EM Works for Pronoun Anaphora Resolution

Eugene Charniak and Micha Elsner
Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University
Providence, RI 02912
{ec,melsner}@cs.brown.edu

Abstract

We present an algorithm for pronoun-anaphora (in English) that uses Expectation Maximization (EM) to learn virtually all of its parameters in an unsupervised fashion. While EM frequently fails to find good models for the tasks to which it is set, in this case it works quite well. We have compared it to several systems available on the web (all we have found so far). Our program significantly outperforms all of them. The algorithm is fast and robust, and has been made publically available for downloading.

1 Introduction

We present a new system for resolving (personal) pronoun anaphora[1]. We believe it is of interest for two reasons. First, virtually all of its parameters are learned via the expectation-maximization algorithm (EM). While EM has worked quite well for a few tasks, notably machine translations (starting with the IBM models 1-5 (Brown et al., 1993), it has not had success in most others, such as part-of-speech tagging (Merialdo, 1991), named-entity recognition (Collins and Singer, 1999) and context-free-grammar induction (numerous attempts, too many to mention). Thus understanding the abilities and limitations of EM is very much a topic of interest. We present this work as a positive data-point in this ongoing discussion.

Secondly, and perhaps more importantly, is the system's performance. Remarkably, there are very few systems for actually *doing* pronoun anaphora available on the web. By emailing the corpora-list the other members of the list pointed us to four. We present a head to head evaluation and find that our performance is significantly better than the competition.

[1]The system, the Ge corpus, and the model described here can be downloaded from http://bllip.cs.brown.edu/download/emPronoun.tar.gz.

2 Previous Work

The literature on pronominal anaphora is quite large, and we cannot hope to do justice to it here. Rather we limit ourselves to particular papers and systems that have had the greatest impact on, and similarity to, ours.

Probably the closest approach to our own is Cherry and Bergsma (2005), which also presents an EM approach to pronoun resolution, and obtains quite successful results. Our work improves upon theirs in several dimensions. Firstly, they do not distinguish antecedents of non-reflexive pronouns based on syntax (for instance, subjects and objects). Both previous work (cf. Tetreault (2001) discussed below) and our present results find these distinctions extremely helpful. Secondly, their system relies on a separate preprocessing stage to classify non-anaphoric pronouns, and mark the gender of certain NPs (Mr., Mrs. and some first names). This allows the incorporation of external data and learning systems, but conversely, it requires these decisions to be made sequentially. Our system classifies non-anaphoric pronouns jointly, and learns gender without an external database. Next, they only handle third-person pronouns, while we handle first and second as well. Finally, as a demonstration of EM's capabilities, its evidence is equivocal. Their EM requires careful initialization — sufficiently careful that the EM version only performs 0.4% better than the initialized program alone. (We can say nothing about relative performance of their system vs. ours since we have been able to access neither their data nor code.)

A quite different unsupervised approach is Kehler et al. (2004a), which uses self-training of a discriminative system, initialized with some con-

servative number and gender heuristics. The system uses the conventional ranking approach, applying a maximum-entropy classifier to pairs of pronoun and potential antecedent and selecting the best antecedent. In each iteration of self-training, the system labels the training corpus and its decisions are treated as input for the next training phase. The system improves substantially over a Hobbs baseline. In comparison to ours, their feature set is quite similar, while their learning approach is rather different. In addition, their system does not classify non-anaphoric pronouns,

A third paper that has significantly influenced our work is that of (Haghighi and Klein, 2007). This is the first paper to treat all noun phrase (NP) anaphora using a generative model. The success they achieve directly inspired our work. There are, however, many differences between their approach and ours. The most obvious is our use of EM rather than theirs of Gibbs sampling. However, the most important difference is the choice of training data. In our case it is a very large corpus of parsed, but otherwise unannotated text. Their system is trained on the ACE corpus, and requires explicit annotation of all "markables" — things that are or have antecedents. For pronouns, only anaphoric pronouns are so marked. Thus the system does not learn to recognize non-anaphoric pronouns — a significant problem. More generally it follows from this that the system only works (or at least works with the accuracy they achieve) when the input data is so marked. These markings not only render the non-anaphoric pronoun situation moot, but also significantly restrict the choice of possible antecedent. Only perhaps one in four or five NPs are markable (Poesio and Vieira, 1998).

