A Statistical Approach to Anaphora Resolution

Niyu Ge, John Hale and **Eugene Charniak**
Dept. of Computer Science,
Brown University,
[nge|jth|ec]@cs.brown.edu

Abstract

This paper presents an algorithm for identifying pronominal anaphora and two experiments based upon this algorithm. We incorporate multiple anaphora resolution factors into a statistical framework — specifically the distance between the pronoun and the proposed antecedent, gender/number/animaticity of the proposed antecedent, governing head information and noun phrase repetition. We combine them into a single probability that enables us to identify the referent. Our first experiment shows the relative contribution of each source of information and demonstrates a success rate of 82.9% for all sources combined. The second experiment investigates a method for unsupervised learning of gender/number/animaticity information. We present some experiments illustrating the accuracy of the method and note that with this information added, our pronoun resolution method achieves 84.2% accuracy.

1 Introduction

We present a statistical method for determining pronoun anaphora. This program differs from earlier work in its almost complete lack of hand-crafting, relying instead on a very small corpus of Penn Wall Street Journal Tree-bank text (Marcus et al., 1993) that has been marked with co-reference information. The first sections of this paper describe this program: the probabilistic model behind it, its implementation, and its performance.

The second half of the paper describes a method for using (portions of) the aforementioned program to learn automatically the typical gender of English words, information that is itself used in the pronoun resolution program. In particular, the scheme infers the gender of a referent from the gender of the pronouns that refer to it and selects referents using the pronoun anaphora program. We present some typical results as well as the more rigorous results of a blind evaluation of its output.

2 A Probabilistic Model

There are many factors, both syntactic and semantic, upon which a pronoun resolution system relies. (Mitkov (1997) does a detailed study on factors in anaphora resolution.) We first discuss the training features we use and then derive the probability equations from them.

The first piece of useful information we consider is the distance between the pronoun and the candidate antecedent. Obviously the greater the distance the lower the probability. Secondly, we look at the syntactic situation in which the pronoun finds itself. The most well studied constraints are those involving reflexive pronouns. One classical approach to resolving pronouns in text that takes some syntactic factors into consideration is that of Hobbs (1976). This algorithm searches the parse tree in a left-to-right, breadth-first fashion that obeys the major reflexive pronoun constraints while giving a preference to antecedents that are closer to the pronoun. In resolving inter-sentential pronouns, the algorithm searches the previous sentence, again in left-to-right, breadth-first order. This implements the observed preference for subject position antecedents.

Next, the actual words in a proposed noun-phrase antecedent give us information regarding the gender, number, and animaticity of the proposed referent. For example:

> ***Marie*** *Giraud carries historical significance as one of the last women to be executed in France.* ***She*** *became an abortionist because it enabled her to*

buy jam, cocoa and other war-rationed goodies.

Here it is helpful to recognize that "Marie" is probably female and thus is unlikely to be referred to by "he" or "it". Given the words in the proposed antecedent we want to find the probability that it is the referent of the pronoun in question. We collect these probabilities on the training data, which are marked with reference links. The words in the antecedent sometimes also let us test for number agreement. Generally, a singular pronoun cannot refer to a plural noun phrase, so that in resolving such a pronoun any plural candidates should be ruled out. However a singular noun phrase can be the referent of a plural pronoun, as illustrated by the following example:

> *"I think if I tell **Viacom** I need more time, **they** will take 'Cosby' across the street," says the general manager of a network affiliate.*

It is also useful to note the interaction between the head constituent of the pronoun ρ and the antecedent. For example:

> *A Japanese company might make television picture tubes in Japan, assemble the TV sets in Malaysia and export them to Indonesia.*

Here we would compare the degree to which each possible candidate antecedent (*A Japanese company, television picture tubes, Japan, TV sets,* and *Malaysia* in this example) could serve as the direct object of "export". These probabilities give us a way to implement selectional restriction. A canonical example of selectional restriction is that of the verb "eat", which selects food as its direct object. In the case of "export" the restriction is not as clearcut. Nevertheless it can still give us guidance on which candidates are more probable than others.

The last factor we consider is referents' *mention count.* Noun phrases that are mentioned repeatedly are preferred. The training corpus is marked with the number of times a referent has been mentioned up to that point in the story. Here we are concerned with the probability that a proposed antecedent is correct given that it has been repeated a certain number of times. In effect, we use this probability information to identify the topic of the segment with the belief that the topic is more likely to be referred to by a pronoun. The idea is similar to that used in the centering approach (Brennan et al., 1987) where a continued topic is the highest-ranked candidate for pronominalization.

