Detecting Anaphoricity and Antecedenthood for Coreference Resolution

Detección de la anaforicidad y de los antecedentes para la resolución de la correferencia

Olga Uryupina

Institute of Linguistics, Russian Academy of Science

B. Kislovky per. 1/12, Moscow

uryupina@gmail.com

Resumen: La mayoría de sistemas de resolución de la correferencia (CR) intentan resolver todos los candidatos a anáfora comparándolos a todos los antecedentes candidatos precedentes hasta que se encuentra el correcto. En este estudio se investigan las posibilidades de identificar las anáforas y antecedentes improbables. Se evalúa nuestra aproximación con el corpus MUC-7.

Palabras clave: Correferencia, anaforicidad

Abstract: Most coreference resolution (CR) systems try to resolve all "candidate anaphors" by comparing them to all preceding "candidate antecedents" until the correct one is found. In this study we investigate possibilities to identify unlikely anaphors and antecedents. We evaluate our approach on the MUC-7 corpus.

Keywords: Coreference, Anaphoricity

1 Introduction

Most coreference resolution (CR) systems try to resolve all "candidate anaphors" by comparing them to all preceding "candidate antecedents" until the correct one is found. Not all noun phrases in a document, however, participate in coreference relations, and, even if they do, they often can only be anaphors or antecedents, but not both. Present study investigates possibilities to automatically reduce the pool of anaphors and antecedents by filtering out unlikely candidates.

In some cases, we can determine if a markable could potentially be an anaphor or an antecedent by looking at its structure and surrounding context. Consider the following example:

(1) Shares in [Loral Space]$_1$ will be distributed to Loral shareholders. [The new company]$_{2,ante=1}$ will start life with [no debt]$_3$ and $700 million in cash. [Globalstar]$_4$ still needs to raise [$600 million]$_5$, and Schwartz said that [the company]$_{6,ante=4}$ would try to raise [the money]$_{7,ante=5}$ in [the debt market]$_8$.

The third markable, "no debt" can be neither an anaphor, nor an antecedent. We can tell that by looking at its structure – with the determiner "no", this description does not re-fer to any entity. The second, sixth and seventh markables are all definite descriptions and therefore are likely to be anaphoric. The eighth markable, "the debt market" is a definite NP, but it is a uniquely referring description and thus it might as well be non anaphoric. Finally, the fifth markable, "$600 million" is a possible antecedent (and is indeed mentioned again as "the money" later), but not a very likely anaphor.

Most CR systems, including, for example, the algorithm of Soon, Ng, and Lim (2001) try to resolve all "candidate anaphors" by comparing them to all preceding "candidate antecedents" until the correct one is found. Such approaches require substantial amount of processing: in the worst case one has to check $n(n-1)/2$ candidate pairs, where n is the total number of markables found by the system. Moreover, spurious coreference links may appear when, for example, a non-anaphoric description is resolved to some preceding markable.

Vieira and Poesio (2000) have shown that such an exhaustive search is not needed, because many noun phrases are not anaphoric at all: more than 50% of definite NPs in their corpus have no prior referents. Obviously, this number is even higher if one takes into

account all the other types of NPs – for example, only 30% of our (automatically extracted) markables are anaphoric.

We can conclude that a coreference resolution engine might benefit from a pre-filtering algorithm for identifying non-anaphoric and non-antecedent descriptions. First, we save much processing time by discarding at least half of the markables. Second, the prefiltering module is expected to improve the system's precision by discarding spurious candidates.

In Section 2 we briefly summarise theoretical research on anaphoricity and referentiality and discuss the related applications. Note that theoretical studies focus on referentiality, whereas we will consider a related task of detecting antecedenthood (this will be described in details below). In Section 3 we experiment on learning anaphoricity and antecedenthood filters from the MUC data. In Section 4 we incorporate the anaphoricity and antecedenthood classifiers into a baseline no-prefiltering coreference resolution system to see if such prefiltering modules help.

