A Statistical Model of Nominal Anaphora

Keith B. Hall

May 21, 2001

Abstract

In this paper, we present an analysis of nominal anaphora along with a probabilistic model of nominal anaphora resolution. Throughout this work we focus on non-pronominal anaphora. The goal of this work is to present a viable probabilistic model that is consistent with the ideas of theoretical linguistics and the results of experimental psycholinguistics. Where applicable, we present research which motivated the model derivation. This work differs from previous work in that it is completely statistical and isolates the task of non-pronominal noun-phrase coreference. Finally, we present the experimental results of an implementation of the proposed probabilistic model.

1 Introduction

1.1 Why noun-phrase coreference?

Semantic interpretation of a discourse is undoubtedly dependent on the identification discourse entities and the interactions between these entities. The problem addressed in this work is that of identifying references to unique discourse entities. The most common representation for an entity in natural languages is the noun-phrase. People use noun-phrases to introduce and refer to the entities of a discourse. In order to introduce a unique entity, we may provide extra information within a noun-phrase that uniquely defines the discourse entity. Later references to this discourse entity often vary from previous references. Information may be deleted or added to ease the resolution of the reference. There are also stylistic factors that govern the use of noun-phrases. These factors vary by domain (e.g. newspaper text vs. literature) and by speaker.

Example 1 presents a typical example of coreference found in the Wall Street Journal. In this work we will not be dealing with pronominal coreference (the circles).

Much of the theoretical linguistics and psycholinguistics research has focuses on pronominal anaphora. Chomsky's Government and Binding Theory Chomsky (1981) as well as other (Reinhart, 1981) presents a syntactic theory for pronominal anaphora. Later work (Reinhart, 1983) found problems with this strictly semantic theory. In a response to these problems many other theories

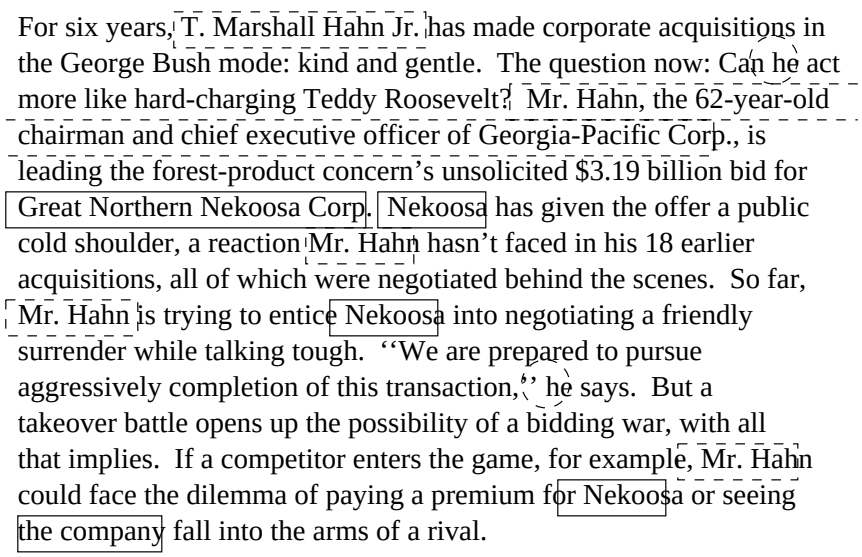

Figure 1: Example of coreference classes from the Wall Street Journal. Two of the coreferent classes are identified by the boxes/circles. The dashed boxes/circles indicate elements in the *Mr. Hahn* coreference class. Solid boxes mark the noun-phrases in the *Nekoosa* coreference class.

of anaphora have been proposed (Ariel, 1990; Ward et al., 1991; Sag & Hankamer, 1984; Hudson-D'Zmura & Tanenhaus, 1988). These theories are based on meta-linguistic (non-grammatical) features such as topicality. Additionally, these features are more suited to explaining non-pronominal coreference.

Another area of research that motivates the nominal anaphora resolution problem is that of discourse modeling. KampKamp and Reyle (1993) and others have proposed well-defined models of discourse. In this model, as with all models of discourse, they make the assumption that discourse entities can be identified.

1.2 Computational models of coreference

Although the models proposed by the computational community are at times orthogonal to theoretical or psychological models, recent computational work on anaphora has been related to, but not always consistent with, these other bodies of research.

Current work on modeling anaphora either focuses specifically on pronominal coreference or attempts to model all nominal coreference. Research presented at Message Understanding Conference (MUC) has introduced a number of non-statistical models. In general, systems presented at MUC-7 are dependent on hand crafted rules and knowledge-bases(Baldwin et al., 1997; Fukumoto et al., 1997; Garigliano et al., 1997). A dependency on world-knowledge makes these systems practical for restricted domains but unrealistic as broad coverage mod-

els. To date, the most successful nominal coreference system is the clustering model of Cardie and Wagstaff (1999).

Hand built models of pronominal coreference have been shown to perform much better than the models of full nominal coreference(Mitkov, 1998). Though, the highest accuracy comes from a simple statistical model (Ge et al., 1998; Ge, 2000).