Given the above possible sources of information, we arrive at the following equation, where $F(\rho)$ denotes a function from pronouns to their antecedents:

$$F(\rho) = \arg\max_a P(A(\rho) = a|\rho, h, \vec{W}, t, l, s_\rho, \vec{d}, \vec{M})$$

where $\mathrm{A}(\rho)$ is a random variable denoting the referent of the pronoun ρ and a is a proposed antecedent. In the conditioning events, h is the head constituent above ρ, $\vec{W}$ is the list of candidate antecedents to be considered, t is the type of phrase of the proposed antecedent (always a noun-phrase in this study), l is the type of the head constituent, s_ρ describes the syntactic structure in which ρ appears, $\vec{d}$ specifies the distance of each antecedent from ρ and $\vec{M}$ is the number of times the referent is mentioned. Note that $\vec{W}$, $\vec{d}$, and $\vec{M}$ are vector quantities in which each entry corresponds to a possible antecedent. When viewed in this way, a can be regarded as an index into these vectors that specifies which value is relevant to the particular choice of antecedent.

This equation is decomposed into pieces that correspond to all the above factors but are more statistically manageable. The decomposition makes use of Bayes' theorem and is based on certain independence assumptions discussed below.

$$\begin{aligned}
&P(A(\rho) = a|\rho, h, \vec{W}, t, l, s_\rho, \vec{d}, \vec{M}) \\
&= \frac{P(a|\vec{M})P(\rho, h, \vec{W}, t, l, s_\rho, \vec{d}|a, \vec{M})}{P(\rho, h, \vec{W}, t, l, s_\rho, \vec{d}|\vec{M})} \qquad (1)\\
&\propto P(a|\vec{M})P(\rho, h, \vec{W}, t, l, s_\rho, \vec{d}|a, \vec{M}) \qquad (2)\\
&= P(a|\vec{M})P(s_\rho, \vec{d}|a, \vec{M}) \\
&\qquad P(\rho, h, \vec{W}, t, l|a, \vec{M}, s_\rho, \vec{d}) \qquad (3)\\
&= P(a|\vec{M})P(s_\rho, \vec{d}|a, \vec{M}) \\
&\qquad P(h, t, l|a, \vec{M}, s_\rho, \vec{d}) \\
&\qquad P(\rho, \vec{W}|a, \vec{M}, s_\rho, \vec{d}, h, t, l) \qquad (4)\\
&\propto P(a|\vec{M})P(s_\rho, \vec{d}|a, \vec{M})
\end{aligned}$$

$$
\begin{aligned}
& P(\rho, \vec{W}|a, \vec{M}, s_\rho, \vec{d}, h, t, l) && (5)\\
= \; & P(a|\vec{M}) P(s_\rho, \vec{d}|a, \vec{M}) \\
& P(\vec{W}|a, \vec{M}, s_\rho, \vec{d}, h, t, l) \cdot && (6)\\
& P(\rho|a, \vec{M}, s_\rho, \vec{d}, h, t, l, \vec{W}) \\
\propto \; & P(a|\vec{M}) P(d_H|a) P(\vec{W}|h, t, l, a) \\
& P(\rho|w_a) && (7)
\end{aligned}
$$

Equation (1) is simply an application of Bayes' rule. The denominator is eliminated in the usual fashion, resulting in equation (2). Selectively applying the chain rule results in equations (3) and (4). In equation (4), the term $P(h,t,l|a,\vec{M},s_\rho,\vec{d})$ is the same for every antecedent and is thus removed. Equation (6) follows when we break the last component of (5) into two probability distributions. In equation (7) we make the following independence assumptions:

- Given a particular choice of the antecedent candidates, the distance is independent of distances of candidates other than the antecedent (and the distance to non-referents can be ignored):

  $$P(s_\rho, \vec{d}|a, \vec{M}) \propto P(s_\rho, d_a|a, \vec{M})$$

- The syntactic structure s_ρ and the distance from the pronoun d_a are independent of the number of times the referent is mentioned. Thus

  $$P(s_\rho, d_a|a, \vec{M}) = P(s_\rho, d_a|a)$$

  Then we combine s_ρ and d_a into one variable d_H, *Hobbs distance*, since the Hobbs algorithm takes both the syntax and distance into account.

