Resolving and Generating Definite Anaphora by Modeling Hypernymy using Unlabeled Corpora

Nikesh Garera and **David Yarowsky**
Department of Computer Science
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218, USA
{ngarera,yarowsky}@cs.jhu.edu

Abstract

We demonstrate an original and successful approach for both resolving and generating definite anaphora. We propose and evaluate unsupervised models for extracting hypernym relations by mining cooccurrence data of definite NPs and potential antecedents in an unlabeled corpus. The algorithm outperforms a standard WordNet-based approach to resolving and generating definite anaphora. It also substantially outperforms recent related work using pattern-based extraction of such hypernym relations for coreference resolution.

1 Introduction

Successful resolution and generation of definite anaphora requires knowledge of hypernym and hyponym relationships. For example, determining the antecedent to the definite anaphor "*the drug*" in text requires knowledge of what previous noun-phrase candidates could be drugs. Likewise, generating a definite anaphor for the antecedent "*Morphine*" in text requires both knowledge of potential hypernyms (e.g. "*the opiate*", "*the narcotic*", "*the drug*", and "*the substance*"), as well as selection of the most appropriate level of generality along the hypernym tree in context (i.e. the "natural" hypernym anaphor). Unfortunately existing manual hypernym databases such as WordNet are very incomplete, especially for technical vocabulary and proper names. WordNets are also limited or non-existent for most of the world's languages. Finally, WordNets also do not include notation of the "natural" hypernym level for anaphora generation, and using the immediate parent performs quite poorly, as quantified in Section 5.
In first part of this paper, we propose a novel approach for *resolving* definite anaphora involving hyponymy relations. We show that it performs substantially better than previous approaches on the task of antecedent selection. In the second part we demonstrate how this approach can be successfully extended to the problem of *generating* a natural definite NP given a specific antecedent.
In order to explain the antecedent selection task for definite anaphora clearly, we provide the following example taken from the LDC Gigaword corpus (Graff *et al.*, 2005).

(1)...*pseudoephedrine* is found in an allergy treatment, which was given to Wilson by a doctor when he attended Blinn junior college in Houston. In a unanimous vote, the Norwegian sports confederation ruled that Wilson had not taken **the drug** to enhance his performance...

In the above example, the task is to resolve the definite NP *the drug* to its correct antecedent *pseudoephedrine*, among the potential antecedents <*pseudoephedrine, allergy, blinn, college, houston, vote, confederation, wilson*>. Only *Wilson* can be ruled out on syntactic grounds (Hobbs, 1978). To be able to resolve the correct antecedent from the remaining potential antecedents, the system requires the knowledge that *pseudoephedrine* is a *drug*. Thus, the problem is to create such a knowledge source and apply it to this task of antecedent selection. A total of 177 such anaphoric examples

Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X),
pages 37–44, New York City, June 2006. ©2006 Association for Computational Linguistics

were extracted randomly from the LDC Gigaword corpus and a human judge identified the correct antecedent for the definite NP in each example (given a context of previous sentences).[1] Two human judges were asked to perform the same task over the same examples. The agreement between the judges was 92% (of all 177 examples), indicating a clearly defined task for our evaluation purposes.

We describe an unsupervised approach to this task that extracts examples containing definite NPs from a large corpus, considers all head words appearing before the definite NP as potential antecedents and then filters the noisy <antecedent, definite-NP> pair using Mutual Information space. The co-occurence statistics of such pairs can then be used as a mechanism for detecting a hypernym relation between the definite NP and its potential antecedents. We compare this approach with a WordNet-based algorithm and with an approach presented by Markert and Nissim (2005) on resolving definite NP coreference that makes use of lexico-syntactic patterns such as *'X and Other Ys'* as utilized by Hearst (1992).