Statistical models have been quite successful when applied to natural language processing tasks. Currently, statistical (hidden Markov model) taggers and parsers achieve the highest accuracy for these tasks (Charniak, 2000; Collins, 1997). Though a variety of statistical techniques are used, the models tend to be simpler than their non-statistical counterparts. This is definitely the case in the pronominal coreference model presented by Ge et al. (1998); Ge (2000). In this work, the pronominal coreference problem is presented under a Bayesian setting. In fact, the model we present has many similarities. We will describe the Bayesian framework while presenting our model.

1.3 Coreference resolution rather than generation

We note here that our model is a model of coreference resolution. That is to say that we are not attempting to model the generation of noun-phrases based on the preceding discourse. This should be more obvious as we describe the model. The features we use may be good indicators for identifying coreferent noun-phrases, but would not be sufficient in generating a noun-phrase.

1.4 Organization of this paper

The remaining sections present the proposed model as well as the motivation behind this particular formulation. Beginning with the feature selection, we provide both an explanation of the intuition which motivated the feature as well as references to the theoretical and phsycolinguistic work which suggest these features are appropriate. Following this, we present the statistical formulation of the proposed model. This includes a brief introduction to Bayesian inference and pointers to further readings on the topic.

The final sections discuss an implementation of the proposed model of coreference. This includes an experimental evaluation, and evaluation of evaluation metrics and results. We also investigate the nature of the errors generated by our coreference classifier.

2 Modeling nominal coreference

This model of nominal coreference can be broken in to two interdependent topics: feature selection and the statistical model. The model itself is clearly dependent on the selected features, but the reverse dependency is not as obvious. In the following section we present the model features. The dependency on our choice of model should be clear after reading this section.

We have chosen to design a model which adheres to an online processing paradigm. We are not suggesting that anaphora resolution is processed in a serial manner. What we are stating is that there are points, specifically when a noun-phrase has been read [1], where a decision about coreference relationships can be made.

Algorithm 1 Online processing of anaphora

```
for all noun-phrase do
  best_antecedent = nil
  for all antecedent < noun_phrase do
    if antecedent is more likely coreferent than best_antecedent then
      best_antecedent = antecedent
    end if
  end for
  if we have confidence in best_antecedent then
    make_coreferent(noun-phrase, antecedent)
  end if
end for
```

Algorithm 1 presents the framework for our coreference model. We process each noun-phrase looking for the antecedent which is most similar. There are two critical points to note here. First, the antecedents in algorithm 1 in our model are not antecedent noun-phrases, but classes of noun-phrases. A coreference link between a noun-phrase and an antecedent class creates a new class containing the elements of the antecedent class and the additional noun-phrase. When we begin processing, antecedent classes are made up of single noun-phrases. The second point to note regards the final if statement. Within a discourse not all noun-phrases will be anaphoric(there is only a single reference to the discourse entity). This will be most common in short discourses such as newspaper articles. Thus, there are two passes to the decision making process: determine whether the current noun-phrase is coreferent and if so, which antecedent class it is coreferent with.

3 Feature definitions

As mentioned above, algorithm 1 performs two decision-making tasks. In designing the coreference model we must identify features which assist in solving each of these tasks. In the following description of the model features we describe how the features impact these decisions.

The model features are made up from five feature classes. These classes are: head-nouns, determiners, open-class words, distance, and class size. In the following subsections we describe each of these features and the motivations behind their inclusion in the coreference model. The specific feature variables

[1] Throughout this work we will use the term reading to cover both reading a text or hearing a spoken statement.

are presented in detail in section 4.3 where they are included in the statistical model.

Eugene, the following sections have example statistics that I need to extract. Let me know if these examples seem interesting (of course, the stats will make them more interesting).

3.1 Head-nouns

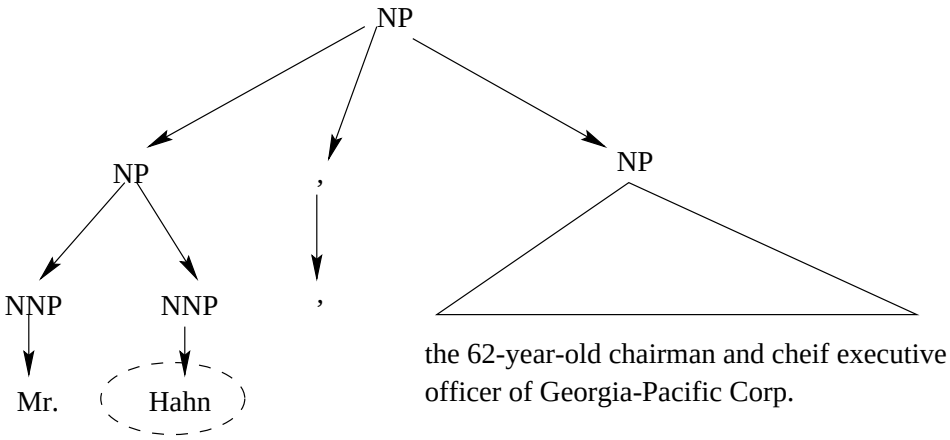

Figure 2: The head-noun is identified by the dashed circle.

