Distributional Identification of Non-Referential Pronouns

Shane Bergsma
Department of Computing Science
University of Alberta
Edmonton, Alberta
Canada, T6G 2E8
bergsma@cs.ualberta.ca

Dekang Lin
Google, Inc.
1600 Amphitheatre Parkway
Mountain View
California, 94301
lindek@google.com

Randy Goebel
Department of Computing Science
University of Alberta
Edmonton, Alberta
Canada, T6G 2E8
goebel@cs.ualberta.ca

Abstract

We present an automatic approach to determining whether a pronoun in text refers to a preceding noun phrase or is instead *non-referential*. We extract the surrounding textual context of the pronoun and gather, from a large corpus, the distribution of words that occur within that context. We learn to reliably classify these distributions as representing either referential or non-referential pronoun instances. Despite its simplicity, experimental results on classifying the English pronoun *it* show the system achieves the highest performance yet attained on this important task.

1 Introduction

The goal of coreference resolution is to determine which noun phrases in a document refer to the same real-world entity. As part of this task, coreference resolution systems must decide which pronouns refer to preceding noun phrases (called antecedents) and which do not. In particular, a long-standing challenge has been to correctly classify instances of the English pronoun *it*. Consider the sentences:

(1) You can make it in advance.

(2) You can make it in Hollywood.

In sentence (1), *it* is an anaphoric pronoun referring to some previous noun phrase, like "the sauce" or "an appointment." In sentence (2), *it* is part of the idiomatic expression "make it" meaning "succeed." A coreference resolution system should find an antecedent for the first *it* but not the second. Pronouns that do not refer to preceding noun phrases are called *non-anaphoric* or *non-referential* pronouns.

The word *it* is one of the most frequent words in the English language, accounting for about 1% of tokens in text and over a quarter of all third-person pronouns.[1] Usually between a quarter and a half of *it* instances are non-referential (e.g. Section 4, Table 3). As with other pronouns, the preceding discourse can affect *it*'s interpretation. For example, sentence (2) can be interpreted as referential if the preceding sentence is "You want to make a movie?" We show, however, that we can reliably classify a pronoun as being referential or non-referential based solely on the local context surrounding the pronoun.

We do this by turning the context into patterns and enumerating all the words that can take the place of *it* in these patterns. For sentence (1), we can extract the context pattern "make * in advance" and for sentence (2) "make * in Hollywood," where "*" is a wildcard that can be filled by any token. Non-referential distributions tend to have the word *it* filling the wildcard position. Referential distributions occur with many other noun phrase fillers. For example, in our n-gram collection (Section 3.4), "make it in advance" and "make them in advance" occur roughly the same number of times (442 vs. 449), indicating a referential pattern. In contrast, "make it in Hollywood" occurs 3421 times while "make them in Hollywood" does not occur at all.

These simple counts strongly indicate whether another noun can replace the pronoun. Thus we can computationally distinguish between a) pronouns that refer to nouns, and b) all other instances: including those that have no antecedent, like sentence (2),

[1] e.g. http://ucrel.lancs.ac.uk/bncfreq/flists.html

Proceedings of ACL-08: HLT, pages 10–18,
Columbus, Ohio, USA, June 2008.

and those that refer to sentences, clauses, or implied topics of discourse. Beyond the practical value of this distinction, Section 3 provides some theoretical justification for our binary classification.

Section 3 also shows how to automatically extract and collect counts for context patterns, and how to combine the information using a machine learned classifier. Section 4 describes our data for learning and evaluation, *It-Bank*: a set of over three thousand labelled instances of the pronoun *it* from a variety of text sources. Section 4 also explains our comparison approaches and experimental methodology. Section 5 presents our results, including an interesting comparison of our system to human classification given equivalent segments of context.

