Joint Determination of Anaphoricity and Coreference Resolution using Integer Programming

Pascal Denis and **Jason Baldridge**
Department of Linguistics
University of Texas at Austin
{denis,jbaldrid}@mail.utexas.edu

Abstract

Standard pairwise coreference resolution systems are subject to errors resulting from their performing anaphora identification as an implicit part of coreference resolution. In this paper, we propose an integer linear programming (ILP) formulation for coreference resolution which models anaphoricity and coreference as a joint task, such that each local model informs the other for the final assignments. This joint ILP formulation provides f-score improvements of 3.7-5.3% over a base coreference classifier on the ACE datasets.

1 Introduction

The task of coreference resolution involves imposing a partition on a set of entity mentions in a document, where each partition corresponds to some entity in an underlying discourse model. Most work treats coreference resolution as a binary classification task in which each decision is made in a pairwise fashion, independently of the others (McCarthy and Lehnert, 1995; Soon et al., 2001; Ng and Cardie, 2002b; Morton, 2000; Kehler et al., 2004).

There are two major drawbacks with most systems that make pairwise coreference decisions. The first is that identification of anaphora is done *implicitly* as part of the coreference resolution. Two common types of errors with these systems are cases where: (i) the system mistakenly identifies an antecedent for non-anaphoric mentions, and (ii) the system does not try to resolve an actual anaphoric mention. To reduce such errors, Ng and Cardie (2002a) and Ng (2004) use an *anaphoricity* classifier –which has the sole task of saying whether or not *any* antecedents should be identified for each mention– as a filter for their coreference system. They achieve higher performance by doing so; however, their setup uses the two classifiers in a cascade. This requires careful determination of an anaphoricity threshold in order to not remove too many mentions from consideration (Ng, 2004). This sensitivity is unsurprising, given that the tasks are co-dependent.

The second problem is that most coreference systems make each decision independently of previous ones in a greedy fashion (McCallum and Wellner, 2004). Clearly, the determination of membership of a particular mention into a partition should be conditioned on how well it matches the entity as a whole. Since independence between decisions is an unwarranted assumption for the task, models that consider a more global context are likely to be more appropriate. Recent work has examined such models; Luo et al. (2004) using Bell trees, and McCallum and Wellner (2004) using conditional random fields, and Ng (2005) using rerankers.

In this paper, we propose to recast the task of coreference resolution as an optimization problem, namely an integer linear programming (ILP) problem. This framework has several properties that make it highly suitable for addressing the two aforementioned problems. The first is that it can utilize existing classifiers; ILP performs global inference based on their output rather than formulating a

Proceedings of NAACL HLT 2007, pages 236–243,
Rochester, NY, April 2007.

new inference procedure for solving the basic task. Second, the ILP approach supports inference over multiple classifiers, without having to fiddle with special parameterization. Third, it is much more efficient than conditional random fields, especially when long-distance features are utilized (Roth and Yih, 2005). Finally, it is straightforward to create categorical global constraints with ILP; this is done in a declarative manner using inequalities on the assignments to indicator variables.

This paper focuses on the first problem, and proposes to model anaphoricity determination and coreference resolution as a joint task, wherein the decisions made by each locally trained model are mutually constrained. The presentation of the ILP model proceeds in two steps. In the first, intermediary step, we simply use ILP to find a global assignment based on decisions made by the coreference classifier alone. The resulting assignment is one that maximally agrees with the decisions of the classifier, that is, where *all and only* the links predicted to be coreferential are used for constructing the chains. This is in contrast with the usual clustering algorithms, in which a *unique* antecedent is typically picked for each anaphor (e.g., the most probable or the most recent). The second step provides the joint formulation: the coreference classifier is now combined with an anaphoricity classifier and constraints are added to ensure that the ultimate coreference and anaphoricity decisions are mutually consistent. Both of these formulations achieve significant performance gains over the base classifier. Specifically, the joint model achieves f-score improvements of 3.7-5.3% on three datasets.