The head-noun of a noun-phrase is defined as the noun which provides the most information about the noun-phrase. The standard method used to identify head-nouns is as follows. For a given noun-phrase, we choose the rightmost noun of the leftmost embedded noun-phrase. In Figure 2 we see that *Hahn* is identified as the head-noun. We call the head-noun using this standard method, the simple head-noun.

$P(+anaphoric$|head-noun = department$)$	0.884260
$P(+anaphoric$|head-noun = cars$)$	0.3530520
$P(+anaphoric$|head-noun = law$)$	0.8000780
$P(+anaphoric$|head-noun = councils$)$	0.3530520
$P(+anaphoric$|head-noun = money$)$	0.3530520
$P(+anaphoric$|head-noun = spokesman$)$	0.303890
$P(+anaphoric$|head-noun = bill$)$	0.7180480

Table 1: Example statistics for head-nouns.

We include two features in our model that make use of head-nouns. The first feature uses the head-noun of the current noun-phrase. This feature identifies head-nouns that are more likely to be found in coreferent relationships (indicated

by a $+anaphoric$) than others. In table 1 we present example statistics for this features. This examples, as well as other presented throughout this paper have been extracted from the Wall Street Journal training data explained in section 5.

$P(+corefMatch \mid +A, +\text{head-noun match})$	0.143963
$P(+corefMatch \mid +A, -\text{head-noun match})$	0.968753

Table 2: Example statistics for pairs of head-nouns.

The other head-noun feature used is based on pairs of noun-phrases. We wish to identify the likelihood of a coreferent match given the head-nouns of two noun-phrases match. In table 2 we present example statistics of this sort. We will use the abbreviated convention of a + meaning a boolean variable is true and a − meaning it is false. We will also abbreviate the variable indicating the current noun-phrase is anaphoric. We use the variable A to indicate this fact.

3.1.1 Bad head-nouns

For the task of identifying coreference relationships between noun-phrases, the simple head-noun is often useless. We believe a head-noun-like term would be very useful in identifying coreference relationships. In the following table we present a number of noun-phrases, the head-phrase and the simple head-noun. Here the head-phrase is simply a set of words that we feel are unique identifiers for the noun-phrase.

[ht] **noun-phrase**	**head-phrase**	**simple head-noun**
Perpetual preferred shares	Perpetual preferred	shares
Toronto cable television	Toronto cable television	television
General Motor Corp.'s beleaguered Buick division	Buick division	division
the vice president	vice president	president
the president	president	president
American Express Co.	American Express	Co.

In the last example we find the most problematic type of simple head-noun. This head-noun, although providing information about the noun-phrase, does little in identifying the discourse entity. We call these head-nouns **bad head-nouns**.

We have developed a simple yet sufficient technique to find these **bad head-nouns**. We note that bad head-nouns are high frequency words. Furthermore, we observer than when a bad head-noun is used, the noun to the left of the simple head-noun is more informative.

Given a current noun-phrase and some antecedent, we define the following two values. The first value we call potential matches. A potential match is when either: the simple-head nouns match, or the noun to the left of the current

noun-phrases simple head-noun matches. The second value we call the bad head match. If the current noun-phrase and antecedent noun-phrase matched only when choosing the noun to the left of the simple head noun, then this was a bad head match. We collect these counts for the head of every noun-phrase as compared with all antecedents.

Algorithm 2 Collecting counts for bad head-nouns

```
for all noun-phrase do
  best_antecedent = nil
  left_current = noun to left of current head-noun
  for all antecedent < noun_phrase do
    left_antecedent = noun to left of antecedent head-noun
    if left_current = antecedent head-noun then
      increment CountBad(current head-noun)
      increment CountMatched(current head-noun)
    end if
    if left_antecedent = current head-noun then
      increment CountBad(antecedent head-noun)
      increment CountMatched(antecedent head-noun)
    end if
    if antecedent head-noun = current head-noun then
      increment CountMatched(current head-noun)
      increment CountMatched(antecedent head-noun)
    end if
  end for
end for
```

Algorithm 2 collects the counts of cases where a noun to the left of the head-noun would have been a match with some other head-noun.

$$P^*(hisbad|matchwaspossible) \approx \frac{CountBad(h)}{CountMatched(h)} \quad (1)$$

We then threshold this value (not really a probability) for nouns with high counts. Table 3 is a list of the bad head nouns which occurred over 40,000 times in the BLLIP '99 corpus, a corpus of the Wall Street Journal articles from 1987, 1988, and part of 1989.

3.2 Determiners

Determiners, most notably the definite article *the* and indefinite article *a*, provide some clues for object reference. Baker(Baker, 1996) points out that in simple cases, the *a* is used to **register** a discourse entity and the *the* indicates references to previously registered items.