- The words in the antecedent depend only on the parent constituent h, the type of the words t, and the type of the parent l. Hence

  $$P(\vec{W}|a, \vec{M}, s_\rho, \vec{d}, h, t, l) = P(\vec{W}|h, t, l, a)$$

- The choice pronoun depends only on the words in the antecedent, i.e.

  $$P(\rho|a, \vec{M}, s_\rho, \vec{d}, h, t, l, \vec{W}) = P(\rho|a, \vec{W})$$

- If we treat a as an index into the vector $\vec{W}$, then $(a, \vec{W})$ is simply the ath candidate in the list $\vec{W}$. We assume the selection of the pronoun is independent of the candidates other than the antecedent. Hence

  $$P(\rho|a, \vec{W}) = P(\rho|w_a)$$

Since $\vec{W}$ is a vector, we need to normalize $P(\vec{W}|h,t,l,a)$ to obtain the probability of each element in the vector. It is reasonable to assume that the antecedents in $\vec{W}$ are independent of each other; in other words, $P(w_{a+1}|w_a,h,t,l,a) = P(w_{a+1}|h,t,l,a)$. Thus,

$$P(\vec{W}|h, t, l, a) = \prod_{i=1}^{n} P(w_i|h, t, l, a)$$

where

$$P(w_i|h, t, l, a) = P(w_i|t) \text{ if } i \neq a$$

and

$$P(w_i|h, t, l, a) = P(w_a|h, t, l) \text{ if } i = a$$

Then we have,

$$P(\vec{W}|h, t, l, a) = P(w_1|t) \ldots P(w_a|h, t, l) \ldots P(w_n|t)$$

To get the probability for each candidate, we divide the above product by:

$$
\begin{aligned}
& P(\vec{W}|h, t, l, a) \\
& \quad \propto \frac{P(w_1|t) \ldots P(w_a|h, t, l) \ldots P(w_n|t)}{P(w_1|t) \ldots P(w_a|t) \ldots P(w_n|t)} \\
& \quad = \frac{P(w_a|h, t, l)}{P(w_a|t)}
\end{aligned}
$$

Now we arrive at the final equation for computing the probability of each proposed antecedent:

$$
\begin{aligned}
& P(A(\rho) = w_a) && (8)\\
& \quad \propto \; P(d_H|a) P(\rho|w_a) \frac{P(w_a|h, t, l)}{P(w_a|t)} P(a|m_a)
\end{aligned}
$$

We obtain $P(d_H|a)$ by running the Hobbs algorithm on the training data. Since the training corpus is tagged with reference information, the probability $P(\rho|w_a)$ is easily obtained. In building a statistical parser for the Penn Tree-bank various statistics have been collected

(Charniak, 1997), two of which are $P(w_a|h,t,l)$ and $P(w_a|t,l)$. To avoid the sparse-data problem, the heads h are clustered according to how they behave in $P(w_a|h,t,l)$. The probability of w_a is then computed on the basis of h's cluster $\mathbf{c}(h)$. Our corpus also contains referents' repetition information, from which we can directly compute $P(a|m_a)$. The four components in equation (8) can be estimated in a reasonable fashion. The system computes this product and returns the antecedent w_a for a pronoun ρ that maximizes this probability. More formally, we want the program to return our antecedent function $F(\rho)$, where

$$\begin{aligned} &F(\rho) \\ &= \arg\max_a P(A(\rho)=a|\rho,h,\vec{W},t,l,s_\rho,\vec{d},\vec{M}) \\ &= \arg\max_{w_a} P(d_H|a)P(\rho|w_a) \\ &\qquad \frac{P(w_a|h,t,l)}{P(w_a|t,l)}P(a|m_a) \end{aligned}$$

3 The Implementation

We use a small portion of the Penn Wall Street Journal Tree-bank as our training corpus. From this data, we collect the three statistics detailed in the following subsections.

3.0.1 The Hobbs algorithm

The Hobbs algorithm makes a few assumptions about the syntactic trees upon which it operates that are not satisfied by the tree-bank trees that form the substrate for our algorithm. Most notably, the Hobbs algorithm depends on the existence of an $\bar{N}$ parse-tree node that is absent from the Penn Tree-bank trees. We have implemented a slightly modified version of Hobbs algorithm for the Tree-bank parse trees. We also transform our trees under certain conditions to meet Hobbs' assumptions as much as possible. We have not, however, been able to duplicate exactly the syntactic structures assumed by Hobbs.