2 Related Work

The difficulty of non-referential pronouns has been acknowledged since the beginning of computational resolution of anaphora. Hobbs (1978) notes his algorithm does not handle pronominal references to sentences nor cases where *it* occurs in time or weather expressions. Hirst (1981, page 17) emphasizes the importance of detecting non-referential pronouns, "lest precious hours be lost in bootless searches for textual referents." Müller (2006) summarizes the evolution of computational approaches to non-referential *it* detection. In particular, note the pioneering work of Paice and Husk (1987), the inclusion of non-referential *it* detection in a full anaphora resolution system by Lappin and Leass (1994), and the machine learning approach of Evans (2001).

There has recently been renewed interest in non-referential pronouns, driven by three primary sources. First of all, research in coreference resolution has shown the benefits of modules for general noun anaphoricity determination (Ng and Cardie, 2002; Denis and Baldridge, 2007). Unfortunately, these studies handle pronouns inadequately; judging from the decision trees and performance figures, Ng and Cardie (2002)'s system treats all pronouns as anaphoric by default. Secondly, while most pronoun resolution evaluations simply exclude non-referential pronouns, recent unsupervised approaches (Cherry and Bergsma, 2005; Haghighi and Klein, 2007) must deal with all pronouns in unrestricted text, and therefore need robust modules to automatically handle non-referential instances. Finally, reference resolution has moved beyond written text into in spoken dialog. Here, non-referential pronouns are pervasive. Eckert and Strube (2000) report that in the Switchboard corpus, only 45% of demonstratives and third-person pronouns have a noun phrase antecedent. Handling the common non-referential instances is thus especially vital.

One issue with systems for non-referential detection is the amount of language-specific knowledge that must be encoded. Consider a system that jointly performs anaphora resolution and word alignment in parallel corpora for machine translation. For this task, we need to identify non-referential anaphora in multiple languages. It is not always clear to what extent the features and modules developed for English systems apply to other languages. For example, the detector of Lappin and Leass (1994) labels a pronoun as non-referential if it matches one of several syntactic patterns, including: "It is **Cogv-ed** that **Sentence**," where **Cogv** is a "cognitive verb" such as *recommend*, *think*, *believe*, *know*, *anticipate*, etc. Porting this approach to a new language would require not only access to a syntactic parser and a list of cognitive verbs in that language, but the development of new patterns to catch non-referential pronoun uses that do not exist in English.

Moreover, writing a set of rules to capture this phenomenon is likely to miss many less-common uses. Alternatively, recent machine-learning approaches leverage a more general representation of a pronoun instance. For example, Müller (2006) has a feature for "distance to next complementizer (*that*, *if*, *whether*)" and features for the tokens and part-of-speech tags of the context words. Unfortunately, there is still a lot of implicit and explicit English-specific knowledge needed to develop these features, including, for example, lists of "seem" verbs such as *appear*, *look*, *mean*, *happen*. Similarly, the machine-learned system of Boyd et al. (2005) uses a set of "idiom patterns" like "*on the face of it*" that trigger binary features if detected in the pronoun context. Although machine learned systems can flexibly balance the various indicators and contra-indicators of non-referentiality, a particular feature is only useful if it is relevant to an example in limited labelled training data.

Our approach avoids hand-crafting a set of spe-

cific indicator features; we simply use the distribution of the pronoun's context. Our method is thus related to previous work based on Harris (1985)'s distributional hypothesis.[2] It has been used to determine both word and syntactic path similarity (Hindle, 1990; Lin, 1998a; Lin and Pantel, 2001). Our work is part of a trend of extracting other important information from statistical distributions. Dagan and Itai (1990) use the distribution of a pronoun's context to determine which candidate antecedents can fit the context. Bergsma and Lin (2006) determine the likelihood of coreference along the syntactic path connecting a pronoun to a possible antecedent, by looking at the distribution of the path in text. These approaches, like ours, are ways to inject sophisticated "world knowledge" into anaphora resolution.

