A probabilistic model for noun-phrase coreference

Keith Hall and Eugene Charniak
Dept. of Computer Science,
Brown University,
[kh|ec]@cs.brown.edu

Paper ID Code: NAACL-17
Topic Area(s): Discourse, Dialogue, and Pragmatics
Keywords: coreference, anaphora, noun-phrase, corpus, statistical
Submission to NAACL or ANLP (you must choose only one)? NAACL
Word Count: 3082
Under consideration for other conferences (specify)? None

Abstract

Acknowledging recent successes in applying probabilistic methods to pronominal reference, we extend the approach to common noun-phrase coreference. We have found this to be a more difficult problem and address the differences. The model presented in this paper explores a restricted domain where no world knowledge is available. However, we achieve an accuracy of 65.2% when run on articles from the Penn Wall Street Journal Treebank.

1 Introduction

Discourse semantic models are constrained by the requirement that unique referents within a discourse must be identified. This demand emphasizes the need to model noun-phrases coreference relationships. For example, Kamp's(1993) Discourse Representation Theory presupposes that a unique label be assigned to every object and that each reference of an object within a discourse is identified by this label. Noun-phrase coreference specifically addresses the problem of determining the groups of noun-phrases which refer to the same object (and would be assigned the same referent label).

Recently, statistical models have been applied to a variety of language semantic problems. One of these models addressed a problem closely related to common noun-phrase coreference, pronominal reference(Ge et al., 1998). Initially, we assumed the two problems to be similar enough that we could use the same features as presented in the model for pronoun resolution. Common noun-phrase coreference turns out to be a more complicated problem. The crucial difference stems from the use of pronouns in a text. In general, pronouns are used to refer to an antecedent referent and we can easily identify the cases where a pronoun does not refer to any antecedent. In contrast, we observed only around 20% of the common noun-phrases in our corpus to have antecedent referents (the first noun-phrase of every class is included in the remaining 80%). Furthermore, we have considered noun-phrase similarity features which depend upon the makeup of the antecedent class and not only on the noun-phrases contained within the class. Modeling the reference relationship between a single entity and an entire class has been found useful for similar tasks as in (Vilain et al., 1995), (Cardie and Wagstaff, 1999), and (Kehler, 1997).

Ultimately, we believe the problem of noun-phrase coreference will not be solved without the use of world knowledge. In the current work, we have chosen to use, in isolation, the superficial attributes of the noun-phrases and coreferent classes of noun-phrases. There has been some success in modeling the related problem of pronominal reference in such knowledge poor situations(Mitkov, 1998), (Ge et al., 1998). Previous work on full noun-phrase coreference have generally been dependent on external sources of world knowledge or semantic networks.

2 Model Architecture

We first present a discussion of the motivations behind the model features. While describing the model, we often refer to the *current* noun-phrase, which is the noun-phrase being processed. We also refer to the *antecedent* classes and *antecedent* noun-phrases, those classes (and the noun-phrases within) which have previously been processed. We begin by segmenting the features into three logical groups: features which address the structure of an antecedent equivalence class, features operating on pairs of noun-phrases (a current and antecedent noun-phrase) and features describing lexical components of a single noun-phrase.

Class-based features provide information that is not otherwise embedded within the noun-phrases. We experimented with two features of this sort: class size and distance. The class size is simply a count of the noun-phrases within the class. Distance, we define as the number of word tokens between the current noun-phrases and the last noun-phrase in the antecedent class. Neither of these features proved to be helpful when used on their own, but class size provides an important conditioning event for other features (we discuss this when describing the pair-wise determiner feature).

The next set of features, and presumedly the most important, are the noun-phrases comparative features. These are attributes based on the comparison of the current noun-phrases and an antecedent noun-phrase. Given a pair of noun-phrases, we exploit three contrastive relationships. The first of these is the *content* similarity, where we observe the number of the current noun-phrase's open-class words that were matched in the antecedent. Next we consider a comparison of the logical head nouns. The logical head noun is the noun close to the head noun position which is most descriptive (we describe the selection of logical head nouns in 2.3). An example of the logical head noun selection is shown in Figure 1[1]. Another noun-phrase comparative features, also based on open-class words, is relative size. We observed a preference for noun-phrases of an antecedent class becoming smaller (or staying the same size) as they occur later in the article. A good example of this phenomena occurs with appositions following proper names. The first referent to a person often includes an apposition, whereas later reference drop the additional information.

