

DOES DISCOURSE-NEW DETECTION HELP DEFINITE DESCRIPTION RESOLUTION?

Massimo Poesio,[†] Mijail A. Kabadjov[†], Renata Vieira,^{*}
Rodrigo Goulart^{*} and Olga Uryupina,[§]

[†]University of Essex, Computer Science and Cognitive Science (UK)

[§]Universität des Saarlandes, Computerlinguistik (Germany)

^{*}Unisinos, Computação Aplicada (Brazil)

Abstract

Recent results by Ng and Cardie (2002a) challenge the idea—advanced, e.g., by Vieira and Poesio (2000)—that including methods for identifying discourse new descriptions in an anaphoric resolver leads to better results. In previous work we analyzed the literature on discourse-new detection that followed Vieira and Poesio’s original work, and identified a set of features that appear to be present in all this work. In this paper, we discuss the results obtained by augmenting an existing anaphora resolution system with a statistical discourse-new detector using these features. Our DN classifier achieves $F=90.6$ on the discourse new classification task; incorporating this classifier into our existing system results in a significant improvement (from $F=79.6$ to $F=82.4$ on hand-annotated text).

1 Introduction

Although many theories of definiteness and many anaphora resolution algorithms are based on the assumption that definite descriptions are anaphoric, in fact in most corpora at least half of definite descriptions are DISCOURSE-NEW (Prince, 1992) like those underlined below (both examples are the first sentences of texts from the Penn Treebank).

- (1) a. Toni Johnson pulls a tape measure across the front of what was once a stately Victorian home.
- b. The Federal Communications Commission allowed American Telephone & Telegraph Co. to continue offering discount phone services

for large-business customers and said it would soon re-examine its regulation of the long-distance market.

Vieira and Poesio (2000) proposed an algorithm for definite description resolution that incorporates a number of heuristics for detecting discourse-new (henceforth: DN) descriptions. But whereas the inclusion of detectors for non-anaphoric pronouns (e.g., *It* in *It's raining*) in algorithms such as Lapin and Leass' (1994) has been shown to result in clear improvements in precision, the improvements to anaphoric DD resolution (as opposed to classification) observed by Vieira and Poesio were rather small. In fact, Ng and Cardie (2002a) challenged the motivation for the inclusion of such detectors, reporting no improvements or even worse performance. Poesio *et al.* (2004b) re-examined the literature on the topic and proposed a new set of features for the task, taking advantage of the improved techniques for DN detection developed by Bean and Riloff (1999) and Uryupina (2003). In this work we report the results of machine learning experiments evaluating whether this set of features would improve the performance of the GUITAR system (Poesio and Alexandrov-Kabadjov, 2004).

2 Detecting Discourse-New Definite Descriptions

2.1 Vieira and Poesio

Poesio and Vieira (1998) carried out corpus studies indicating that in corpora like the Wall Street Journal portion of the Penn Treebank (Marcus *et al.*, 1993), around 52% of DDs are discourse-new (Prince, 1992), and another 15% or so are bridging references, for a total of about 66-67% first-mention. These results led Vieira and Poesio to propose a definite description resolution algorithm incorporating independent heuristic strategies for recognizing DN definite descriptions (Vieira, 1998; Vieira and Poesio, 2000). The heuristics proposed by Vieira and Poesio assumed a parsed input (the Penn Treebank) and aimed at identifying five categories of DDs licensed to occur as first mention on semantic or pragmatic grounds on the basis of work on definiteness including Loebner's account (1987):

1. So-called SEMANTICALLY FUNCTIONAL descriptions (Loebner, 1987). This class included descriptions with modifiers like *first* or *best* that turned a possibly sortal predicate into a function (as in *the first person to cross the Pacific on a row boat*); as well as descriptions with predicates like *fact* or *belief* followed by a *that*-clause with the function of specifying the fact or belief under question.

2. Descriptions serving as disguised PROPER NAMES, such as *The Federal Communications Commission* or *the Iran-Iraq war*.
3. PREDICATIVE descriptions, i.e., descriptions semantically functioning as predicates. These include descriptions occurring in appositive position (as in *Glenn Cox, the president of Phillips Petroleum*) and in certain copular constructions (as in *the man most likely to gain custody of all this is a career politician named Dinkins*).
4. Descriptions ESTABLISHED (i.e., turned into functions in context) by restrictive modification, particularly by establishing relative clauses (Loebner, 1987) and prepositional phrases, as in *The hotel where we stayed last night was pretty good*.
5. LARGER SITUATION definite descriptions (Hawkins, 1978), i.e., definite descriptions like *the sun*, *the pope* or *the long distance market* which denote uniquely on the grounds of shared knowledge about the situation (these are Loebner’s ‘situational functions’). Vieira and Poesio’s system had a small list of such definites.

These heuristics were used as features of decision trees attempting to classify DDs as anaphoric, bridging or discourse new. These decision trees included both DN detection tests and attempts to find an antecedent for such DDs. Both hand-coded decision trees and automatically acquired ones (using ID3, (Quinlan, 1986)) were used for the task of two-way classification into discourse-new and anaphoric. Vieira and Poesio found only small differences between the hand-coded and automatically acquired decision tree, both in the order of the tests and in the performance. The hand-coded decision tree executes in the following order:

1. Try the DN heuristics with the highest accuracy (recognition of some types of semantically functional DDs using special predicates, and of potentially predicative DDs occurring in appositions);
2. Otherwise, attempt to resolve the DD as direct anaphora;
3. Otherwise, attempt the remaining DN heuristics in the order: proper names, descriptions established by relatives and PPs, proper name modification, predicative DDs occurring in copular constructions.

If none of these tests succeeds, the algorithm can either leave the DD unclassified, or classify it as DN. The automatically learned decision tree attempts

direct anaphora resolution first. The overall results on the 195 DDs on which the automatically trained decision tree was tested are shown in Table 1. The baseline is the result achieved by classifying every DD as discourse-new—with 99 discourse-new DDs out of 195, this means a precision of 50.8%. Two results are shown for the hand-coded decision tree: in one version, the system doesn’t attempt to classify all DDs; in the other, all unclassified DDs are classified as discourse-new. This is the best version of the system.

Version of the System	P	R	F
Baseline	50.8	100	67.4
Discourse-new detection only	69	72	70
Hand-coded DT: partial	62	85	71.7
Hand-coded DT: total	77	77	77
ID3	75	75	75

Table 1: Summary of the results obtained by Vieira and Poesio

2.2 Bean and Riloff

Bean and Riloff (1999) developed a system for identifying discourse-new DDs¹ that incorporates, in addition to syntax-based heuristics aimed at recognizing predicative and established DDs using postmodification heuristics similar to those used by Vieira and Poesio, additional techniques for mining from corpora unfamiliar DDs including proper names, larger situation, and semantically functional. Two of the techniques proposed by Bean and Riloff are particularly worth noticing. The SENTENCE-ONE (S1) EXTRACTION heuristic identifies as discourse-new every DD found in the first sentence of a text. More general patterns can then be extracted from the DDs initially found by s1-extraction, using the EXISTENTIAL HEAD PATTERN method which, e.g., would extract **the N+ Government** from *the Salvadoran Government* and *the Guatemalan Government*. The DEFINITE ONLY (DO) list contained NPs like *the National Guard* or *the FBI* with a high DEFINITE PROBABILITY, i.e., whose nominal complex has been encountered at least 5 times with the definite article, but never with the indefinite. In the algorithm proposed by Bean and Riloff, as in the machine-learned decision tree obtained by Vieira and Poesio, a (simplified) direct anaphora test is tried first, followed by DN detectors in decreasing order of accuracy.

Bean and Riloff trained their system on 1600 articles from MUC-4, and tested it on 50 texts. The s1 extraction methods produced 849 DDs; the DO list contained 65 head nouns and 321 full NPs. The overall results are shown in Table 2; the baseline are the results obtained when classifying all

¹Bean and Riloff use the term EXISTENTIAL for these DDs.