3 Methodology

3.1 Definition

Our approach distinguishes contexts where pronouns cannot be replaced by a preceding noun phrase (non-noun-referential) from those where nouns can occur (noun-referential). Although coreference evaluations, such as the MUC (1997) tasks, also make this distinction, it is not necessarily used by all researchers. Evans (2001), for example, distinguishes between "clause anaphoric" and "pleonastic" as in the following two instances:

(3) The paper reported that it had snowed. *It* was obvious. (*clause anaphoric*)

(4) *It* was obvious that it had snowed. (*pleonastic*)

The word *It* in sentence (3) is considered referential, while the word *It* in sentence (4) is considered non-referential.[3] From our perspective, this interpretation is somewhat arbitrary. One could also say that the *It* in both cases refers to the clause "that it had snowed." Indeed, annotation experiments using very fine-grained categories show low annotation reliability (Müller, 2006). On the other hand, there is no debate over the importance nor the definition of distinguishing pronouns that refer to nouns from those that do not. We adopt this distinction for our work, and show it has good inter-annotator reliability (Section 4.1). We henceforth refer to non-noun-referential simply as non-referential, and thus consider the word *It* in both sentences (3) and (4) as non-referential.

Non-referential pronouns are widespread in natural language. The es in the German "Wie geht es Ihnen" and the il in the French "S'il vous plaît" are both non-referential. In pro-drop languages that may omit subject pronouns, there remains the question of whether an omitted pronoun is referential (Zhao and Ng, 2007). Although we focus on the English pronoun *it*, our approach should differentiate any words that have both a structural and a referential role in language, e.g. words like *this*, *there* and *that* (Müller, 2007). We believe a distributional approach could also help in related tasks like identifying the generic use of *you* (Gupta et al., 2007).

3.2 Context Distribution

Our method extracts the context surrounding a pronoun and determines which other words can take the place of the pronoun in the context. The extracted segments of context are called *context patterns*. The words that take the place of the pronoun are called *pattern fillers*. We gather pattern fillers from a large collection of n-gram frequencies. The maximum size of a context pattern depends on the size of n-grams available in the data. In our n-gram collection (Section 3.4), the lengths of the n-grams range from unigrams to 5-grams, so our maximum pattern size is five. For a particular pronoun in text, there are five possible 5-grams that span the pronoun. For example, in the following instance of *it*:

... said here Thursday that it is unnecessary to continue ...

We can extract the following 5-gram patterns:

said here Thursday that *
here Thursday that * is
Thursday that * is unnecessary
that * is unnecessary to
* is unnecessary to continue

Similarly, we extract the four 4-gram patterns. Shorter n-grams were not found to improve performance on development data and hence are not extracted. We only use context within the current sentence (including the beginning-of-sentence and end-of-sentence tokens) so if a pronoun occurs near a sentence boundary, some patterns may be missing.

[2] Words occurring in similar contexts have similar meanings

[3] The *it* in "it had snowed" is, of course, non-referential.

Pattern Filler Type	String
#1: 3rd-person pron. sing.	*it/its*
#2: 3rd-person pron. plur.	*they/them/their*
#3: any other pronoun	*he/him/his/*, *I/me/my*, etc.
#4: infrequent word token	⟨*UNK*⟩
#5: any other token	*

Table 1: Pattern filler types

We take a few steps to improve generality. We change the patterns to lower-case, convert sequences of digits to the # symbol, and run the Porter stemmer[4] (Porter, 1980). To generalize rare names, we convert capitalized words longer than five characters to a special *NE* tag. We also added a few simple rules to stem the irregular verbs *be*, *have*, *do*, and *said*, and convert the common contractions *'nt*, *'s*, *'m*, *'re*, *'ve*, *'d*, and *'ll* to their most likely stem.

We do the same processing to our n-gram corpus. We then find all n-grams matching our patterns, allowing any token to match the wildcard in place of *it*. Also, other pronouns in the pattern are allowed to match a corresponding pronoun in an n-gram, regardless of differences in inflection and class.

We now discuss how to use the distribution of pattern fillers. For identifying non-referential *it* in English, we are interested in how often *it* occurs as a pattern filler versus other *nouns*. However, determining part-of-speech in a large n-gram corpus is not simple, nor would it easily extend to other languages. Instead, we gather counts for five different classes of words that fill the wildcard position, easily determined by string match (Table 1). The third-person plural *they* (#2) reliably occurs in patterns where referential *it* also resides. The occurrence of *any other pronoun* (#3) guarantees that at the very least the pattern filler is a noun. A match with the infrequent word token ⟨UNK⟩ (#4) (explained in Section 3.4) will likely be a noun because nouns account for a large proportion of rare words in a corpus. Gathering *any other token* (#5) also mostly finds nouns; inserting another part-of-speech usually results in an unlikely, ungrammatical pattern.

