In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), pp. 373-380

Corpus-Based Identification of Non-Anaphoric Noun Phrases

David L. Bean and **Ellen Riloff**
Department of Computer Science
University of Utah
Salt Lake City, Utah 84112
{bean,riloff}@cs.utah.edu

Abstract

Coreference resolution involves finding antecedents for anaphoric discourse entities, such as definite noun phrases. But many definite noun phrases are not anaphoric because their meaning can be understood from general world knowledge (e.g., "the White House" or "the news media"). We have developed a corpus-based algorithm for automatically identifying definite noun phrases that are non-anaphoric, which has the potential to improve the efficiency and accuracy of coreference resolution systems. Our algorithm generates lists of non-anaphoric noun phrases and noun phrase patterns from a training corpus and uses them to recognize non-anaphoric noun phrases in new texts. Using 1600 MUC-4 terrorism news articles as the training corpus, our approach achieved 78% recall and 87% precision at identifying such noun phrases in 50 test documents.

1 Introduction

Most automated approaches to coreference resolution attempt to locate an antecedent for every potentially coreferent discourse entity (DE) in a text. The problem with this approach is that a large number of DE's may not have antecedents. While some discourse entities such as pronouns are almost always referential, definite descriptions[1] may not be. Earlier work found that nearly 50% of definite descriptions had no prior referents (Vieira and Poesio, 1997), and we found that number to be even higher, 63%, in our corpus. Some non-anaphoric definite descriptions can be identified by looking for syntactic clues like attached prepositional phrases or restrictive relative clauses. But other definite descriptions are non-anaphoric because readers understand their meaning due to common knowledge. For example, readers of this paper will probably understand the real world referents of "the F.B.I.," "the White House," and "the Golden Gate Bridge." These are instances of definite descriptions that a coreference resolver does not need to resolve because they each fully specify a cognitive representation of the entity in the reader's mind.

One way to address this problem is to create a list of all non-anaphoric NPs that could be used as a filter prior to coreference resolution, but hand coding such a list is a daunting and intractable task. We propose a corpus-based mechanism to identify non-anaphoric NPs automatically. We will refer to non-anaphoric definite noun phrases as *existential* NPs (Allen, 1995). Our algorithm uses statistical methods to generate lists of existential noun phrases and noun phrase patterns from a training corpus. These lists are then used to recognize existential NPs in new texts.

2 Prior Research

Computational coreference resolvers fall into two categories: systems that make no attempt to identify non-anaphoric discourse entities prior to coreference resolution, and those that apply a filter to discourse entities, identifying a subset of them that are anaphoric. Those that do not practice filtering include decision tree models (Aone and Bennett, 1996), (McCarthy and Lehnert, 1995) that consider all possible combinations of potential anaphora and referents. Exhaustively examining all possible combinations is expensive and, we believe, unnecessary.

Of those systems that apply filtering prior to coreference resolution, the nature of the filtering varies. Some systems recognize when an anaphor and a candidate antecedent are incompatible. In SRI's probabilistic model (Kehler,

[1] In this work, we define a definite description to be a noun phrase beginning with *the*.

The ARCE battalion command has reported that about 50 peasants of various ages have been kidnapped by terrorists of **the Farabundo Marti National Liberation Front** [**FMLN**] in San Miguel Department. According to that garrison, **the mass kidnapping** took place on 30 December in San Luis de la Reina. **The source** added that **the terrorists** forced **the individuals**, who were taken to an unknown location, out of their residences, presumably to incorporate them against their will into clandestine groups.

Figure 1: Anaphoric and Non-Anaphoric NPs (definite descriptions highlighted.)

1997), a pair of extracted templates may be removed from consideration because an outside knowledge base indicates contradictory features. Other systems look for particular constructions using certain trigger words. For example, pleonastic[2] pronouns are identified by looking for modal adjectives (e.g. "necessary") or cognitive verbs (e.g. "It is thought that...") in a set of patterned constructions (Lappin and Leass, 1994), (Kennedy and Boguraev, 1996).