2 Related work

There is a rich tradition of work using lexical and semantic resources for anaphora and coreference resolution. Several researchers have used WordNet as a lexical and semantic resource for certain types of bridging anaphora (Poesio *et al.*, 1997; Meyer and Dale, 2002). WordNet has also been used as an important feature in machine learning of coreference resolution using supervised training data (Soon *et al.*, 2001; Ng and Cardie, 2002). However, several researchers have reported that knowledge incorporated via WordNet is still insufficient for definite anaphora resolution. And of course, WordNet is not available for all languages and is missing inclusion of large segments of the vocabulary even for covered languages. Hence researchers have investigated use of corpus-based approaches to build a WordNet like resource automatically (Hearst, 1992; Caraballo, 1999; Berland and Charniak, 1999). Also, several researchers have applied it to resolving different types of bridging anaphora (Clark, 1975). Poesio *et al.* (2002) have proposed extracting lexical knowledge about part-of relations using Hearst-style patterns and applied it to the task of resolving bridging references. Poesio *et al.* (2004) have suggested using Google as a source of computing lexical distance between antecedent and definite NP for mereological bridging references (references referring to parts of an object already introduced). Markert *et al.* (2003) have applied relations extracted from lexico-syntactic patterns such as *'X and other Ys'* for Other-Anaphora (referential NPs with modifiers *other* or *another*) and for bridging involving meronymy.

There has generally been a lack of work in the existing literature for automatically building lexical resources for definite anaphora resolution involving hyponyms relations such as presented in Example (1). However, this issue was recently addressed by Markert and Nissim (2005) by extending their work on Other-Anaphora using lexico syntactic pattern *'X and other Y's* to antecedent selection for definite NP coreference. However, our task is more challenging since the anaphoric definite NPs in our test set include only hypernym anaphors without including the much simpler cases of headword repetition and other instances of string matching. For direct evaluation, we also implemented their corpus-based approach and compared it with our models on identical test data.

We also describe and evaluate a mechanism for combining the knowledge obtained from WordNet and the six corpus-based approaches investigated here. The resulting models are able to overcome the weaknesses of a WordNet-only model and substantially outperforms any of the individual models.

3 Models for Lexical Acquisition

3.1 TheY-Model

Our algorithm is motivated by the observation that in a discourse, the use of the definite article ("the") in a non-deictic context is primarily licensed if the concept has already been mentioned in the text. Hence a sentence such as "The drug is very expensive" generally implies that either the word *drug* itself was previously mentioned (e.g. "He is taking a new drug for his high cholesterol.") or a hyponym of *drug* was

[1] The test examples were selected as follows: First, all the sentences containing definite NP "*The Y*" were extracted from the corpus. Then, the sentences containing instances of anaphoric definite NPs were kept and other cases of definite expressions (like existential NPs "*The White House*","*The weather*") were discarded. From this anaphoric set of sentences, 177 sentence instances covering 13 distinct hypernyms were randomly selected as the test set and annotated for the correct antecedent by human judges.

previously mentioned (e.g. "He is taking Lipitor for his high cholesterol."). Because it is straightforward to filter out the former case by string matching, the residual instances of the phrase "the drug" (without previous mentions of the word "drug" in the discourse) are likely to be instances of hypernymic definite anaphora. We can then determine which nouns earlier in the discourse (e.g. *Lipitor*) are likely antecedents by unsupervised statistical co-occurrence modeling aggregated over the entire corpus. All we need is a large corpus without any anaphora annotation and a basic tool for noun tagging and NP head annotation. The detailed algorithm is as follows:

1. Find each sentence in the training corpus that contains a definite NP (*'the Y'*) and does not contain *'a Y'*, *'an Y'* or other instantiations of Y[2] appearing before the definite NP within a fixed window.[3]

2. In the sentences that pass the above definite NP and *a/an test*, regard all the head words (X) occurring in the current sentence before the definite NP and the ones occurring in previous two sentences as potential antecedents.
3. Count the frequency c(X,Y) for each pair obtained in the above two steps and pre-store it in a table.[4] The frequency table can be modified to give other scores for pair(X,Y) such as standard TF-IDF and Mutual Information scores.
4. Given a test sentence having an anaphoric definite NP Y, consider the nouns appearing before Y within a fixed window as potential antecedents. Rank the candidates by their precomputed co-occurence measures as computed in Step 3.