[1] Note that in Figure 1, the noun-phrase: "The Voice Of America" is not marked, instead the larger noun-phrase is. We adopted this rule, marking the *larger* noun-phrase, for our annotations.

```
(NP#-8021~1
    (NP (DT the)
        (NAC (NNP Voice)
            (PP (IN of)
                (NP (NNP America) )))
        (NNS offices) )
    (PP-LOC (IN in)
        (NP#-8010~1 (NNP Washington) )))
```

Figure 1: Head noun of an noun-phrase. Annotations are in the form #-RefNumber~CountNum.

Linguists have suggested rules, in English, for the use of the definite and indefinite article. Baker(1996) describes a *registration* process whereby the introduction of a referent is accompanied by the indefinite article and later references use the definite article. We attempt to capture this phenomena through a lexically-based determiner feature. We estimate the distribution of all pairs of determiners conditioned on whether the antecedent was the first reference or not. Consider the case in which we observe the noun-phrase *the company*. This noun-phrase is likely to be in a class with a noun-phrase containing a *the*, or a noun-phrase containing an *a* if that noun-phrase is the first in the antecedent class[2].

The last set of features are those capturing noun-phrase-specific characteristics. Though these features do not aide in determining the best antecedent class (as they are identical for every antecedent), they do influence whether the noun-phrases refers to some antecedent or not. Again, we consider the logical head noun and the determiner. In our annotated corpus of Wall Street Journal articles, we found a number of logical head nouns which are less likely to be in coreference relationships. For example, the nouns *year* and *%*[3] seldomly refer to a previous entity. We call these nouns *non-gregarious* as they tend not to group together. Similarly, we found determiners that have this same negative affinity for coreferent classes. Quantitative determiners such as *none* and *some* are rarely found in coreferent classes (in our corpus). We include a feature for these determiners which mimics that of the non-gregarious head noun feature.

[2] Note that the conditioning event is dependent on class size. For example, the probability that the antecedent noun-phrase is the first given the class size is one is 1.0.

[3] % is marked as a noun in the Penn Treebank.

2.1 Probabilistic Model

We define an objective function, $F(N_\nu)$, which returns the antecedent class that the current noun-phrase (denoted N_ν) refers to. If the current noun-phrase does not refer to any antecedent, $F(N_\nu)$ returns $\emptyset$. We compute this function in two passes. The first-pass determines whether the noun-phrase refers to any antecedent. The second pass determines the most likely antecedent class that N_ν refers to. We discuss the first-pass is section 2.2.

Here, we discuss the second pass. Given a noun-phrase, E_ν, and a vector of the antecedent classes, $\vec{\mathcal{E}}$, we solve for the following equation[4].

$$\begin{aligned}
&\Pr(C_j|E_\nu, \vec{\mathcal{E}}, \mathcal{C}) \\
&= \frac{\Pr(C_j|\mathcal{C}) \Pr(\vec{\mathcal{E}}, E_\nu|C_j)}{\Pr(\vec{\mathcal{E}}, E_\nu)}
\end{aligned}$$

In practice we have found it easier to maximize the posterior odds(Pearl, 1988).

