High-precision Identification of Discourse New and Unique Noun Phrases

Olga Uryupina
Computational Linguistics, Saarland University
Building 17
Postfach 15 11 50
66041 Saarbrücken, Germany
ourioupi@coli.uni-sb.de

Abstract

Coreference resolution systems usually attempt to find a suitable antecedent for (almost) every noun phrase. Recent studies, however, show that many definite NPs are not anaphoric. The same claim, obviously, holds for the indefinites as well.

In this study we try to learn automatically two classifications, $\pm discourse\_new$ and $\pm unique$, relevant for this problem. We use a small training corpus (MUC-7), but also acquire some data from the Internet. Combining our classifiers sequentially, we achieve 88.9% precision and 84.6% recall for discourse new entities.

We expect our classifiers to provide a good prefiltering for coreference resolution systems, improving both their speed and performance.

1 Introduction

Most coreference resolution systems proceed in the following way: they first identify all the possible markables (for example, noun phrases) and then check one by one candidate pairs $(markable_i, markable_j)$, trying to find out whether the members of those pairs can be coreferent. As the final step, the pairs are ranked using a scoring algorithm in order to find an appropriate partition of all the markables into coreference classes.

Those approaches require substantial processing: in the worst case one has to check $\frac{n(n-1)}{2}$ candidate pairs, where n is the total number of markables found by the system. However, R. Vieira and M. Poesio have recently shown in (Vieira and Poesio, 2000) that such an exhaustive search is not needed, because many noun phrases are not anaphoric at all — about 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, indefinites are almost always non-anaphoric.

We can conclude that a coreference resolution engine might benefit a lot from a pre-filtering algorithm for identifying non-anaphoric entities. First, we save much processing time by discarding at least half of the markables. Second, we can hope to reduce the number of mistakes: without pre-filtering, our coreference resolution system might misclassify a discourse new entity as coreferent to some previous one.

However, such a pre-filtering can also decrease the system's performance if too many anaphoric NPs are classified as discourse new: as those NPs are not processed by the main coreference resolution module at all, we cannot find correct antecedents for them. Therefore, we are interested in an algorithm with a good precision, possibly sacrificing its recall to a reasonable extent. V. Ng and C. Cardie analysed in (Ng and Cardie, 2002) the impact of such a prefiltering on their coreference resolution engine. It turned out that an automatically induced $\pm discourse\_new$ classifier did not help to improve the overall performance and even decreased it. However, when more NPs were considered anaphoric (that is, the precision for the $+discourse\_new$ class

increased and the recall decreased), the prefiltering resulted in improving the coreference resolution.

Several algorithms for identifying discourse new entities have been proposed in the literature. R. Vieira and M. Poesio 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.

The system described in (Bean and Riloff, 1999) also makes use of syntactic heuristics. But in addition the authors mine discourse new entities from the corpus. Four types of entities can be classified as non-anaphoric:

1. having specific syntactic structure,
2. appearing in the first sentence of some text in the training corpus,
3. exhibiting the same pattern as several expressions of type (2),
4. appearing in the corpus at least 5 times and always with the definite article (*"definites-only"*).

Using various combinations of these methods, D. Bean and E. Riloff achieved an accuracy for definite non-anaphoric NPs of about $81 - 82\%$ (F-measure), with various combinations of precision and recall.[1] This algorithm, however, has two limitations. First, one needs a corpus consisting of many small texts. Otherwise it is impossible to find enough non-anaphoric entities of type (2) and, hence, to collect enough patterns for the entities of type (3). Second, for an entity to be recognized as "definite-only", it should be found in the corpus at least 5 times. This automatically results in the data sparseness problem, excluding many infrequent nouns and NPs.

[1]Bean and Riloff's *non-anaphoric* NPs do not correspond to our *+discourse_new* ones, but rather to the union of our *+discourse_new* and *+unique* classes.

In our approach we use machine learning to identify non-anaphoric noun-phrases. We combine syntactic heuristics with the "definite probability". Unlike Bean and Riloff, we model definite probability using the Internet instead of the training corpus itself. This helps us to overcome the data sparseness problem to a large extent. As it has been shown recently in (Keller et al., 2002), Internet counts produce reliable data for linguistic analysis, correlating well with corpus counts and plausibility judgements.