We begin by presenting the basic coreference classifier and anaphoricity classifier and their performance, including an upperbound that shows the limitation of using them in a cascade. We then give the details of our ILP formulations and evaluate their performance with respect to each other and the base classifier.

2 Base models: coreference classifier

The classification approach tackles coreference in two steps by: (i) estimating the probability, $P_C(\text{COREF}|\langle i,j\rangle)$, of having a coreferential outcome given a pair of mentions $\langle i,j\rangle$, and (ii) applying a selection algorithm that will single out a unique candidate out of the subset of candidates i for which the probability $P_C(\text{COREF}|\langle i,j\rangle)$ reaches a particular value (typically .5).

We use a maximum entropy model for the coreference classifier. Such models are well-suited for coreference, because they are able to handle many different, potentially overlapping learning features without making independence assumptions. Previous work on coreference using maximum entropy includes (Kehler, 1997; Morton, 1999; Morton, 2000). The model is defined in a standard fashion as follows:

$$P_C(\text{COREF}|\langle i,j\rangle) = \frac{exp(\sum_{k=1}^{n} \lambda_k f_k(\langle i,j\rangle, \text{COREF}))}{Z(\langle i,j\rangle)} \quad (1)$$

$Z(\langle i,j\rangle)$ is a normalization factor over both outcomes (COREF and $\neg$COREF). Model parameters are estimated using maximum entropy (Berger et al., 1996). Specifically, we estimate parameters with the limited memory variable metric algorithm implemented in the Toolkit for Advanced Discriminative Modeling[1] (Malouf, 2002). We use a Gaussian prior with a variance of 1000 — no attempt was made to optimize this value.

Training instances for the coreference classifier are constructed based on pairs of mentions of the form $\langle i,j\rangle$, where j and i are the descriptions for an anaphor and one of its candidate antecedents, respectively. Each such pair is assigned either a label COREF (i.e. a positive instance) or a label $\neg$COREF (i.e. a negative instance) depending on whether or not the two mentions corefer. In generating the training data, we followed the method of (Soon et al., 2001) creating for each anaphor: (i) a *positive instance* for the pair $\langle i,j\rangle$ where i is the closest antecedent for j, and (ii) a *negative instance* for each pair $\langle i,k\rangle$ where k intervenes between i and j.

Once trained, the classifier is used to create a set of coreferential links for each test document; these links in turn define a partition over the entire set of mentions. In the system of Soon et. al. (2001) system, this is done by pairing each mention j with each preceding mention i. Each test instance $\langle i,j\rangle$ thus

[1] Available from `tadm.sf.net`.

formed is then evaluated by the classifier, which returns a probability representing the likelihood that these two mentions are coreferential. Soon et. al. (2001) use "Closest-First" selection: that is, the process terminates as soon as an antecedent (i.e., a test instance with probability $> .5$) is found or the beginning of the text is reached. Another option is to pick the antecedent with the best overall probability (Ng and Cardie, 2002b).

Our features for the coreference classifier fall into three main categories: (i) features of the anaphor, (ii) features of antecedent mention, and (iii) relational features (i.e., features that describe properties which hold between the two mentions, e.g. distance). This feature set is similar (though not equivalent) to that used by Ng and Cardie (2002a). We omit details here for the sake of brevity — the ILP systems we employ here could be equally well applied to many different base classifiers using many different feature sets.

3 Base models: anaphoricity classifier

As mentioned in the introduction, coreference classifiers such as that presented in section 2 suffer from errors in which (a) they assign an antecedent to a non-anaphor mention or (b) they assign no antecedents to an anaphoric mention. Ng and Cardie (2002a) suggest overcoming such failings by augmenting their coreference classifier with an anaphoricity classifier which acts as a filter during model usage. Only the mentions that are deemed anaphoric are considered for coreference resolution. Interestingly, they find a degredation in performance. In particular, they obtain significant improvements in precision, but with larger losses in recall (especially for proper names and common nouns). To counteract this, they add *ad hoc* constraints based on string matching and extended mention matching which force certain mentions to be resolved as anaphors regardless of the anaphoricity classifier. This allows them to improve overall f-scores by 1-3%. Ng (2004) obtains f-score improvements of 2.8-4.5% by tuning the anaphoricity threshold on held-out data.

