Associative Anaphora Resolution: A Web-Based Approach

Razvan Bunescu
Department of Computer Sciences
University of Texas at Austin
Austin, TX 78712-1188
razvan@cs.utexas.edu

Abstract

We present a novel approach to solving definite descriptions in unrestricted text based on searching the web for a particular type of lexico-syntactic patterns. Using statistics on these patterns, we intend to recover the antecedents for a predefined subset of definite descriptions occurring in two types of anaphoric relations: identity anaphora and associative anaphora. Preliminary results obtained with this method are promising and compare well with other methods.

1 Introduction

Definite descriptions (noun phrases beginning with a definite article) have been extensively studied in linguistics, philosophy and computational linguistics and different authors have proposed different schemes for classifying possible uses of definite descriptions. The terminology that we use here is based on that introduced by Hawkins in (Hawkins, 1978) and is simplified by the fact that we are concerned with two types only:

- **Identity Anaphora** This is the same as Hawkins' *anaphoric* use, and subsumes definite descriptions that refer to the same entity as a previous phrase (antecedent) in the discourse. Examples:
  1. Fred was discussing an *interesting book* in his class. I went to discuss *the book* with him afterwards.
  2. Fred was wearing *trousers*. *The pants* had a big patch on them.
- **Associative Anaphora** This corresponds to Hawkins' *associative anaphoric* use, and refers to definite descriptions whose referent is uniquely identifiable based on general knowledge about associations with entities evoked by antecedents. Examples:
  1. Bill found himself in the middle of a *forest*. *The trees* were tall and sturdy.
  2. Tacos and burritos are the meat of the menu in this Mexican *restaurant*. *The atmosphere* is extremely laid back and *the service* is too.

In his analysis of the associative anaphoric use, Hawkins introduces the terms *trigger* and *associate* to refer to the antecedent and its associated definite description. We will extend the denotation of these two terms to cover also the case of identity anaphora. Consequently, in the above examples we have five *trigger:associate* pairs: *an interesting book: the book*, *trousers: the pants*, *forest: the trees*, *restaurant: the atmosphere*, *restaurant: the service*. The types of relations involved in associative anaphora can be very diverse, from meronymy as in *forest: the trees*, to attributes as in *car: the price*, to complex relationships as in *Auschwitz: the victims*. Therefore, we will concern ourselves with identifying *trigger:associate* pairs, without establishing the exact type of their association.

Extracting the triggers of anaphoric definite descriptions is not an easy task. In one experiment

(Poesio et al., 1997), the authors exploited the WordNet lexical database (Miller, 1991) in order to account for the commonsense knowledge that humans seem to employ when solving definite descriptions. Another approach (Poesio et al., 1998) tried to address the incompleteness of the information hand-coded in WordNet by hypothesizing a semantic priming effect between a trigger and its associate which could be detected automatically using a lexical clustering algorithm similar to that described in (Lund et al., 1995). In (Meyer and Dale, 2002) lexico-syntactic patterns have been used to mine *lexical associative axioms* from a corpus of 2000 encyclopaedia articles. These associations were further generalized based on WordNet and their performance was evaluated on a set of five anaphoric heads.

Our method, as described in the following sections, combines the power of a different type of lexico-syntactic patterns with the huge coverage offered by the world wide web.

2 The Method

Given a pair of nouns n_t:n_a occurring in the same document, we want to detect how likely it is that the two nouns are in a *trigger:associate* relationship. To accomplish this, we plug the two nouns in the pattern from Figure 1, as if n_t were ending a sentence and n_a were at the beginning of the next sentence in a definite description followed by one of the verbs *is/are, was/were, has/have, had, may, might, can, could, should, would*. For each of the three instantiated patterns, a search engine will return the number of matching documents, and, based on these numbers, we shall derive a measure of the degree of association between the two nouns.

Q(n_t,n_a) = "[n_t]. The [n_a] [verb]"
Q(n_t) = "[n_t]."
Q(n_a) = "The [n_a] [verb]"

Figure 1: Phrase Patterns

Our intuition is that ordered pairs of nouns occurring in identity or associative anaphora will get a high degree of association in this kind of pattern, whereas unrelated nouns should get a low value for the same association measure. The rationale behind this pattern is based on the following observations:

