# CS224n – Final Project

**Joseph Baker-Malone and David Christie**

## 1. Theory and Approach

In this project, we attempted to build an anaphora resolver for parsed sentences – i.e. a system that takes a tree of parsed sentences, and links each pronoun to its antecedent noun phrase. We based our approach on the work of Ge et. al [1], which was in turn based on the work of Hobbs [2]. We limit our testing to pronouns and do not attempt to resolve references of non-pronoun words. We also limit our limit our search to pronouns referring to words occurring no more than one sentence earlier in the text. This limitation is lessened somewhat by our reference-assigning method, discussed in the next section.

### 1.1 Corpus

For our training and test data, we used the Brown Laboratory for Linguistic Information Processing (BLLIP) corpus, which contains parsed sentences as well as co-referential information. This corpus is the output of Ge et al's parsing and anaphora resolution algorithm, and thus may not be as accurate as a hand-parsed and tagged corpus. However, it was the only such parsed and tagged corpus available to us, and after evaluation we found it to be accurate enough for our purposes.

As is often the case in such corpora, all phrases referring to the same entity are labeled with the same tag. We only attempt to perform anaphora resolution on pronouns, and so for our purposes we consider each pronoun to refer to the previous non-pronoun phrase with the same tag. While this may be a somewhat nonstandard approach, we believe it is appropriate and useful in this situation. It allows us to limit the search space for each resolution, and improves the performance of the Hobbs algorithm especially. Additionally, it allows us to perform resolution on a larger percentage of the pronouns in the testing corpus from only 58% to over 73%.

### 1.2 The Hobbs Algorithm

The first step to determining the antecedent of a pronoun is to collect a list of possible antecedent candidates. We followed Ge's footsteps in this respect using the Hobbs algorithm. Hobbs algorithm starts by "intelligently" walking through the tree structure of the sentence containing the pronoun, looking for noun phrases that could be possible antecedents, and adding them to a list of candidates. For details of the 9 steps in the Hobbs algorithm, we will refer the reader to reference [2], or to the commented Java code in the method Hobbs.RunHobbs. But in a nutshell, the algorithm iteratively uses multiple breadth-first searches in the tree. It first handles reflexive pronouns by looking at the same depth in the tree where the pronoun resides. It then systematically works its way up the tree, proposing possible NP or S candidate nodes.

In the original Hobbs formulation simply walked through the tree in this manner, and proposed the first antecedent it encountered which matched gender/number qualifications. Ge however uses a more probabilistic approach, which we adopted. He defined a metric called the "Hobbs distance ($d_H$)", which is simply stating that an antecedent is the $n^{th}$ candidate that the Hobbs algorithm found (e.g. if the antecedent was

the second node that the Hobbs algorithm found, it would have $d_H$=2). He then trained on the data to find $P(d_H)$, which he calculated as follows:

$$P(d_H = n) = \frac{\#\,antecedents\ with\ d_H = n}{\#\,antecedents}$$

Note that this is clearly a probability distribution as it equals unity when summed across all $d_H$. He then computed other probabilities (gender, mention count, etc) separately, and multiplicatively computed a total probability. The following sections describe these various factors we used (following the footsteps of Ge).

**1.3 Gender/Animaticity/Number Probability**

We incorporated gender/number information in the same way that Ge did. We first divided the pronouns into various "gender buckets" (male, female, inanimate, unknown plural, and unknown singular). We then collected frequency information on the training data between each pronoun and the individual words in the antecedent NP. The probability of a pronoun (bucketed by gender) given a word in its antecedent is thus:

$$P(p \mid a) = (1-\varepsilon)\frac{\#\,occurences\ of\ a\ in\ antecedent\ for\ p}{\#\,occurences\ of\ a}$$

And the probability given to a pair not previously seen (an UNK) is:

$$P(UNK \mid a) = \varepsilon$$

We can show this to be a probability distribution as follows:

$$\sum_{p \cup UNK} P(p \mid a) = P(UNK \mid a) + (1-\varepsilon)\sum_{p} \frac{\#\,occurences\ of\ a\ in\ antecedent\ for\ p}{\#\,occurences\ of\ a}$$

$$= \varepsilon + (1-\varepsilon)\frac{\#\,occurences\ of\ a}{\#\,occurences\ of\ a} = 1$$

Note the smoothing for the case where we haven't yet seen the word a paired with pronoun gender bucket p (in our implementation, we used $\varepsilon = 0.2$, chosen experimentally). This equation leaves us with the question of which word (a) in the antecedent to choose when testing. We tried computing probabilities on all of the words in the noun phrase (a = string of words), but due to data sparseness, this actually made our accuracy a bit worse. So we took the solution of Ge, and incorporated the Dunning algorithm[3,4] to pick the word in the antecedent phrase that has the highest log likelihood when paired with the pronoun. A summary of the Dunning algorithm and our implementation of it is given in the following section.