The rest of the paper is organised as follows: first we discuss our NPs classification. In Section 3, we describe briefly various data sources we used. Section 4 provides an explanation of our learning strategy and evaluation results. The approach is summarised in Section 5.

2 NP Classification

In our study we follow mainly E. Prince's classification of NPs (Prince, 1981). Prince distinguishes between the discourse and the hearer givenness.The resulting taxonomy is summarised below:

- **brand new** NPs introduce entities which are both discourse and hearer new (*"a bus"*), subclass 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 situational 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"*).

For our present study we do not need such an elaborate classification. Moreover, various experiments of Vieira and Poesio show that even humans have difficulties distinguishing, for example, between inferrables and new NPs, or trying to find an anchor for an inferrable. So, we developed a simple taxonomy following the main Prince's distinction between the discourse and the hearer givenness.

First, we distinguish between discourse new and discourse old entities. An entity is considered discourse old ($-discourse\_new$) if it refers to an object or a person mentioned in the previous discourse. For example, in *"The Navy is considering a new ship that [..] The Navy would like to spend about $ 200 million a year on the arsenal ship.."* the first occurrence of *"The Navy"* and *"a new ship"* are classified as $+discourse\_new$, whereas the second occurrence of *"The Navy"* and *"the arsenal ship"* are classified as $-discourse\_new$. It must be noted that many researchers, in particular, Bean and Riloff, would consider the second *"the Navy"* non-anaphoric, because it fully specifies its referent and does not require information on the first NP to be interpreted successfully. However, we think that a link between two instances of *"the Navy"* can be very helpful, for example, in the Information Extraction task. Therefore we treat those NPs as discourse old. Our $-discourse\_new$ class corresponds to Prince's textually evoked NPs.

Second, we distinguish between uniquely and non-uniquely referring expressions. Uniquely referring expressions ($+unique$) fully specify their referents and can be successfully interpreted without any local supportive context. Main part of the $+unique$ class constitute entities, known to the hearer (reader) already at the moment when she starts processing the text, for example *"The Mount Everest"*. In addition, an NP (unknown to the reader in the very beginning) is considered unique if it fully specifies its referent due to its own content only and thus can be added as it is (maybe, for a very short time) to the reader's World knowledge base after the processing of the text, for example, *"John Smith, chief executive of John Smith Gmbh"* or *"the fact that John Smith is a chief executive of John Smith Gmbh"*. In Prince's terms our $+unique$ class corresponds to the *unused* and, partially, *new*. In our Navy example (cf. above) both occurrences of *"The Navy"* are considered $+unique$, whereas *"a new ship"* and *"the arsenal ship"* are classified as $-unique$.

3 Data

In our research we use 20 texts from the MUC-7 corpus (Hirschman and Chinchor, 1997). The texts were parsed by E. Charniak's parser (Charniak, 2000). Parsing errors were not corrected manually. After this preprocessing step we have 20 lists of noun phrases.

There are discrepancies between our lists and the MUC-7 annotations. First, we consider only noun phrases, whereas MUC-7 takes into account more types of entities (for example, *"his"* in *"his position"* should be annotated according to the MUC-7 scheme, but is not included in our lists). Second, the MUC-7 annotation identifies only markables, participating in some coreference chain. Our lists are produced automatically and thus include all the NPs.

We annotated automatically 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 original MUC-7 corpus, and
- it has an antecedent in the MUC-7 corpus (even if this antecedent does not correspond to any NP in our corpus).

In addition, we annotated our NPs manually as $\pm unique$. The following expressions were considered $+unique$:

- fully specifying the referent without any local or global context (*the chairman of Microsoft Corporation*, *1998*, or *Washington*). We do not take homonymy into account, so, for example, *Washington* is annotated as $+unique$ although it can refer to many different entities: various persons, cities, counties, towns, islands, a state, the government and many others.
- time expressions that can be interpreted uniquely once some starting time point (global context) is specified. The MUC-7 corpus consists of New York Times News Service articles. Obviously, they were designed to be read on some particular day. Thus, for a reader of such

a text, the expressions *on Thursday* or *tomorrow* fully specify their referents. Moreover, the information on the starting time point can be easily extracted from the header of the text.