A more recent system (Vieira and Poesio, 1997) recognizes a large percentage of non-anaphoric definite noun phrases (NPs) during the coreference resolution process through the use of syntactic cues and case-sensitive rules. These methods were successful in many instances, but they could not identify them all. The existential NPs that were missed were existential to the reader, not because they were modified by particular syntactic constructions, but because they were part of the reader's general world knowledge.

Definite noun phrases that do not need to be resolved because they are understood through world knowledge can represent a significant portion of the existential noun phrases in a text. In our research, we found that existential NPs account for 63% of all definite NPs, and 24% of them could not be identified by syntactic or lexical means. This paper details our method for identifying existential NPs that are understood through general world knowledge. Our system requires no hand coded information and can recognize a larger portion of existential NPs than Vieira and Poesio's system.

3 Definite NP Taxonomy

To better understand what makes an NP anaphoric or non-anaphoric, we found it useful to classify definite NPs into a taxonomy. We first classified definite NPs into two broad categories, referential NPs, which have prior referents in the texts, and existential NPs, which do not. In Figure 1, examples of referential NPs are **"the mass kidnapping," "the terrorists"** and **"the individuals."**, while examples of existential NPs are **"the ARCE battalion command"** and **"the Farabundo Marti National Liberation Front."** (The full taxonomy can be found in Figure 2.)

We should clarify an important point. When we say that a definite NP is existential, we say this because it completely specifies a cognitive representation of the entity in the reader's mind. That is, suppose "the F.B.I." appears in both sentence 1 and sentence 7 of a text. Although there may be a cohesive relationship between the noun phrases, because they both completely specify independently, we consider them to be non-anaphoric.

- Definite Noun Phrases
  - Referential
  - Existential
    - Independent
      - Syntactic
      - Semantic
    - Associative

Figure 2: Definite NP Taxonomy

We further classified existential NPs into two categories, independent and associative, which are distinguished by their need for context. Independent existentials can be understood in isolation. Associative existentials are inherently associated with an event, action, object or other context[3]. In a text about a basketball game, for example, we might find "the score," "the hoop" and "the bleachers." Although they may

[2]Pronouns that are semantically empty, e.g. "It is clear that...."

[3]Our taxonomy mimics Prince's (Prince, 1981) in that our independent existentials roughly equate to her *new* class, our associative existentials to her *inferable* class, and our referentials to her *evoked* class.

not have direct antecedents in the text, we understand what they mean because they are all associated with basketball games. In isolation, a reader would not necessarily understand the meaning of "the score" because context is needed to disambiguate the intended word sense and provide a complete specification.

Because associative NPs represent less than 10% of the existential NPs in our corpus, our efforts were directed at automatically identifying independent existentials. Understanding how to identify independent existential NPs requires that we have an understanding of why these NPs are existential. We classified independent existentials into two groups, semantic and syntactic. Semantically independent NPs are existential because they are understood by readers who share a collective understanding of current events and world knowledge. For example, we understand the meaning of "the F.B.I." without needing any other information. Syntactically independent NPs, on the other hand, gain this quality because they are modified structurally. For example, in "the man who shot Liberty Valence," "the man" is existential because the relative clause uniquely identifies its referent.

4 Mining Existential NPs from a Corpus

Our goal is to build a system that can identify independent existential noun phrases automatically. In the previous section, we observed that "existentialism" can be granted to a definite noun phrase either through syntax or semantics. In this section, we introduce four methods for recognizing both classes of existentials.

4.1 Syntactic Heuristics

We began by building a set of syntactic heuristics that look for the structural cues of restrictive premodification and restrictive postmodification. Restrictive premodification is often found in noun phrases in which a proper noun is used as a modifier for a head noun, for example, "the U.S. president." "The president" itself is ambiguous, but "the U.S. president" is not. Restrictive postmodification is often represented by restrictive relative clauses, prepositional phrases, and appositives. For example, "the president of the United States" and "the president who governs the U.S." are existential due to a prepositional phrase and a relative clause, respectively.

We also developed syntactic heuristics to recognize referential NPs. Most NPs of the form "the <number> <noun>" (e.g., "the 12 men") have an antecedent, so we classified them as referential. Also, if the head noun of the NP appeared earlier in the text, we classified the NP as referential.