**1.3.1 Dunning Algorithm**

The idea behind the Dunning algorithm is to treat comparisons of word pairs as a binomial process (like the flipping of a coin). Basically, a pair of words (say an antecedent word A and a pronoun B) are independent if p(A | B) = p(A | ~B) = p(A) (i.e. probability of a same if B present, not present, or no information about B is given). Thus we can test for independence of A and B by checking for this equality. Of course, we know that the words in general are not independent, but we're looking for pairs that have high correlation (by what appears to be low independence). To this end, we calculate the

frequencies of pairs of pronoun words/antecedent words. Thus, for each pair (a, p) of antecedent word a and pronoun p, we compute the following frequencies:

$A = \varepsilon + (1-\varepsilon)*(\#\text{ pairs with a and p}) / (\#\text{ pairs})$
$B = \varepsilon + (1-\varepsilon)*(\#\text{ pairs without a but with p}) / (\#\text{ pairs})$
$C = \varepsilon + (1-\varepsilon)*(\#\text{ pairs with a but without p}) / (\#\text{ pairs})$
$D = \varepsilon + (1-\varepsilon)*(\#\text{ pairs without a or p}) / (\#\text{ pairs})$

Where we have $\varepsilon$=0.1 for smoothing. We then compute the log likelihood as follows:

likelihood = A*log(A) + B*log(B) + C*log(C) + D*log(D)
-(A+B)*log(A+B) – (A+C)*log(A+C)
-(B+D)*log(B+D) – (C+D)*log(C+D)
+(A+B+C+D)*log(A+B+C+D)

When testing an antecedent NP containing multiple nouns, we choose the noun (a) with the highest log likelihood according to the Dunning formulation and then compute P(p | a) for this noun (a).

**1.4 Other Statistics**

Ge used two other pieces of information in resolving pronouns. He uses information on the head constituent above each pronoun, which he obtained by parsing the sentences. However, the gains from this information were minimal (just over 2%).

Additionally, Each NP was labeled with the number of times it had been mentioned up to that point in the story. The idea here was that the more times a NP is used in a story, the more likely that it is a "hot topic" in the segment, and thus more likely to be referred to by a pronoun. Unfortunately we did implement such a mention-count statistic due to pressures of time, despite the fact that it increased the accuracy of Ge's tagger by 5%. However, Ge provides information on the accuracy of his algorithm without mention count and head information, so we are able to compare our results to his.

**1.5 Final Criteria**

When resolving a pronoun in the test data, we start by using the Hobbs algorithm to collect 15 possible antecedents. We pick the antecedent that maximizes the following quantity:

$P(d_H)P(p|a)$

Where a is the word chosen by running the Dunning algorithm on the antecedent candidate, p is the gender/number bucket of the pronoun in question, and $d_H$ is the Hobbs distance of the candidate (1 to 15).