[4]Adapted from the Bow-toolkit (McCallum, 1996). Our method also works without the stemmer; we simply truncate the words in the pattern at a given maximum length (see Section 5.1). With simple truncation, all the pattern processing can be easily applied to other languages.

Pattern	Filler Counts			
	#1	#2	#3	#5
sai here *NE* that *	84	0	291	3985
here *NE* that * be	0	0	0	93
NE that * be unnecessari	0	0	0	0
that * be unnecessari to	16726	56	0	228
* be unnecessari to continu	258	0	0	0

Table 2: 5-gram context patterns and pattern-filler counts for the Section 3.2 example.

Table 2 gives the stemmed context patterns for our running example. It also gives the n-gram counts of pattern fillers matching the first four filler types (there were no matches of the ⟨UNK⟩ type, #4).

3.3 Feature Vector Representation

There are many possible ways to use the above counts. Intuitively, our method should identify as non-referential those instances that have a high proportion of fillers of type #1 (i.e., the word *it*), while labelling as referential those with high counts for other types of fillers. We would also like to leverage the possibility that some of the patterns may be more predictive than others, depending on where the wildcard lies in the pattern. For example, in Table 2, the cases where the *it*-position is near the beginning of the pattern best reflect the non-referential nature of this instance. We can achieve these aims by ordering the counts in a feature vector, and using a labelled set of training examples to learn a classifier that optimally weights the counts.

For classification, we define non-referential as positive and referential as negative. Our feature representation very much resembles Table 2. For each of the five 5-gram patterns, ordered by the position of the wildcard, we have features for the logarithm of counts for filler types #1, #2, ... #5. Similarly, for each of the four 4-gram patterns, we provide the log-counts corresponding to types #1, #2, ... #5 as well. Before taking the logarithm, we smooth the counts by adding a fixed number to all observed values. We also provide, for each pattern, a feature that indicates if the pattern is not available because the *it*-position would cause the pattern to span beyond the current sentence. There are twenty-five 5-gram, twenty 4-gram, and nine indicator features in total.

Our classifier should learn positive weights on the type #1 counts and negative weights on the other types, with higher absolute weights on the more predictive filler types and pattern positions. Note that leaving the pattern counts unnormalized automatically allows patterns with higher counts to contribute more to the prediction of their associated instances.

3.4 N-Gram Data

We now describe the collection of n-grams and their counts used in our implementation. We use, to our knowledge, the largest publicly available collection: the Google Web 1T 5-gram Corpus Version 1.1.[5] This collection was generated from approximately 1 trillion tokens of online text. In this data, tokens appearing less than 200 times have been mapped to the ⟨UNK⟩ symbol. Also, only n-grams appearing more than 40 times are included. For languages where such an extensive n-gram resource is not available, the n-gram counts could also be taken from the page-counts returned by an Internet search engine.

4 Evaluation

4.1 Labelled *It* Data

We need labelled data for training and evaluation of our system. This data indicates, for every occurrence of the pronoun *it*, whether it refers to a preceding noun phrase or not. Standard coreference resolution data sets annotate all noun phrases that have an antecedent noun phrase in the text. Therefore, we can extract labelled instances of *it* from these sets. We do this for the dry-run and formal sets from MUC-7 (1997), and merge them into a single data set.

Of course, full coreference-annotated data is a precious resource, with the pronoun *it* making up only a small portion of the marked-up noun phrases. We thus created annotated data specifically for the pronoun *it*. We annotated 1020 instances in a collection of Science News articles (from 1995-2000), downloaded from the Science News website. We also annotated 709 instances in the WSJ portion of the DARPA TIPSTER Project (Harman, 1992), and 279 instances in the English portion of the Europarl Corpus (Koehn, 2005).