This method, then, consists of two groups of syntactic heuristics. The first group, which we refer to as the rule-in heuristics, contains seven heuristics that identify restrictive premodification or postmodification, thus targeting existential NPs. The second group, referred to as the rule-out heuristics, contains two heuristics that identify referential NPs.

4.2 Sentence One Extractions (S1)

Most referential NPs have antecedents that precede them in the text. This observation is the basis of our first method for identifying semantically independent NPs. If a definite NP occurs in the first sentence[4] of a text, we assume the NP is existential. Using a training corpus, we create a list of presumably existential NPs by collecting the first sentence of every text and extracting all definite NPs that were not classified by the syntactic heuristics. We call this list the S1 extractions.

4.3 Existential Head Patterns (EHP)

While examining the S1 extractions, we found many similar NPs, for example "the Salvadoran Government," "the Guatemalan Government," and "the U.S. Government." The similarities indicate that some head nouns, when premodified, represent existential entities. By using the S1 extractions as input to a pattern generation algorithm, we built a set of Existential Head Patterns (EHPs) that identify such constructions. These patterns are of the form "the <x+>[5] <noun1 ...nounN>" such as "the <x+> government" or "the <x+> Salvadoran government." Figure 3 shows the algorithm for creating EHPs.

[4]Many of the texts we used were newspaper articles and all headers, including titles and bylines, were stripped before processing.

[5]<x+> = one or more words

1. For each NP of more than two words, build a candidate pattern of the form "the <x+> headnoun." Example: if the NP was "the new Salvadoran government," the candidate pattern would be "the <x+> government."
2. Apply that pattern to the corpus, count how many times it matches an NP.
3. If possible, grow the candidate pattern by inserting the word to the left of the headnoun, e.g. the candidate pattern now becomes "the <x+> Salvadoran government."
4. Reapply the pattern to the corpus, count how many times it matches an NP. If the new count is less that the last iteration's count, stop and return the prior pattern. If the new count is equal to the last iteration's count, return to step 3. This iterative process has the effect of recognizing compound head nouns.

Figure 3: EHP Algorithm

```
If the NP was identified via the S1 or EHP methods:
  Is its definite probability above an upper threshold?
    Yes: Classify as existential.
    No: Is its definite probability above a lower threshold?
      Yes: Is its sentence-number less than or equal to an early allowance threshold?
        Yes : Classify as existential.
        No : Leave unclassified (allow later methods to apply).
      No : Leave unclassified (allow later methods to apply).
```

Figure 4: Vaccine Algorithm

4.4 Definite-Only List (DO)

It also became clear that some existentials never appear in indefinite constructions. "The F.B.I.," "the contrary," "the National Guard" are definite NPs which are rarely, if ever, seen in indefinite constructions. The chances that a reader will encounter "an F.B.I." are slim to none. These NPs appeared to be perfect candidates for a corpus-based approach. To locate "definite-only" NPs we made two passes over the corpus. The first pass produced a list of every definite NP and its frequency. The second pass counted indefinite uses of all NPs cataloged during the first pass. Knowing how often an NP was used in definite and indefinite constructions allowed us to sort the NPs, first by the probability of being used as a definite (its *definite probability*), and second by definite-use frequency. For example, "the contrary" appeared high on this list because its head noun occurred 15 times in the training corpus, and every time it was in a definite construction. From this, we created a definite-only list by selecting those NPs which occurred at least 5 times and only in definite constructions.

Examples from the three methods can be found in the Appendix.

4.5 Vaccine

Our methods for identifying existential NPs are all heuristic-based and therefore can be incorrect in certain situations. We identified two types of common errors.

1. An incorrect S1 assumption. When the S1 assumption fails, i.e. when a definite NP in the first sentence of a text is truly referential, the referential NP is added to the S1 list. Later, an Existential Head Pattern may be built from this NP. In this way, a single misclassified NP may cause multiple noun phrases to be misclassified in new texts, acting as an "infection" (Roark and Charniak, 1998).
2. Occasional existentialism. Sometimes an NP is existential in one text but referential in another. For example, "the guerrillas" often refers to a set of counter-government forces that the reader of an El Salvadoran newspaper would understand. In some cases, however, a particular group of guerrillas was mentioned previously in the text ("A group of FMLN rebels attacked the capital..."), and later references to "the guerrillas" referred to this group.