Since we consider *all* head words preceding the definite NP as potential correct antecedents, the raw frequency of the pair (X,Y) can be very noisy. This can be seen clearly in Table 1, where the first column shows the top potential antecedents of definite NP *the drug* as given by raw frequency. We normalize the raw frequency using standard TF-IDF and Mutual Information scores to filter the noisy pairs.[5] In Table 2, we report our results for antecedent selection using Raw frequency c(X,Y), TF-IDF [6] and MI in isolation. *Accuracy* is the fraction of total examples that were assigned the correct antecedent and *Accuracy*$_{tag}$ is the same excluding the examples that had POS tagging errors for the correct antecedent.[7] *Av Rank* is the rank of the true antecedent averaged over the number of test examples.[8] Based on the above experiment, the rest of this paper assumes Mutual Information scoring technique for TheY-Model.

Rank	Raw freq	TF-IDF	**MI**
1	today	kilogram	**amphetamine**
2	police	heroin	**cannabis**
3	kilogram	police	**cocaine**
4	year	cocaine	**heroin**
5	heroin	today	**marijuana**
6	dollar	trafficker	**pill**
7	country	officer	**hashish**
8	official	amphetamine	**tablet**

Table 1: A sample of ranked hyponyms proposed for the definite NP ***The drug*** by TheY-Model illustrating the differences in weighting methods.

	Acc	Acc$_{tag}$	Av Rank
MI	**0.531**	**0.577**	**4.82**
TF-IDF	0.175	0.190	6.63
Raw Freq	0.113	0.123	7.61

Table 2: Results using different normalization techniques for the TheY-Model in isolation. (60 million word corpus)

[2] While matching for both *'the Y'* and *'a/an Y'*, we also account for Nouns getting modified by other words such as adjectives. Thus *'the Y'* will still match to *'the green and big Y'*.

[3] Window size was set to two sentences, we also experimented with a larger window size of five sentences and the results obtained were similar.

[4] Note that the count c(X,Y) is asymmetric

[5] Note that $MI(X,Y) = log\ \frac{P(X,Y)}{P(X)P(Y)}$ and this is directly proportional to $P(Y|X) = \frac{c(X,Y)}{c(X)}$ for a fixed Y. Thus, we can simply use this conditional probability during implementation since the definite NP Y is fixed for the task of antecedent selection.

[6] For the purposes of TF-IDF computation, document frequency df(X) is defined as the number of unique definite NPs for which X appears as an antecedent.

[7] Since the POS tagging was done automatically, it is possible for any model to miss the correct antecedent because it was not tagged correctly as a noun in the first place. There were 14 such examples in the test set and none of the model variants can find the correct antecdent in these instances.

[8] Knowing average rank can be useful when a n-best ranked list from coreference task is used as an input to other downstream tasks such as information extraction.

	Acc	Acc_{tag}	Av Rank
TheY+WN	**0.695**	**0.755**	3.37
WordNet	0.593	0.644	**3.29**
TheY	0.531	0.577	4.82

Table 3: Accuracy and Average Rank showing combined model performance on the antecedent selection task. Corpus Size: 60 million words.

3.2 WordNet-Model (WN)

Because WordNet is considered as a standard resource of lexical knowledge and is often used in coreference tasks, it is useful to know how well corpus-based approaches perform as compared to a standard model based on the WordNet (version 2.0).[9] The algorithm for the WordNet-Model is as follows:
Given a definite NP Y and its potential antecedent X, choose X if it occurs as a hyponym (either direct or indirect inheritance) of Y. If multiple potential antecedents occur in the hierarchy of Y, choose the one that is closest in the hierarchy.