Statistical model features identifying the current noun-phrase's determiner and the interaction between the current noun-phrase's determiner and the antecedent's determiner capture these interactions. Table 4 presents examples

head-noun	Matched	Bad matches	P^*
ltd	44285	31653	0.714757
spokesman	68612	48392	0.705299
co	102388	69976	0.683439
officials	139379	91173	0.654137
executives	46525	29001	0.623342
inc	273227	162519	0.594813
corp	227156	133937	0.589626
unit	89941	41148	0.4575
stock	83542	37114	0.444256

Table 3: Bad head-nouns collected from the BLLIP '99 corpus

$P(+A \mid det_{current} = \text{No determiner})$	0.204222
$P(+A \mid det_{current} = \text{a})$	0.137967
$P(+A \mid det_{current} = \text{the})$	0.510887
$P(+coref Match \mid +A, +\text{first antecedent}, det_{current} = \text{a}, det_{antecedent} = \text{a})$	0.666195
$P(+coref Match \mid +A, -\text{first antecedent}, det_{current} = \text{a}, det_{antecedent} = \text{a})$	0.745273

Table 4: Example statistics for determiners.

statistics in much the same manner as we presented the head-noun statistics in the previous sections. Note that we have an additional feature here (+first antecedent) that indicates whether the antecedent is the first noun-phrase in the antecedent class.

3.3 Open-class words

Results from psycholinguistic research show there is a clear difference in the comprehension of open-class vs. closed-class words [2]. Child language acquisition experiments show that children learn these differences within their first two years(Shi et al., 1999; Shi et al., 1998). This phenomena is observed across languages, but are more distinctive in languages such as English. An interesting observation pertaining to our work is made in ShiShi et al. (1998) is that "functional items tend to be syntactically predictable and semantically light, carrying little information load."

In our model we ignore almost all closed-class words (we do use determiners as mentioned above). What remains are the open-class words. We assume no knowledge about the structure of these words The model simply assumes that matching open-class words provide clues to coreference.

Another observation we have made regarding open-class words is that later references to the same discourse entity use fewer words than earlier references.

[2] Open-class words are also known as lexical tokens. Closed-class are also known as grammatical or functional tokens.

Ariel's work on Accessibility Theory(Ariel, 1990) provides some justification for this observation, suggesting an economical system where smaller noun-phrases are preferred if they offer enough information. Discourse entities that are mentioned more will need less information to identify them as they are more *accessible*.

$P(+corefMatch \mid +A, -ocLEQ, -openMatch)$	0.27132
$P(+corefMatch \mid +A, +ocLEQ, -openMatch)$	0.13202
$P(+corefMatch \mid +A, +ocLEQ, +openMatch)$	0.89626
$P(+corefMatch \mid +A, +ocLEQ, +openMatch, +headMatch)$	0.0217256
$P(+corefMatch \mid +A, +ocLEQ, +openMatch, +headMatch)$	0.666720
$P(+corefMatch \mid +A, +ocLEQ, +openMatch, +headMatch)$	0.966748

Table 5: Example statistics for open-class matches.

Table 5 presents some example open-class match statistics. Note that the last few examples are the combination of open-class Matches and head-noun Matches. These combined features provide more information than when used independently.

3.4 Distance

The distance between two noun-phrases would seem to have some impact on the likelihood of the noun-phrases being coreferent. Ariel (Ariel, 1990) suggests distance as a factor in identifying coreference. Specifically, Ariel states that the choice of anaphor is dependent on the distance. In our model, we include a distance feature which is simple the surface distance between two noun-phrases.

Factors that confound this feature include discourse size and noun-phrase types. The experimental corpus we used is made up of short discourse (newspaper articles). We expect this to distort the distance effect. All noun-phrases in this corpus are annotated with coreference markings. In many cases these noun-phrases are objects of prepositions, sometimes embedded in a company name (i.e. *New York* in *The Electric Company of New York*. We believe the distance effect to be a more semantic feature that effects subject and objects but not these embedded objects.

$P(+corefMatch \mid log_2(distance) = 1)$	0.281293
$P(+corefMatch \mid log_2(distance) = 3)$	0.385731
$P(+corefMatch \mid log_2(distance) = 6)$	0.268674

Table 6: Example statistics for distance between noun-phrases and antecedent classes.

The example statistics in table 6 are based on distance buckets. These buckets help prevent data sparseness problems.

3.5 Class Size

The final feature used in our model is that of class size. The intuition here is simply that most coreference class do not grow past a particular size. This limit is clearly a domain dependent attribute. We will expect to see a single class grow larger. We assume this class is related to the topic of the discourse. Most importantly, there is a trend for most classes to stay the same size.

$P(+corefMatch \mid classSize = 2)$	0.284516
$P(+corefMatch \mid classSize = 5)$	0.491489
$P(+corefMatch \mid classSize = 15)$	0.699609
$P(+corefMatch \mid classSize = 20)$	0.803242

Table 7: Example statistics for class size.

The example statistics in table 7 show the class size effect in our training corpus.