The task for the anaphoricity determination component is the following: one wants to decide for each mention i in a document whether i is anaphoric or not. That is, this task can be performed using a simple binary classifier with two outcomes: ANAPH and $\neg$ANAPH. The classifier estimates the conditional probabilities $P(\text{ANAPH}|i)$ and predicts ANAPH for i when $P(\text{ANAPH}|i) > .5$.

We use the following model for our anaphoricity classifier:

$$P_A(\text{ANAPH}|i) = \frac{exp(\sum_{k=1}^{n} \lambda_k f_k(i, \text{ANAPH}))}{Z(i)} \quad (2)$$

This model is trained in the same manner as the coreference classifier, also with a Gaussian prior with a variance of 1000.

The features used for the anaphoricity classifier are quite simple. They include information regarding (1) the mention itself, such as the number of words and whether it is a pronoun, and (2) properties of the potential antecedent set, such as the number of preceding mentions and whether there is a previous mention with a matching string.

4 Base model results

This section provides the performance of the pairwise coreference classifier, both when used alone (COREF-PAIRWISE) and when used in a cascade where the anaphoricity classifier acts as a filter on which mentions should be resolved (AC-CASCADE). In both systems, antecedents are determined in the manner described in section 2.

To demonstrate the inherent limitations of cascading, we also give results for an oracle system, ORACLE-LINK, which assumes *perfect linkage*. That is, it always picks the correct antecedent for an anaphor. Its only errors are due to being unable to resolve mentions which were marked as non-anaphoric (by the imperfect anaphoricity classifier) when in fact they were anaphoric.

We evaluate these systems on the datasets from the ACE corpus (Phase 2). This corpus is divided into three parts, each corresponding to a different genre: newspaper texts (NPAPER), newswire texts (NWIRE), and broadcasted news transcripts (BNEWS). Each of these is split into a `train` part and a `devtest` part. Progress during the development phase was determined by using cross-validation on only the training set for the NPAPER

System	BNEWS			NPAPER			NWIRE		
	R	P	F	R	P	F	R	P	F
COREF-PAIRWISE	54.4	77.4	63.9	58.1	80.7	67.6	53.8	78.2	63.8
AC-CASCADE	51.1	79.7	62.3	53.7	79.0	63.9	53.0	81.8	64.3
ORACLE-LINK	69.4	100	82.0	71.2	100	83.1	66.7	100	80.0

Table 1: Recall (R), precision (P), and f-score (F) on the three ACE datasets for the basic coreference system (COREF-PAIRWISE), the anaphoricity-coreference cascade system (AC-CASCADE), and the oracle which performs perfect linkage (ORACLE-LINK). The first two systems make strictly local pairwise coreference decisions.

section. No human-annotated linguistic information is used in the input. The corpus text was preprocessed with the OpenNLP Toolkit[2] (i.e., a sentence detector, a tokenizer, a POS tagger, and a Named Entity Recognizer).

In our experiments, we consider only the *true* ACE mentions. This is because our focus is on evaluating pairwise local approaches versus the global ILP approach rather than on building a full coreference resolution system. It is worth noting that previous work tends to be vague in both these respects: details on mention filtering or providing performance figures for markable identification are rarely given.

Following common practice, results are given in terms of recall and precision according to the standard model-theoretic metric (Vilain et al., 1995). This method operates by comparing the equivalence classes defined by the resolutions produced by the system with the gold standard classes: these are the two "models". Roughly, the scores are obtained by determining the minimal perturbations brought to one model in order to map it onto the other model. Recall is computed by trying to map the predicted chains onto the true chains, while precision is computed the other way around. We test significant differences with paired t-tests ($p < .05$).