$$\begin{aligned}
&\frac{\Pr(C_j|E_\nu, \vec{\mathcal{E}}, \mathcal{C})}{\Pr(\neg C_j|E_\nu, \vec{\mathcal{E}}, \mathcal{C})} && (1)\\
&= \frac{\Pr(C_j|\mathcal{C})}{\Pr(\neg C_j|\mathcal{C})} \frac{\Pr(E_\nu|C_j)}{\Pr(E_\nu|\neg C_j)} \frac{\Pr(\vec{\mathcal{E}}|E_\nu, C_j)}{\Pr(\vec{\mathcal{E}}|E_\nu, \neg C_j)} && (2)\\
&\propto \frac{\Pr(\vec{\mathcal{E}}|E_\nu, C_j)}{\Pr(\vec{\mathcal{E}}|E_\nu, \neg C_j)} && (3)\\
&= \frac{\Pr(\vec{\mathcal{E}}_j|E_\nu, C_j)}{\Pr(\vec{\mathcal{E}}_j|E_\nu, \neg C_j)} \prod_{\vec{\mathcal{E}}_{i\{i \neq j\}}} \frac{\Pr(\vec{\mathcal{E}}_i|E_\nu, C_j)}{\Pr(\vec{\mathcal{E}}_i|E_\nu, \neg C_j)} && (4)
\end{aligned}$$

In equation 3, we ignore $\Pr(E_\nu|C_j)$ and $\Pr(C_j|\mathcal{C})$ as they remain constant for all C_j. In equation 4 we are assuming that the probability of a noun-phrase occurring in one class is independent of the probability of a noun-phrase occurring in different classes (given we know which is the correct class).

We cannot make the same independence assumption for the noun-phrases within a class as we have for the noun-phrases in different classes. Assume that the following noun-phrases are in the same equivalence class.

[4] Antecedent noun-phrases which do not have any coreference links are considered singleton classes

the law, a 1948 law, the law, the statute

We can see that the probability of observing the third noun-phrase in this class is dependent on the previously observed noun-phrases. In particular, the having observed two noun-phrases with the head word *law* makes observing a noun-phrase in this class with the head word *law* more probable.

In order to model this dependency exactly, we would need to make each noun-phrase dependent on the previously selected noun-phrases. A class with n noun-phrases is expressed in the following equation $\vec{\mathcal{E}}_{j_k}$ represents the k^{th} noun-phrases of the j^{th} antecedent class (denoted C_j).

$$\begin{aligned}
&\frac{\Pr(\vec{\mathcal{E}}_j, E_\nu|C_j)}{\Pr(\vec{\mathcal{E}}_j, E_\nu|\neg C_j)} && (5)\\
&= \frac{\Pr(\vec{\mathcal{E}}_{j_1}|E_\nu, C_j)}{\Pr(\vec{\mathcal{E}}_{j_1}|E_\nu, \neg C_j)} \frac{\Pr(\vec{\mathcal{E}}_{j_2}|\vec{\mathcal{E}}_{j_1}, E_\nu, C_j)}{\Pr(\vec{\mathcal{E}}_{j_2}|\vec{\mathcal{E}}_{j_1}, E_\nu, \neg C_j)} \cdots \frac{\Pr(\vec{\mathcal{E}}_{j_n}|\vec{\mathcal{E}}_{j_1}, \cdots, \vec{\mathcal{E}}_{j_{(n-1)}}, E_\nu, C_j)}{\Pr(\vec{\mathcal{E}}_{j_n}|\vec{\mathcal{E}}_{j_1}, \cdots, \vec{\mathcal{E}}_{j_{(n-1)}}, E_\nu, \neg C_j)} && (6)\\
&\approx \frac{\Pr(\vec{\mathcal{E}}_{j_1}|E_\nu, C_j)}{\Pr(\vec{\mathcal{E}}_{j_1}|E_\nu, \neg C_j)} && (7)
\end{aligned}$$

In equation 7 we assume that all but the first term in equation 6 will be close to 1, so we ignore them. That is, we assume that once we know one or more noun-phrases in an antecedent class, knowing N_ν is a member of the class adds little information.

$$\Pr(\vec{\mathcal{E}}_{j_2}|\vec{\mathcal{E}}_{j_1}, E_\nu, C_j) \approx \Pr(\vec{\mathcal{E}}_{j_2}|\vec{\mathcal{E}}_{j_1}, E_\nu)$$

$$\Pr(\vec{\mathcal{E}}_{j_2}|\vec{\mathcal{E}}_{j_1}, E_\nu, \neg C_j) \approx \Pr(\vec{\mathcal{E}}_{j_2}|\vec{\mathcal{E}}_{j_1}, E_\nu)$$