Once we have the trees in the proper form (to the degree this is possible) we run Hobbs' algorithm repeatedly for each pronoun until it has proposed n (= 15 in our experiment) candidates. The ith candidate is regarded as occurring at "Hobbs distance" $d_H = i$. Then the probability $P(d_H = i|a)$ is simply:

$$\begin{aligned} &P(d_H = i|a) \\ &= \frac{|\ \textit{correct antecedent at Hobbs distance } i\ |}{|\ \textit{correct antecedents}\ |} \end{aligned}$$

We use $|\ x\ |$ to denote the number of times x is observed in our training set.

3.1 The gender/animaticity statistics

After we have identified the correct antecedents it is a simple counting procedure to compute $P(\rho|w_a)$ where w_a is in the correct antecedent for the pronoun ρ (Note the pronouns are grouped by their gender):

$$P(\rho|w_a) = \frac{|\ w_a \textit{in the antecedent for } \rho\ |}{|\ w_a\ |}$$

When there are multiple relevant words in the antecedent we apply the likelihood test designed by Dunning (1993) on all the words in the candidate NP. Given our limited data, the Dunning test tells which word is the most informative, call it w_j, and we then use $P(\rho|w_j)$.

3.1.1 The mention count statistics

The referents range from being mentioned only once to begin mentioned 120 times in the training examples. Instead of computing the probability for each one of them we group them into "buckets", so that m_a is the bucket for the number of times that a is mentioned. We also observe that the position of a pronoun in a story influences the mention count of its referent. In other words, the nearer the end of the story a pronoun occurs, the more probable it is that its referent has been mentioned several times. We measure position by the sentence number, j. The method to compute this probability is:

$$P(a|m_a,j) = \frac{|\ a \textit{ is antecedent}, m_a, j\ |}{|\ m_a, j\ |}$$

(We omitted j from equations (1-7) to reduce the notational load.)

3.2 Resolving Pronouns

After collecting the statistics on the training examples, we run the program on the test data. For any pronoun we collect n(= 15 in the experiment) candidate antecedents proposed by Hobbs' algorithm. It is quite possible that a word appears in the test data that the program never saw in the training data and fow which it hence has no $P(\rho|w_a)$ probability. In this case

Probability Model	Percent Correct	Standard Deviation	Significance Level
P(d_H)	65.3%	0.061	-
$P(\rho \vert w_a)$	75.7%	0.039	< 0.005
$P(w_a \vert h,t,l)$	77.9%	0.046	> 0.1
$P(a \vert m_a)$	82.9%	0.042	> 0.01 < 0.025

Table 1: Cross-validation: incremental results

we simply use the prior probability of the pronoun $P(\rho)$. From the parser project mentioned earlier, we obtain the probability $\frac{P(w_a|h,t,l)}{P(w_a|t,l)}$. Finally, we extract the mention count number associated with each candidate NP, which is used to obtain $P(a|m_a)$. The four probabilities are multiplied together. The procedure is repeated for each proposed NP in $\vec{W}$ and the one with the highest combined probability is selected as the antecedent.

4 The Experiment

The algorithm has two modules. One collects the statistics on the training corpus required by equation (8) and the other uses these probabilities to resolve pronouns in the test corpus.

Our data consists of 93,931 words (3975 sentences) and contains 2477 pronouns, 1371 of which are singular (*he, she* and *it*). The corpus is manually tagged with reference indices and referents' repetition numbers. The result presented here is the accuracy of the program in finding antecedents for *he, she*, and *it* and their various forms (e.g. *him*, *his*, *himself*, etc.) The cases where "it" is merely a dummy subject in a cleft sentence (example 1) or has conventional unspecified referents (example 2) are excluded from computing the precision:

- Example 1: ***It*** is very hard to justify paying a silly price for Jaguar if an out-and-out bidding war were to start now.
- Example 2: ***It*** is raining.

We performed a ten-way cross-validation where we reserved 10% of the corpus for testing and used the remaining 90% for training. Our preliminary results are shown in the last line of Table 1.

We are also interested in finding the relative importance of each probability (i.e. each of the four factors in equation (8) in pronoun resolution. To this end, we ran the program "incrementally", each time incorporating one more probability. The results are shown in Table 1 (all obtained from cross-validation). The last column of Table 1 contains the *p-values* for testing the statistical significance of each improvement.

Due to relatively large differences between Tree-bank parse trees and Hobbs' trees, our Hobbs' implementation does not yield as high an accuracy as it would have if we had had perfect Hobbs' tree representations. Since the Hobbs' algorithm serves as the base of our scheme, we expect the accuracy to be much higher with more accurately transformed trees. We also note that the very simple model that ignores syntax and takes the last mentioned noun-phrase as the referent performs quite a bit worse, about 43% correct. This indicates that syntax does play a very important role in anaphora resolution.