The anaphoricity classifier has an average accuracy of 80.2% on the three ACE datasets (using a threshold of .5). This score is slightly lower than the scores reported by Ng and Cardie (2002a) for another data set (MUC).

Table 1 summarizes the results, in terms of recall (R), precision (P), and f-score (F) on the three ACE data sets. As can be seen, the AC-CASCADE system generally provides slightly better precision at the expense of recall than the COREF-PAIRWISE system, but the performance varies across the three datasets. The source of this variance is likely due to the fact that we applied a uniform anaphoricity threshold of .5 across all datasets; Ng (2004) optimizes this threshold for each of the datasets: .3 for BNEWS and NWIRE and .35 for NPAPER. This variance reinforces our argument for determining anaphoricity and coreference jointly.

The limitations of the cascade approach are also shown by the oracle results. Even if we had a system that can pick the correct antecedents for all truly anaphoric mentions, it would have a maximum recall of roughly 70% for the different datasets.

5 Integer programming formulations

The results in the previous section demonstrate the limitations of a cascading approach for determining anaphoricity and coreference with separate models. The other thing to note is that the results in general provide a lot of room for improvement — this is true for other state-of-the-art systems as well. The integer programming formulation we provide here has qualities which address both of these issues. In particular, we define two objective functions for coreference resolution to be optimized with ILP. The first uses only information from the coreference classifier (COREF-ILP) and the second integrates both anaphoricity and coreference in a joint formulation (JOINT-ILP). Our problem formulation and use of ILP are based on both (Roth and Yih, 2004) and (Barzilay and Lapata, 2006).

For solving the ILP problem, we use lp_solve, an open-source linear programming solver which implements the simplex and the Branch-and-Bound

[2] Available from `opennlp.sf.net`.

methods.[3] In practice, each test document is processed to define a distinct ILP problem that is then submitted to the solver.

5.1 COREF-ILP: coreference-only formulation

Barzilay and Lapata (2006) use ILP for the problem of aggregation in natural language generation: clustering sets of propositions together to create more concise texts. They cast it as a set partitioning problem. This is very much like coreference, where each partition corresponds to an entity in a discourse model.

COREF-ILP uses an objective function that is based on *only* the coreference classifier and the probabilities it produces. Given that the classifier produces probabilities $p_C = P_C(\text{COREF}|i,j)$, the assignment cost of commiting to a coreference link is $c^C_{\langle i,j\rangle} = -log(p_C)$. A complement assignment cost $\overline{c}^C_{\langle i,j\rangle} = -log(1-p_C)$ is associated with choosing not to establish a link. In what follows, M denotes the set of mentions in the document, and P the set of possible coreference links over these mentions (i.e., $P = \{\langle i,j\rangle | \langle i,j\rangle \in M \times M \text{ and } i < j\}$). Finally, we use indicator variables $x_{\langle i,j\rangle}$ that are set to 1 if mentions i and j are coreferent, and 0 otherwise. The objective function takes the following form:

$$min \sum_{\langle i,j\rangle \in P} c^C_{\langle i,j\rangle} \cdot x_{\langle i,j\rangle} + \overline{c}^C_{\langle i,j\rangle} \cdot (1 - x_{\langle i,j\rangle}) \quad (3)$$

subject to:

$$x_{\langle i,j\rangle} \in \{0,1\} \qquad \forall \langle i,j\rangle \in P \quad (4)$$

This is essentially identical to Barzilay and Lapata's objective function, except that we consider only pairs in which the i precedes the j (due to the structure of the problem). Also, we minimize rather than maximize due to the fact we transform the model probabilities with $-log$ (like Roth and Yih (2004)).

This preliminary objective function merely guarantees that ILP will find a global assignment that maximally agrees with the decisions made by the coreference classifier. Concretely, this amounts to taking all (and only) those links for which the classifier returns a probability above .5. This formulation does not yet take advantage of information from a classifier that specializes in anaphoricity; this is the subject of the next section.

[3] Available from `http://lpsolve.sourceforge.net/`.