To address these problems, we developed a *vaccine*. It was clear that we had a number of infections in our S1 list, including "the base," "the

For every definite NP in a text

1. Apply syntactic RuleOutHeuristics, if any fired, classify the NP as referential.
2. Look up the NP in the S1 list, if found, classify the NP as existential (unless stopped by vaccine).
3. Look up the NP in the DO list, if found, classify the NP as existential.
4. Apply all EHPs, if any apply, classify the NP as existential (unless stopped by vaccine).
5. Apply syntactic RuleInHeuristics, if any fired, classify the NP as existential.
6. If the NP is not yet classified, classify the NP as referential.

Figure 5: Existential Identification Algorithm

individuals," "the attack," and "the banks." We noticed, however, that many of these incorrect NPs also appeared near the bottom of our definite/indefinite list, indicating that they were often seen in indefinite constructions. We used the definite probability measure as a way of detecting errors in the S1 and EHP lists. If the definite probability of an NP was above an upper threshold, the NP was allowed to be classified as existential. If the definite probability of an NP fell below a lower threshold, it was not allowed to be classified by the S1 or EHP method. Those NPs that fell between the two thresholds were considered occasionally existential.

Occasionally existential NPs were handled by observing where the NPs first occurred in the text. For example, if the first use of "the guerrillas" was in the first few sentences of a text, it was usually an existential use. If the first use was later, it was usually a referential use because a prior definition appeared in earlier sentences. We applied an early allowance threshold of three sentences – occasionally existential NPs occuring under this threshold were classified as existential, and those that occurred above were left unclassified. Figure 4 details the vaccine's algorithm.

5 Algorithm & Training

We trained and tested our methods on the Latin American newswire articles from MUC-4 (MUC-4 Proceedings, 1992). The training set contained 1,600 texts and the test set contained 50 texts. All texts were first parsed by SUNDANCE, our heuristic-based partial parser developed at the University of Utah.

We generated the S1 extractions by processing the first sentence of all training texts. This produced 849 definite NPs. Using these NPs as input to the existential head pattern algorithm, we generated 297 EHPs. The DO list was built by using only those NPs which appeared at least 5 times in the corpus and 100% of the time as definites. We generated the DO list in two iterations, once for head nouns alone and once for full NPs, resulting in a list of 65 head nouns and 321 full NPs[6].

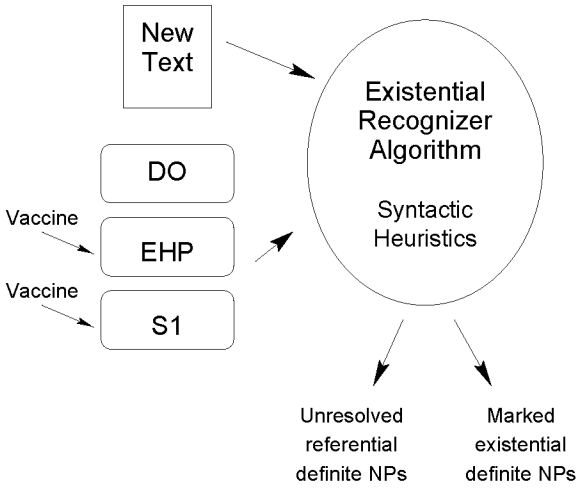

Figure 6: Recognizing Existential NPs

Once the methods had been trained, we classified each definite NP in the test set as referential or existential using the algorithm in Figure 5. Figure 6 graphically represents the main elements of the algorithm. Note that we applied vaccines to the S1 and EHP lists, but not to the DO list because gaining entry to the DO list is much more difficult — an NP must occur at least 5 times in the training corpus, and every time it must occur in a definite construction.

[6]The full NP list showed best performance using parameters of 5 and 75%, not the 5 and 100% used to create the head noun only list.