DDS as discourse-new. Results on different corpora should not be compared

Method	R	P
Baseline	100	72.2
Syntactic Heuristics	43	93.1
Synt. Heuristics + S1	66.3	84.3
Synt. Heuristics + EHP	60.7	87.3
Synt. Heuristics + DO	69.2	83.9
Synt. Heuristics + S1 + EHP + DO	81.7	82.2
Synt. Heuristics + S1 + EHP + DO + V	79.1	84.5

Table 2: Discourse-new prediction results by Bean and Riloff

directly; we will nevertheless observe that the Bean / Riloff system achieves a precision comparable to that obtained with the partial hand-coded decision tree used by Vieira and Poesio, but a much better recall.

2.3 Ng and Cardie

Ng and Cardie (2002a) directly investigate the question of whether employing a discourse-new prediction component improves the performance of a coreference resolution system (specifically, the system discussed in (Ng and Cardie, 2002b)). Ng and Cardie’s work differs from the work discussed so far in that their system attempts to deal with all types of NPs, not just definite descriptions. The discourse-new detectors proposed by Ng and Cardie are statistical classifiers taking as input 37 features and trained using either C4.5 (Quinlan, 1993) or RIPPER (Cohen, 1995). The 37 features of a candidate anaphoric expression specify much, although not all, of the information proposed in previous work. Ng and Cardie’s discourse-new predictor was trained and tested over the MUC-6 and MUC-7 coreference data sets, achieving accuracies of 86.1% and 84%, respectively, against a baseline of 63.8% and 73.2%, respectively. Inspection of the top parts of the decision tree produced with the MUC-6 suggests that `head_match` is the most important feature, followed by the features specifying NP type, the `alias` feature, and the features specifying the structure of definite descriptions.

Ng and Cardie discuss two architectures for the integration of a DN detector in a coreference system. In the first architecture, the DN detector is run first, and the coreference resolution algorithm is run only if the DN detector classifies that NP as anaphoric. In the second architecture, the system first computes `str_match` and `alias`, and runs the anaphoric resolver if any of them is Y; otherwise, it proceeds as in the first architecture. The results obtained on the MUC-6 data with the baseline anaphoric resolver, the

anaphoric resolver augmented by a DN detector as in the first architecture, and as in the second architecture (using C4.5), are shown in Table 3. The results for all NPs, pronouns only, proper names only, and common nouns only are shown.²

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline (no DN detector)	70.3	58.3	63.8	65.5	58.2	61.6
Pronouns	17.9	66.3	28.2	10.2	62.1	17.6
Proper names	29.9	84.2	44.1	27.0	77.7	40.0
Common nouns	25.2	40.1	31.0	26.6	45.2	33.5
DN detector runs first	57.4	71.6	63.7	47.0	77.1	58.4
Pronouns	17.9	67.0	28.2	10.2	62.1	17.6
Proper names	26.6	89.2	41.0	21.5	84.8	34.3
Common nouns	15.4	56.2	24.2	13.8	77.5	23.4
Same head runs first	63.4	68.3	65.8	59.7	69.3	64.2
Pronouns	17.9	67.0	28.2	10.2	62.1	17.6
Proper names	27.4	88.5	41.9	26.1	84.7	40.0
Common nouns	20.5	53.1	29.6	21.7	59.0	31.7

Table 3: Three anaphoric resolvers evaluated by Ng and Cardie.

As indicated in Table 3, running the DN detector first leads to worse results; this is because the detector misclassifies a number of anaphoric NPs as non-anaphoric. However, looking first for a same-head antecedent leads to a statistically significant improvement over the results of the baseline anaphoric resolver. This confirms the finding both of Vieira and Poesio and of Bean and Riloff that the direct anaphora should be called very early.