5.2 JOINT-ILP: joint anaphoricity-coreference formulation

Roth and Yih (2004) use ILP to deal with the joint inference problem of named entity and relation identification. This requires labeling a set of named entities in a text with labels such as *person* and *location*, and identifying relations between them such as *spouse_of* and *work_for*. In theory, each of these tasks would likely benefit from utilizing the information produced by the other, but if done as a cascade will be subject to propogation of errors. Roth and Yih thus set this up as problem in which each task is performed separately; their output is used to assign costs associated with indicator variables in an objective function, which is then minimized subject to constraints that relate the two kinds of outputs. These constraints express qualities of what a global assignment of values for these tasks must respect, such as the fact that the arguments to the *spouse_of* relation must be entities with *person* labels. Importantly, the ILP objective function encodes not only the best label produced by each classifier for each decision; it utilizes the probabilities (or scores) assigned to each label and attempts to find a global optimum (subject to the constraints).

The parallels to our anaphoricity/coreference scenario are straightforward. The anaphoricity problem is like the problem of identifying the type of entity (where the labels are now ANAPH and ¬ANAPH), and the coreference problem is like that of determining the relations between mentions (where the labels are now COREF or ¬COREF).

Based on these parallels, the JOINT-ILP system brings the two decisions of anaphoricity and coreference together by including both in a single objective function and including constraints that ensure the *consistency* of a solution for both tasks. Let c^A_j and $\overline{c}^A_j$ be defined analogously to the coreference classifier costs for $p_A = P_A(\text{ANAPH}|j)$, the probability the anaphoricity classifier assigns to a mention j being anaphoric. Also, we have indicator variables y_j that are set to 1 if mention j is anaphoric and 0 otherwise. The objective function takes the following

form:

$$min \sum_{\langle i,j\rangle \in P} c^C_{\langle i,j\rangle} \cdot x_{\langle i,j\rangle} + \overline{c}^C_{\langle i,j\rangle} \cdot (1 - x_{\langle i,j\rangle}) + \sum_{j \in M} c^A_j \cdot y_j + \overline{c}^A_j \cdot (1 - y_j) \quad (5)$$

subject to:

$$x_{\langle i,j\rangle} \in \{0,1\} \qquad \forall \langle i,j\rangle \in P \quad (6)$$

$$y_j \in \{0,1\} \qquad \forall j \in M \quad (7)$$

The structure of this objective function is very similar to Roth and Yih's, except that we do not utilize constraint costs in the objective function itself. Roth and Yih use these to make certain combinations impossible (like a *location* being an argument to a *spouse_of* relation); we enforce such effects in the constraint equations instead.

The joint objective function (5) does not constrain the assignment of the $x_{\langle i,j\rangle}$ and y_j variables to be consistent with one another. To enforce consistency, we add further constraints. In what follows, M_j is the set of all mentions preceding mention j in the document.

Resolve only anaphors: if a pair of mentions $\langle i,j\rangle$ is coreferent ($x_{\langle i,j\rangle}$=1), then mention j must be anaphoric (y_j=1).

$$x_{\langle i,j\rangle} \leq y_j \qquad \forall \langle i,j\rangle \in P \quad (8)$$

Resolve anaphors: if a mention is anaphoric (y_j=1), it *must* be coreferent with at least one antecedent.

$$y_j \leq \sum_{i \in M_j} x_{\langle i,j\rangle} \qquad \forall j \in M \quad (9)$$

Do not resolve non-anaphors: if a mention is non-anaphoric (y_j=0), it should have no antecedents.

$$y_j \geq \frac{1}{|M_j|} \sum_{i \in M_j} x_{\langle i,j\rangle} \qquad \forall j \in M \quad (10)$$

These constraints thus directly relate the two tasks. By formulating the problem this way, the decisions of the anaphoricity classifier are not taken on faith as they were with AC-CASCADE. Instead, we optimize over consideration of both possibilities in the objective function (relative to the probability output by the classifier) while ensuring that the final assignments respect the signifance of what it is to be anaphoric or non-anaphoric.