There are also several papers which treat coference as an unsupervised clustering problem (Cardie and Wagstaff, 1999; Angheluta et al., 2004). In this literature there is no generative model at all, and thus this work is only loosely connected to the above models.

Another key paper is (Ge et al., 1998). The data annotated for the Ge research is used here for testing and development data. Also, there are many overlaps between their formulation of the problem and ours. For one thing, their model is generative, although they do not note this fact, and (with the partial exception we are about to mention) they obtain their probabilities from hand annotated data rather than using EM. Lastly, they learn their gender information (the probability of that a pronoun will have a particular gender given its antecedent) using a truncated EM procedure. Once they have derived all of the other parameters from the training data, they go through a larger corpus of unlabeled data collecting estimated counts of how often each word generates a pronoun of a particular gender. They then normalize these probabilities and the result is used in the final program. This is, in fact, a single iteration of EM.

Tetreault (2001) is one of the few papers that use the (Ge et al., 1998) corpus used here. They achieve a very high 80% correct, but this is given hand-annotated number, gender and syntactic binding features to filter candidate antecedents and also ignores non-anaphoric pronouns.

We defer discussion of the systems against which we were able to compare to Section 7 on evaluation.

3 Pronouns

We briefly review English pronouns and their properties. First we only concern ourselves with "personal" pronouns: "I", "you", "he", "she", "it", and their variants. We ignore, e.g., relative pronouns ("who", "which", etc.), deictic pronouns ("this", "that") and others.

Personal pronouns come in four basic types:

subject "I", "she", etc. Used in subject position.

object "me", "her" etc. Used in non-subject position.

possessive "my" "her", and

reflexive "myself", "herself" etc. Required by English grammar in certain constructions — e.g., "I kicked myself."

The system described here handles all of these cases.

Note that the type of a pronoun is not connected with its antecedent, but rather is completely determined by the role it plays in it's sentence.

Personal pronouns are either anaphoric or non-anaphoric. We say that a pronoun is anaphoric when it is coreferent with another piece of text in the same discourse. As is standard in the field we distinguish between a referent and an antecedent. The referent is the thing in the world that the pronoun, or, more generally, noun phrase (NP), denotes. Anaphora on the other hand is a relation be-

tween pieces of text. It follows from this that non-anaphoric pronouns come in two basic varieties — some have a referent, but because the referent is not mentioned in the text[2] there is no anaphoric relation to other text. Others have no referent (*expletive* or *pleonastic* pronouns, as in "It seems that ..."). For the purposes of this article we do not distinguish the two.

Personal pronouns have three properties other than their type:

person first ("I","we"), second ("you") or third ("she","they") person,

number singular ("I","he") or plural ("we", "they"), and

gender masculine ("he"), feminine ("she") or neuter ("they").

These are critical because it is these properties that our generative model generates.

4 The Generative Model

Our generative model ignores the generation of most of the discourse, only generating a pronoun's person, number,and gender features along with the governor of the pronoun and the syntactic relation between the pronoun and the governor. (Informally, a word's governor is the head of the phrase above it. So the governor of both "I" and "her" in "I saw her" is "saw".

We first decide if the pronoun is anaphoric based upon a distribution p(anaphoric). (Actually this is a bit more complex, see the discussion in Section 5.3.) If the pronoun is anaphoric we then select a possible antecedent. Any NP in the current or two previous sentences is considered. We select the antecedent based upon a distribution p(anaphora|context). The nature of the "context" is discussed below. Then given the antecedent we generative the pronoun's person according to p(person|antecedent), the pronoun's gender according to p(gender|antecedent), number, p(number|antecedent) and governor/relation-to-governor from p(governor/relation|antecedent).