We see a significant improvement after the word knowledge is added to the program. The $P(\rho|w_a)$ probability gives the system information about gender and animaticity. The contribution of this factor is quite significant, as can be seen from Table 1. The impact of this probability can be seen more clearly from another experiment in which we tested the program (using just Hobbs distance and gender information) on the training data. Here the program can be thought of having "perfect" gender/animaticity knowledge. We obtained a success rate of 89.3%. Although this success rate overstates the effect, it is a clear indication that knowledge of a referent's gender and animaticity is essential to anaphora resolution.

We hoped that the knowledge about the governing constituent would, like gender and animaticity, make a large contribution. To our surprise, the improvement is only about 2.2%. This is partly because selection restrictions are not clearcut in many cases. Also, some head verbs are too general to restrict the selection of any NP. Examples are "is" and "has", which appear frequently in Wall Street Journal: these verbs are not "selective" enough and the associated probability is not strong enough to rule out erroneous candidates. Sparse data also causes a problem in this statistic. Consequently, we observe a relatively small enhancement to the system.

The *mention* information gives the system some idea of the story's focus. The more frequently an entity is repeated, the more likely it is to be the topic of the story and thus to be a candidate for pronominalization. Our results show that this is indeed the case. References by pronouns are closely related to the topic or the center of the discourse. NP repetition is one simple way of approximately identifying the topic. The more accurately the topic of a segment can be identified, the higher the success rate we expect an anaphora resolution system can achieve.

5 Unsupervised Learning of Gender Information

The importance of gender information as revealed in the previous experiments caused us to consider automatic methods for estimating the probability that nouns occurring in a large corpus of English text deonote inanimate, masculine or feminine things. The method described here is based on simply counting co-occurrences of pronouns and noun phrases, and thus can employ any method of analysis of the text stream that results in referent/pronoun pairs (cf. (Hatzivassiloglou and McKeown, 1997) for another application in which no explicit indicators are available in the stream). We present two very simple methods for finding referent/pronoun pairs, and also give an application of a salience statistic that can indicate how confident we should be about the predictions the method makes. Following this, we show the results of applying this method to the 21-million-word 1987 Wall Street Journal corpus using two different pronoun reference strategies of varying sophistication, and evaluate their performance using honorifics as reliable gender indicators.

The method is a very simple mechanism for harvesting the kind of gender information present in discourse fragments like "Kim slept. She slept for a long time." Even if Kim's gender was unknown before seeing the first sentence, after the second sentence, it is known.

The probability that a referent is in a particular gender class is just the relative frequency with which that referent is referred to by a pronoun ρ that is part of that gender class. That is, the probability of a referent *ref* being in gender class gc_i is

$$P(ref \in gc_i) = \frac{|\text{ refs to } ref \text{ with } \rho \in gc_i \,|}{\sum_j |\text{ refs to } ref \text{ with } \rho \in gc_j \,|} \quad (9)$$

In this work we have considered only three gender classes, masculine, feminine and inanimate, which are indicated by their typical pronouns, HE, SHE, and IT. However, a variety of pronouns indicate the same class: Plural pro-

pronoun	*gender class*
he,himself,him,his	HE
she,herself,her,hers	SHE
it,itself,its	IT

nouns like "they" and "us" reveal no gender information about their referent and consequently aren't useful, although this might be a way to learn pluralization in an unsupervised manner.

In order to gather statistics on the gender of referents in a corpus, there must be some way of identifying the referents. In attempting to bootstrap lexical information about referents' gender, we consider two strategies, both completely blind to any kind of semantics.

One of the most naive pronoun reference strategies is the "previous noun" heuristic. On the intuition pronouns closely follow their referents, this heuristic simply keeps track of the last noun seen and submits that noun as the referent of any pronouns following. This strategy is certainly simple-minded but, as noted earlier, it achieves an accuracy of 43%.

In the present system, a statistical parser is used (see (Charniak, 1997)) simply as a tagger. This apparent parser overkill is a control to ensure that the part-of-speech tags assigned to words are the same when we use the previous noun heuristic and the Hobbs algorithm, to which we wish to compare the previous noun method. In fact, the only part-of-speech tags necessary are those indicating nouns and pronouns.