Method Tested	Recall	Precision
0. Baseline	100%	72.2%
1. Syntactic Heuristics	43.0%	93.1%
2. Syntactic Heuristics + S1	66.3%	84.3%
3. Syntactic Heuristics + EHP	60.7%	87.3%
4. Syntactic Heuristics + DO	69.2%	83.9%
5. Syntactic Heuristics + S1 + EHP	79.9%	82.2%
6. Syntactic Heuristics + S1 + EHP + DO	81.7%	82.2%
7. Syntactic Heuristics + S1 + EHP + DO + $V_a(70/25)$	77.7%	86.6%
8. Syntactic Heuristics + S1 + EHP + DO + $V_b(50/25)$	79.1%	84.5%

Figure 7: Evaluation Results

To evaluate the performance of our algorithm, we hand-tagged each definite NP in the 50 test texts as a syntactically independent existential, a semantically independent existential, an associative existential or a referential NP. Figure 8 shows the distribution of definite NP types in the test texts. Of the 1,001 definite NPs tested, 63% were independent existentials, so removing these NPs from the coreference resolution process could have substantial savings. We measured the accuracy of our classifications using recall and precision metrics. Results are shown in Figure 7.

478	Independent existential, syntactic	48%
153	Independent existential, semantic	15%
92	Associative existential	9%
270	Referential	28%
1001	Total	

Figure 8: NP Distribution

As a baseline measurement, we considered the accuracy of classifying every definite NP as existential. Given the distribution of definite NP types in our test set, this would result in recall of 100% and precision of 72%. Note that we are more interested in high measures of precision than recall because we view this method to be the precursor to a coreference resolution algorithm. Incorrectly removing an anaphoric NP means that the coreference resolver would never have a chance to resolve it, on the other hand, non-anaphoric NPs that slip through can still be ruled as non-anaphoric by the coreference resolver.

We first evaluated our system using only the syntactic heuristics, which produced only 43% recall, but 92% precision. Although the syntactic heuristics are a reliable way to identify existential definite NPs, they miss 57% of the true existentials.

6 Evaluation

We expected the S1, EHP, and DO methods to increase coverage. First, we evaluated each method independently (on top of the syntactic heuristics). The results appear in rows 2-4 of Figure 7. Each method increased recall to between 61-69%, but decreased precision to 84-87%. All of these methods produced a substantial gain in recall at some cost in precision.

Next, we tried combining the methods to make sure that they were not identifying exactly the same set of existential NPs. When we combined the S1 and EHP heuristics, recall increased to 80% with precision dropping only slightly to 82%. When we combined all three methods (S1, EHP, and DO), recall increased to 82% without any corresponding loss of precision. These experiments show that these heuristics substantially increase recall and are identifying different sets of existential NPs.

Finally, we tested our vaccine algorithm to see if it could increase precision without sacrificing much recall. We experimented with two variations: V_a used an upper definite probability threshold of 70% and V_b used an upper definite probability threshold of 50%. Both variations used a lower definite probability threshold of 25%. The results are shown in rows 7-8 of Figure 7. Both vaccine variations increased precision by several percentage points with only a slight drop in recall.

In previous work, the system developed by Vieria & Poesio achieved 74% recall and 85% precision for identifying "larger situation and unfamiliar use" NPs. This set of NPs does not correspond exactly to our definition of existential NPs because we consider associative NPs

to be existential and they do not. Even so, our results are slightly better than their previous results. A more equitable comparison is to measure our system's performance on only the independent existential noun phrases. Using this measure, our algorithm achieved 81.8% recall with 85.6% precision using V_a, and achieved 82.9% recall with 83.5% precision using V_b.

7 Conclusions

We have developed several methods for automatically identifying existential noun phrases using a training corpus. It accomplishes this task with recall and precision measurements that exceed those of the earlier Vieira & Poesio system, while not exploiting full parse trees, appositive constructions, hand-coded lists, or case sensitive text[7]. In addition, because the system is fully automated and corpus-based, it is suitable for applications that require portability across domains. Given the large percentage of non-anaphoric discourse entities handled by most coreference resolvers, we believe that using a system like ours to filter existential NPs has the potential to reduce processing time and complexity and improve the accuracy of coreference resolution.