To generate a non-anaphoric third person singular "it" we first guess that the non-anaphoric pronouns is "it" according to p("it"|non-anaphoric). and then generate the governor/relation according to p(governor/relation|non-anaphoric-it);

Lastly we generate any other non-anaphoric pronouns and their governor with a fixed probability p(other). (Strictly speaking, this is mathematically invalid, since we do not bother to normalize over all the alternatives; a good topic for future research would be exploring what happens when we make this part of the model truly generative.)

One inelegant part of the model is the need to scale the p(governor/rel|antecedent) probabilities. We smooth them using Kneser-Ney smoothing, but even then their dynamic range (a factor of 10^6) greatly exceeds those of the other parameters. Thus we take their nth root. This n is the last of the model parameters.

5 Model Parameters

5.1 Intuitions

All of our distributions start with uniform values. For example, gender distributions start with the probability of each gender equal to one-third. From this it follows that on the first EM iteration all antecedents will have the same probability of generating a pronoun. At first glance then, the EM process might seem to be futile. In this section we hope to give some intuitions as to why this is not the case.

As is typically done in EM learning, we start the process with a much simpler generative model, use a few EM iterations to learn its parameters, and gradually expose the data to more and more complex models, and thus larger and larger sets of parameters.

The first model only learns the probability of an antecedent generating the pronoun given what sentence it is in. We train this model through four iterations before moving on to more complex ones.

As noted above, all antecedents initially have the same probability, but this is not true after the first iteration. To see how the probabilities diverge, and diverge correctly, consider the first sentence of a news article. Suppose it starts "President Bush announced that he ..." In this situation there is only one possible antecedent, so the expectation that "he" is generated by the NP in the same sentence is 1.0. Contrast this with the situation in the third and subsequent sentences. It is only then that we have expectation for sentences two back generating the pronoun. Furthermore, typically by this point there will be, say, twenty NPs to share the

[2] Actually, as in most previous work, we only consider referents realized by NPs. For more general approaches see Byron (2002).

probability mass, so each one will only get an increase of 0.05. Thus on the first iteration only the first two sentences have the power to move the distributions, but they do, and they make NPs in the current sentence very slightly more likely to generate the pronoun than the sentence one back, which in turn is more likely than the ones two back.

This slight imbalance is reflected when EM readjusts the probability distribution at the end of the first iteration. Thus for the second iteration everyone contributes to subsequent imbalances, because it is no longer the case the all antecedents are equally likely. Now the closer ones have higher probability so forth and so on.

To take another example, consider how EM comes to assign gender to various words. By the time we start training the gender assignment probabilities the model has learned to prefer nearer antecedents as well as ones with other desirable properties. Now suppose we consider a sentence, the first half of which has no pronouns. Consider the gender of the NPs in this half. Given no further information we would expect these genders to distribute themselves accord to the prior probability that any NP will be masculine, feminine, etc. But suppose that the second half of the sentence has a feminine pronoun. Now the genders will be skewed with the probability of one of them being feminine being much larger. Thus in the same way these probabilities will be moved from equality, and should, in general be moved correctly.

5.2 Parameters Learned by EM

Virtually all model parameters are learned by EM. We use the parsed version of the North-American News Corpus. This is available from the (McClosky et al., 2008). It has about 800,000 articles, and 500,000,000 words.

The least complicated parameter is the probability of gender given word. Most words that have a clear gender have this reflected in their probabilities. Some examples are shown in Table 1. We can see there that EM gets "Paul", "Paula", and "Wal-mart" correct. "Pig" has no obvious gender in English, and the probabilities reflect this. On the other hand "Piggy" gets feminine gender. This is no doubt because of "Miss Piggy" the puppet character. "Waist" the program gets wrong. Here the probabilities are close to gender-of-pronoun priors. This happens for a (comparatively small) class of pronouns that, in fact, are probably never an antecedent, but are nearby random pronouns. Because of their non-antecedent proclivities, this sort of mistake has little effect.