2.4 Uryupina

Uryupina (2003) trained two separate classifiers (using RIPPER, (Cohen, 1995)): a DN detector and a UNIQUENESS DETECTOR, i.e., a classifier that determines whether an NP refers to a unique object. This is useful to identify proper names (like *1998*, or *the United States of America*), semantic definites (like *the chairman of Microsoft*) and larger situation definite descriptions (like *the pope*). Both classifiers use the same set of 32 features. The features of an NP encode, first, of all, string-level information: e.g., whether the NP contains capitalized words, digits, or special symbols. A second group of features specifies syntactic information: whether the NP is postmodified, and whether it contains an apposition. Two types of appositions are distinguished, with and without commas. CONTEXT features

²It’s not clear to us why the overall performance of the algorithm is much better than the performance on the three individual types of anaphoric expressions considered—i.e., which other anaphoric expressions are handled by the coreference resolver.

specify the distance between the NP and the previous NP with the same head, if any. Finally, Uryupina’s system computes four features specifying the NP’s definite probability. Unlike the definite probability used by Bean and Riloff, these features are computed from the Web, using Altavista. From each NP, its head H and entire NP without determiner Y are determined, and four ratios are then computed:

$$\frac{\# \text{ "the Y" }}{\# Y}, \frac{\# \text{ "the Y" }}{\# \text{ "aY" }}, \frac{\# \text{ "the H" }}{\# H}, \frac{\# \text{ "the H" }}{\# \text{ "aH" }}.$$

The classifiers were tested on 20 texts from MUC-7 (a subset of the second data set used by Ng and Cardie), parsed by Charniak’s parser. 19 texts were used for training and for tuning RIPPER’s parameters, one for testing. The results for the discourse new detection task are shown in Table 4, separating the results for all NPs and definite NPs only, and the results without definite probabilities and including them.

	Features	P	R	F
All NPs	String+Syn+Context	87.9	86.0	86.9
	All	88.5	84.3	86.3
Def NPs	String+Syn+Context	82.5	79.3	80.8
	All	84.8	82.3	83.5

Table 4: Results of Uryupina’s discourse new classifier

The result to note is that both of Uryupina’s classifiers work very well. Her results also show that the definite probability helps somewhat the discourse new detector, but is especially useful for the uniqueness detector, as one would expect on the basis of Loebner’s discussion.

3 A Preliminary Test

Vieira and Poesio did not test their system without DN-detection, but Ng and Cardie’s results indicate that DN detection does improve results, if not dramatically, provided that the `same_head` test is run first—although their DN detector does not improve results for pronouns, the one category for which detection of non-anaphoricity has been shown to be essential (Lappin and Leass, 1994). In order to evaluate how much improvement can we expect by just improving the DN detector, we did a preliminary evaluation with a reimplementaion of Vieira and Poesio’s algorithm which does not include a discourse-new detector, running over treebank text as the original algorithm.

GUITAR (Poesio and Alexandrov-Kabadjov, 2004) is a general-purpose anaphoric resolver that includes an implementation of the Vieira / Poesio

algorithm for definite descriptions and of Mitkov’s algorithm for pronoun resolution (Mitkov, 1998). It is implemented in Java, takes its input in XML format and returns as output its input augmented with the anaphoric relations it has discovered. GUITAR has been implemented in such a way as to be fully *modular*, making it possible, for example, to test alternative DD resolution methods. It includes a pre-processor incorporating a chunker so that it can run over both hand-parsed and raw text.

A version of GUITAR without DN detection was evaluated on the GNOME corpus (Poesio, 2000; Poesio *et al.*, 2004a), which contains 623 definite descriptions, 195 of which are anaphoric. The results obtained by GUITAR running over hand-annotated text are shown in Table 5.³

Total	Res	Corr	NM	WM	SM	R	P	F
574	574	457	38	27	52	79.6	79.6	79.6

Table 5: Evaluation of GUITAR without DN detection.

GUITAR without a DN recognizer takes 198 DDs (Res) to be anaphoric. Of these, 119 are resolved correctly; overall, 457 DDs (**Corr** column) are correctly classified and resolved. Failure to identify discourse-new DDs is the main problem: of the 198 DDs GUITAR attempts to resolve, only 27 are incorrectly resolved (WM)—almost twice that number (52) are Spurious Matches (SM), i.e., discourse-new DDs incorrectly interpreted as anaphoric. The system can’t find an antecedent for 38 of the 184 anaphoric DDs (NM)—the antecedent may be out of segment, or the match may require lexical / commonsense knowledge. Precision and recall overall are quite high (F=79.6), which makes this performance hard to improve upon, although on anaphoric DDs the results are not quite as good (P=60.8, R=64.6, F=62.4). When endowed with a perfect DN detector—i.e., if the SM column could be reduced to 0—GUITAR could achieve P=R=88.7.