2   Related Work

In this section, we present an overview of theoretical studies of referentiality (Karttunen, 1976) and anaphoricity (Prince, 1981). We also discuss relevant computational approaches (Bean and Riloff, 1999; Ng and Cardie, 2002; Uryupina, 2003; Vieira and Poesio, 2000; Byron and Gegg-Harrison, 2004).

Karttunen (1976) points out that in some cases an NP, in particular an indefinite one, does not refer to any entity:

(2) Bill doesn't have [a car].

Obviously, (2) does not imply the existence of any specific "car". In Karttunen's terms, the NP "a car" does not establish a discourse referent and therefore it cannot participate in any coreference chain – none of the alternatives in (3) can follow (2):

(3) A.[It] is black.
B.[The car] is black.
C.[Bill's car] is black.

Karttunen (1976) identifies several factors affecting referential status of NPs, including modality, negation, or nonfactive verbs. He argues that an extensive analysis of the phenomenon requires sophisticated inference: "In order to decide whether or not a nonspe-

cific indefinite NP is to be associated with a referent, a text-interpreting device must be able to assign a truth value to the proposition represented by the sentence in which the NP appears. It must be sensitive to the semantic properties of verbs that take sentential complements; distinguish between assertion, implication, and presupposition; and finally, it must distinguish what exists for the speaker from what exists only for somebody else".

Byron and Gegg-Harrison (2004) present an algorithm for identifying "nonlicensing" NPs based on Karttunen's theory of referentiality. Their approach relies on a handcrafted heuristic, encoding some of (Karttunen, 1976) factors. In the present study we represent this information as features for machine learning.

Numerous theories of anaphoricity, especially for definite descriptions, have been proposed in the literature. We point the reader to Vieira (1998) for an extensive overview and comparison of the major theoretic studies in the field. The theories aim at interpreting (definite) descriptions by relating them to the linguistic and situational context and, more specifically, to their antecedents.

From this perspective, an NP may be *given* (related to the preceding discourse) or *new* (introducing an independent entity). The theories of anaphoricity provide different detailed subclassifications of given and new descriptions. For example, Prince (1981) distinguishes between the discourse and the hearer givenness. This results in the following taxonomy:

- *brand new* NPs introduce entities which are both discourse and hearer new ("a bus"), some of them, *brand new anchored* NPs, contain explicit link to some given discourse entity ("a guy I work with"),

- *unused* NPs introduce discourse new, but hearer old entities ("Noam Chomsky"),

- evoked NPs introduce entities already present in the discourse model and thus discourse and hearer old: *textually evoked* NPs refer to entities which have already been mentioned in the previous discourse ("he" in "A guy I worked with says he knows your sister"), whereas *situationally evoked* are known for situ-

ational reasons ("you" in "Would you have change of a quarter?"),

- *inferrables* are not discourse or hearer old, however, the speaker assumes the hearer can infer them via logical reasoning from evoked entities or other inferrables ("the driver" in "I got on a bus yesterday and the driver was drunk"), *containing inferrables* make this inference link explicit ("one of these eggs").

Linguistic theories, including (Prince, 1981), focus on anaphoric usages of definite descriptions (either evoked or inferrables). Recent corpus studies (Poesio and Vieira, 1998) have revealed, however, that more than 50% of (definite) NPs in newswire texts are not anaphoric. These findings have motivated recent approaches to automatic identification of discourse new vs. old NPs.

Several algorithms for identifying discourse-new markables have been proposed in the literature, especially for definite descriptions. Vieira and Poesio (2000) use hand-crafted heuristics, encoding syntactic information. For example, the noun phrase "the inequities of the current land-ownership system" is classified by their system as discourse new, because it contains the restrictive postmodification "of the current land-ownership system". This approach leads to 72% precision and 69% recall for definite discourse-new NPs on their corpus. Palomar and Muñoz (2000) propose a related algorithm for Spanish.