3.3 Combination: TheY+WordNet Model

Most of the literature on using lexical resources for definite anaphora has focused on using individual models (either corpus-based or manually build resources such as WordNet) for antecedent selection. Some of the difficulties with using WordNet is its limited coverage and its lack of empirical ranking model. We propose a combination of TheY-Model and WordNet-Model to overcome these problems. Essentially, we rerank the hypotheses found in WordNet-Model based on ranks of TheY-model or use a backoff scheme if WordNet-Model does not return an answer due to its limited coverage. Given a definite NP Y and a set of potential antecedents Xs the detailed algorithm is specified as follows:

1. *Rerank with TheY-Model:* Rerank the potential antecedents found in the WordNet-Model table by assiging them the ranks given by TheY-Model. If TheY-Model does not return a rank for a potential antecedent, use the rank given by the WordNet-Model. Now pick the top ranked antecedent after reranking.

2. *Backoff:* If none of the potential antecedents were found in the WordNet-Model then pick the correct antecedent from the ranked list of The-Y model. If none of the models return an answer then assign ranks uniformly at random.

The above algorithm harnesses the strength of WordNet-Model to identify good hyponyms and the strength of TheY-model to identify which are more likely to be used as an antecedent. Note that this combination algorithm can be applied using any corpus-based technique to account for poor-ranking and low-coverage problems of WordNet and the Sections 3.4, 3.5 and 3.6 will show the results for backing off to a Hearst-style hypernym model. Table 4 shows the decisions made by TheY-model, WordNet-Model and the combined model for a sample of test examples. It is interesting to see how both the models mutually complement each other in these decisions. Table 3 shows the results for the models presented so far using a 60 million word training text from the Gigaword corpus. The combined model results in a substantially better accuracy than the individual WordNet-Model and TheY-Model, indicating its strong merit for the antecedent selection task.[10]

3.4 OtherY-Model$_{freq}$

This model is a reimplementation of the corpus-based algorithm proposed by Markert and Nissim (2005) for the equivalent task of antecedent selection for definite NP coreference. We implement their approach of using the lexico-syntactic pattern *X and* $A*$ *other* $B*$ $Y\{pl\}$ for extracting (X,Y) pairs.The $A*$ and $B*$ allow for adjectives or other modifiers to be placed in between the pattern. The model presented in their article uses the raw frequency as the criteria for selecting the antecedent.

3.5 OtherY-Model$_{MI}$(normalized)

We normalize the OtherY-Model using Mutual Information scoring method. Although Markert and Nissim (2005) report that using Mutual Information performs similar to using raw frequency, Table 5 shows that using Mutual Information makes a substantial impact on results using large training corpora relative to using raw frequency.

[9] We also computed the accuracy using a weaker baseline, namely, selecting the closest previous headword as the correct antecedent. This recency based baseline obtained a low accuracy of 15% and hence we used the stronger WordNet based model for comparison purposes.

[10] The claim is statistically significant with a $p < 0.01$ obtained by sign-test

Summary	Keyword (Def. Ana)	True Antecedent	TheY Choice	Truth Rank	WordNet Choice	Truth Rank	TheY+WN Choice	Truth Rank
Both correct	metal	gold	gold	1	gold	1	gold	1
	sport	soccer	soccer	1	soccer	1	soccer	1
TheY-Model helps	drug	steroid	steroid	1	NA	NA	steroid	1
	drug	azt	azt	1	medication	2	azt	1
WN-Model helps	instrument	trumpet	king	10	trumpet	1	trumpet	1
	drug	naltrexone	alcohol	14	naltrexone	1	naltrexone	1
Both incorrect	weapon	bomb	artillery	3	NA	NA	artillery	3
	instrument	voice	music	9	NA	NA	music	9

Table 4: A sample of output from different models on antecedent selection (60 million word corpus).

3.6 Combination: TheY+OtherY$_{MI}$ Model

Our two corpus-based approaches (TheY and OtherY) make use of different linguistic phenomena and it would be interesting to see whether they are complementary in nature. We used a similar combination algorithm as in Section 3.3 with the WordNet-Model replaced with the OtherY-Model for hypernym filtering, and we used the noisy TheY-Model for reranking and backoff. The results for this approach are showed as the entry TheY+OtherY$_{MI}$ in Table 5. We also implemented a combination (OtherY+WN) of Other-Y model and WordNet-Model by replacing TheY-Model with OtherY-Model in the algorithm described in Section 3.3. The respective results are indicated as OtherY+WN entry in Table 5.