Of course, these results are obtained assuming perfect parsing. For a fairer comparison with the results of Ng and Cardie we report in Table 6 the results for both pronouns and definite descriptions obtained by running GUITAR off raw text (anaphoric NPs only), although such comparisons across corpora are of course only indicative. The two aspects to notice are that the performance deterioration for DDs is not that bad (F=56.4, as opposed to F=62.4); and that the system without DN detector achieves a much higher performance on DDs in this corpus than Ng and Cardie’s system without a DN detector achieved on theirs (F=33.5 on common nouns), from which we

³For comparison purposes, the results over the subset of 574 DDs used in the subsequent experiments are displayed. Both anaphoric and discourse-new descriptions are included; a DD that the system does not resolve is considered classified as a DN.

can conclude that improving the performance of our system will be harder.

	R	P	F
Pronouns	65.5	63.0	64.2
DDs	56.7	56.1	56.4

Table 6: Evaluation of the GUITAR system without DN detection off raw text

4 A New Discourse-New Detector

The results of the discussion above led to the development of a discourse new detector operating in tandem with GUITAR as follows.

Architecture This detector works in two steps. For each DD,

1. The direct anaphora resolution algorithm from (Vieira and Poesio, 2000) is run, as implemented in GUITAR. This algorithm attempts to find an head-matching antecedent, taking into account premodification and segmentation. The results of GUITAR (i.e., whether an antecedent was found) is used as one of the input features in step 2.
2. A classifier is used to classify the DD as anaphoric (in which case the antecedents identified in 1. are also returned) or discourse-new.

Features The classifier uses features attempting to capture five types of information (all the features are normalized in the range [-1,1]):

Anaphora A single feature, `direct-anaphora`, specifying the distance of the (same-head) antecedent from the DD, if any (values: `none`, `zero`, `one`, `more`)

Predicative NPs Two boolean features:

- `apposition`, if the DD occurs in apposition; position;
- `copular`, if the DD occurs in a copular construction.

Proper Names Three boolean features:

- `c-head`: whether the head is capitalized;
- `c-premod`: whether one of the premodifiers is capitalized;
- `S1`: whether the DD occurs in the first sentence of a Web page.

Functionality The four definite probabilities used by Uryupina (computed accessing the Web), plus a **superlative** feature specifying if one of the premodifiers is a superlative (info extracted from POS tags).

Establishing relative A single feature, specifying whether NP is postmodified, and by a relative clause or a prepositional phrase;

Text Position Whether the DD occurs in the title, the first sentence, or the first paragraph.

5 Evaluation

We carried out a series of experiments to evaluate the DN detector proposed above (only preliminary results were discussed in (Poesio *et al.*, 2004b)).

Data The plan is to use three corpora for the evaluation, including texts from different genres. The GNOME corpus includes pharmaceutical leaflets and museum 'labels' (i.e., descriptions of museum objects and of the artists that realized them). As said above, the corpus contains 623 definite descriptions. The Vieira and Poesio corpus (henceforth, VPC) used in (Vieira and Poesio, 2000) consists of 14 news articles from the Penn Treebank, containing a total of 1400 DDs. The corpus has now been converted to XML, and anaphoric information for all types of NPs according to the GNOME scheme has been added, but the annotation is still being revised. Finally, we plan to test the system on the MUC-7 data used by Ng and Cardie and by Uryupina, which contains about 3,000 anaphoric expressions in total (not all DDs), but whose annotation is in many senses problematic.⁴ For the moment only the experiments with the GNOME corpus have been completed.

Machine learning algorithms We tested several learning algorithms included in the Weka 3.4 library (<http://www.cs.waikato.ac.nz/~ml/>) including Weka's implementation of the decision tree classifier C4.5, of the multi-layer perceptron (MLP) algorithm, and of a Support Vector Machine.