A single annotator (A_1) labelled all three data sets, while two additional annotators not connected with the project (A_2 and A_3) were asked to separately re-annotate a portion of each, so that inter-annotator agreement could be calculated. A_1 and A_2 agreed on 96% of annotation decisions, while A_1-A_3, and A_2-A_3, agreed on 91% and 93% of decisions, respectively. The *Kappa* statistic (Jurafsky and Martin, 2000, page 315), with P(E) computed from the confusion matrices, was a high 0.90 for A_1-A_2, and 0.79 and 0.81 for the other pairs, around the 0.80 considered to be good reliability. These are, perhaps surprisingly, the only known *it*-annotation-agreement statistics available for written text. They contrast favourably with the low agreement seen on categorizing *it* in spoken dialog (Müller, 2006).

Data Set	Number of *It*	% Non-Referential
Europarl	279	50.9
Sci-News	1020	32.6
WSJ	709	25.1
MUC	129	31.8
Train	1069	33.2
Test	1067	31.7
Test-200	200	30.0

Table 3: Data sets used in experiments.

We make all the annotations available in *It-Bank*, an online repository for annotated *it*-instances.[6] *It-Bank* also allows other researchers to distribute their *it* annotations. Often, the full text of articles containing annotations cannot be shared because of copyright. However, sharing just the sentences containing the word *it*, randomly-ordered, is permissible under fair-use guidelines. The original annotators retain their copyright on the annotations.

We use our annotated data in two ways. First of all, we perform cross-validation experiments on each of the data sets individually, to help gauge the difficulty of resolution on particular domains and volumes of training data. Secondly, we randomly distribute all instances into two main sets, a training set and a test set. We also construct a smaller test set, *Test-200*, containing only the first 200 instances in the *Test* set. We use *Test-200* for human experiments and error analysis (Section 5.2). Table 3 summarizes all the sets used in the experiments.

[5] Available from the LDC as LDC2006T13

[6] www.cs.ualberta.ca/~bergsma/ItBank/. *It-Bank* also contains an additional 1,077 examples used as development data.

4.2 Comparison Approaches

We represent feature vectors exactly as described in Section 3.3. We smooth by adding 40 to all counts, equal to the minimum count in the n-gram data. For classification, we use a maximum entropy model (Berger et al., 1996), from the logistic regression package in Weka (Witten and Frank, 2005), with all default parameter settings. Results with our distributional approach are labelled as DISTRIB. Note that our maximum entropy classifier actually produces a *probability* of non-referentiality, which is thresholded at 50% to make a classification.

As a baseline, we implemented the non-referential *it* detector of Lappin and Leass (1994), labelled as LL in the results. This is a *syntactic* detector, a point missed by Evans (2001) in his criticism: the patterns are robust to intervening words and modifiers (e.g. "it was *never* thought *by the committee* that...") provided the sentence is parsed correctly.[7] We automatically parse sentences with Minipar, a broad-coverage dependency parser (Lin, 1998b).

We also use a separate, extended version of the LL detector, implemented for large-scale non-referential detection by Cherry and Bergsma (2005). This system, also for Minipar, additionally detects instances of *it* labelled with Minipar's pleonastic category *Subj*. It uses Minipar's named-entity recognition to identify time expressions, such as "it was midnight," and provides a number of other patterns to match common non-referential *it* uses, such as in expressions like "darn it," "don't overdo it," etc. This extended detector is labelled as MINIPL (for Minipar pleonasticity) in our results.

Finally, we tested a system that combines the above three approaches. We simply add the LL and MINIPL decisions as binary features in the DISTRIB system. This system is called COMBO in our results.

4.3 Evaluation Criteria

We follow Müller (2006)'s evaluation criteria. Precision (P) is the proportion of instances that we label as non-referential that are indeed non-referential. Recall (R) is the proportion of true non-referentials that we detect, and is thus a measure of the coverage of the system. F-Score (F) is the geometric average of precision and recall; it is the most common non-referential detection metric. Accuracy (Acc) is the percentage of instances labelled correctly.