4 Further Anaphora Resolution Results

Table 5 summarizes results obtained from all the models defined in Section 3 on three different sizes of training unlabeled corpora (from Gigaword corpus). The models are listed from high accuracy to low accuracy order. The OtherY-Model performs particularly poorly on smaller data sizes, where coverage of the Hearst-style patterns maybe limited, as also observed by Berland and Charniak (1999). We further find that the Markert and Nissim (2005) OtherY-Model and our MI-based improvement do show substantial relative performance growth at increased corpus sizes, although they still underperform our basic TheY-Model at all tested corpus sizes. Also, the combination of corpus-based models (TheY-Model+OtherY-model) does indeed performs better than either of them in isolation. Finally, note that the basic TheY-algorithm still does

	Acc	Acc$_{tag}$	Av Rank
60 million words			
TheY+WN	**0.695**	**0.755**	3.37
OtherY$_{MI}$+WN	0.633	0.687	**3.04**
WordNet	0.593	0.644	3.29
TheY	0.531	0.577	4.82
TheY+OtherY$_{MI}$	0.497	0.540	4.96
OtherY$_{MI}$	0.356	0.387	5.38
OtherY$_{freq}$	0.350	0.380	5.39
230 million words			
TheY+WN	**0.678**	**0.736**	3.61
OtherY$_{MI}$+WN	0.650	0.705	**2.99**
WordNet	0.593	0.644	3.29
TheY+OtherY$_{MI}$	0.559	0.607	4.50
TheY	0.519	0.564	4.64
OtherY$_{MI}$	0.503	0.546	4.37
OtherY$_{freq}$	0.418	0.454	4.52
380 million words			
TheY+WN	**0.695**	**0.755**	3.47
OtherY$_{MI}$+WN	0.644	0.699	**3.03**
WordNet	0.593	0.644	3.29
TheY+OtherY$_{MI}$	0.554	0.601	4.20
TheY	0.537	0.583	4.26
OtherY$_{MI}$	0.525	0.571	4.20
OtherY$_{freq}$	0.446	0.485	4.36

Table 5: Accuracy and Average Rank of Models defined in Section 3 on the antecedent selection task.

relatively well by itself on smaller corpus sizes, suggesting its merit on resource-limited languages with smaller available online text collections and the unavailability of WordNet. The combined models of WordNet-Model with the two corpus-based approaches still significantly ($p < 0.01$) outperform any of the other individual models.[11]

5 Generation Task

Having shown positive results for the task of antecedent selection, we turn to a more difficult task, namely *generating* an anaphoric definite NP given a nominal antecedent. In Example (1), this would correspond to generating *"the drug"* as an anaphor knowing that the antecedent is *pseudoephedrine*. This task clearly has many applications: current generation systems often limit their anaphoric usage to pronouns and thus an automatic system that does well on hypernymic definite NP generation can directly be helpful. It also has strong potential application in abstractive summarization where rewriting a fluent passage requires a good model of anaphoric usage.

There are many interesting challenges in this problem: first of all, there maybe be multiple acceptable choices for definite anaphor given a particular antecedent, complicating automatic evaluation. Second, when a system generates a definite anaphora, the space of potential candidates is essentially unbounded, unlike in antecdent selection, where it is limited only to the number of potential antecedents in prior context. In spite of the complex nature of this problem, our experiments with the human judgements, WordNet and corpus-based approaches show a simple feasible solution. We evaluate our automatic approaches based on exact-match agreement with definite anaphora actually used in the corpus (accuracy) and also by agreement with definite anaphora predicted independently by a human judge in an absence of context.