Results We left aside 8% of the corpus for parameter tuning and ran a 10-fold cross validation over the rest of the corpus. The best overall results for DN classification were obtained by training it with the multi-layer

⁴E.g., in copular phrases like *Higgins was the president of DD*, *Higgins* would be marked as coreferential with *the president of DD*.

perceptron, default configuration. Our classifier achieves an accuracy of 85.8, and F=90.2 for discourse-new detection (against a baseline of F=67.9 when every DD is classified as discourse-new, the majority). The results with and without DN-detection are shown in Table 7. The improvement with DN detection is significant at the .1 level ($t=1.74, \geq 1.34$) but not quite at the .05 level (for which a value greater than 1.83 would be required)—however, the performance is getting pretty close to the upper bound. With a decision tree classifier, F=81.2 is obtained, slightly worse than with the MLP and slightly better than without DN classification, but not significantly so.

Version of the System	P	R	F
Without DN detection	79.6	79.6	79.6
With DN detection	82.4	82.4	82.4

Table 7: Anaphora resolution with and without DN detection

Error Analysis The most important features used by all decision trees are the result of anaphora resolution (the system never attempts to classify a DD as anaphoric unless this is suggested by the direct anaphora feature), the presence of relative postmodification, and the values of the definite probabilities. Examples of ‘spurious matches’—DDs incorrectly classified as anaphoric—dixed by the DN detector include *the title of cabinet maker and sculptor to Louis XIV, King of France* and *The other half of the plastic*.

6 Discussion and Conclusions

These preliminary results suggest that DN detection leads to clear improvements, at least with parsed text. In subsequent work we will test our DN classifiers on raw text, as well as on the texts used by Ng and Cardie.

Acknowledgments

Mijail A. Kabadjov is supported by Conacyt. Renata Vieira and Rodrigo Goulart are partially supported by CNPq.

References

- Bean, D. L. and Riloff, E. (1999). Corpus-based identification of non-anaphoric noun phrases. In *Proc. of the 37th ACL*, pages 373–380, University of Maryland. ACL.
- Cohen, W. (1995). Fast effective rule induction. In *Proc. of ICML*.

- Hawkins, J. A. (1978). *Definiteness and Indefiniteness*. Croom Helm, London.
- Lappin, S. and Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**(4), 535–562.
- Loebner, S. (1987). Definites. *Journal of Semantics*, **4**, 279–326.
- Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.
- Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. In *Proc. of the 18th COLING*, pages 869–875, Montreal.
- Ng, V. and Cardie, C. (2002a). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proc. of 19th COLING*.
- Ng, V. and Cardie, C. (2002b). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Meeting of the ACL*.
- Poesio, M. (2000). Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In *Proc. of the 2nd LREC*, pages 211–218, Athens.
- Poesio, M. and Alexandrov-Kabadjov, M. (2004). A general-purpose, off the shelf anaphoric resolver. In *Proc. of LREC*, Lisbon.
- Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, **24**(2), 183–216. Also available as Research Paper CCS-RP-71, Centre for Cognitive Science, University of Edinburgh.
- Poesio, M., Stevenson, R., Di Eugenio, B., and Hitzeman, J. M. (2004a). Centering: A parametric theory and its instantiations. *Computational Linguistics*, **30**(3), 309–363.
- Poesio, M., Uryupina, O., Vieira, R., Alexandrov-Kabadjov, M., and Goulart, R. (2004b). Discourse-new detectors for definite description resolution: A survey. In *Proc. of ACL Workshop on Reference Resolution*, Barcelona.
- Prince, E. F. (1992). The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund-raising text*, pages 295–325. John Benjamins.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, **1**(1), 81–106.
- Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann, San Mateo, CA.
- Uryupina, O. (2003). High-precision identification of discourse-new and unique noun phrases. In *Proc. of the ACL 2003 Student Workshop*, pages 80–86.
- Vieira, R. (1998). *Definite Description Resolution in Unrestricted Texts*. Ph.D. thesis, University of Edinburgh, Centre for Cognitive Science.
- Vieira, R. and Poesio, M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, **26**(4).