- expressions, denoting political or administrative objects (for example, *"the Army"*). Although such expressions do not fully specify their referents without an appropriate global context (many countries have *armies*), in an U.S. newspaper they can be interpreted uniquely.

Overall, we have 3710 noun phrases. 2628 of them were annotated as $+discourse\_new$ and 1082 — as $-discourse\_new$. 2651 NPs were classified as $-unique$ and 1059 — as $+unique$. We provide these data to a machine learning system (Ripper).

Another source of data for our experiments is the World Wide Web. To model "definite probability" for a given NP, we construct various phrases, for example, "*the* NP", and send them to the AltaVista search engine. Obtained counts (number of pages worldwide written in English and containing the phrases) are used to calculate values for several "definite probability" features (see Section 4.1 below). We do not use morphological variants in this study.

4 Identifying Discourse New and Unique Expressions

In our experiments we want to learn both classifications $\pm discourse\_new$ and $\pm unique$ automatically. However, not every learning algorithm would be appropriate due to the specific requirements we have. First, we need an algorithm that does not always require all the features to be specified. For example, we might want to calculate "definite probability" for a definite NP, but not for a pronoun. We also don't want to decide a priori, which features are important and which ones are not in any particular case. This requirement rules out such approaches as Memory-based Learning, Naive Bayes, and many others. On the contrary, algorithms, providing tree- or rule-based classifications (for example, C4.5 and Ripper) would fulfil our first requirement ideally.

Second, we want to control precision-recall trade-off, at least for the $\pm discourse\_new$ task. For these reasons we have finally chosen the Ripper learner (Cohen, 1995).

4.1 Features

Our feature set consists currently of 32 features. They can be divided into three groups:

1. **Syntactic Features.** We encode part of speech of the head word and type of the determiner. Several features contain information on the characters, constituting the NP's string (digits, capital and low case letters, special symbols). We use several heuristics for restrictive postmodification. Two types of appositions are identified: with and without commas (*"Rupert Murdoch, News Corp.'s chairman and chief executive officer,"* and *"News Corp.'s chairman and chief executive officer Rupert Murdoch"*). In the MUC-7 corpus, appositions of the latter type are usually annotated as a whole. Charniak's parser, however, analyses these constructions as two NPs (*['News Corp.'s chairman and chief executive officer] [Rupert Murdoch]*). Therefore those cases require special treatment.

2. **Context Features.** For every NP we calculate the distance (in NPs and in sentences) to the previous NP with the same head if such an NP exists. Obtaining values for these features does not require exhaustive search when heads are stored in an appropriate data structure, for example, in a trie.

3. **"Definite probability" features.** Suppose X is a noun phrase, Y is the same noun phrase without a determiner, and H is its head. We obtain Internet counts for "Det Y" and "Det H", where Det stays for *"the"*, *"a(n)"*, or the empty string. Then the following ratios are used as features:

$$\frac{\#\text{“}the\ Y\text{”}}{\#Y}, \frac{\#\text{“}the\ H\text{”}}{\#H},$$

$$\frac{\#\text{“}the\ Y\text{”}}{\#\text{“}a\ Y\text{”}}, \frac{\#\text{“}the\ H\text{”}}{\#\text{“}a\ H\text{”}}$$

We expect our NPs to behave w.r.t. the "definite probability" as follows: pronouns and long proper names are seldom used with any article:

	Features	P	R	F
All the entities	All	88.5	84.3	86.3
	Synt+Context	87.9	86	86.9
Definite NPs only	All	84.8	82.3	83.5
	Synt+Context	82.5	79.3	80.8

Table 1: Precision, Recall, and F-score for the $+discourse\_new$ class

Optimisation	Features	P	R	F
Best prec.	All	**95.0**	83.5	88.9
	Synt+Cont.	94.0	84.0	88.7
Best recall	All	87.2	**97.0**	**91.8**
	Synt+Cont.	86.7	96.0	91.1
Best accur.	All	87.8	**96.6**	92.0
	Synt+Cont.	87.7	95.6	91.5