- Relatively many *trigger:associate* pairs appear in this kind of configuration (we ignore the verb for the time being). Three of our five examples follow this pattern; the literature on the anaphora phenomena is rife with this type of examples, and, most important, there is a slight tendency in natural language towards employing associated nouns (in the two anaphoric senses discussed in the introduction) when using this pattern.
- Definite descriptions, when used in identity or associative anaphora, do not need an establishing modifier, relative clause or prepositional phrase, as the hearer can already identify their referent. We enforce this in our pattern as follows: there is no modifier between the definite article and the potential associate, and the presence of a subordinate relative clause or attached prepositional phrase after the associate is prohibited by the immediate presence of a generic verb. This is the reason for placing a verb after the associate in the pattern.
- The particular set of verbs used in the pattern was determined by the need to exclude any possible association between the verb and the two nouns. Therefore, we have selected verbs which we thought were general enough to preclude such associations: the verb *"to be"* and modal verbs at different tenses. The generality of these verbs was also instrumental in increasing the number of hits returned by the search engine.
- The final form of this pattern was highly constrained by the fact that we needed it to conform to one of the query formats accepted by the current search engines. The only available query formats emphasizing word co-occurrences in short spans of text are the "Phrase" and "NEAR" queries. With a "Phrase" type of query, a search engine looks for documents containing the exact specified phrase. With a "NEAR" type of

query, the search engine will return documents containing both specified words or phrases within a constant number of words of each other. However, because the *trigger:associate* relation is generally asymmetric while the "NEAR" type of query (shown in Figure 2) is symmetric, we hypothesized that the phrase pattern should give a better performance (the experiments will validate this as will be seen later in Section 3).

There is also the issue of time complexity. We could, for instance, drop the verb from the phrase pattern, download each matching document and filter out those documents in which the second noun from the pattern had a subordinate relative clause or an attached prepositional phrase. This however would be very time consuming and would require too much network traffic.

Q(n_t,n_a) = "[n_t]" NEAR "the [n_a]"
Q(n_t) = "[n_t]"
Q(n_a) = "the [n_a]"

Figure 2: NEAR Patterns

The exact method for computing the degree of anaphoric association between a potential trigger noun n_t and a potential associate noun n_a is based on the information-theoretic measure of pointwise mutual information (Church and Hanks, 1990; Manning and Schütze, 1999). If we denote with *N* the total number of web documents indexed by a search engine, and with *D(query)* the number of web documents returned by the same search engine with the given query, then the degree of association (when we use either the Phrase pattern or the NEAR pattern) will be given by $I(n_t, n_a)$ computed as in Figure 3.

$$I(n_t, n_a) = \log_2 \frac{P(Q(n_t,n_a))}{P(Q(n_t))P(Q(n_a))}$$

$$\approx \log_2 \frac{\frac{D(Q(n_t,n_a))}{N}}{\frac{D(Q(n_t))}{N} * \frac{D(Q(n_a))}{N}}$$

$$= \log_2 \frac{N * D(Q(n_t,n_a))}{D(Q(n_t)) * D(Q(n_a))}$$

Figure 3: Pointwise Mutual Information

For a particular pair of nouns n_t:n_a, we collect the number of hits that the Altavista[1] search engine returns for the three phrase queries illustrated in Figure 1 and compute $I(n_t, n_a)$ accordingly. If there is no association (in the anaphoric sense) between n_t and n_a, then the numerator and denominator tend to have the same value, and the pointwise mutual information will be close to 0. On the other hand, if n_t and n_a can be related in associative anaphora, then there will be a dependence between the events involved in the probabilities from Figure 3, which will result in a higher value for $I(n_t, n_a)$.

3 Experimental Evaluation

We tested our method on the first 32 documents from the Brown section of the Treebank corpus (Marcus et al., 1994) (cf01 to cf32). For each document, we created a list of potential associates consisting of definite descriptions with only one noun, with the additional constraint that no prepositional phrase or relative phrase was attached to them (in this way we focused our method on those definite descriptions that were most susceptible to be anaphoric). We have also excluded from the set of possible associates all definite descriptions whose head noun had occurred before in the document (as is the case with the first example of identity anaphora in Section 1). The resulting list of potential associates contains 686 definite descriptions, and the task becomes that of identifying the trigger (if such a trigger exists) for each of them.

All 686 potential associates were annotated by hand as belonging to one or more of the following 6 classes:

1. If a definite noun phrase is anaphoric and it has one or more trigger nouns (it is possible to have more than one trigger noun for the same associate), then the definite noun phrase is annotated as an associate with the corresponding list of triggers.

2. This class contains anaphoric definite noun phrases for which the trigger is not a noun (for instance, the trigger may be a verb, or even an entire phrase).

[1] URL: `http://www.altavista.com`

3. Some definite noun phrases like *"the world"*, *"the moon"*, *"the earth"*, *"the balkans"*, *"the past"*, *"the future"* or *"the pope"* (the larger situation uses in Hawkins' classification) have a well known referent based on the common knowledge shared by speaker and hearer. These were included in a separate class. By reification, many definite descriptions can enter this class, and in our test set we have examples such as *"the brain"*, *"the eyes"*, *"the eye"*, *"the street"*, etc.