[7] Our approach, on the other hand, would seem to be susceptible to such intervening material, if it pushes indicative context tokens out of the 5-token window.

System	P	R	F	Acc
LL	**93.4**	21.0	34.3	74.5
MINIPL	66.4	49.7	56.9	76.1
DISTRIB	81.4	71.0	75.8	85.7
COMBO	81.3	**73.4**	**77.1**	**86.2**

Table 4: *Train*/*Test*-split performance (%).

5 Results

5.1 System Comparison

Table 4 gives precision, recall, F-score, and accuracy on the *Train*/*Test* split. Note that while the LL system has high detection precision, it has very low recall, sharply reducing F-score. The MINIPL approach sacrifices some precision for much higher recall, but again has fairly low F-score. To our knowledge, our COMBO system, with an F-Score of 77.1%, achieves the highest performance of any non-referential system yet implemented. Even more importantly, DISTRIB, which requires only minimal linguistic processing and no encoding of specific indicator patterns, achieves 75.8% F-Score. The difference between COMBO and DISTRIB is not statistically significant, while both are significantly better than the rule-based approaches.[8] This provides strong motivation for a "light-weight" approach to non-referential *it* detection – one that does not require parsing or hand-crafted rules and – is easily ported to new languages and text domains.

Since applying an English stemmer to the context words (Section 3.2) reduces the portability of the distributional technique, we investigated the use of more portable pattern abstraction. Figure 1 compares the use of the stemmer to simply truncating the words in the patterns at a certain maximum length. Using no truncation (Unaltered) drops the F-Score by 4.3%, while truncating the patterns to a length of four only drops the F-Score by 1.4%, a difference which is not statistically significant. Simple truncation may be a good option for other languages where stemmers are not readily available. The optimum

[8] All significance testing uses McNemar's test, $p<0.05$

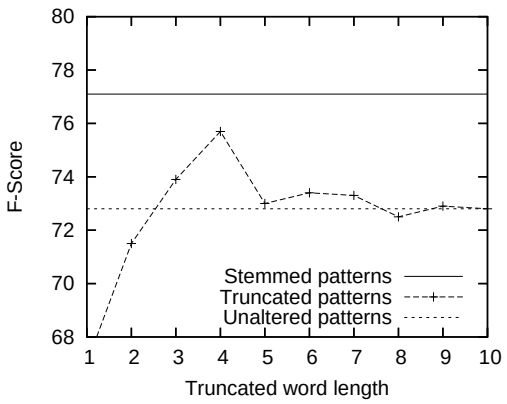

Figure 1: Effect of pattern-word truncation on non-referential *it* detection (COMBO system, *Train*/*Test* split).

System	*Europl.*	*Sci-News*	*WSJ*	*MUC*
LL	44.0	39.3	21.5	13.3
MINIPL	70.3	61.8	22.0	50.7
DISTRIB	79.7	77.2	**69.5**	68.2
COMBO	76.2	**78.7**	68.1	65.9
COMBO4	**83.6**	76.5	67.1	**74.7**

Table 5: 10-fold cross validation F-Score (%).

truncation size will likely depend on the length of the base forms of words in that language. For real-world application of our approach, truncation also reduces the table sizes (and thus storage and look-up costs) of any pre-compiled *it*-pattern database.

Table 5 compares the 10-fold cross-validation F-score of our systems on the four data sets. The performance of COMBO on *Europarl* and *MUC* is affected by the small number of instances in these sets (Section 4, Table 3). We can reduce data fragmentation by removing features. For example, if we only use the length-4 patterns in COMBO (labelled as COMBO4), performance increases dramatically on *Europarl* and *MUC*, while dipping slightly for the larger *Sci-News* and *WSJ* sets. Furthermore, selecting just the three most useful filler type counts as features (#1,#2,#5), boosts F-Score on *Europarl* to 86.5%, 10% above the full COMBO system.

5.2 Analysis and Discussion

In light of these strong results, it is worth considering where further gains in performance might yet be found. One key question is to what extent a limited context restricts identification performance. We first tested the importance of the pattern length by using only the length-4 counts in the DISTRIB system (*Train*/*Test* split). Surprisingly, the drop in F-Score was only one percent, to 74.8%. Using only the length-5 counts drops F-Score to 71.4%. Neither are statistically significant; however there seems to be diminishing returns from longer context patterns.