Word	Male	Female	Neuter
paul	0.962	0.002	0.035
paula	0.003	0.915	0.082
pig	0.445	0.170	0.385
piggy	0.001	0.853	0.146
wal-mart	0.016	0.007	0.976
waist	0.380	0.155	0.465

Table 1: Words and their probabilities of generating masculine, feminine and neuter pronouns

antecedent	p(singular|antecedent)
Singular	0.939048
Plural	0.0409721
Not NN or NNP	0.746885

Table 2: The probability of an antecedent generation a singular pronoun as a function of its number

Next consider p(number|antecedent), that is the probability that a given antecedent will generate a singular or plural pronoun. This is shown in Table 2. Since we are dealing with parsed text, we have the antecedent's part-of-speech, so rather than the antecedent we get the number from the part of speech: "NN" and "NNP" are singular, "NNS" and "NNPS" are plural. Lastly, we have the probability that an antecedent which is not a noun will have a singular pronoun associated with it. Note that the probability that a singular antecedent will generate a singular pronoun is not one. This is correct, although the exact number probably is too low. For example, "IBM" may be the antecedent of both "we" and "they", and vice versa.

Next we turn to p(person|antecedent), predicting whether the pronoun is first, second or third person given its antecedent. We simplify this by noting that we know the person of the antecedent (everything except "I" and "you" and their variants are third person), so we compute p(person|person). Actually we condition on one further piece of information, if either the pronoun or the antecedent is being quoted. The idea is that an "I" in quoted material may be the same person as "John Doe" outside of quotes, if Mr. Doe is speaking. Indeed, EM picks up on this as is illustrated in Tables 3 and 4. The first gives the situation when neither antecedent nor pronoun is within a quotation. The high numbers along the

	Person of Pronoun		
Person of Ante	First	Second	Third
First	0.923	0.076	0.001
Second	0.114	0.885	0.001
Third	0.018	0.015	0.967

Table 3: Probability of an antecedent generating a first,second or third person pronoun as a function of the antecedents person

	Person of Pronoun		
Person of Ante	First	Second	Third
First	0.089	0.021	0.889
Second	0.163	0.132	0.705
Third	0.025	0.011	0.964

Table 4: Same, but when the antecedent is in quoted material but the pronoun is not

diagonal (0.923, 0.885, and 0.967) show the expected like-goes-to-like preferences. Contrast this with Table 4 which gives the probabilities when the antecedent is in quotes but the pronoun is not. Here we see all antecedents being preferentially mapped to third person (0.889, 0.705, and 0.964).

We save p(antecedent|context) till last because it is the most complicated. Given what we know about the context of the pronoun not all antecedent positions are equally likely. Some important conditioning events are:

- the exact position of the sentence relative to the pronoun (0, 1, or 2 sentences back),
- the position of the head of the antecedent within the sentence (bucketed into 6 bins). For the current sentence position is measured backward from the pronoun. For the two previous sentences it is measure forward from the start of the sentence.
- syntactic positions — generally we expect NPs in subject position to be more likely antecedents than those in object position, and those more likely than other positions (e.g., object of a preposition).
- position of the pronoun — for example the subject of the previous sentence is very likely to be the antecedent if the pronoun is very early in the sentence, much less likely if it is at the end.
- type of pronoun — reflexives can only be bound within the same sentence, while subject and object pronouns may be anywhere. Possessives may be in previous sentences but this is not as common.
- type of antecedent. Intuitively other pronouns and proper nouns are more likely to be antecedents than common nouns and NPs headed up by things other than nouns.