Bean and Riloff (1999) make use of syntactic heuristics, but also mine additional patterns for discourse-new markables from corpus data. Using various combinations of these methods, (Bean and Riloff, 1999) achieve an F-measure for existential NPs of about 81–82% on the MUC-4 data.[1]

In an earlier paper (Uryupina, 2003) we have proposed a web-based algorithm for identifying discourse-new and unique NPs. Our approach helps overcome the data sparseness problem of Bean and Riloff (1999) by relying on Internet counts.

The above-mentioned algorithms for automatic detection of discourse-new and non-referential descriptions are helpful for interpreting NPs, accounting for documents information structure.However, it is not a priori clear whether such approaches are useful for coreference resolution. On the one hand, discarding discourse-new and/or non-referential NPs from the pool of candidate anaphors and antecedents, we can drastically narrow down the algorithm's search space. This reduces the processing time and makes candidate re-ranking much easier. On the other hand, errors, introduced by automatic anaphoricity or referentiality detectors, may propagate and thus deteriorate the performance of a coreference resolution engine.

Ng and Cardie (2002) have shown that an automatically induced detector of non-anaphoric descriptions leads to performance losses for their coreference resolution engine, because too many anaphors are misclassified as discourse-new. To deal with the problem, they have augmented their discourse-new classifier with several precision-improving heuristics. In our web-based study (Uryupina, 2003) we have tuned machine learning parameters to obtain a classifier with a better precision level. In a later study, Ng (2004) relies on held-out data to optimise relevant learning parameters and to decide on the possible system architecture.

Byron and Gegg-Harrison (2004) report ambivalent results concerning the importance of a referentiality detector for pronominal coreference. On the one hand, the incorporation of referentiality prefiltering in several pronoun resolution algorithms does not yield any significant precision gains. On the other hand, such a prefiltering significantly reduced the systems' processing time.

To summarise, several algorithms for detecting non-referring or non-anaphoric descriptions have been proposed in the literature. These studies revealed two major problems. First, it is necessary to identify and represent relevant linguistic factors affecting the referentiality or anaphoricity status of an NP. Second, incorporating error-prone automatic modules for identifying discourse-new or non-referential descriptions into a coreference resolution engine is a non-trivial task of its own: when not properly optimised, such modules may lead to performance losses. We will address these two problems in the following sections.

---

[1] (Bean and Riloff, 1999) *existential* class contains not only brand new NPs, but also all mentions (including anaphoric) of unique descriptions, such as "the pope" or "the FBI".

3 Identifying Non-anaphors and Non-antecedents

Corpus studies (Poesio and Vieira, 1998) suggest that human annotators are able to successfully distinguish between anaphoric (discourse old) and non-anaphoric (discourse-new) descriptions. This motivates the present experiment: using machine learning techniques we try to automatically detect probable anaphors and antecedents. In our next experiment (Section 4) we will incorporate our anaphoricity and referentiality classifiers into a coreference resolution system.

3.1 Data

We use the MUC-7 corpus in our experiment. We have automatically extracted noun phrases using Charniak's parser (Charniak, 2000) and C&C NE-tagging system (Curran and Clark, 2003).

We have automatically annotated our NPs as $\pm discourse\_new$ using the following simple rule: an NP is considered $-discourse\_new$ if and only if it is marked in the corpus and has an antecedent.

Extracting referentiality information from coreference annotated data is by far less trivial. By definition (Karttunen, 1976), non-referential descriptions cannot be antecedents for any subsequent NPs. Consider, however, the following example:

(7) There was [no listing]₁ for [the company]₂ in [Wilmington]₃.

In (7), the NP "no listing" is not referential and, therefore, cannot be an antecedent for any subsequent markable. Both "the company" and "Wilmington", on the contrary, are referential and could potentially be re-mentioned. However, this does not happen, as the document ends with the next sentence. By looking at coreference annotated data, we can only say whether an NP is an antecedent, but, if it is not, we cannot decide if it is referential (as "the company" or "Wilmington") or not (as "no listing"). Consequently, we cannot automatically induce referentiality annotation from coreference data.