System	P	R	F	Acc
DISTRIB	80.0	73.3	76.5	86.5
COMBO	80.7	76.7	78.6	87.5
Human-1	92.7	63.3	75.2	87.5
Human-2	84.0	70.0	76.4	87.0
Human-3	72.2	86.7	78.8	86.0

Table 6: Evaluation on *Test-200* (%).

Another way to view the limited context is to ask, given the amount of context we have, are we making optimum use of it? We answer this by seeing how well humans can do with the same information. As explained in Section 3.2, our system uses 5-gram context patterns that together span from four-to-the-left to four-to-the-right of the pronoun. We thus provide these same nine-token windows to our human subjects, and ask them to decide whether the pronouns refer to previous noun phrases or not, based on these contexts. Subjects first performed a dry-run experiment on separate development data. They were shown their errors and sources of confusion were clarified. They then made the judgments unassisted on the final *Test-200* data. Three humans performed the experiment. Their results show a range of preferences for precision versus recall, with both F-Score and Accuracy on average below the performance of COMBO (Table 6). Foremost, these results show that our distributional approach is already getting good leverage from the limited context information, around that achieved by our best human.

It is instructive to inspect the twenty-five *Test-200* instances that the COMBO system classified incorrectly, given human performance on this same set. Seventeen of the twenty-five COMBO errors were also made by one or more human subjects, suggesting system errors are also mostly due to limited context. For example, one of these errors was for the context: "it takes an astounding amount..." Here, the non-referential nature of the instance is not apparent without the infinitive clause that ends the sentence: "... of time to compare very long DNA sequences

with each other."

Six of the eight errors unique to the COMBO system were cases where the system falsely said the pronoun was non-referential. Four of these could have referred to entire sentences or clauses rather than nouns. These confusing cases, for both humans and our system, result from our definition of a referential pronoun: pronouns with verbal or clause antecedents are considered non-referential (Section 3.1). If an antecedent verb or clause is replaced by a nominalization (*Smith researched...* to *Smith's research*), a referring pronoun, in the same context, becomes referential. When we inspect the probabilities produced by the maximum entropy classifier (Section 4.2), we see only a weak bias for the non-referential class on these examples, reflecting our classifier's uncertainty. It would likely be possible to improve accuracy on these cases by encoding the presence or absence of preceding nominalizations as a feature of our classifier.

Another false non-referential decision is for the phrase "... machine he had installed it on." The *it* is actually referential, but the extracted patterns (e.g. "he had install * on") are nevertheless usually filled with *it*.[9] Again, it might be possible to fix such examples by leveraging the preceding discourse. Notably, the first noun-phrase before the context is the word "software." There is strong compatibility between the pronoun-parent "install" and the candidate antecedent "software." In a full coreference resolution system, when the anaphora resolution module has a strong preference to link *it* to an antecedent (which it should when the pronoun is indeed referential), we can override a weak non-referential probability. Non-referential *it* detection should not be a pre-processing step, but rather part of a globally-optimal configuration, as was done for general noun phrase anaphoricity by Denis and Baldridge (2007).

The suitability of this kind of approach to correcting some of our system's errors is especially obvious when we inspect the probabilities of the maximum entropy model's output decisions on the *Test-200* set. Where the maximum entropy classifier makes mistakes, it does so with less confidence than when it classifies correct examples. The average predicted probability of the incorrect classifications is 76.0% while the average probability of the correct classifications is 90.3%. Many incorrect decisions are ready to switch sides; our next step will be to use features of the preceding discourse and the candidate antecedents to help give them a push.

[9] This example also suggests using filler counts for the word "the" as a feature when *it* is the last word in the pattern.