Obviously a much superior strategy would be to apply the anaphora-resolution strategy from previous sections to finding putative referents. However, we chose to use only the Hobbs distance portion thereof. We do not use the "mention" probabilities $P(a|m_a)$, as they are not given in the unmarked text. Nor do we use the gender/animiticity information gathered from the much smaller hand-marked text, both because we were interested in seeing what unsupervised learning could accomplish, and because we were concerned with inheriting strong biases from the limited hand-marked data. Thus our second method of finding the pronoun/noun co-occurrences is simply to parse the text and then assume that the noun-phrase at Hobbs distance one is the antecedent.

Given a pronoun resolution method and a corpus, the result is a set of pronoun/referent pairs. By collating by referent and abstracting away to the gender classes of pronouns, rather than individual pronouns, we have the relative frequencies with which a given referent is referred to by pronouns of each gender class. We will say that the gender class for which this relative frequency is the highest is the gender class to which the referent most probably belongs.

However, any syntax-only pronoun resolution strategy will be wrong some of the time – these methods know nothing about discourse boundaries, intentions, or real-world knowledge. We would like to know, therefore, whether the pattern of pronoun references that we observe for a given referent is the result of our supposed "hypothesis about pronoun reference" – that is, the pronoun reference strategy we have provisionally adopted in order to gather statistics – or whether the result of some other unidentified process.

This decision is made by ranking the referents by log-likelihood ratio, termed salience, for each referent. The likelihood ratio is adapted from Dunning (1993, page 66) and uses the raw frequencies of each pronoun class in the corpus as the null hypothesis, $\Pr(gc0_i)$ as well as $\Pr(ref \in gc_i)$ from equation 9.

$$\begin{aligned}&\text{salience}(ref)\\ &= -2\log\left(\frac{\prod_i gc0_i^{|\,gc_i\,|}}{\prod_i \mathrm{P}(ref \in gc_i)^{|\,ref,\,gc_i\,|}}\right)\end{aligned}$$

Making the unrealistic simplifying assumption that references of one gender class are completely independent of references for another classes[1], the likelihood function in this case is just the product over all classes of the probabilities of each class of reference to the power of the number of observations of this class.

6 Evaluation

We ran the program on 21 million words of Wall Street Journal text. One can judge the program informally by simply examining the results and determining if the program's gender decisions are correct (occasionally looking at the text for difficult cases). Figure 1 shows the 43 noun phrases with the highest salience figures (run using the Hobbs algorithm). An examination of these show that all but three are correct. (The three mistakes are "husband," "wife," and "years." We return to the significance of these mistakes later.)

As a measure of the utility of these results, we also ran our pronoun-anaphora program with these statistics added. This achieved an accuracy rate of 84.2%. This is only a small improvement over what was achieved without the data. We believe, however, that there are ways to improve the accuracy of the learning method and thus increase its influence on pronoun anaphora resolution.

Finally we attempted a fully automatic direct test of the accuracy of both pronoun methods for gender determination. To that end, we devised a more objective test, useful only for scoring the subset of referents that are names of people. In particular, we assume that any noun-phrase with the honorifics "Mr.", "Mrs." or "Ms." may be confidently assigned to gender classes HE, SHE, and SHE, respectively. Thus we compute precision as follows:

$$\text{precision} = \frac{|\ r \text{ attrib. as } \textsc{he} \wedge \text{ Mr.} \in r\ | + |\ r \text{ attrib. as } \textsc{she} \wedge \text{ Mrs. or Ms.} \in r\ |}{|\ \text{Mr., Mrs., or Ms.} \in r\ |}$$

Here r varies over referent types, not tokens.

The precision score computed over all phrases containing any of the target honorifics are 66.0%

[1] In effect, this is the same as admitting that a referent can be in different gender classes across different observations.

Word	count(salience)	p(he)	p(she)	p(it)
COMPANY	7052(1629.39)	0.0764	0.0060	0.9174
WOMAN	250(368.267)	0.172	0.708	0.12
PRESIDENT	931(356.539)	0.8206	0.0139	0.1654
GROUP	1096(287.319)	0.0602	0.0054	0.9343
MR. REAGAN	534(270.8)	.882022	0.0037	0.1142
MAN	441(202.102)	0.8480	0.0385	0.1133
PRESIDENT REAGAN	455(194.928)	0.8439	0.0043	0.1516
GOVERNMENT	1220(194.187)	0.1172	0.0122	0.8704
U.S.	969(188.468)	0.1021	0.0041	0.8937
BANK	816(161.23)	0.0955	0.0073	0.8970
MOTHER	113(161.204)	0.3008	0.6548	0.0442
COL. NORTH	258(158.692)	0.9263	0.0077	0.0658
MOODY	383(152.405)	0.0078	0.0052	0.9869
SPOKESWOMAN	139(145.627)	0.1223	0.5827	0.2949
MRS. AQUINO	73(142.223)	0.0958	0.8356	0.0684
MRS. THATCHER	68(128.306)	0.0735	0.8235	0.1029
GM	513(119.664)	0.0779	0.0038	0.9181
PLAN	514(111.134)	0.0856	0.0058	0.9085
MR. GORBACHEV	205(108.776)	0.8926	0.0048	0.1024
JUDGE BORK	212(108.746)	0.8820	0	0.1179
HUSBAND	91(107.438)	0.3626	0.5714	0.0659
JAPAN	450(100.727)	0.0755	0.0111	0.9133
AGENCY	476(97.4016)	0.0840	0.0147	0.9012
WIFE	153(93.7485)	0.6143	0.2875	0.0980
DOLLAR	621(90.8963)	0.1304	0.0096	0.8599
STANDARD POOR	200(90.1062)	0	0	1
FATHER	146(89.4178)	0.8082	0.1438	0.0479
UTILITY	242(87.1821)	0.0247	0	0.9752
MR. TRUMP	129(86.5345)	0.9457	0.0077	0.0465
MR. BAKER	187(84.2796)	0.8556	0.0053	0.1390
IBM	316(82.4361)	0.0696	0	0.9303
MAKER	224(82.252)	0.0223	0	0.9776
YEARS	1055(82.1632)	0.5298	0.0815	0.3886
MR. MEESE	166(82.1007)	0.8734	0	0.1265
BRAZIL	285(79.7311)	0.0596	0	0.9403
SPOKESMAN	665(78.3441)	0.6075	0.0045	0.3879
MR. SIMON	105(72.6446)	0.9523	0	0.0476
DAUGHTER	47(71.3863)	0.2340	0.7021	0.0638
FORD	249(71.3603)	0.0562	0	0.9437
MR. GREENSPAN	120(68.7807)	0.9083	0	0.0916
AT&T	198(67.9668)	0.0252	0.0050	0.9696
MINISTER	125(67.7475)	0.864	0.064	0.072
JUDGE	239(67.5899)	0.7154	0.0836	0.2008

Figure 1: Top 43 noun phrases according to salience

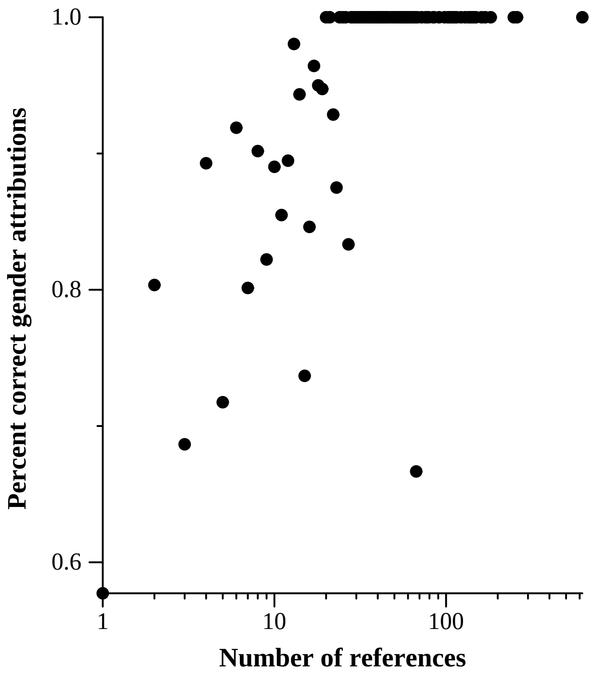

Figure 2: Precision using honorific scoring scheme with syntactic Hobbs algorithm

for the last-noun method and 70.3% for the Hobbs method.

There are several things to note about these results. First, as one might expect given the already noted superior performance of the Hobbs scheme over last-noun, Hobbs also performs better at determining gender. Secondly, at first glance,the 70.3% accuracy of the Hobbs method is disappointing, only slightly superior to the 65.3% accuracy of Hobbs at finding correct referents. It might have been hoped that the statistics would make things considerably more accurate.

In fact, the statistics do make things considerably more accurate. Figure 2 shows average accuracy as a function of number of references for a given referent. It can be seen that there is a significant improvement with increased referent count. The reason that the average over all referents is so low is that the counts on referents obey Zipf's law, so that the mode of the distribution on counts is one. Thus the 70.3% overall accuracy is a mix of relatively high accuracy for referents with counts greater than one, and relatively low accuracy for referents with counts of exactly one.

7 Previous Work

The literature on pronoun anaphora is too extensive to summarize, so we concentrate here on corpus-based anaphora research.

Aone and Bennett (1996) present an approach to an automatically trainable anaphora resolution system. They use Japanese newspaper articles tagged with discourse information as training examples for a machine-learning algorithm which is the C4.5 decision-tree algorithm by Quinlan (1993). They train their decision tree using *(anaphora, antecedent)* pairs together with a set of feature vectors. Among the 66 features are lexical, syntactic, semantic, and positional features. Their Machine Learning-based Resolver (MLR) is trained using decision trees with 1971 anaphoras (excluding those referring to multiple discontinuous antecedents) and they report an average success rate of 74.8%.

Mitkov (1997) describes an approach that uses a set of factors as constraints and preferences. The constraints rule out implausible candidates and the preferences emphasize the selection of the most likely antecedent. The system is not entirely "statistical" in that it consists of various types of rule-based knowledge — syntactic, semantic, domain, discourse, and heuristic. A statistical approach is present in the discourse module only where it is used to determine the probability that a noun (verb) phrase is the center of a sentence. The system also contains domain knowledge including the domain concepts, specific list of subjects and verbs, and topic headings. The evaluation was conducted on 133 paragraphs of annotated Computer Science text. The results show an accuracy of 83% for the 512 occurrences of *it*.

Lappin and Leass (1994) report on a (essentially non-statistical) approach that relies on salience measures derived from syntactic structure and a dynamic model of attentional state. The system employs various constraints for NP-pronoun non-coreference within a sentence. It also uses person, number, and gender features for ruling out anaphoric dependence of a pronoun on an NP. The algorithm has a sophisticated mechanism for assigning values to several salience parameters and for computing global salience values. A blind test was conducted on manual text containing 360 pronoun occurrences; the algorithm successfully identified the antecedent of the pronoun in 86% of these pronoun occurrences. The addition of a module that contributes statistically measured lexical preferences to the range of factors the algorithm considers improved the performance by 2%.

8 Conclustion and Future Research

We have presented a statistical method for pronominal anaphora that achieves an accuracy of 84.2%. The main advantage of the method is its essential simplicity. Except for implementing the Hobbs referent-ordering algorithm, all other system knowledge is imbedded in tables giving the various component probabilities used in the probability model.

We believe that this simplicity of method will translate into comparative simplicity as we improve the method. Since the research described herein we have thought of other influences on anaphora resolution and their statistical correlates. We hope to include some of them in future work.

Also, as indicated by the work on unsupervised learning of gender information, there is a growing arsenal of learning techniques to be applied to statistical problems. Consider again the three high-salience words to which our unsupervised learning program assigned incorrect gender: "husband", "wife", and "years." We suspect that had our pronoun-assignment method been able to use the topic information used in the complete method, these might well have been decided correctly. That is, we suspect that "husband", for example, was decided incorrectly because the topic of the article was the woman, there was a mention of her "husband," but the article kept on talking about the woman and used the pronoun "she." While our simple program got confused, a program using better statistics might not have. This too is a topic for future research.

9 Acknowledgements

The authors would like to thank Mark Johnson and other members of the Brown NLP group for many useful ideas and NSF and ONR for support (NSF grants IRI-9319516 and SBR-9720368, ONR grant N0014-96-1–0549).

References

Chinatsu Aone and Scott William Bennett, 1996. *Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies*, pages 302–315. Springer.

Susan E. Brennan, Marilyn Walker Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proc. 25th Annual Meeting of the ACL*, pages 155–162. Association of Computational Linguistics.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th National Conference on Artificial Intelligence*, Menlo Park, CA. AAAI Press/MIT Press.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), March.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proc. 35th Annual Meeting of the ACL*, pages 174–181. Association of Computational Linguistics.

Jerry R. Hobbs. 1976. Pronoun resolution. Technical Report 76-1, City College, New York.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, pages 535–561.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19:313–330.

Ruslan Mitkov. 1997. Factors in anaphora resolution: they are not the only things that matter. a case study based on two different approaches. In *Proceedings of the ACL '97/EACL '97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution.*

J. Ross Quinlan. 1993. *C4.5 Programs for Machine Learning.* Morgan Kaufmann Publishers.