Table 2: Precision, Recall, and F-score for the $-unique$ class

"he" was found on the Web 44681672 times, *"the he"* — 134978 times (0.3%), and *"a he"* — 154204 times (0.3%). Uniques (including short proper names) and plural non-uniques are used with the definite article much more often than with the indefinite one: *"government"* was found 23197407 times, *"the government"* — 5539661 times (23.9%), and *"a government"* — 1109574 times (4.8%). Singular non-unique expressions are used only slightly (if at all) more often with the definite article: *"retailer"* was found 1759272 times, *"the retailer"* — 204551 times (11.6%), and *"a retailer"* — 309392 times (17.6%).

4.2 Discourse New entities

We use Ripper to learn the $\pm discourse\_new$ classification from the feature representations described above. The experiment is designed in the following way: one text is reserved for testing (we do not want to split our texts and always process them as a whole). The remaining 19 texts are first used to optimise Ripper parameters — class ordering, possibility of negative tests, hypothesis simplification, and minimal number of training examples to be covered by a rule. We perform 5-fold cross-validation on these 19 texts in order to find the settings with the best precision for the $+discourse\_new$ class. These settings are then used to train Ripper on all the 19 files and test on the reserved one. The whole procedure is repeated for all the 20 test files and the average precision and recall are calculated. The parameter "Loss Ratio" (ratio of the cost of a false negative to the cost of a false positive) is adjusted separately — we decreased it as much as possible (to 0.3) to have a classification with a good precision and a reasonable recall.

The automatically induced classifier includes, for example, the following rules:

R2: (applicable to such NPs as "you")
IF an NP is a pronoun,
CLASSIFY it as discourse old.

R14: (applicable to such NPs as "Mexico" or "the Shuttle")
IF an NP has no premodifiers,
is more often used with "the" than with "a(n)"
(the ratio is between 2 and 10),
and a same head NP is found within the 18-NPs window,
CLASSIFY it as discourse old.

The performance is shown in table 1.

4.3 Uniquely Referring Expressions

Although the "definite probability" features could not help us much to classify NPs as $\pm discourse\_new$, we expect them to be useful for identifying unique expressions.

We conducted a similar experiment trying to learn a $\pm unique$ classifier. The only difference was in the optimisation strategy: as we did not know a priori, what was more important, we looked for settings with the best precision for non-uniques, recall for non-uniques, and overall accuracy (number of correctly classified items of both classes) separately. The results are summarised in table 2.

4.4 Combining two approaches

Unique and non-unique NPs demonstrate different behaviour w.r.t. the coreference: discourse entities are seldom introduced by vague descriptions and then referred to by fully specifying NPs. Therefore

	P	R	F
Uniques	85.2	68.8	76.1
Non-uniques	90.4	88.9	89.6
All	88.9	84.6	86.7

Table 3: Accuracy of $\pm discourse\_new$ classification for unique and non-unique NPs separately, all the features are used

we can expect a unique NP to be discourse new, if obvious checks for coreference fail. The "obvious checks" include in our case looking for same head expressions and appositive constructions, both of them requiring only constant time.

On the other hand, unique expressions always have the same or similar form: *"The Navy"* can be either discourse new or discourse old. Non-unique NPs, on the contrary, look differently when introducing entities (for example, *"a company"* or *"the company that ..."*) and referring to the previous ones (*"it"* or *"the company"* without postmodifiers). Therefore our syntactic features should be much more helpful when classifying non-uniques as $\pm discourse\_new$.

To investigate this difference we conducted another experiment. We split our data into two parts — $+uniques$ and $-uniques$. Then we learn the $\pm discourse\_new$ classification for both parts separately as described in section 4.2. Finally the rules are combined, producing a classifier for all the NPs. The results are summarised in table 3.

4.5 Discussion

As far as the $\pm discourse\_new$ task is concerned, our system performed slightly, if at all, better with the definite probability features than without them: the improvement in precision (our main criterion) is compensated by the loss in recall. However, when only definite NPs are taken into account, the improvement becomes significant. It's not surprising, as these features bring much more information for definites than for other NPs.