Part of Speech	pron	proper	common
	0.094	0.057	0.032
Word Position	bin 0	bin 2	bin 5
	0.111	0.007	0.0004
Syntactic Type	subj	other	object
	0.068	0.045	0.037

Table 5: Geometric mean of the probability of the antecedent when holding everything expect the stated feature of the antecedent constant

All told this comes to 2592 parameters (3 sentences, 6 antecedent word positions, 3 syntactic positions, 4 pronoun positions, 3 pronoun types, and 4 antecedent types). It is impossible to say if EM is setting all of these correctly. There are too many of them and we do not have knowledge or intuitions about most all of them. However, all help performance on the development set, and we can look at a few where we do have strong intuitions. Table 5 gives some examples. The first two rows are devoted to the probabilities of particular kind of antecedent (pronouns, proper nouns, and common nouns) generating a pronoun, holding everything constant except the type of antecedent. The numbers are the geometric mean of the probabilities in each case. The probabilities are ordered according to, at least my, intuition with pronoun being the most likely (0.094), followed by proper nouns (0.057), followed by common nouns (0.032), a fact also noted by (Haghighi and Klein, 2007). When looking at the probabilities as a function of word position again the EM derived probabilities accord with intuition, with bin 0 (the closest) more likely than bin 2 more likely than bin 5. The last two lines have the only case where we have found the EM probability not in accord with our intuitions. We would have expected objects of verbs to be more likely to generate a pronoun than the catch-all "other" case. This proved not to be the case. On the other hand, the two are much closer in probabilities than any of the other, more intuitive, cases.

5.3 Parameters Not Set by EM

There are a few parameters not set by EM.

Several are connected with the well known syntactic constraints on the use of reflexives. A simple version of this is built in. Reflexives must have an antecedent in same sentence, and generally cannot be coreferent-referent with the subject of the sentence.

There are three system parameters that we set by hand to optimize performance on the development set. The first is n. As noted above, the distribution p(governor/relation|antecedent) has a much greater dynamic range than the other probability distributions and to prevent it from, in essence, completely determining the answer, we take its nth root. Secondly, there is a probability of generating a non-anaphoric "it". Lastly we have a probability of generating each of the other non-monotonic pronouns along with (the nth root of) their governor. These parameters are 6, 0.1, and 0.0004 respectively.

6 Definition of Correctness

We evaluate all programs according to Mitkov's "resolution etiquette" scoring metric (also used in Cherry and Bergsma (2005)), which is defined as follows: if N is the number of non-anaphoric pronouns correctly identified, A the number of anaphoric pronouns correctly linked to their antecedent, and P the total number of pronouns, then a pronoun-anaphora program's percentage correct is $\frac{N+A}{P}$.

Most papers dealing with pronoun coreference use this simple ratio, or the variant that ignores non-anaphoric pronouns. It has appeared under a number of names: *success* (Yang et al., 2006), *accuracy* (Kehler et al., 2004a; Angheluta et al., 2004) and *success rate* (Tetreault, 2001). The other occasionally-used metric is the MUC score restricted to pronouns, but this has well-known problems (Bagga and Baldwin, 1998).

To make the definition perfectly concrete, however, we must resolve a few special cases. One is the case in which a pronoun x correctly says that it is coreferent with another pronoun y. However, the program misidentifies the antecedent of y. In this case (sometimes called *error chaining* (Walker, 1989)), both x and y are to be scored as wrong, as they both end up in the wrong coreferential chain. We believe this is, in fact, the standard (Mitkov, personal communication), although there are a few papers (Tetreault, 2001; Yang et al., 2006) which do the opposite and many which simply do not discuss this case.

One more issue arises in the case of a system attempting to perform complete NP anaphora[3]. In these cases the coreferential chains they create may not correspond to any of the original chains. In these cases, we call a pronoun correctly resolved if it is put in a chain including at least one correct non-pronominal antecedent. This definition cannot be used in general, as putting all NPs into the same set would give a perfect score. Fortunately, the systems we compare against do not do this – they seem more likely to over-split than under-split. Furthermore, if they do take some inadvertent advantage of this definition, it helps them and puts our program at a possible disadvantage, so it is a more-than-fair comparison.