4 Statistical model

4.1 Bayes models

Most classifications problems can be rephrased as a probabilistic maximization problem. In the general case we wish to find the most probable class i given some observed evidence $\vec{E}$.

$$\arg\max_i P(i|\vec{E})$$

Bayes rule [3] allows us the flexibility to calculate $P(i|\vec{E})$ through the reversed conditional $P(\vec{E}|i)$.

$$P(i|\vec{E}) = P(i)\frac{P(\vec{E}|i)}{P(\vec{E})}$$

The first term is know as the prior and is used to capture any *a priori* knowledge about the class i. The second term is known as the likelihood ratio.

The power of Bayes rule is most obvious when attempting to estimate the distribution $P(i|\vec{E})$. Given some training data, we are unlikely to have seen all classes i with all possible combinations of the evidence. However, we are much more likely to have observed the evidence for all possible classes.

A Naive Bayes model is one which assumes all conditional variables are conditionally independent. In other words, we are able to calculate the distribution as follows:

$$P(i|\vec{E}) = P(i)\frac{\prod_{E_j \in \vec{E}} P(E_j|i)}{P(\vec{E})}$$

[3] Bayes rules is a direct observation following from the rules of conditional probability(Feller, 1950)

The model presented in this paper is a Bayesian model which has limited conditional independence. The dependency structure of the evidence is defined by the model builder.

There are a number of references which provide a more thorough explanation of Bayesian inference. Pearl(Pearl, 1988) presents a comprehensive overview of Bayesian inference. Manning and Schtze (Manning & Schütze, 1999), Charniak(Charniak, 1993), and Jelinek (Jelinek, 1997) provide an overview of these approaches as applied to natural language processing and speech recognition. Finally, Mitchell (Mitchell, 1997) presents an introduction to Bayesian models in the general context of machine learning.

4.2 Estimating multinomial distributions

Once we have formulated our model in the Bayesian framework we must estimate the conditional distributions over the evidence variables. Typically in natural language processing we assume these variables are multinomially distributed. Informally, this means that the variables take on values, where the likelihood of taking on a particular value is determined by a specific probability (the probabilities for all values must sum to 1).

Let's look at an example related to our model. The determiner feature has a finite number of values. We can express the distribution of the value of the determiner random variable D conditioned on the coreference variable C as follows:

$$\begin{aligned} P(D = the|C = true) &= .434 \\ P(D = a|C = true) &= .029 \\ P(D = this|C = true) &= .012 \\ &\cdots \\ \sum_{d \in D} P(D = d|C = true) &= 1 \end{aligned}$$

The maximum likelihood estimate (MLE) for a multinomial distribution is maximized by relative frequencies given a large enough sample. For example, the estimate for the the first probability above is simple:

$$P(D = the|C = true) = \frac{\#(D = the, C = true)}{\#(C = true)}$$

Here the #() function is a raw count from training examples. ($\#(a, b)$ means the count examples where a and b are true.)

The first problem with this estimate arises when a value is not observed in the training data. For example, if the determiner *this* never occurs in our training data but does occur in the test data. There are a number of techniques to handle these cases (Manning & Schütze, 1999; Jelinek, 1997; Charniak, 1993). The simplest of these is the Laplace estimate. In short, this estimate assigns a small amount of probability mass to the unobserved instances.

The second problem with this estimate is that is is maximized for the training data. We would prefer an estimate which has a high generalization accuracy. There are a number of techniques known as smoothing which attempt to distribute the probability mass more smoothly over the variables. In this work, we use a linear interpolation back-off model where appropriate.

4.3 Coreference model

We now present the statistical model for noun-phrase coreference. The model follows from the discussion of Bayesian models above. Recall that there are actually two tasks at hand. The first is to determiner whether a noun-phrase is coreferent with any antecedent class. We call this the first-pass decision(though actually performed after the second-pass). If the noun-phrase does have a coreferent antecedent, the second-pass involves identifying that coreferent antecedent. We begin by deriving the second-pass equation and show how it can be used to determiner the first-pass decision.

4.3.1 Best antecedent class

From the pseudo-code of algorithm 1 we make a classification decision for each noun-phrase. We will refer to the index n of the noun-phrase N_n to mean the noun-phrase itself. Given $\{N_1, N_2, \cdots, N_{n-1}\}$ and N_n, we wish to find the best antecedent class $c \in C$ where the set C is a partition of $\{N_1, N_2, \cdots, N_{n-1}\}$ representing the coreferent classes up to the noun-phrase indexed at n. We introduce the boolean variable A whose value is true iff the noun-phrase at index n is coreferent with one of the antecedents in C. To state this probabilistically we want to maximize the following:

$$\arg\max_c P(C = c, A = true|\mathcal{E}) \tag{2}$$

The evidence vector $\mathcal{E}$ contains the evidence from the current noun-phrase n (E_n), the antecedent classes ($\{E_1, E_2, \cdots, E_{n-1}\}$), and other observations from the discourse (E_d). In our model we have combined any information from E_d into E_n and the $\{E_1, E_2, \cdots, E_{n-1}\}$.