6 Conclusion

We have presented an approach to detecting non-referential pronouns in text based on the distribution of the pronoun's context. The approach is simple to implement, attains state-of-the-art results, and should be easily ported to other languages. Our technique demonstrates how large volumes of data can be used to gather world knowledge for natural language processing. A consequence of this research was the creation of *It-Bank*, a collection of thousands of labelled examples of the pronoun *it*, which will benefit other coreference resolution researchers.

Error analysis reveals that our system is getting good leverage out of the pronoun context, achieving results comparable to human performance given equivalent information. To boost performance further, we will need to incorporate information from preceding discourse. Future research will also test the distributional classification of other ambiguous pronouns, like *this*, *you*, *there*, and *that*. Another avenue of study will look at the interaction between coreference resolution and machine translation. For example, if a single form in English (e.g. *that*) is separated into different meanings in another language (e.g., Spanish demonstrative *ese*, nominal reference *ése*, abstract or statement reference *eso*, and complementizer *que*), then aligned examples provide automatically-disambiguated English data. We could extract context patterns and collect statistics from these examples like in our current approach. In general, jointly optimizing translation and coreference is an exciting and largely unexplored research area, now partly enabled by our portable non-referential detection methodology.

Acknowledgments

We thank Kristin Musselman and Christopher Pinchak for assistance preparing the data, and we thank Google Inc. for sharing their 5-gram corpus. We gratefully acknowledge support from the Natural Sciences and Engineering Research Council of Canada, the Alberta Ingenuity Fund, and the Alberta Informatics Circle of Research Excellence.

References

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *COLING-ACL*, pages 33–40.

Adrianne Boyd, Whitney Gegg-Harrison, and Donna Byron. 2005. Identifying non-referential *it*: a machine learning approach incorporating linguistically motivated patterns. In *ACL Workshop on Feature Engineering for Machine Learning in NLP*, pages 40–47.

Colin Cherry and Shane Bergsma. 2005. An expectation maximization approach to pronoun resolution. In *CoNLL*, pages 88–95.

Ido Dagan and Alan Itai. 1990. Automatic processing of large corpora for the resolution of anaphora references. In *COLING*, volume 3, pages 330–332.

Pascal Denis and Jason Baldridge. 2007. Joint determination of anaphoricity and coreference using integer programming. In *NAACL-HLT*, pages 236–243.

Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Richard Evans. 2001. Applying machine learning toward an automatic classification of *it*. *Literary and Linguistic Computing*, 16(1):45–57.

Surabhi Gupta, Matthew Purver, and Dan Jurafsky. 2007. Disambiguating between generic and referential "you" in dialog. In *ACL Demo and Poster Sessions*, pages 105–108.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *ACL*, pages 848–855.

Donna Harman. 1992. The DARPA TIPSTER project. *ACM SIGIR Forum*, 26(2):26–28.

Zellig Harris. 1985. Distributional structure. In J.J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. Oxford University Press, New York.

Donald Hindle. 1990. Noun classification from predicate-argument structures. In *ACL*, pages 268–275.

Graeme Hirst. 1981. *Anaphora in Natural Language Understanding: A Survey*. Springer Verlag.

Jerry Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(311):339–352.

Daniel Jurafsky and James H. Martin. 2000. *Speech and language processing*. Prentice Hall.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*, pages 79–86.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Dekang Lin. 1998a. Automatic retrieval and clustering of similar words. In *COLING-ACL*, pages 768–773.

Dekang Lin. 1998b. Dependency-based evaluation of MINIPAR. In *LREC Workshop on the Evaluation of Parsing Systems*.

Andrew Kachites McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow.

MUC-7. 1997. Coreference task definition (v3.0, 13 Jul 97). In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Christoph Müller. 2006. Automatic detection of non-referential *It* in spoken multi-party dialog. In *EACL*, pages 49–56.

Christoph Müller. 2007. Resolving *It*, *This*, and *That* in unrestricted multi-party dialog. In *ACL*, pages 816–823.

Vincent Ng and Claire Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *COLING*, pages 730–736.

Chris D. Paice and Gareth D. Husk. 1987. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun "it". *Computer Speech and Language*, 2:109–132.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, second edition.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of Chinese zero pronouns: A machine learning approach. In *EMNLP*, pages 541–550.