6 Joint Results

Table 2 summarizes the results for these different systems. Both ILP systems are significantly better than the baseline system COREF-PAIRWISE. Despite having lower precision than COREF-PAIRWISE, the COREF-ILP system obtains very large gains in recall to end up with overall f-score gains of 4.3%, 4.2%, and 3.0% across BNEWS, NPAPER, and NWIRE, respectively. The fundamental reason for the increase in recall and drop in precision is that COREF-ILP can posit multiple antecedents for each mention. This is an extra degree of freedom that allows COREF-ILP to cast a wider net, with a consequent risk of capturing incorrect antecedents. Precision is not completely degraded because the optimization performed by ILP utilizes the pairwise probabilities of mention pairs as weights in the objective function to make its assignments. Thus, highly improbable links are still heavily penalized and are not chosen as coreferential.

The JOINT-ILP system demonstrates the benefit ILP's ability to support joint task formulations. It produces significantly better f-scores by regaining some of the ground on precision lost by COREF-ILP. The most likely source of the improved precision of JOINT-ILP is that weights corresponding to the anaphoricity probabilities and constraints (8) and (10) reduce the number of occurrences of non-anaphors being assigned antecedents. There are also improvements in recall over COREF-ILP for NPAPER and NWIRE. A possible source of this difference is constraint (9), which ensures that mentions which are considered anaphoric must have at least one antecedent.

Compared to COREF-PAIRWISE, JOINT-ILP dramatically improves recall with relatively small losses in precision, providing overall f-score gains of 5.3%, 4.9%, and 3.7% on the three datasets.

7 Related Work

As was just demonstrated, ILP provides a principled way to model dependencies between anaphoricity decisions and coreference decisions. In a similar manner, this framework could also be used to capture dependencies among coreference decisions themselves. This option —which we will leave for future work— would make such an approach akin to

System	BNEWS			NPAPER			NWIRE		
	R	P	F	R	P	F	R	P	F
COREF-PAIRWISE	54.4	77.4	63.9	58.1	80.7	67.6	53.8	78.2	63.8
COREF-ILP	62.2	75.5	68.2	67.1	77.3	71.8	60.1	74.8	66.8
JOINT-ILP	62.1	78.0	69.2	68.0	77.6	72.5	60.8	75.8	67.5

Table 2: Recall (R), precision (P), and f-score (F) on the three ACE datasets for the basic coreference system (COREF-PAIRWISE), the coreference only ILP system (COREF-ILP), and the joint anaphoricity-coreference ILP system (JOINT-ILP). All f-score differences are significant ($p < .05$).

a number of recent global approaches.

Luo et al. (2004) use Bell trees to represent the search space of the coreference resolution problem (where each leaf is possible partition). The problem is thus recast as that of finding the "best" path through the tree. Given the rapidly growing size of Bell trees, Luo et al. resort to a beam search algorithm and various pruning strategies, potentially resulting in picking a non-optimal solution. The results provided by Luo et al. are difficult to compare with ours, since they use a different evaluation metric.

Another global approach to coreference is the application of Conditional Random Fields (CRFs) (McCallum and Wellner, 2004). Although both are global approaches, CRFs and ILP have important differences. ILP uses separate local classifiers which are learned without knowledge of the output constraints and are then integrated into a larger inference task. CRFs estimate a global model that directly uses the constraints of the domain. This involves heavy computations which cause CRFs to generally be slow and inefficient (even using dynamic programming). Again, the results presented in McCallum and Wellner (2004) are hard to compare with our own results. They only consider proper names, and they only tackled the task of identifying the correct antecedent only for mentions which have a true antecedent.