As mentioned above, we will now apply Bayes rule to reformulate the equation.

$$\begin{aligned}
&\arg\max_c P(C = c, A = true|E_n, E_1, E_2, \cdots, E_{n-1}) \\
&= P(C = c, A = true)\frac{P(E_n, E_1, E_2, \cdots, E_{n-1}|C = c, A = true)}{P(E_n, E_1, E_2, \cdots, E_{n-1})} \qquad (3)\\
&= P(c)\frac{P(E_n|c)P(E_1, E_2, \cdots, E_{n-1}|E_n, c)}{P(E_n|c)P(E_1, E_2, \cdots, E_{n-1}|E_n)} \qquad (4)
\end{aligned}$$

Equation 4 introduces some shorthand that should make the equations more readable. We simply use the variable c to indicate $(C = c, A = true)$. Also,

in this equation we have chosen to break apart the evidence for the current noun-phrase E_n and the antecedents $\{E_1, E_2, \cdots, E_{n-1}\}$.

Our first assumption is that the evidence for one antecedent classes is independent of another. This is not a conditional independence assumption but an absolute independence assumption. This allows us some flexibility in factoring the numerator and denominator.

$$\begin{aligned} & P(c)\frac{P(E_n, E_1, E_2, \cdots, E_{n-1}|c)}{P(E_n, E_1, E_2, \cdots, E_{n-1})} \\ = \quad & P(c)\frac{P(E_c|E_n, c)\prod_{i\neq c}^{n-1} P(E_i|E_n, c)}{P(E_c|E_n)\prod_{i\neq c}^{n-1} P(E_i|E_n)} \end{aligned} \tag{5}$$

This new equation separates the positive evidence and the negative evidence. E_c is the evidence for antecedent class that we are interested in. The second half of the numerator (and the denominator) represents the negative evidence.

We will now decompose the equation $\frac{P(E_c|E_n,c)}{P(E_c|E_n)}$ into the our model features. The same decomposition can be done for the negative evidence terms.

$\mathcal{H}$ - Head nouns of N_c and N_n match

$\mathcal{D}_c$ - Determiner of N_c

$\mathcal{D}_n$ - Determiner of N_n

$\mathcal{O}$ - All open-class words in N_n found in N_c

$\mathcal{S}$ - Number of open-class words in $N_n \leq$ number open class words in N_c

Δ - Number of noun-phrases between N_n and A_c

Γ - Number of noun-phrases in A_c

$$\begin{aligned} & \frac{P(E_c|E_n, c)}{P(E_c|E_n)} \\ = \quad & \frac{P(\mathcal{H}, \mathcal{D}_c, \mathcal{O}, \mathcal{S}, \Delta, \Gamma|\mathcal{D}_n, c)}{P(\mathcal{H}, \mathcal{D}_c, \mathcal{O}, \mathcal{S}, \Delta, \Gamma|\mathcal{D}_n)} \end{aligned} \tag{6}$$

Note that many of our features are functions of two noun-phrases, yet the parameters are the current noun-phrase N_n and the antecedent class A_c. The noun-phrases within an antecedent class are clearly dependent upon one another. For example, if we see the head-noun *president* in one noun-phrase, the likelihood of seeing it in another noun-phrase within the same class is expected to be higher than seeing the head-noun *president* without class information. Unfortunately, modeling this dependency would require a training corpus that contained all possible coreference chains.

$$\begin{aligned} & \frac{P(E_c|E_n, c)}{P(E_c|E_n)} \\ = \quad & \frac{P(E_{c,1}|E_n, c)P(E_{c,2}|E_{c,1}, E_n, c)\cdots}{P(E_{c,1}|E_n)P(E_{c,2}|E_{c,1}, E_n)} \end{aligned} \tag{7}$$

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

Table 8: Hand-annotated coreference corpus information

Instead, we choose to simplify this process by evaluating equation 6 using the antecedent noun-phrases within the antecedent class. We will approximate this equation with the maximal probability over the antecedents.

$$\frac{P(E_c|E_n,c)}{P(E_c|E_n)} \approx \max_{i \in A_c} \frac{P(E_{c,i}|E_n,c)}{P(E_{c,i}|E_n)} \quad (8)$$

We take some liberties with equation 6 and insert the functions of the evidence without accounting for the dependency. In other words, substituting $P(\mathcal{H}|c)$ for i $P(h(E_c) = h(E_n)|h(E_n), c)$. We choose this backed-off version primarily due to sparse-data constraints. The same approximations can be made for the negative evidence ratio.

4.3.2 First-pass decision

Recall that the first-pass decision is whether any antecedent class is coreferent with the current noun-phrase. We do this simply by thresholding the probability that maximized equation 2.

$$P(e|\mathcal{E}) \quad (9)$$

$$\approx \max_c P(c|\mathcal{E}) \quad (10)$$

In practice we find that setting this parameter to a training-data tuned value gives adequate results.

5 Experimental analysis

We now focus on the empirical performance analysis of an implementation of the statistical coreference model described above. We implemented this model in C++ and have tested it on the Penn Treebank(Mitchell, 1997), a human-parsed set of Wall Street Journal articles. Note that the input need not be prepared by a human. We would expect similar results with data that was parsed automatically by a high accuracy parser such as that found in (Charniak, 2000). A small subset of the Penn Treebank was hand annotated with correct coreference information. This small subset is referred to as our corpus for the remainder of the paper. Table 8 presents statistics about the hand-annotated corpus, a relatively small corpus.