For the $\pm unique$ classification our definite probability features were more important, leading to significantly better results compared to the case when only syntactic and context features were used. Although the improvement is only about 0.5%, it must be taken into account that overall figures are high: 1% improvement on 90% and on 70% accuracy is not the same. We conducted the t-test to check the significance of these improvements, using weighted means and weighted standard deviations, as all the texts have different sizes. Table 2 shows in bold performance measures (precision, recall, or F-score) that improve significantly ($p < 0.05$) when we use the definite probability features.

As our third experiment shows, non-unique entities can be classified very reliably into $\pm discourse\_new$ classes. Uniques, however, have shown quite poor performance, although we expected them to be resolved successfully by heuristics for appositions and same heads. Such a low performance is mainly due to the fact that many objects can be referred to by very similar, but not the same unique NPs: *"Lockheed Martin Corp."*, *"Lockheed Martin"*, and *"Lockheed"*, for example, introduce the same object. We hope to improve the accuracy by developing more sophisticated matching rules for unique descriptions.

Although uniques currently perform poorly, the overall classification still benefits from the sequential processing (identify $\pm uniques$ first, then learn $\pm discourse\_new$ classifiers for uniques and non-uniques separately, and then combine them). And we hope to get a better overall accuracy once our matching rules are improved.

5 Conclusion and Future Work

We have implemented a system for automatic identification of discourse new and unique entities. To learn the classification we use a small training corpus (MUC-7). However, much bigger corpus (the WWW, as indexed by AltaVista) is used to obtain values for some features. Combining heuristics and Internet counts we are able to achieve 88.9% precision and 84.6% recall for discourse new entities.

Our system can also reliably classify NPs as $\pm uniquely\_referring$. The accuracy of this classification is about 89–92% with various precision/recall combinations. The classifier provide useful information for coreference resolution in general, as $+unique$ and $-unique$ descriptions exhibit different behaviour w.r.t. the anaphoricity. This fact is partially reflected by the performance of our se-

quential classifier (table 3): the context information is not sufficient to determine whether a unique NP is a first-mention or not, one has to develop sophisticated names matching techniques instead.

We expect our algorithms to improve both the speed and the performance of the main coreference resolution module: once many NPs are discarded, the system can proceed quicker and make fewer mistakes (for example, almost all the parsing errors were classified by our algorithm as $+discourse\_new$).

Some issues are still open. First, we need sophisticated rules to compare unique expressions. At the present stage our system looks only for full matches and for same head expressions. Thus, *"China and Taiwan"* and *"Taiwan"* (or *"China"*, depending on the rules one uses for coordinates' heads) have much better chances to be considered coreferent, than *"World Trade Organisation"* and *"WTO"*.

We also plan to conduct more experiments on the interaction between the $\pm discourse\_new$ and $\pm unique$ classifications, treating, for example, time expressions as $-unique$, or exploring the influence of various optimisation strategies for $\pm uniques$ on the overall performance of the sequential classifier.

Finally, we still have to estimate the impact of our pre-filtering algorithm on the overall coreference resolution performance. Although we expect the coreference resolution system to benefit from the $\pm discourse\_new$ and $\pm unique$ classifiers, this hypothesis has to be verified.

References

David L. Bean and Ellen Riloff. 1999. Corpus-based Identification of Non-Anaphoric Noun Phrases. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 373–380.

Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*, 132–139.

William W. Cohen. 1995. Fast effective rule induction. *Proceedings of the 12th International Conference on Machine Learning (ICML-95)*, 115–123.

Lynette Hirschman and Nancy Chinchor. 1997. MUC-7 Coreference Task Definition. *Message Understanding Conference Proceedings.*

Frank Keller, Maria Lapata, and Olga Ourioupina. 2002. Using the Web to Overcome Data Sparseness. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, 230–237.

Vincent Ng and Claire Cardie. 2002. Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution. *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING-2002)*, 730–736.

Ellen F. Prince. 1981. Toward a Taxonomy of given-new information. *Radical Pragmatics*, 223–256.

Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–594.