8 Acknowledgments

This research is supported in part by the National Science Foundation under grants IRI-9509820 and IRI-9704240.

References

James Allen. 1995. *Natural Language Understanding.* Benjamin/Cummings Press, Redwood City, CA.

Chinatsu Aone and Scott William Bennett. 1996. Applying Machine Learning to Anaphora Resolution. In *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Understanding*, pages 302–314. Springer-Verlag, Berlin.

Andrew Kehler. 1997. Probabilistic coreference in information extraction. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97).*

Christopher Kennedy and Branimir Boguraev. 1996. Anaphor for everyone: Pronomial anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96).*

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

Joseph F. McCarthy and Wendy G. Lehnert. 1995. Using Decision Trees for Coreference Resolution. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 1050–1055.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press.

Brian Roark and Eugene Charniak. 1998. Noun-phrase co-occurence statistics for semi-automatic semantic lexcon construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics.*

R. Vieira and M. Poesio. 1997. Processing definite descriptions in corpora. In S. Botley and M. McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora.* UCL Press.

[7] Case sensitive text can have a significant positive effect on performance because it helps to identify proper nouns. Proper nouns can then be used to look for restrictive premodification, something that our system cannot take advantage of because the MUC-4 corpus is entirely in uppercase.

Appendix

Examples from the S1, EHP, & DO lists.

S1 Extractions	Existential Head Patterns	Definite-Only NPs
THE FMLN TERRORISTS	THE <X+> NATIONAL CAPITOL	THE STATE DEPARTMENT
THE NATIONAL CAPITOL	THE <X+> AFFAIR	THE PAST 16 YEARS
THE FMLN REBELS	THE <X+> ATTACKS	THE CENTRAL AMERICAN UNIVERSITY
THE NATIONAL REVOLUTIONARY NETWORK	THE <X+> AUTHORITIES	THE MEDIA
THE PAVON PRISON FARM	THE <X+> INSTITUTE	THE 6TH INFRANTRY BRIGADE
THE FMLN TERRORIST LEADERS	THE <X+> GOVERNMENT	THE PAST FEW HOURS
THE CUSCATLAN RADIO NETWORK	THE <X+> COMMUNITY	THE U.N. SECRETARY GENERAL
THE PAVON REHABILITATION FARM	THE <X+> STRUCTURE	THE PENTAGON
THE PLO	THE <X+> PATROL	THE CONTRARY
THE TELA AGREEMENTS	THE <X+> BORDER	THE MRTA
THE SALVADORAN ARMY	THE <X+> SQUARE	THE CARIBBEAN
THE COLOMBIAN GUERRILLA MOVEMENTS	THE <X+> COMMAND	THE USS
THE COLOMBIAN ARMY	THE <X+> SENATE	THE DRUG TRAFFICKING MAFIA
THE RELIGIOUS MONTHLY MAGAZINE 30 GIORNI	THE <X+> NETWORK	THE MAQUILIGUAS
THE REVOLUTIONARY LEFT	THE <X+> LEADERS	THE MAYORSHIP
THE PERUVIAN ARMY	THE <X+> RESULT	THE SANDINISTS
THE CENTRAL AMERICAN PEOPLES	THE <X+> SECURITY	THE LATTER
THE GUATEMALAN ARMY	THE <X+> CRIMINALS	THE WOUNDED
THE BUSINESS SECTOR	THE <X+> HOSPITAL	THE SAME
THE HONDURAN ARM	THE <X+> CENTER	THE CITIZENRY
THE ANTICOMMUNIST ACTION ALLIANCE	THE <X+> REPORTS	THE KREMLIN
THE DEMOCRATIC SYSTEM	THE <X+> ELN	THE BEST
THE U.S.	THE <X+> AGREEMENTS	THE NEXT
THE BUSH ADMINISTRATION	THE <X+> CONSTITUTION	THE MEANTIME
THE CATHOLIC CHURCH	THE <X+> PEOPLES	THE COUNTRYSIDE
THE WAR	THE <X+> EMBASSY	THE NAVY