5.1 Experimental organization

The experimental process used here is similar to that used in many natural language processing and machine learning tasks (see (Mitchell, 1997). We use ten-fold cross-validation. This means that we first split the data into ten equally sized sets. Then we perform the experiment using nine of these sets as training data and one set as test data. We run the experiment ten times, each time using a different set as the test set. All scores presented are the average over all ten experiments.

As mentioned above, we also need to tune the first-pass threshold. We do this during the training phase mentioned above. Once we have trained our classifier, we run it on the training data with different values for the first-pass threshold. Currently we are use a coarse linear increment. We choose the value that maximizes the evaluation metric over the training data.

5.2 Evaluation metrics

Evaluating the results of our classifier is done by comparing the response (the results of our classifier) to the key (the hand-annotated data).

5.2.1 Vilain

Currently, the most common evaluation metric is the Vilain scoring algorithm(Vilain et al., 1995). This algorithm considers the response to be a permutation of the key. For example, the noun-phrases of a single class in the key may be found in many classes in the response. We count the number of links needed to create the key class. Let $c(S) = (|S| - 1)$ be the number of correct links for class S. Let $m(S) = (|p(S)| - 1)$ be the number of missing links where $|p(S)|$ is the number of partitions S is broken into in the response. Then Vilain defines recall for the class S to be as follows.

$$\frac{c(S) - m(S)}{c(S)} \tag{11}$$

$$= \frac{(|S| - 1) - (|p(S)| - 1)}{|S| - 1} \tag{12}$$

$$= \frac{|S| - |p(S)|}{|S| - 1} \tag{13}$$

Recall for all the classes is simply a sum of the individual links:

$$R = \frac{\sum_S (|S| - |p(S)|}{\sum_S (|S| - 1)} \tag{14}$$

Precision is calculated by reversing the role of the key and response.

One problem that arises from the Vilain metric is the tendency to reward large classes in the response. We have found that putting all noun-phrases in a single class results in a relatively high score. For example, in one of our cross

validated test sets we places all noun-phrases in coreference class. The recall is clearly 100%, but surprisingly the precision is 20% producing an F-measure around 34% [4].

Due to this unwanted outcome we have experimented with other evaluation metrics.

5.2.2 B-cubed

Bagga and Baldwin(Bagga & Baldwin,) suggest an alternative to the Vilain metric. This technique is similar to that of Vilain except the size of the class is used to weigh the score. The recall for a key coreference class S_i is defined as:

$$R_i = 1 - \frac{\sum_{j=1}^{m} |P_{i,j}| \times (|S_i| - |P_{i,j}|)}{|S_i|^2} \tag{15}$$

Where P_j is the the union of the j^{th} partition of S_i (created by the response) and S_i. In other words, we are only interested in the partitions of S_i created by the response. Note that calculating the scores for all coreference classes is not as easy as with the Vilain metric. The totals can be calculated as an average or a weighted average of the individual class scores.

5.2.3 Predications

Another scoring technique that we have tried is to measure the number of false predications. We consider all coreference relationships to be predications. For examples, a key S with $|S|$ noun-phrases will have $((|S|-1)\times|S|)/2$ predications. In the same manner as Vilain, we consider the response as a permutation of the key class S. The recall for a particular coreference class S_i is as follows (borrowing the notation from the section on B-cubed scoring).

$$R_i = \frac{\sum_{j=1}^{m}((|P_{i,j}| - 1) \times |P_{i,j}|)/2}{((|S_i| - 1) \times |S_i|)/2} \tag{16}$$

Scores for multiple classes can be calculated similarly to the method suggested in the previous section on B-cubed scoring.

5.2.4 Classification accuracy

Finally, we consider an evaluation of the specific task: classifying the current noun-phrase correctly given the preceding discourse context. First, we consider each noun-phrase decision as a separate classification task. In this evaluation we are only interested in how well we classify the noun-phrase (and not how well the antecedents were classified). We do this by using the preceding discourse context from the key, including all coreference information. Running our classifier using this evidence provides a useful evaluation of the classification performance.

Eugene, I plan on rerunning the experiments with the best model so far and reporting scores using all metrics.

[4] The F-measure is simply a weighted average $\frac{2*P*R}{P+R}$.

5.3 Results

	Precision	Recall	Geometric Mean	F Measure
Open-class match	54.3	53.8	54.0	54.0
Open-class match & Head-match	65.3	61.5	63.4	63.3
Full Model	69.4	61.5	65.3	65.2

Table 9: Experimental results (Vilain)

As mentioned previously, we ran our implementation of the proposed model on a small hand-annotated corpus. Table 9 presents the current scores for our model using different sets of features. The base-line model simply uses open-class comparative information. Note that the full model (excluding distance) performs approximately 11% better than the baseline (a 24% reduction in error).

5.4 Error analysis

The results found in table 9 are encouraging but not spectacular. There is a lot of room for improvement. We have performed an error analysis on the 34.8% of the errors.