We have derived an approximation which avoids estimating the statistics for the dependent noun-phrase probabilities and can improve it by defining the notion of noun-phrase *informativeness*. That is, a noun-phrase is *informative* if the phrase contains particular features (specifically, the model features which we describe below). Each feature has a relative weight assigned, thereby making some features more important than others. Ensuring the first noun-phrase is the most *informative*, we order the noun-phrases by their weights (this will be the noun-phrase in equation 7). We approximate this process by assuming that the likelihood ratio in equation 7 will be highest for the most *informative* noun-phrase. For each noun-phrase, we calculate the likelihood ratio and pick the noun-phrase which produced the highest value.

Next, we break-out equation 7 to consider the evidence from section 2.

$$
\begin{aligned}
& \frac{\Pr(\vec{\mathcal{E}}_{j_k} | E_\nu, C_j)}{\Pr(\vec{\mathcal{E}}_{j_k} | E_\nu, \neg C_j)} && (8) \\
& \quad = \frac{\Pr(\mathcal{M}, \mathcal{O}, \mathcal{L}, H_\nu, D_\nu, D_\mu, \mathcal{F}_j, \mathcal{S}_j | C_j)}{\Pr(\mathcal{M}, \mathcal{O}, \mathcal{L}, H_\nu, D_\nu, D_\mu, \mathcal{F}_j, \mathcal{S}_j | \neg C_j)} && (9) \\
& \quad = \frac{\Pr(\mathcal{M} | C_j, \mathcal{O})}{\Pr(\mathcal{M} | \neg C_j, \mathcal{O})} \frac{\Pr(\mathcal{O} | C_j)}{\Pr(\mathcal{O} | \neg C_j)} \frac{\Pr(\mathcal{L} | C_j)}{\Pr(\mathcal{L} | \neg C_j)} \frac{\Pr(D_\mu | C_j, D_\nu, \mathcal{F}_i)}{\Pr(D_\mu | \neg C_j, D_\nu, \mathcal{F}_i)} \frac{\Pr(\mathcal{F}_i | C_j)}{\Pr(\mathcal{F}_i | \neg C_j)} \frac{\Pr(\mathcal{S}_i | C_j)}{\Pr(\mathcal{S}_i | \neg C_j)} && (10)
\end{aligned}
$$

In equation 9, $\mathcal{M}$ indicates the head of $\vec{\mathcal{E}}_{j_i}$ matched the head of E_ν. $\mathcal{O}$ is a binary variable, taking the value *true* when all open-class words in the current noun-phrases occurred somewhere in the antecedent. The $\mathcal{L}$ feature is *true* when E_ν has less than or equal the number of open-class words than the antecedent $\vec{\mathcal{E}}_{j_k}$. H_ν is the head word of E_ν, D_ν the determiner of the current noun-phrase, and D_μ is the determiner of the antecedent noun-phrase, $\vec{\mathcal{E}}_{j_k}$. If the antecedent noun-phrases is the first in its class, the $\mathcal{F}_j$ features takes the value *true*. Finally, $\mathcal{S}_j$ represents the count of noun-phrases within the j^{th} antecedent class.

Equation 10 is the equation we use to select the best antecedent class given the current noun-phrase. We have made a number of independence assumptions. The logical head nouns are included in the open-class word matching comparison, thereby making the dependency between their probabilities obvious. As mentioned previously, the comparative determiner feature is dependent on whether the antecedent noun-phrase is the first in its class or not. We assume independence of all other features.