5.1 Human experiment

We extracted a total of 103 <true antecedent, definite NP> pairs from the set of test instances used in the resolution task. Then we asked a human judge (a native speaker of English) to predict a parent class of the antecedent that could act as a good definite anaphora choice in general, independent of a particular context. Thus, the actual corpus sentence containing the antecedent and definite NP and its context was not provided to the judge. We took the predictions provided by the judge and matched them with the actual definite NPs used in the corpus. The agreement between corpus and the human judge was 79% which can thus be considered as an upper bound of algorithm performance. Table 7 shows a sample of decisions made by the human and how they agree with the definite NPs observed in the corpus. It is interesting to note the challenge of the sense variation and figurative usage. For example, *"corruption"* is refered to as a *"tool"* in the actual corpus anaphora, a metaphoric usage that would be difficult to predict unless given the usage sentence and its context. However, a human agreement of 79% indicate that such instances are relatively rare and the task of predicting a definite anaphor without its context is viable. In general, it appears from our experiements that humans tend to select from a relatively small set of parent classes when generating hypernymic definite anaphora. Furthermore, there appears to be a relatively context-independent concept of the "natural" level in the hypernym hierarchy for generating anaphors. For example, although <*"alkaloid", "organic compound", "compound", "substance", "entity"*> are all hypernyms of *"Pseudoephederine"* in WordNet, *"the drug"* appears to be the preferred hypernym for definite anaphora in the data, with the other alternatives being either too specific or too general to be natural. This natural level appears to be difficult to define by rule. For example, using just the immediate parent hypernym in the WordNet hierarchy only achieves 4% match with the corpus data for definite anaphor generation.

5.2 Algorithms

The following sections presents our corpus-based algorithms as more effective alternatives.

[11] Note that syntactic co-reference candidate filters such as the Hobbs algorithm were not utilized in this study. To assess the performance implications, the Hobbs algorithm was applied to a randomly selected 100-instance subset of the test data. Although the Hobbs algorithm frequently pruned at least one of the coreference candidates, in only 2% of the data did such candidate filtering change system output. However, since both of these changes were improvements, it could be worthwhile to utilize Hobbs filtering in future work, although the gains would likely be modest.

	Agreement w/ human judge	Agreement w/ corpus
TheY+OtherY+WN	**47%**	**46%**
OtherY +WN	43%	43%
TheY+WN	42%	37%
TheY +OtherY	39%	36%
OtherY	39%	36%
WordNet	4%	4%
Human judge	100%	79%
Corpus	79%	100%

Table 6: Agreement of different generation models with human judge and with definite NP used in the corpus.

Antecedent	Corpus Def Ana	Human Choice	TheY+OtherY +WN
racing	sport	sport	sport
azt	drug	drug	drug
missile	weapon	weapon	weapon
alligator	animal	animal	animal
steel	metal	metal	metal
osteporosis	disease	disease	condition
grenade	device	weapon	device
baikonur	site	city	station
corruption	tool	crime	activity

Table 7: Sample of decisions made by human judge and our best performing model (TheY+OtherY+WN) on the generation task.

5.2.1 Individual Models

For the corpus-based approaches, the TheY-Model and OtherY-Model were trained in the same manner as for the antecedent selection task. The only difference was that in the generation case, the frequency statistics were reversed to provide a hypernym given a hyponym. Additionally, we found that raw frequency outperformed either TF-IDF or Mutual Information and was used for all results in Table 6.

The stand-alone WordNet model is also very simple: Given an antecedent, we lookup its direct hypernym (using first sense) in the WordNet and use it as the definite NP, for lack of a better rule for preferred hypernym location.

5.2.2 Combining corpus-based approaches and WordNet

Each of the corpus-based approaches was combined with WordNet resulting in two different models as follows: Given an antecedent X, the corpus-based approach looks up in its table the hypernym of X, for example Y, and only produces Y as the output if Y also occurs in the WordNet as hypernym. Thus WordNet is used as a filtering tool for detecting viable hypernyms. This combination resulted in two models: *'TheY+WN'* and *'OtherY+WN'*.

We also combined all the three approaches, *'TheY'*, *'OtherY'* and WordNet resulting in a single model *'TheY+OtherY+WN'*. This was done as follows: We first combine the models *'TheY'* and *'OtherY'* using a backoff model. The first priority is to use the hypernym from the model *'OtherY'*, if not found then use the hypernym from the model *'TheY'*. Given a definite NP from the backoff model, apply the WordNet filtering technique, specifically, choose it as the correct definite NP if it also occurs as a hypernym in the WordNet hierarchy of the antecedent.