We find that the majority of our errors (over 80%) are caused by incorrect first-pass decisions. This is not surprising considering only 27% of the noun-phrases are found in coreferent classes. These errors are split fairly evenly between assigning a non-coreferent noun-phrase to a class and failing to assign a coreferent noun-phrase to a class.

The results reported above are for models which excluded the distance feature. We have found the distance feature to degrade the performance of the system. Although we believe the distance feature should provide useful information, we have been unable to alleviate the problem. One possible cause for the performance degradation is that the distance feature we are currently using is a poor indicator of coreference. If the distance were completely independent of coreference we would not expect the performance degradation, rather we would expect a null effect. However, if there is a correlation between the distance and coreference, but the distance feature is dependent on other noun-phrase attributes, it is possible the current model would suffer. For example, the distance between coreferent items may be dependent on whether the coreference class is the topic or not. In our current model, the non-topic events dominate the statistics and could negatively impact decisions about topical classes. At the same time, the distance feature may not improve decisions about non-topical classes. Nonetheless, we intend to perform a further analysis of the distance feature and consequent performance degradation.

6 Further endeavors

Currently we have evaluated the performance of our model on our Wall Street Journal corpus. We plan to experiment with other corpora to determine the generalization performance of the model. The most obvious corpus to use is the Message Understanding Conference (MUC) Corpus. MUC coordinates a competition between researchers on a variety of semantic tasks. One of these tasks in the nominal coreference task. Testing our system on the MUC corpus will also provide a more competitive evaluation of the system.

Another competitive evaluation technique is to compare the performance to a different type of model. We are currently working on an experiment using a Support Vector Machine (SVM) algorithm(Vapnik, 1995; Burges, 1998). There is some evidence that discriminative methods such as SVM perform better on some classification tasks. We will use our classification metric to compare these two systems.

Finally, there is much room for model development. As noted in the discussion about the distance feature, there are many uncertainties about the current set of features. Initially, we plan to identify the strengths of each feature and identify any need for additional conditioning information. Following this we will look for other features. One obvious source of information are the grammatical cues. For example, we may find that noun-phrases modified by prepositional phrases are never modified by prepositional phrases headed by different prepositions.

7 Conclusion

We have presented a statistical model for nominal coreference. Specifically, we have looked at the non-pronominal cases which are more difficult than pronominal anaphora, especially in the absence of world knowledge. We have motivated our model design through the results from the theoretical linguistics and psycholinguistics literature.

In order to evaluate this model, we presented the experimental results of an implementation of the model. The results are encouraging and suggest the model provides a good basis for continued work in this direction.

References

Ariel, M. (1990). *Accessing noun-phrase antecedents*. Routledge, London.

Bagga, A., & Baldwin, B. Algorithms for scoring coreference chains. *Unpublished manuscript from UPenn FTP*.

Baker, C. L. (1996). *English syntax*. MIT Press.

Baldwin, B., Morton, T., Bagga, A., Baldridge, J., Chandraseker, R., Dimitriadis, A., Snyder, K., & Wolska, M. (1997). Description of the upenn camp

system as used for coreference. *Proceedings of the Seventh Message Understanding Conference*.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Knowledge Discovery and Data Mining*, *2*.

Cardie, C., & Wagstaff, K. (1999). Noun phrase coreference as clustering. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (1999)*, 82–89.

Charniak, E. (1993). *Statistical language learning*. MIT Press.

Charniak, E. (2000). A maximum-entropy-inspired parser. *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics. Seattle, WA.*, 132–139.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Forris.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. *Proceedings of the 35th Annual Meeting of the ACL. Madrid*.

Feller, W. (1950). *An introduction to probability theory and its applications*. Wiley.

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

Garigliano, R., Urbanowicz, A., & Nettleton, D. J. (1997). University of durham: Description of the lolita system as used in muc-7. *Proceedings of the Seventh Message Understanding Conference*.

Ge, N. (2000). An approach to anaphoric pronouns. *PhD Thesis, Brown University, May 2000*.

Ge, N., Hale, J., & Charniak, E. (1998). A statistical approach to anaphora resolution. *Proceedings of the Sixth Workshop on Very Large Corpora. Montreal, Canada. ACL SIGDAT.*, 161–170.

Hudson-D'Zmura, S., & Tanenhaus, M. (1988). Discourse constraints on anaphoric processing. *Centering Theory in Discourse*.

Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT Press.

Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Kluwer Academic Publishers.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

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

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems:networks of plausible inference*. Morgan Kaufman Publishers.

Reinhart, T. (1981). Definite np anaphora and c-command domains. *Linguistic Inquiry*, *12*, 605–635.

Reinhart, T. (1983). Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy*, *6*, 47–88.

Sag, I., & Hankamer, J. (1984). Toward a theory of anaphoric processing. 325–346.

Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language*, *25*, 169– 201.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*, B11–B21.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag, New York.

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

Ward, G., Sproat, R., & McKoon, G. (1991). A pragmatic analysis of so-called anaphoric islands. *Language*, *67*, 439–474.