For our main task, coreference resolution, we are not exactly interested in the referential vs. non-referential distinction. We would rather like to know how likely it is for a markable to be an antecedent. Therefore, instead of a referentiality detector in the strict sense, we need a $\pm ante$ labelling: an NP is considered $+ante$, if it is annotated in MUC-7 and

is an antecedent for some subsequent markable. We have therefore changed the scope of the present experiment to detecting antecedenthood – the probability for a markable to be an antecedent.

In the present experiment, we rely on 30 MUC-7 "dry-run" documents for training. For testing, we use the validation (3 MUC-7 "train" documents) and testing (20 MUC-7 "formal test" documents) sets. This results in 5028 noun phrases for training and 976/3375 for the validation/testing data. 3325 training instances were annotated as $+discourse\_new/-ante$ and 1703 – as $-discourse\_new/+ante$[2] (613/2245 and 363/1130 for testing). All the performance figures reported below are for $+discourse\_new$ and $-ante$ classes.

3.2 Features

We encode our markables with feature vectors, representing different linguistic factors: surface, syntactic, semantic, salience, same-head, and (Karttunen, 1976) properties.

Surface features encode the most shallow properties of an NP, such as its length, amount of upper and lower case characters and digits etc. Syntactic features include POS tags, number and person values, determiner and pre- and post-modification. Semantic features encode gender ans semantic class properties. Salience features encode various rankings within a sentence and a paragraph according to the linear order of the NPs and their grammatical role.

"Same-head" features represent coreference knowledge on a very simplistic level. The boolean feature `same_head_exists` shows if there exists a markable in the preceding discourse with the same head as the given NP, and the continuous feature `same_head_distance` encodes the distance to this markable. Obtaining values for these features does not require exhaustive search when heads are stored in an appropriate data structure, for example, in a trie. The motivation for "same-head" features comes from (Vieira and Poesio, 2000) and (Poesio et al., 2004): they show that anaphoricity detectors might benefit from an early inclusion of a simplified coreference check.

---

[2]As each anaphor is linked to exactly one antecedent according to the MUC-7 annotation guidelines, there is a one-to-one correspondence between $-discourse\_new$ and $+ante$ classes.

The last group encodes the referentiality-related factors investigated by Karttunen (1976) and Byron and Gegg-Harrison (2004): apposition, copula, negation, modal constructions, determiner, grammatical role, and semantic class. The values are extracted from the parser's and the NE-tagger's output.

Altogether we have 49 features: 12 surface, 20 syntactic, 3 semantic, 10 salience, 2 "same-head", and 7 of Karttunen's constructions, corresponding to 123 boolean/continuous features.

3.3 Identifying discourse-new markables

As a baseline for our experiments we use the major class labelling: all markables are classified as $+discourse\_new$. This results in F-scores of 79.9% and 77.2% for the testing and validation data. This baseline can be used as a comparison point for $\pm discourse\_new$ detectors. However, it has no practical relevance for our main task, coreference resolution: if we classify all the markables as $+discourse\_new$ and, consequently, discard them, the system would not even try to resolve any anaphors. In all the tables in this paper we show significant improvements over the baseline for $p < 0.05 / p < 0.01$ by $*/**$ and significant losses – by $\dagger / \dagger\dagger$.

We have trained the SVMlight classifier for $\pm discourse\_new$ descriptions. Its performance is summarised in Table 1. Compared to the baseline, the recall goes down (the baseline classifies everything as $+discourse\_new$, showing the recall level of 100%), but the precision improves significantly. This results in an F-score improvement of 5-8%, corresponding to 23-38% relative error reduction.

Among different feature groups, surface, salience, and (Karttunen, 1976) factors show virtually no performance gain over the baseline. Surface features are too shallow. Salience and (Karttunen, 1976)-motivated features have primarily been designed to account for the probability of a markable being an antecedent, not an anaphor. Based on semantic features alone, the classifier does not perform different from the baseline – although, by bringing the recall and precision values closer together, the F-score improves, the precision is still low.