A third global approach is offered by Ng (2005), who proposes a global reranking over partitions generated by different coreference systems. This approach proceeds by first generating 54 candidate partitions, which are each generated by a different system. These different coreference systems are obtained as combinations over three different learners (C4.5, Ripper, and Maxent), three sampling methods, two feature sets (Soon et al., 2001; Ng and Cardie, 2002b), and three clustering algorithms (Best-First, Closest-First, and aggressive-merge). The features used by the reranker are of two types: (i) *partition-based* features which are here simple functions of the local features, and (ii) *method-based* features which simply identify the coreference system used for generating the given partition. Although this approach leads to significant gains on the both the MUC and the ACE datasets, it has some weaknesses. Most importantly, the different systems employed for generating the different partitions are all instances of the local classification approach, and they all use very similar features. This renders them likely to make the same types of errors.

The ILP approach could in fact be integrated with these other approaches, potentially realizing the advantages of multiple global systems, with ILP conducting their interactions.

8 Conclusions

We have provided two ILP formulations for resolving coreference and demonstrated their superiority to a pairwise classifier that makes its coreference assignments greedily.

In particular, we have also shown that ILP provides a natural means to express the use of both anaphoricity classification and coreference classification in a single system, and that doing so provides even further performance improvements, specifically f-score improvements of 5.3%, 4.9%, and 3.7% over the base coreference classifier on the ACE datasets.

With ILP, it is not necessary to carefully control the anaphoricity threshold. This is in stark contrast to systems which use the anaphoricity classifier as a filter for the coreference classifier in a cascade setup.

The ILP objective function incorporates the probabilities produced by both classifiers as weights on variables that indicate the ILP assignments for those tasks. The indicator variables associated with those assignments allow several constraints between the tasks to be straightforwardly stated to ensure consistency to the assignments. We thus achieve large improvements with a simple formulation and no fuss. ILP solutions are also obtained very quickly for the objective functions and constraints we use.

In future work, we will explore the use of global constraints, similar to those used by (Barzilay and Lapata, 2006) to improve both precision and recall. For example, we expect transitivity constraints over coreference pairs, as well as constraints on the entire partition (e.g., the number of entities in the document), to help considerably. We will also consider linguistic constraints (e.g., restrictions on pronouns) in order to improve precision.

Acknowledgments

We would like to thank Ray Mooney, Rohit Kate, and the three anonymous reviewers for their comments. This work was supported by NSF grant IIS-0535154.

References

Regina Barzilay and Mirella Lapata. 2006. Aggregation via set partitioning for natural language generation. In *Proceedings of the HLT/NAACL*, pages 359–366, New York, NY.

A. Berger, S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

A. Kehler, D. Appelt, L. Taylor, and A. Simma. 2004. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of HLT/NAACL*, pages 289–296.

Andrew Kehler. 1997. Probabilistic coreference in information extraction. In *Proceedings of EMNLP*, pages 163–173.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, , and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell tree. In *Proceedings of the ACL*.

Robert Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Workshop on Natural Language Learning*, pages 49–55, Taipei, Taiwan.

Andrew McCallum and Ben Wellner. 2004. Conditional models of identity uncertainty with application to noun coreference. In *Proceedings of NIPS*.

Joseph F. McCarthy and Wendy G. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of IJCAI*, pages 1050–1055.

Thomas Morton. 1999. Using coreference for question answering. In *Proceedings of ACL Workshop on Coreference and Its Applications*.

Thomas Morton. 2000. Coreference for NLP applications. In *Proceedings of ACL*, Hong Kong.

Vincent Ng and Claire Cardie. 2002a. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of COLING*.

Vincent Ng and Claire Cardie. 2002b. Improving machine learning approaches to coreference resolution. In *Proceedings of ACL*, pages 104–111.

Vincent Ng. 2004. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of ACL*.

Vincent Ng. 2005. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of ACL*.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of CoNLL*.

Dan Roth and Wen-tau Yih. 2005. Integer linear programming inference for conditional random fields. In *Proceedings of ICML*, pages 737–744.

W. Soon, H. Ng, and D. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings fo the 6th Message Understanding Conference (MUC-6)*, pages 45–52, San Mateo, CA. Morgan Kaufmann.