2.2 First-pass Selection (ϵ-test)

The objective function, $F(N_\nu)$ is defined in two steps. This first step determines whether the noun-phrase refers to any antecedent class(we call this the first-pass selection).

$$
F(N_\nu) = \begin{cases} \emptyset & : \frac{\Pr(\mathcal{C} | E_\nu, \vec{\mathcal{E}})}{\Pr(\neg \mathcal{C} | E_\nu, \vec{\mathcal{E}})} \leq \epsilon \\ \arg\max_j \frac{\Pr(C_j | E_\nu, \vec{\mathcal{E}})}{\Pr(\neg C_j | E_\nu, \vec{\mathcal{E}})} & : \frac{\Pr(\mathcal{C} | E_\nu, \vec{\mathcal{E}})}{\Pr(\neg \mathcal{C} | E_\nu, \vec{\mathcal{E}})} > \epsilon \end{cases} \tag{11}
$$

$\mathcal{C}$ indicates that the current noun-phrase refers to some antecedent referent. As in equation 1, we use the posterior odds as our evaluation metric.

$$\begin{aligned} & \frac{\Pr(\mathcal{C}|E_\nu,\vec{\mathcal{E}})}{\Pr(\neg\mathcal{C}|E_\nu,\vec{\mathcal{E}})} && (12) \\ & = \frac{\Pr(\mathcal{C})}{\Pr(\neg\mathcal{C})}\frac{\Pr(E_\nu|\mathcal{C})}{\Pr(E_\nu|\neg\mathcal{C})}\sum_{C_j}\frac{\Pr(C_j|\mathcal{C})}{\Pr(\neg C_j|\neg\mathcal{C})}\frac{\Pr(\vec{\mathcal{E}}|E_\nu,C_j)}{\Pr(\vec{\mathcal{E}}|E_\nu,\neg C_j)} && (13) \end{aligned}$$

The objective function is defined with one parameter, the epsilon. We choose as our epsilon the value which maximizes the geometric mean of the precision and recall(the scoring algorithm is defined in Vilain(1995)), when run on the training data.

2.3 Uninformative Head Nouns

When selecting a logical head, we introduced the phenomena of uninformative head nouns. The simplest algorithm for finding the syntactic head noun is to choose the rightmost noun of the base noun-phrase. In the following example, that algorithm would choose *Corp.* as the head noun.

Georgia-Pacific Corp.

Unfortunately, the word *Corp.* says little about the specific object referred to by the noun-phrase. Instead, we would like to select *Georgia-Pacific* as the logical head noun. We do this by learning that *Corp.* is an uninformative noun and selecting the noun to the left of it. We learn *uninformative* nouns by examining noun-phrases within coreferent classes whose head nouns did not match. If the a noun to the left of either head words would have allowed a match, we treat this as a positive occurrence for the original head noun. We perform this evaluation over our training data and calculate the relative frequency for the head nouns. We empirically set a threshold and define those nouns with a frequency over the threshold to be uninformative.

Tokens	14722
Noun-phrases	3314
Non-singleton Classes	210
Noun-phrases in Non-singleton Classes	907
Sentences	556

Table 1: Corpus Statistics

3 Results

We use a small set of articles (40) from the Penn Wall Street Journal Treebank which have been annotated with coreference markings (Table 1 provides relevant information on our corpus). We then partition this data into ten groups, whereby we perform a ten-fold cross-validation. In both statistics collection and classification, we process the text incrementally. Classification decisions are made on-line. For each noun-phrases we consider all antecedent classes including singleton classes.

	Precision	Recall	Geometric Mean	F Measure
Content Similarity Baseline	54.3	53.8	54.0	54.0
Content Similarity Head-match	65.3	61.5	63.4	63.3
Full Model	69.4	61.5	65.3	65.2

Table 2: Results

The results of our system are reported in Table 2. We have included two simpler models for comparison. In Table 3 we evaluate each of the model components based on their improvement over the model with the content similarity and head noun matching features. Both class size and distance affected the score negatively and were not included in the full model as reported in Table 2. Additionally, we found that approximately 85% of the errors are first-pass errors[5].

4 Related Work

We present the following subset of the work on noun-phrase coreference due to its relevance to the model we have proposed. Cardie and Wagstaff(1999) address the coreference problem as a clustering task. They

[5] First-pass errors are those where the noun-phrases had no antecedent referent in the key, but we assigned an antecedent, visa versa.