7 Evaluation

To develop and test our program we use the dataset annotated by Niyu Ge (Ge et al., 1998). This consists of sections 0 and 1 of the Penn treebank. Ge marked every personal pronoun and all noun phrases that were coreferent with these pronouns. We used section 0 as our development set, and section 1 for testing. We reparsed the sentences using the Charniak and Johnson parser (Charniak and Johnson, 2005) rather than using the gold-parses that Ge marked up. We hope thereby to make the results closer to those a user will experience. (Generally the gold trees perform about 0.005 higher than the machine parsed version.) The test set has 1119 personal pronouns of which 246 are non-anaphoric. Our selection of this dataset, rather than the widely used MUC-6 corpus, is motivated by this large number of pronouns.

We compared our results to four currently-available anaphora programs from the web. These four were selected by sending a request to a commonly used mailing list (the "corpora-list") asking for such programs. We received four leads: JavaRAP, Open-NLP, BART and GuiTAR. Of course, these systems represent the best available work, not the state of the art. We presume that more recent supervised systems (Kehler et al., 2004b; Yang et al., 2004; Yang et al., 2006) per-

[3]Of course our system does not attempt NP coreference resolution, nor does JavaRAP. The other three comparison systems do.

form better. Unfortunately, we were unable to obtain a comparison unsupervised learning system at all.

Only one of the four is explicitly aimed at personal-pronoun anaphora — RAP (Resolution of Anaphora Procedure) (Lappin and Leass, 1994). It is a non-statistical system originally implemented in Prolog. The version we used is JavaRAP, a later reimplementation in Java (Long Qiu and Chua, 2004). It only handles third person pronouns.

The other three are more general in that they handle all NP anaphora. The GuiTAR system (Poesio and Kabadjov, 2004) is designed to work in an "off the shelf" fashion on general text GUITAR resolves pronouns using the algorithm of (Mitkov et al., 2002), which filters candidate antecedents and then ranks them using morphosyntactic features. Due to a bug in version 3, GUITAR does not currently handle possessive pronouns.GUITAR also has an optional discourse-new classification step, which cannot be used as it requires a discontinued Google search API.

OpenNLP (Morton et al., 2005) uses a maximum-entropy classifier to rank potential antecedents for pronouns. However despite being the best-performing (on pronouns) of the existing systems, there is a remarkable lack of published information on its innards.

BART (Versley et al., 2008) also uses a maximum-entropy model, based on Soon et al. (2001). The BART system also provides a more sophisticated feature set than is available in the basic model, including tree-kernel features and a variety of web-based knowledge sources. Unfortunately we were not able to get the basic version working. More precisely we were able to run the program, but the results we got were substantially lower than any of the other models and we believe that the program as shipped is not working properly.

Some of these systems provide their own preprocessing tools. However, these were bypassed, so that all systems ran on the Charniak parse trees (with gold sentence segmentation). Systems with named-entity detectors were allowed to run them as a preprocess. All systems were run using the models included in their standard distribution; typically these models are trained on annotated news articles (like MUC-6), which should be relatively similar to our WSJ documents.

System	Restrictions	Performance
GuiTAR	No Possessives	0.534
JavaRap	Third Person	0.529
Open-NLP	None	0.593
Our System	None	0.686

Table 6: Performance of Evaluated Systems on Test Data

The performance of the remaining systems is given in Table 6. The two programs with restrictions were only evaluated on the pronouns the system was capable of handling.

These results should be approached with some caution. In particular it is possible that the results for the systems other than ours are underestimated due to errors in the evaluation. Complications include the fact all of the four programs all have different output conventions. The better to catch such problems the authors independently wrote two scoring programs.

Nevertheless, given the size of the difference between the results of our system and the others, the conclusion that ours has the best performance is probably solid.

8 Conclusion

We have presented a generative model of pronoun-anaphora in which virtually all of the parameters are learned by expectation maximization. We find it of interest first as an example of one of the few tasks for which EM has been shown to be effective, and second as a useful program to be put in general use. It is, to the best of our knowledge, the best-performing system available on the web. To down-load it, go to (to be announced).