## 2. Data

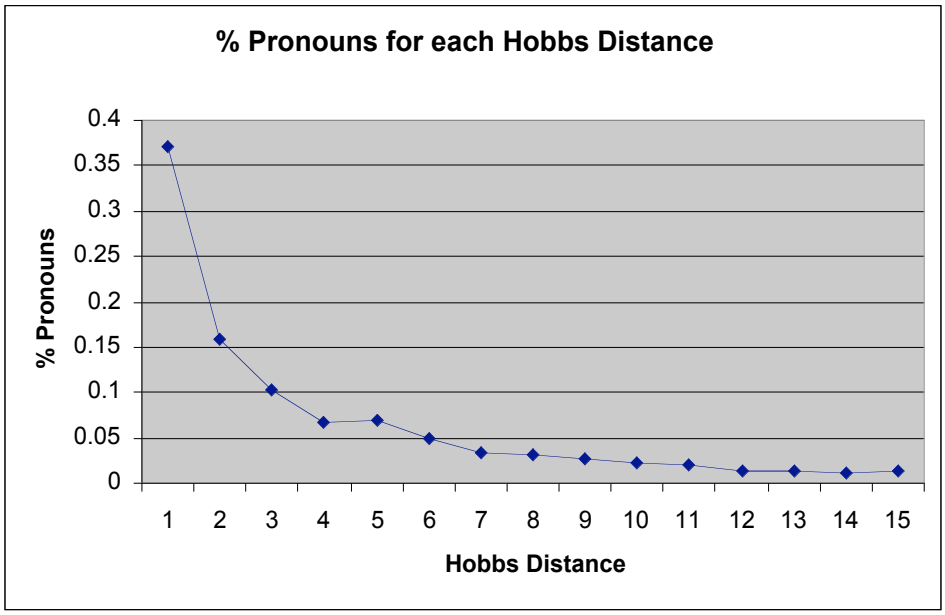


Figure 1: Plot of $P(d_H)$ – the probability of a pronoun having a given Hobbs distance from its antecedent.

Phrase 1: India's pretext was that **it** wanted a joint treaty…

Phrase 2: On Monday, entertainment giant Fuji Sankei Communications Group said it acquired 25 of Virgin Music Group from London's Virgin Atlantic Group for about 155.5 million. Fuji Sankei said **it** intends to strengthen its U.S. presence through its stake in Virgin.

Phrase 3: Sony's acquisition of CBS Records last year boosted sales and improved working conditions there, **he** said.

Phrase 4: Meanwhile, several traders and analysts said seasonal factors were helping to firm prices in the crude oil market. "**We're** looking for continued strength through October," said Peter Beutel of Elders Futures Inc.

Table 1: Some specific "errors." Bold text is the pronoun being evaluated, single underline indicates our result, and double underline indicates the "correct" result.

| Antecedent Phrase | Dunning Word |
|---|---|
| (NP (NNP Baylor) (NNP College)) | Baylor |
| (NP-SBJ (DT the) (NNS STS) (NN strategy)) | strategy |
| (NP-SBJ (NN rating) (NNS agencies)) | agencies |

Table 2: A few examples of words picked from a noun phrase using the Dunning Maximum Likelihood algorithm.

| | Accuracy |
|---|---|
| **Hobbs** | 47% |
| **Hobbs + Gender** | 57% |

Table 3: Resulting accuracy from using various components of pronoun resolution. Results shown for test data containing 2644 pronouns where each pronoun is no more than one sentence away from its antecedent.

# 3. Analysis

## 3.1 Accuracy

The first thing to notice is that our accuracy is considerably lower than Ge's result (who had 65% with Hobbs alone, and 75% with Hobbs+Gender). Ge used a small portion of the Penn Treebank for his experiments, that was hand-labeled with co-reference information. Unfortunately this corpus was not available to us. Instead used the BLLIP Treebank, which is in fact the output of Ge's parsing and tagging algorithm on a (partially non-overlapping) set of the WSJ text. One hypothesis for why our results were lower than Ge's was that the BLLIP Treebank often contained an antecedent, and then had pronouns refer to it over the next several sentences (i.e. there wasn't always close proximity between the antecedent and pronoun, which caused Hobbs algorithm to yield poor results). In fact, most of the time, there were strings of two to three pronouns in a row that referred to the same antecedent. The following section describes an algorithm we put in place to try to address this issue.

### 3.1.1 Learn-As-You-Go Algorithm

Given that analysis of the test data showed that often times, several pronouns in a row refer to the same antecedent phrase, we decided to implement a "learn as you go" algorithm to exploit this. The first pronoun in one of these "chains" is easy to identify, because it is in close proximity to its antecedent (small Hobbs distance). Thus, upon identifying an antecedent, we store the relative frequency with which that NP is associated with a particular pronoun (e.g. if the word "John" is associated with "he" then it could be likely that if the next pronoun seen is "he" as well, it should be associated with "John"). Thus, we compute the following quantity:

$$P(a \mid p) = \begin{cases} (1-\varepsilon)\dfrac{\#observed(a,p)}{\sum_{p'} \#observed(a,p')} & \#observed(a,p) > 0 \\ \varepsilon & otherwise \end{cases}$$

And we included this probability as another multiplicative factor when determining the most likely antecedent (see section 1.5). Note that the counts (#observed) are collected

for each antecedent/pronoun pair previously identified during testing. This normally wouldn't make sense to train on the testing data, but given that we can identify the first occurrence of a reference easier than the second, third, etc, this idea would seem to have some merit. Unfortunately, no matter what values of $\varepsilon$ we used, we couldn't get an improvement (at best, we got the same results as without this factor). The reason for this is probably due to the fact that not every antecedent is always referred to by the same pronoun. In addition, this method can give misleading results if a pronoun is the same as those occurring before it, but refers to a different antecedent.

