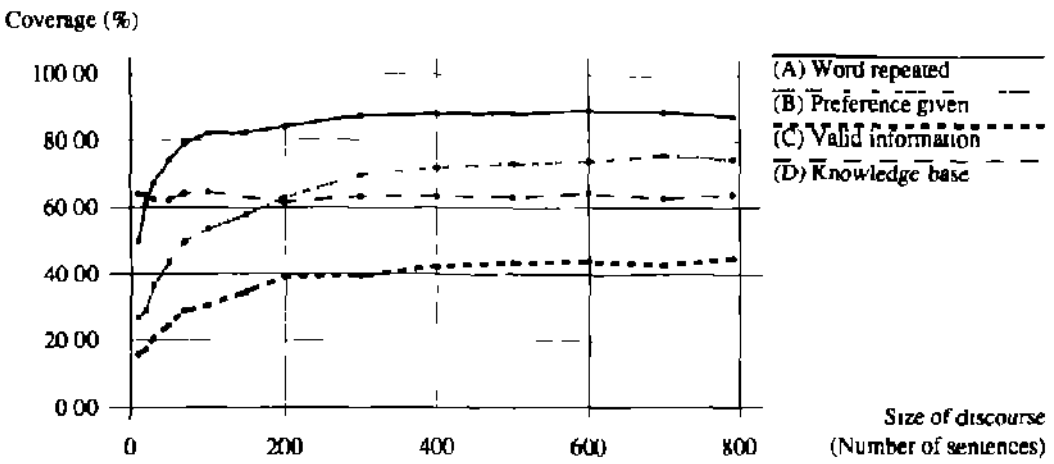

Figure 1 Relation between the coverage of discourse information and the size of the discourse

clauses and to-infinitive rlauses In Figure 1 *(\.)* indicates the percentage of cases in which either the modifu 1 Word or out of the modihee candidate words was found in the discourse information Iu other words at hast one of the words related to a resolution of a structural ambiguit) was repeated iu the discourse in a pert rut age *of* case*, indicated *by* (\), this shows that main of the words were repeated within a relative!; small area of the text To be precise, more than 80'X of the words were used more than once within 100 sentences (B) shows the percentage of cases in which some preference values were gi\en to an\ of the candidate modifiers on the basis of information extracted from the discourse However, the assignment of preference values does not imply that a valid disambiguation has heen earned out since identical Values mav be assigned to different can didates (C) indicates the percentage of cases in which discourse information allowed a preferable candidate to bt dis animated The percentage of cases in which the information store d in the knowledge base provided preference stores is shown *hy* (D) Our knowledge base con tains more than 57 000 ruins cm word to word relation ships extracted from almost 10 000 sentences in various computer manuals and technical documents aud some broader coverage rules such as tin one that prefers a prepositional phrase with *to* to modify a verb *go* \s shown in the figure about 40(\X of structural ambiguities can be resolved with information extracted from the dis course provided the discourse contains more than 200 consecutive sentences

To determine whether <tn input text should consist of (onnecutive sentences wt combined 100 consecutive sentences from throe different parts of the manual exam mod in the previous experiments on coverage - say the