4. Another category is that of definite noun phrases triggered by the discourse but for which we cannot find a trigger in the form of a word or a phrase. Typical examples are most occurrences of noun phrases like *"the problem"*, *"the situation"*, *"the issue"*, *"the question"*, etc.

5. There are cases of definite description use where the hearer/reader cannot infer the referent. This may happen especially at the beginning of documents or in direct speech. One document from our test set begins with the following sentence: "*The food* is wonderful and it is a lot of fun to be here!". These definite descriptions were then tagged accordingly as belonging to a separate category.

6. Yet another class of definite descriptions is that of definite noun phrases occurring inside an idiomatic phrase, such as: *"out of the blue"*, *"on the contrary"*, *"in the making"*, or *"let the cat out of the bag"*.

Some of these classes may overlap, especially the first and the third class. One example is the phrase *"the historian"* used with a reified meaning in one document. It was included in the third class of definite descriptions, nevertheless the same phrase could be viewed as triggered by the preceding noun *"history"* (document cf19 from the test corpus).

The distribution of the 686 potential associates over these 6 categories is shown in Table 1.

Given a definite noun phrase (from the list of potential associates) containing a noun n_a, we

Class	1	2	3	4	5	6
Total	324	29	175	126	42	30

Table 1: Class Statistics

consider each of the preceding 50 nouns as a possible trigger. Consequently we create 50 n_t:n_a pairs and compute for each of them the degree of association as described in Figure 3. We then select the trigger for which we get the highest degree of association. Except in the case when it contains the first noun from the document, a definite description will be associated with its highest ranking trigger noun. By imposing a threshold on the minimum acceptable value for the association measure and by varying this threshold we get the precision-recall graphs from Figure 4. The graph labeled "Phrase" corresponds to our method when using the Phrase pattern from Figure 1. The precision-recall graph labeled "NEAR" stands for the same method when the pattern was changed to the NEAR pattern from Figure 2.

The definite descriptions for which there exists a word or phrase trigger are those from classes 1 and 2, therefore our method was evaluated on the task of extracting exactly this maximal set of associations, starting from the entire set of 686 potential associates. Note that the method was designed to extract associations from the first class only, consequently its performance on extracting only the first type of associations is actually higher.

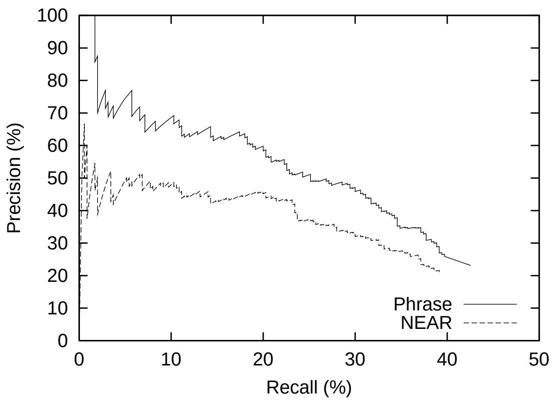

Figure 4: Precision-recall graphs

We've also evaluated the performance of 2 baseline approaches:

1. In the first baseline, for each potential associate we assigned as trigger a random noun from the preceding 50 nouns. The precision was 1.1% at a recall of 2.1%.

2. In the second baseline, each potential associate was considered triggered by the closest preceding noun. The precision was 3.4% at a recall of 6.2%.

4 Discussion of results

The precision-recall graphs show a significant better performance for the Phrase pattern when compared with the NEAR pattern, confirming our intuition about the importance of enforcing two anaphoric features in the statistical pattern: the asymmetry of the *trigger:association* relationship and the fact that associates do not need establishing modifiers or relative clauses.

The method compares well with the approach from (Poesio et al., 1998) where precision and recall on the class of *inferential descriptions* were 22.7%. At the same recall our method has a precision of 53%. However the comparison is complicated by the fact that the test sets are from different sections of Treebank, plus the fact that our set of potential associates was created using a different method which, as described at the beginning of Section 3, did not take into account the manual annotation of the document. The approach presented here also validates the prediction made in the same paper (Poesio et al., 1998), namely that statistical pattern matching may provide a way to trade off between precision and recall by varying a suitable threshold.

The Phrase pattern seems more general than other patterns used for the same task: a pattern based on genitive constructions, for instance, cannot cover *trigger:associate* cases like that of *"couple"* and *"the man"* (a web search for the phrase *"the couple's man"* returns no hits).

In the following paragraphs we'll investigate some of the errors and limitations of our method.

Sometimes the second best trigger was a correct trigger, while the first one was incorrect. It was also closer to the associate noun than the first ranked trigger, which suggests that distance should play a role in the ranking decision.