**3.2 Dunning Algorithm**

Table 2 shows a few antecedent noun phrases, and the single word that the Dunning algorithm chose. Notice that in the first case, the word choice (Baylor) may not be the best pick. Any pronoun will be referring to "college" and thus won't have gender, but Baylor in this case may very well have gender (probably not, because it sounds like a surname, but it's possible that it could in other cases e.g. the College of William and Mary). In the second example however, Dunning seems to have picked the best word (strategy). It's genderless and singular, as would be the pronoun referring to this phrase. Likewise, Dunning seems to have done well with the third choice (agencies). It's plural, and any pronoun referring to this phrase would also be plural (whereas had it picked rafting, the results might have been worse as rafting is singular).

**3.3 Further Error Analysis**

It is interesting to note that in 7.3% of the cases in which a pronoun was mistagged, the phrase identified as the antecedent is a supertree of the correct antecedent, whereas it is a subtree only 3.9% of the time. While it is probabilistically more likely that a supertree wil be found than a subtree, 55% of the time a supertree was found the identified node was a parent or grandparent of the correct node, and when a subtree was found 50% of the nodes were a direct child of the correct node. This indicates that it might be a good idea, once a node is identified, to discount the contribution of the Hobbs algorithm to the score and then reevaluate at least the children and parent of the identified node.

Table 1 shows several specific sentences on which our resolver obtained results different from the tagged sentence. Here is where the benefit of using hand-labeled text is truly evidenced. In phrase 1, we choose the phrase "India's pretext" instead of the correct "India." This is one of the aforementioned cases in which we choose a node directly above the correct node, so perhaps this error could be corrected by taking that into consideration. In phrase 2, we choose (rather surprisingly, due to its distance from the pronoun in question) the phrase "entertainment giant Fuji Sankei Communications Group." While it might have been preferable to leave off "entertainment giant," this seems not a significantly worse choice than the labeled one of "Fuji Sankei" in the second sentence. (While this could have occurred because of our policy of associating pronouns with the previous phrase of the same label, in this case the labels are in fact different.) Phrase 3 shows what is in fact a fairly common occurrence, in which both the corpus and our results are incorrect. Finally, phrase 4 shows a situation in which our results are superior to those in the corpus. This sentence was not unique. There were even cases in which there errors other than in the co-reference information. For example, both

we and the parser made an error on the sentence “Bond Corp. in June 1988 guaranteed the Insurance Commission 2.7 Australian dollars (US $2.10) a share for the Bell Group stake.” This isn’t especially surprising, since the pronoun being resolved was “US.”

In examining errors, it seems that it is often the case that the resolution of plural pronouns, especially when spoken by or in the voice of an individual, are difficult to resolve. For example, even out of context it is possible to resolve the “we” in the sentence “‘In some ways we 're attempting to be the USA Today of the Caribbean’, publisher Tim Forsythe said….” However, there is no clear link between this “we” and “Caribbean Week” at the beginning of the previous sentence, and so it is not surprising that both our parser and the corpus were in error on this sentence. Statistically, while 10% of the pronouns we attempted to resolve were one of “we,” “our,” or “us,” 12% of our errors were made on those pronouns. Of course, these are only errors when compared to the corpus; in fact the rate could be higher or lower than that, depending on the accuracy of the corpus.

## 4. Conclusion

While our final results were hardly top-of-the-line, and in fact were poorer than the simplest implementation by Ge et al, we believe that several important insights can be gained from our work. It is clear that in order to resolve some pronouns, it may be important to chain several inferences together. For example, if a plural pronoun is most strongly linked to a singular person, it is likely that the person is a member of the group referred to, and perhaps there is another context in which the person is closely linked to a group, which is in fact the correct reference. We found that it may be useful to make one step taking distance into account, and then search in a small area around the initial result, possibly finding a better result only slightly farther away. Finally we learned that anaphora resolution is difficult, and that in some cases, as is the case with much of NLP, it may be impossible to perform correctly without some kind of “understanding” of the text.

## 5. References


1. Ge, Hale, and Charniak, 1998, *A Statistical Anaphora Resolution*, Proceedings of the Sixth Workshop on Very Large Corpora.
2. Beatrice Daille and Place Jussieu, 1994, *Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering*, UCREL Technical Papers 5
3. Ted Dunning, 1993, *Accurate Methods for the Statistics of Surprise and Coincidence,* Computational Linguistics 19:1
4. Jerry Hobbs, 1976, *Pronoun Resolution*, Technical Report 76-1, City College, New York.