The two groups with the best precision level are syntactic and "same head" features.

In fact, the classifier based on these features alone (Table 1, last line) achieves almost the same performance level as the one based on all features taken together (no significant difference in precision and recall, χ^2-test).

As we have already mentioned when discussing the baseline, from a coreference resolution perspective, we are interested in a discourse-new detector with a high precision level: each anaphor misclassified as discourse new is excluded from further processing and therefore cannot be resolved. On the contrary, if we misclassify a non-anaphoric entity as discourse old, we still can hope to correctly leave it unresolved by rejecting all the candidate antecedents. Therefore we might want to improve the precision of our discourse-new detector as much as possible, even at the expense of recall.

To increase the precision level, we have chosen another machine learner, Ripper, that allows to control the precision/recall trade-off by manually optimising the LossRatio parameter: by varying the LossRatio from 0.3^3 to 1.0, we obtain different precision and recall values. As in SVM's case, the best performing groups are syntactic and "same head" features. With all the features activated, the precision gets as high as 90% when the LossRatio is low. In Section 4 we will see if this performance is reliable enough to help a coreference resolution engine.

3.4 Identifying non-antecedents

We have trained another family of classifiers to detect non-antecedents. Table 2 shows SVM's performance for the $\pm ante$ task. The major class labelling, $-ante$ serves as a baseline. The classifier's performance is lower than for the $\pm discourse\_new$ task, with only syntactic and semantic features leading to a significant precision improvement over the baseline.

The lower performance level reflects the intrinsic difficulty of the task. When processing a text, the reader has to decide if an encountered description is a re-mention or a new entity to be able to correctly ground it in the discourse model. Therefore we can expect linguistic cues to signal if a markable is $\pm discourse\_new$. For $\pm ante$ descriptions, on the contrary, there is no need for such signals: often an entity is introduced but then never

---

[3]Lower values result in the trivial labelling ("classify everything as discourse old").

mentioned again as the topic changes.

As Table 2 shows, the classifier mostly makes precision errors. For non-antecedents, precision is not as crucial as for non-anaphors: if we erroneously discard a correct antecedent, we still can resolve subsequent anaphors to other markables from the same chain. However, if we misclassify the first markable and discard it from the pool of antecedents, we have no chance to correctly resolve the subsequent anaphors.

Consequently, we would still prefer recall errors over precision errors, although not to such extent as for the $\pm discourse\_new$ classifier. We have trained a family of Ripper classifiers to improve the precision level by decreasing the LossRatio parameter from 1.0 to 0.3. The best observed precision level is 80.4% for the "all features" classifier.

To summarise, the present experiment shows that automatically induced classifiers, both SVM and Ripper-based, can successfully identify unlikely anaphors and antecedents. The performance level (F-score) varies around 75-88% for different test sets (validation vs. testing) and tasks ($\pm discourse\_new$ vs. $\pm ante$).

Features	Recall	Precision	F
Baseline	100	66.52	79.89
All	††93.54	**82.29	87.56
Surface	100	66.52	79.89
Syntactic	††97.37	**71.96	82.76
Semantic	††98.53	*68.89	81.09
Salience	††91.22	*69.26	78.74
Same-head	††84.45	**81.16	82.77
Karttunen's	††91.63	**71.15	80.10
Synt+SH	††89.98	**83.51	86.62

Table 1: An SVM-based anaphoricity detector: performance for the $\pm discourse\_new$ class on the test data (20 MUC-7 "formal" documents).

4 Integrating Anaphoricity and Antecedenthood Prefiltering into a Coreference Resolution Engine

In the previous experiment we have learnt two families of classifiers, detecting unlikely anaphors and antecedents. In this section we incorporate them into a baseline coreference resolution system – an SVM classifier with (Soon, Ng, and Lim, 2001) features.

Features	Recall	Precision	F
Baseline	100	66.52	79.89
All	††95.72	*69.23	80.35
Surface	††94.56	68.50	79.45
Syntactic	††95.72	*69.23	80.35
Semantic	††94.92	*69.41	80.18
Salience	††98.88	67.0	79.88
Same-head	100	66.52	79.89
Karttunen's	††99.29	67.31	80.23

Table 2: An SVM-based antecedenthood detector: performance for the $-ante$ class on the test data (20 MUC-7 "formal" documents).

4.1 Oracle settings

To investigate the relevance of anaphoricity and antecedenthood for coreference resolution, we start by incorporating oracle-based prefiltering into the baseline system. For example, our oracle-based anaphoricity filter discards all the discourse-new markables (according to the MUC-7 coreference chains) from the pool of anaphors.

The impact of our ideal filters on the main system is summarised in Table 3. As expected, by constraining the set of possible anaphors and/or antecedents, we dramatically improve the algorithm's precision. Slightly unexpected, the recall goes down even in the oracle setting. This reflects a peculiarity of the MUC-7 scoring scheme – it strongly favours long chains. Prefiltering modules, on the contrary, split long chains into smaller ones.

Several other studies (Ng and Cardie, 2002; Mitkov, Evans, and Orasan, 2002) have revealed similar problems: existing coreference scoring schemes cannot capture the performance of an anaphoricity classifier.

With precision getting much higher at the cost of a slight recall loss, the ideal $\pm discourse\_new$ and $\pm ante$ detectors improve the baseline coreference engine's performance by up to 10% (F-score).

4.2 Automatically acquired detectors

Getting from the oracle setting to a more realistic scenario, we have combined our baseline system with the $\pm discourse\_new$ and $\pm ante$ detectors we have learnt in our first experiment.

The evaluation has been organised as follows. For a given LossRatio value, we have

Prefiltering	Recall	Precision	F-score
No prefiltering (baseline)	54.5	56.9	55.7
Ideal *discourse_new* detector	49.6	**73.6	59.3
Ideal *ante* detector	54.2	**69.4	60.9
Ideal *discourse_new* and *ante* detectors	52.9	**81.9	64.3

Table 3: Incorporating oracle-based $\pm discourse\_new$ and $\pm ante$ prefiltering into a baseline coreference resolution system: performance on the validation data (3 MUC-7 "train" documents).

learnt a $\pm discourse\_new/ \pm ante$ detector as described above. The detector is then incorporated as a pre-filtering module into the baseline system. This allows us to evaluate the performance level of the main coreference resolution engine (the MUC score) depending on the precision/recall trade-off of the pre-filtering modules.

The results (Figures 1 and 2) show that automatically induced detectors drastically decrease the main system's recall: it goes down to 40% (for $\pm discourse\_new$, $L = 0.8$) or even 33% (for $\pm ante$, $L = 1$). For small L values, the system's recall is slightly lower, and the precision higher than the baseline (both differences are not significant). The resulting F-score for the system with prefiltering is slightly lower than the baseline's performance for small values of the Loss Ratio parameter and then decreases rapidly for $L > 0.5$.

To summarise, the results of the present experiment are ambivalent. On the one hand, ideal detectors bring F-score gains by significantly increasing the system's precision. On the other hand, error-prone automatically induced detectors are not reliable enough to produce a similar precision gain and the system's F-score goes down because of the recall loss, as the baseline's recall is already relatively low. Consequently, a coreference resolution algorithm might profit from an automatic $\pm discourse\_new$ or $\pm ante$ detector if its precision has to be improved, for example, if it mainly makes recall errors or, for a specific application, if a high-precision coreference resolution algorithm is required (as, for example, the CogNIAC system proposed by (Baldwin, 1996)).

5 Conclusion

In this paper we have investigated the possibility of automatically identifying unlikely anaphors and antecedents. As only around 30% of markables in newswire
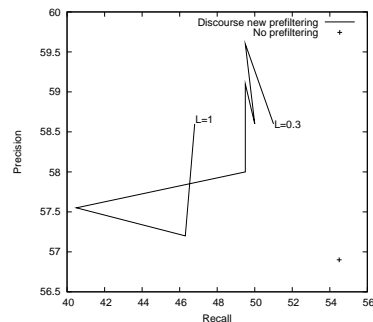

Figure 1: A baseline coreference resolution engine augmented with Ripper-based anaphoricity prefiltering: performance on the validation (3 MUC-7 "train" documents) data for different LossRatio (L) values of prefiltering classifiers.
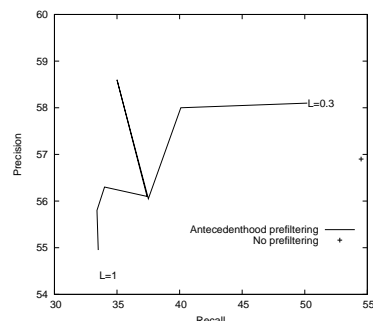

Figure 2: A baseline coreference resolution engine augmented with Ripper-based antecedenthood prefiltering: performance on the validation (3 MUC-7 "train" documents) data for different LossRatio (L) values of prefiltering classifiers.

texts participate in coreference chains, our $\pm discourse\_new$ and $\pm ante$ detectors might significantly constrain the main algorithm's search space, improving its speed and performance.

We have compared different feature groups for the tasks of $\pm discourse\_new$ and $\pm ante$ detection. We have seen that, for both tasks, SVM and Ripper classifiers based on all the investigated features outperform the

baseline. We have also learnt two families of classifiers with different precision/recall trade-offs.

We have incorporated our $\pm discourse\_new$ and $\pm ante$ detectors into a baseline coreference resolution system. We have seen that ideal prefiltering significantly improves the system's precision at the expense of a slight recall loss. This leads to an F-score improvement of up to 10%. Automatically acquired detectors can only moderately improve the system's precision and therefore do not bring any F-score gains.

We still believe, however, that anaphoricity and antecedenthood detectors might help a coreference resolution system with a lower precision and higher recall.

References

Baldwin, Breck. 1996. Cogniac: A high precision pronoun resolution engine. Technical report, University of Pennsylvania.

Bean, David L. and Ellen Riloff. 1999. Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computationa Linguistics*, pages 373–380.

Byron, Donna and Whitney Gegg-Harrison. 2004. Eliminating non-referring noun phrases from coreference resolution. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium*.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 132–139.

Curran, James R. and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the Seventh Conference on Natural Language Learning*, pages 164–167.

Karttunen, Lauri. 1976. Discourse referents. In J. McKawley, editor, *Sytax and Semantics*, volume 7. Academic Press, pages 361–385.

Mitkov, Ruslan, Richard Evans, and Constantin Orasan. 2002. A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In Alexander Gelbukh, editor, *Computational Lin-guistics and Intelligent Text Processing*. Springer, pages 169–187.

Ng, Vincent. 2004. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of the 42nd Annual Meeting of the Association for Computationa Linguistics*.

Ng, Vincent and Claire Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics*.

Palomar, Manuel and Rafael Muñoz. 2000. Definite descriptions in an information extraction systems. In *IBERAMIA-SBIA*, pages 320–328.

Poesio, Massimo, Olga Uryupina, Renata Vieira, Mijail Alexandrov-Kabadjov, and Rodrigo Goulart. 2004. Discourse-new detectors for definite description resolution: a survey and preliminary proposal. In *Proceedings of the Refrence Resolution Workshop at ACL'04*.

Poesio, Massimo and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Prince, Ellen E. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*. Academic Press, pages 223–256.

Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics (Special Issue on Computational Anaphora Resolution)*, 27(4):521–544.

Uryupina, Olga. 2003. High-precision identification of discourse-new and unique noun phrases. In *Proceedings of the ACL'03 Student Workshop*, pages 80–86.

Vieira, Renata. 1998. A review of the linguistic literature on definite descriptions. *Acta Semiotica et Linguistica*, 7:219–258.

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