5.3 Evaluation of Anaphor Generation

We evaluated the resulting algorithms from Section 5.2 on the definite NP prediction task as described earlier. Table 6 shows the agreement of the algorithm predictions with the human judge as well as with the definite NP actually observed in the corpus. It is interesting to see that WordNet by itself performs very poorly on this task since it does not have any word-specific mechanism to choose the correct level in the hierarchy and the correct word sense for selecting the hypernym. However, when combined with our corpus-based approaches, the agreement increases substantially indicating that the corpus-based approaches are effectively filtering the space of hypernyms that can be used as natural classes. Likewise, WordNet helps to filter the noisy hypernyms from the corpus predictions. Thus, this interplay between the corpus-based and WordNet algorithm works out nicely, resulting in the best model being a combination of all three individual models and achieving a substantially better agreement with both the corpus and human judge than any of the individual models. Table 7 shows decisions made by this algorithm on a sample test data.

6 Conclusion

This paper provides a successful solution to the problem of incomplete lexical resources for definite anaphora resolution and further demonstrates how the resources built for resolution can be naturally extended for the less studied task of anaphora generation. We first presented a simple and noisy corpus-based approach based on globally modeling headword co-occurrence around likely anaphoric definite NPs. This was shown to outperform a recent approach by Markert and Nissim (2005) that makes use of standard Hearst-style patterns extracting hypernyms for the same task. Even with a relatively small training corpora, our simple TheY-model was able to achieve relatively high accuracy, making it suitable for resource-limited languages where annotated training corpora and full WordNets are likely not available. We then evaluated several variants of this algorithm based on model combination techniques. The best combined model was shown to exceed 75% accuracy on the resolution task, beating any of the individual models. On the much harder anaphora generation task, where the stand-alone WordNet-based model only achieved an accuracy of 4%, we showed that our algorithms can achieve 35%-47% accuracy on blind exact-match evaluation, thus motivating the use of such corpus-based learning approaches on the generation task as well.

Acknowledgements

Thanks to Charles Schafer for sharing his tools on POS/Headword tagging for the Gigaword corpus.

References

M. Berland and E. Charniak. 1999. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 57–64.

S. Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 120–126.

H. H. Clark. 1975. Bridging. In *Proceedings of the Conference on Theoretical Issues in Natural Language Processing*, pages 169–174.

D. Connoly, J. D. Burger, and D. S. Day. 1997. A machine learning approach to anaphoric reference. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 133–144.

D. Graff, J. Kong, K. Chen, and K. Maeda. 2005. English Gigaword Second Edition. Linguistic Data Consortium, catalog number LDC2005T12.

S. Harabagiu, R. Bunescu, and S. J. Maiorano. 2001. Text and knowledge mining for coreference resolution. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 55–62.

M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545.

J. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

K. Markert and M. Nissim. 2005. Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–402.

K. Markert, M. Nissim, and N. N. Modjeska. 2003. Using the web for nominal anaphora resolution. In *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, pages 39–46.

J. Meyer and R. Dale. 2002. Mining a corpus to support associative anaphora resolution. In *Proceedings of the Fourth International Conference on Discourse Anaphora and Anaphor Resolution*.

V. Ng and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111.

M. Poesio, R. Vieira, and S. Teufel. 1997. Resolving bridging references in unrestricted text. In *Proceedings of the ACL Workshop on Operational Factors in Robust Anaphora*, pages 1–6.

M. Poesio, T. Ishikawa, S. Schulte im Walde, and R. Viera. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proccedings of the Third Conference on Language Resources and Evaluation*, pages 1220–1224.

M. Poesio, R. Mehta, A. Maroudas, and J. Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 143–150.

W. M. Soon, H. T. Ng, and D. C. Y. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

M. Strube, S. Rapp, and C. Müller. 2002. The influence of minimum edit distance on reference resolution. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 312–319.

R. Vieira and M. Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

X. Yang, G. Zhou, J. Su, and C. L. Tan. 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 176–183.