The current system has several obvious limitation. It does not handle cataphora (antecedents occurring after the pronoun), only allows antecedents to be at most two sentences back, does not recognize that a conjoined NP can be the antecedent of a plural pronoun, and has a very limited grasp of pronominal syntax. Perhaps the largest limitation is the programs inability to recognize the speaker of a quoted segment. The result is a very large fraction of first person pronouns are given incorrect antecedents. Fixing these problems would no doubt push the system's performance up several percent.

However the most critical direction for future research is to push the approach to handle full NP

anaphora. Besides being of the greatest importance in its own right, it would also allow us to add one piece of information we currently neglect in our pronominal system — the more times a document refers to an entity the more likely it is to do so again.

9 Acknowledgements

We would like to thank the authors and maintainers of the four systems against which we did our comparison, especially Tom Morton, Mijail Kabadjov and Yannick Versley. Making your system freely available to other researchers is one of the best ways to push the field forward. In addition, we thank three anonymous reviewers.

References

Roxana Angheluta, Patrick Jeuniaux, Rudradeb Mitra, and Marie-Francine Moens. 2004. Clustering algorithms for noun phrase coreference resolution. In *Proceedings of the 7es Journes internationales d'Analyse statistique des Donnes Textuelles*, pages 60–70, Louvain La Neuve, Belgium, March 10–12.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2).

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002)*, pages 80–87, Philadelphia, PA, USA, July 6–12.

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *In Proceedings of EMNLP*, pages 82–89.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine *n*-best parsing and MaxEnt discriminative reranking. In *Proc. of the 2005 Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 173–180.

Colin Cherry and Shane Bergsma. 2005. An Expectation Maximization approach to pronoun resolution. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 88–95, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Michael Collins and Yorav Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP 99)*.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–171, Orlando, Florida. Harcourt Brace.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 848–855. Association for Computational Linguistics.

Andrew Kehler, Douglas Appelt, Lara Taylor, and Aleksandr Simma. 2004a. Competitive self-trained pronoun interpretation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Short Papers*, pages 33–36, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Andrew Kehler, Douglas E. Appelt, Lara Taylor, and Aleksandr Simma. 2004b. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 289–296.

Shalom Lappin and Herber J. Leass. 1994. An algorithm for pronouminal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Min-Yen Kan Long Qiu and Tat-Seng Chua. 2004. A public reference implementation of the RAP anaphora resolution algorithm. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, volume I, pages 291–294.

David McClosky, Eugene Charniak, and MarkJohnson. 2008. *BLLIP North American News Text, Complete*. Linguistic Data Consortium. LDC2008T13.

Bernard Merialdo. 1991. Tagging text with a probabilistic model. In *International Conference on Speech and Signal Processing*, volume 2, pages 801–818.

Ruslan Mitkov, Richard Evans, and Constantin Orăsan. 2002. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexico, February, 17 – 23.

Thomas Morton, Joern Kottmann, Jason Baldridge, and Gann Bierner. 2005. Opennlp: A java-based nlp toolkit. http://opennlp.sourceforge.net.

Massimo Poesio and Mijail A. Kabadjov. 2004. A general-purpos, of-the-shelf anaphora resolution module: implementataion and preliminary evaluation. In *Proceedings of the 2004 international Conference on Language Evaluation and Resources*, pages 663,668.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Joel R. Tetreault. 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.

Yannick Versley, Simone Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. Bart: A modular toolkit for coreference resolution. In *Companion Volume of the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 9–12.

Marilyn A. Walker. 1989. Evaluating discourse processing algorithms. In *ACL*, pages 251–261.

Xiaofeng Yang, Jian Su, Guodong Zhou, and Chew Lim Tan. 2004. Improving pronoun resolution by incorporating coreferential information of candidates. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL2004)*, pages 127–134, Barcelona, Spain, July 21–26.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2006. Kernel-based pronoun resolution with structured syntactic knowledge. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 41–48, Sydney, Australia, July. Association for Computational Linguistics.