	Precision	Recall	Geometric Mean	% Improvement
Content Sim. & Head-match	65.3	61.5	63.4	
Content Sim., Head-match & Determiner	65.7	63.3	64.5	1.7%
Content Sim., Head-match & Distance	55.3	70.7	62.5	-1.4%
Content Sim., Head-match & ClassSize	56.0	67.3	61.4	-3.2%
Content Sim., Head-match & Relative Size	66.7	60.5	63.5	0.2%
Content Sim., Head-match & Non-Gregarious	66.9	61.5	64.1	1.1%

Table 3: Component-wise Improvements

present an unsupervised agglomerative clustering model which achieves a score of 53.6% on the MUC-6 data. Baldwin, et. al. (1997) reports a score of 58.5% on the MUC-7 coreference data set. Fukumoto, et. al.(1997) achieved an accuracy of 41.3% for MUC-7 full noun-phrases coreference. Roberto Garigliano, et. al.(1997) used a system based on an semantic network to achieve 51.5% when tested on MUC-7 full noun-phrase coreference. Ge, Charniak and Hale(1998) presented a statistical model for pronominal reference. This system achieved 84.2% accuracy when trained on a small set of the Penn Treebank articles. Ruslan Mitkov proposed a non-statistical model for pronominal reference where limited knowledge is available (Mitkov, 1998). This system achieved an accuracy of 89.7% when tested on a set of Technical Manuals.

In a recent paper, Bean and Rilof addressed a problem similar to what we have called the *first-pass* selection(Bean and Riloff, 1999). They proposed a model which helps identify non-anaphoric noun-phrases.

5 Conclusions and Future Work

In this paper we suggest a statistical framework for exploring common noun-phrase coreference. We abstained from the use of external knowledge sources to explore the power of the model in the most restrictive environment. Achieving an accuracy of 65.2% supports our assumption that the problem lends itself to a probabilistic approach, though clearly, further research in this direction is necessary. Though we do not achieve the level of performance that statistical pronominal reference models have, we believe common noun-phrase coreference to be a more difficult problem.

The most obvious extension to this work is to incorporate world knowledge. We believe including features similar to those in Cardie and Wagstaff(1999) would increase performance. More importantly though, we must explore the first-pass decision problem and identify features which elucidate the process governing it. Specifically, we will include a feature which captures non-anaphoric attributes similar to those introduced by Bean and Riloff(1999).

References

C. L. Baker. 1996. *English Syntax.* MIT Press.

Breck Baldwin, Tom Morton, Amit Bagga, Jason Baldridge, Raman Chandraseker, Alexis Dimitriadis, Kieran Snyder, and Magdalena Wolska. 1997. Description of the upenn camp system as used for coreference. *Proceedings of the Seventh Message Understanding Conference.*

David L. Bean and Ellen Riloff. 1999. Corpus-based identification of non-anaphoric noun phrases. *Proceedings of The* 37^{th} *Annual Conference of The Association for Computational Linguistics. College Park, Maryland*, pages 373–379.

Clair Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. College Park, Maryland.*

J. Fukumoto, F. Masui, M. Shimohata, and M. Sasaki. 1997. Oki electric industry: Description of the oki system as used for muc-7. *Proceedings of the Seventh Message Understanding Conference.*

Roberto Garigliano, Agnieszka Urbanowicz, and David J. Nettleton. 1997. University of durham: Description of the lolita system as used in muc-7. *Proceedings of the Seventh Message Understanding Conference.*

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. *Proceedings of the Sixth Workshop on Very Large Corpora. Montreal, Canada. ACL SIGDAT.*, pages 161–170.

Hans Kamp and Uwe Reyle. 1993. *From Discourse To Logic.* Kluwer Academin Publishers.

Andrew Kehler. 1997. Probabilistic coreference in information extraction. *Proceedings of the Second Conference on Empirical Methods in NLP (EMNLP-2), Providence, RI.*

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. *Proceedings of the* 18^{th} *International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference. Montreal, Canada.*

Judea Pearl. 1988. *Probabilistic Reasoning In Intelligent Systems:Networks of Plausible Inference.* Morgan Kaufman Publishers.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. *Proceedings of the Sixth Message Understanding Conference (MUC-6). San Francisco, CA.*, pages 45–52.