There were cases when a spurious trigger was ranked very high because of an "unjustified" high number of hits from the search engine. One example is the pair *car:butter* for which the phrase pattern, instantiated to *"car. The butter is"*, returned 7 documents. But all 7 documents contain the same text section called *"Ideas for Caregivers"* with the text *"We need ONE egg. That's a RED car. The butter is in this SQUARE box."* Normally, these 7 documents should be counted as one, however this kind of checking requires significant processing and network traffic (there may be thousands of documents returned by the search engine).

Anaphora resolution may require more complicated reasoning than a simple decision based on a lexical association measure. In one document, *"the Greek"* corefers with *"a member of a greek syndicate"*. The degree of association between *"member"* and *"Greek"*, as computed by our method, is very low, nevertheless the two are coreferent. Another interesting example that couldn't be solved with the current method was the following sequence of phrases (document cf02): *"She was just another freighter from the States"*, *"She was the John Harvey"*, *"John Harvey"*, *"the ship"*. Between the two nouns *"ship"* and *"freighter"* there is a distance of more than 50 nouns, therefore our method could not detect an association between them. One may think of extending this window size and making a more informed decision based on the (presumably already detected) coreference between *"freighter"* and *"John Harvey"*, and the fact that *"John Harvey"* is very close to *"the ship"*, which has a high degree of association with *"freighter"*. In other words, the association should propagate to all coreferent items.

There are special cases of coreference for which the association measure cannot help, cases which, nevertheless, can be solved by a simple approach. An example is with the *"the latter"* constructions, as in *"the decisions of the highest ecclesiastical authority and the natural law . The latter plays a prominent role..."* (document cf15).

Changes in discourse topic may invalidate strong associations between nouns from fragments with different topic. An algorithm for detecting topic changes may prove very useful in restricting the set of potential triggers for a given associate.

Another limitation of our method is the fact that it can be applied for finding *trigger:associate* pairs consisting of single nouns only. To make it applicable to more general noun phrases we need a mechanism for detecting collocations. The importance of detecting collocations is evident if we try to find *"sonata"* as the trigger for *"the first movement"*. If we create a phrase pattern as in Figure 1 based only on the head nouns *"sonata"* and *"movement"*, we get no hits from the search engine. This happens because *"first movement"* is a terminological collocation and it should be used as it is in the phrase pattern. The dependence on a good collocation extraction algorithm shall become more critical if we intend to use our approach on a corpus in which terminological collocations are very common, as is the case with the Wall Street Journal section of the Treebank.

5 Conclusions and Future Work

Web search engines have been used before to help in tasks from natural language processing. In (Turney, 2001), the Altavista search engine was used for recognizing synonyms. A similar approach was followed in (Turney, 2002) in order to detect the semantic orientation of reviews. To our knowledge, our method represents the first attempt to use a web search engine for solving identity and associative anaphora in unrestricted text. Preliminary results are promising and compare well with other methods. There is still room for improvement - a direction to follow is that of enhancing the current method with the ability to detect collocations and use them accordingly in the statistical patterns. In particular, we may follow a web based approach for finding collocations too. Another improvement may come from considering the distance between a definite description and a potential trigger as a factor in the trigger selection process. We also intend to evaluate how much can be gained by using the knowledge encoded in WordNet to boost potential triggers whose relation with the associate is one of synonymy, hyponymy, or meronymy. For this kind of relations there is also the option of using other types of statistical patterns with a potentially higher accuracy, as shown in (Poesio et al., 2002).

References

Kenneth W. Church and Patrick W. Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

John A. Hawkins. 1978. *Definiteness and Indefiniteness*. Humanities Press, Atlantic Highlands, NJ.

K. Lund, C. Burgess, and R. A. Atchley. 1995. Semantic and associative priming in highdimensional semantic space. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, pages 660–665, Mahwah, NJ.

C. D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Josef Meyer and Robert Dale. 2002. Mining a corpus to support associative anaphora resolution. In *4th Discourse Anaphora and Anaphor Resolution Colloquium*, Lisbon, Portugal, september.

George Miller. 1991. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4).

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. In *Proc. of the ACL Workshop on Operational Factors in Robust Anaphora Resolution*, pages 1–6.

Massimo Poesio, Sabine Schulte im Walde, and Chris Brew. 1998. Lexical clustering and definite description interpretation. In *AAAI Spring Symposium on Learning for Discourse*, pages 82–89, Stanford, CA, march.

Massimo Poesio, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of LREC*, Las Palmas, Spain.

Peter D. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proc. of the 12th European Conference on Machine Learning*, Freiburg, Germany, september.

Peter D. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania.