Proceedings of the 42nd Annual Meeting of the Association for
Computational Linguistics (ACL), Barcelona, Spain, July 2004, pp. 152-159.

Learning Noun Phrase Anaphoricity to Improve Coreference Resolution: Issues in Representation and Optimization

Vincent Ng
Department of Computer Science
Cornell University
Ithaca, NY 14853-7501
yung@cs.cornell.edu

Abstract

Knowledge of the anaphoricity of a noun phrase might be profitably exploited by a coreference system to bypass the resolution of non-anaphoric noun phrases. Perhaps surprisingly, recent attempts to incorporate automatically acquired anaphoricity information into coreference systems, however, have led to the degradation in resolution performance. This paper examines several key issues in computing and using anaphoricity information to improve learning-based coreference systems. In particular, we present a new corpus-based approach to anaphoricity determination. Experiments on three standard coreference data sets demonstrate the effectiveness of our approach.

1 Introduction

Noun phrase coreference resolution, the task of determining which noun phrases (NPs) in a text refer to the same real-world entity, has long been considered an important and difficult problem in natural language processing. Identifying the linguistic constraints on when two NPs can co-refer remains an active area of research in the community. One significant constraint on coreference, the *non-anaphoricity* constraint, specifies that a non-anaphoric NP cannot be coreferent with any of its preceding NPs in a given text.

Given the potential usefulness of knowledge of (non-)anaphoricity for coreference resolution, anaphoricity determination has been studied fairly extensively. One common approach involves the design of heuristic rules to identify specific types of (non-)anaphoric NPs such as pleonastic pronouns (e.g., Paice and Husk (1987), Lappin and Leass (1994), Kennedy and Boguraev (1996), Denber (1998)) and definite descriptions (e.g., Vieira and Poesio (2000)). More recently, the problem has been tackled using unsupervised (e.g., Bean and Riloff (1999)) and supervised (e.g., Evans (2001), Ng and Cardie (2002a)) approaches.

Interestingly, existing machine learning approaches to coreference resolution have performed reasonably well without anaphoricity determination (e.g., Soon et al. (2001), Ng and Cardie (2002b), Strube and Müller (2003), Yang et al. (2003)). Nevertheless, there is empirical evidence that resolution systems might further be improved with anaphoricity information. For instance, our coreference system mistakenly identifies an antecedent for many non-anaphoric common nouns in the absence of anaphoricity information (Ng and Cardie, 2002a).

Our goal in this paper is to improve learning-based coreference systems using automatically computed anaphoricity information. In particular, we examine two important, yet largely unexplored, issues in anaphoricity determination for coreference resolution: *representation* and *optimization*.

Constraint-based vs. feature-based representation. How should the computed anaphoricity information be used by a coreference system? From a linguistic perspective, knowledge of non-anaphoricity is most naturally represented as "bypassing" constraints, with which the coreference system bypasses the resolution of NPs that are determined to be non-anaphoric. But for learning-based coreference systems, anaphoricity information can be simply and naturally accommodated into the machine learning framework by including it as a feature in the instance representation.

Local vs. global optimization. Should the anaphoricity determination procedure be developed independently of the coreference system that uses the computed anaphoricity information (local optimization), or should it be optimized with respect to coreference performance (global optimization)? The principle of software modularity calls for local optimization. However, if the primary goal is to improve coreference performance, global optimization appears to be the preferred choice.

Existing work on anaphoricity determination for anaphora/coreference resolution can be characterized along these two dimensions. Interestingly, most existing work employs constraint-based, locally-optimized methods (e.g., Mitkov et al. (2002) and Ng and Cardie (2002a)), leaving the remaining three possibilities largely unexplored. In particular, to our knowledge, there have been no attempts to (1) globally optimize an anaphoricity determination procedure for coreference performance and (2) incorporate anaphoricity into coreference systems as a feature. Consequently, as part of our investigation, we propose a new corpus-based method for achieving global optimization and experiment with representing anaphoricity as a feature in the coreference system.

In particular, we systematically evaluate all four combinations of local vs. global optimization and constraint-based vs. feature-based representation of anaphoricity information in terms of their effectiveness in improving a learning-based coreference system. Results on three standard coreference data sets are somewhat surprising: our proposed globally-optimized method, when used in conjunction with the constraint-based representation, outperforms not only the commonly-adopted locally-optimized approach but also its seemingly more natural feature-based counterparts.

The rest of the paper is structured as follows. Section 2 focuses on optimization issues, discussing locally- and globally-optimized approaches to anaphoricity determination. In Section 3, we give an overview of the standard machine learning framework for coreference resolution. Sections 4 and 5 present the experimental setup and evaluation results, respectively. We examine the features that are important to anaphoricity determination in Section 6 and conclude in Section 7.

2 The Anaphoricity Determination System: Local vs. Global Optimization

In this section, we will show how to build a model of anaphoricity determination. We will first present the standard, locally-optimized approach and then introduce our globally-optimized approach.

2.1 The Locally-Optimized Approach

In this approach, the anaphoricity model is simply a classifier that is trained and optimized independently of the coreference system (e.g., Evans (2001), Ng and Cardie (2002a)).

Building a classifier for anaphoricity determination. A learning algorithm is used to train a classifier that, given a description of an NP in a document, decides whether or not the NP is anaphoric. Each training instance represents a single NP and consists of a set of features that are potentially useful for distinguishing anaphoric and non-anaphoric NPs. The classification associated with a training instance — one of ANAPHORIC or NOT ANAPHORIC — is derived from coreference chains in the training documents. Specifically, a *positive instance* is created for each NP that is involved in a coreference chain but is not the head of the chain. A *negative instance* is created for each of the remaining NPs.

Applying the classifier. To determine the anaphoricity of an NP in a test document, an instance is created for it as during training and presented to the anaphoricity classifier, which returns a value of ANAPHORIC or NOT ANAPHORIC.

2.2 The Globally-Optimized Approach

To achieve global optimization, we construct a parametric anaphoricity model with which we optimize the parameter[1] for *coreference* accuracy on held-out development data. In other words, we tighten the connection between anaphoricity determination and coreference resolution by using the parameter to generate a set of anaphoricity models from which we select the one that yields the best coreference performance on held-out data.

Global optimization for a constraint-based representation. We view anaphoricity determination as a problem of determining how *conservative* an anaphoricity model should be in classifying an NP as (non-)anaphoric. Given a constraint-based representation of anaphoricity information for the coreference system, if the model is too liberal in classifying an NP as non-anaphoric, then many anaphoric NPs will be misclassified, ultimately leading to a deterioration of recall and of the overall performance of the coreference system. On the other hand, if the model is too conservative, then only a small fraction of the truly non-anaphoric NPs will be identified, and so the resulting anaphoricity information may not be effective in improving the coreference system. The challenge then is to determine a "good" degree of conservativeness. As a result, we can design a parametric anaphoricity model whose conservativeness can be adjusted via a *conservativeness parameter*. To achieve global optimization, we can simply tune this parameter to optimize for coreference performance on held-out development data.

Now, to implement this conservativeness-based anaphoricity determination model, we propose two methods, each of which is built upon a different definition of conservativeness.

Method 1: Varying the Cost Ratio

Our first method exploits a parameter present in many off-the-shelf machine learning algorithms for

[1]We can introduce multiple parameters for this purpose, but to simply the optimization process, we will only consider single-parameter models in this paper.

training a classifier — the cost ratio (cr), which is defined as follows.

$$cr := \frac{\text{cost of misclassifying a positive instance}}{\text{cost of misclassifying a negative instance}}$$

Inspection of this definition shows that cr provides a means of adjusting the relative misclassification penalties placed on training instances of different classes. In particular, the larger cr is, the more conservative the classifier is in classifying an instance as negative (i.e., non-anaphoric). Given this observation, we can naturally define the conservativeness of an anaphoricity classifier as follows. We say that classifier A is more conservative than classifier B in determining an NP as non-anaphoric if A is trained with a higher cost ratio than B.

Based on this definition of conservativeness, we can construct an anaphoricity model parameterized by cr. Specifically, the parametric model maps a given value of cr to the anaphoricity classifier trained with this cost ratio. (For the purpose of training anaphoricity classifiers with different values of cr, we use RIPPER (Cohen, 1995), a propositional rule learning algorithm.) It should be easy to see that increasing cr makes the model more conservative in classifying an NP as non-anaphoric. With this parametric model, we can tune cr to optimize for coreference performance on held-out data.

Method 2: Varying the Classification Threshold We can also define conservativeness in terms of the number of NPs classified as non-anaphoric for a given set of NPs. Specifically, given two anaphoricity models A and B and a set of instances I to be classified, we say that A is more conservative than B in determining an NP as non-anaphoric if A classifies fewer instances in I as non-anaphoric than B. Again, this definition is consistent with our intuition regarding conservativeness.

We can now design a parametric anaphoricity model based on this definition. First, we train in a supervised fashion a probablistic model of anaphoricity $P_{\mathrm{A}}(c \mid i)$, where i is an instance representing an NP and c is one of the two possible anaphoricity values. (In our experiments, we use maximum entropy classification (MaxEnt) (Berger et al., 1996) to train this probability model.) Then, we can construct a parametric model making binary anaphoricity decisions from P_{A} by introducing a threshold parameter t as follows. Given a specific t ($0 \leq t \leq 1$) and a new instance i, we define an anaphoricity model M^t_{A} in which $M^t_{\mathrm{A}}(i)$ = NOT ANAPHORIC if and only if $P_{\mathrm{A}}(c = \text{NOT ANAPHORIC} \mid i) \geq t$. It should be easy to see that increasing t yields progressively more conservative anaphoricity models. Again, t can be tuned using held-out development data.

Global optimization for a feature-based representation. We can similarly optimize our proposed conservativeness-based anaphoricity model for coreference performance when anaphoricity information is represented as a feature for the coreference system. Unlike in a constraint-based representation, however, we cannot expect that the recall of the coreference system would increase with the conservativeness parameter. The reason is that we have no control over whether or how the anaphoricity feature is used by the coreference learner. In other words, the behavior of the coreference system is less predictable in comparison to a constraint-based representation. Other than that, the conservativeness-based anaphoricity model is as good to use for global optimization with a feature-based representation as with a constraint-based representation.

We conclude this section by pointing out that the locally-optimized approach to anaphoricity determination is indeed a special case of the global one. Unlike the global approach in which the conservativeness parameter values are tuned based on labeled data, the local approach uses "default" parameter values. For instance, when RIPPER is used to train an anaphoricity classifier in the local approach, cr is set to the default value of one. Similarly, when probabilistic anaphoricity decisions generated via a MaxEnt model are converted to binary anaphoricity decisions for subsequent use by a coreference system, t is set to the default value of 0.5.

3 The Machine Learning Framework for Coreference Resolution

The coreference system to which our automatically computed anaphoricity information will be applied implements the standard machine learning approach to coreference resolution combining classification and clustering. Below we will give a brief overview of this standard approach. Details can be found in Soon et al. (2001) or Ng and Cardie (2002b).

Training an NP coreference classifier. After a pre-processing step in which the NPs in a document are automatically identified, a learning algorithm is used to train a classifier that, given a description of two NPs in the document, decides whether they are COREFERENT or NOT COREFERENT.

Applying the classifier to create coreference chains. Test texts are processed from left to right. Each NP encountered, NP_j, is compared in turn to each preceding NP, NP_i. For each pair, a test instance is created as during training and is presented to the learned coreference classifier, which returns a number between 0 and 1 that indicates the likelihood that the two NPs are coreferent. The NP with the highest coreference likelihood value among the preceding NPs with coreference class values above 0.5 is selected as the antecedent of NP_j; otherwise, no antecedent is selected for NP_j.

4 Experimental Setup

In Section 2, we examined how to construct locally- and globally-optimized anaphoricity models. Recall that, for each of these two types of models, the resulting (non-)anaphoricity information can be used by a learning-based coreference system either as hard bypassing constraints or as a feature. Hence, given a coreference system that implements the two-step learning approach shown above, we will be able to evaluate the four different combinations of computing and using anaphoricity information for improving the coreference system described in the introduction. Before presenting evaluation details, we will describe the experimental setup.

Coreference system. In all of our experiments, we use our learning-based coreference system (Ng and Cardie, 2002b).

Features for anaphoricity determination. In both the locally-optimized and the globally-optimized approaches to anaphoricity determination described in Section 2, an instance is represented by 37 features that are specifically designed for distinguishing anaphoric and non-anaphoric NPs. Space limitations preclude a description of these features; see Ng and Cardie (2002a) for details.

Learning algorithms. For training coreference classifiers and locally-optimized anaphoricity models, we use both RIPPER and MaxEnt as the underlying learning algorithms. However, for training globally-optimized anaphoricity models, RIPPER is always used in conjunction with Method 1 and MaxEnt with Method 2, as described in Section 2.2.

In terms of setting learner-specific parameters, we use default values for all RIPPER parameters unless otherwise stated. For MaxEnt, we always train the feature-weight parameters with 100 iterations of the improved iterative scaling algorithm (Della Pietra et al., 1997), using a Gaussian prior to prevent overfitting (Chen and Rosenfeld, 2000).

Data sets. We use the Automatic Content Extraction (ACE) Phase II data sets.[2] We choose ACE rather than the more widely-used MUC corpus (MUC-6, 1995; MUC-7, 1998) simply because ACE provides much more labeled data for both training and testing. However, our system was set up to perform coreference resolution according to the MUC rules, which are fairly different from the ACE guidelines in terms of the identification of markables as well as evaluation schemes. Since our goal is to evaluate the effect of anaphoricity information on coreference resolution, we make no attempt to modify our system to adhere to the rules specifically designed for ACE.

The coreference corpus is composed of three data sets made up of three different news sources: Broadcast News (BNEWS), Newspaper (NPAPER), and Newswire (NWIRE). Statistics collected from these data sets are shown in Table 1. For each data set, we train an anaphoricity classifier and a coreference classifier on the (same) set of training texts and evaluate the coreference system on the test texts.

	BNEWS	NPAPER	NWIRE
Number of training texts	216	76	130
Number of test texts	51	17	29
Number of training insts (for anaphoricity)	20567	21970	27338
Number of training insts (for coreference)	97036	148850	122168

Table 1: Statistics of the three ACE data sets

5 Evaluation

In this section, we will compare the effectiveness of four approaches to anaphoricity determination (see the introduction) in improving our baseline coreference system.

5.1 Coreference Without Anaphoricity

As mentioned above, we use our coreference system as the baseline system where no explicit anaphoricity determination system is employed. Results using RIPPER and MaxEnt as the underlying learners are shown in rows 1 and 2 of Table 2 where performance is reported in terms of recall, precision, and F-measure using the model-theoretic MUC scoring program (Vilain et al., 1995). With RIPPER, the system achieves an F-measure of 56.3 for BNEWS, 61.8 for NPAPER, and 51.7 for NWIRE. The performance of MaxEnt is comparable to that of RIPPER for the BNEWS and NPAPER data sets but slightly worse for the NWIRE data set.

5.2 Coreference With Anaphoricity

The Constraint-Based, Locally-Optimized (CBLO) Approach. As mentioned before, in constraint-based approaches, the automatically computed non-anaphoricity information is used as

[2]See `http://www.itl.nist.gov/iad/894.01/tests/ace` for details on the ACE research program.

	System Variation		BNEWS				NPAPER				NWIRE			
	Experiments	L	R	P	F	C	R	P	F	C	R	P	F	C
1	No	RIP	57.4	55.3	56.3	-	60.0	63.6	61.8	-	53.2	50.3	51.7	-
2	Anaphoricity	ME	60.9	52.1	56.2	-	65.4	58.6	61.8	-	54.9	46.7	50.4	-
3	Constraint-	RIP	42.5	77.2	54.8	*cr*=1	46.7	79.3	58.8†	*cr*=1	42.1	64.2	50.9	*cr*=1
4	Based,	RIP	45.4	72.8	55.9	*t*=0.5	52.2	75.9	61.9	*t*=0.5	36.9	61.5	46.1†	*t*=0.5
5	Locally-	ME	44.4	76.9	56.3	*cr*=1	50.1	75.7	60.3	*cr*=1	43.9	63.0	51.7	*cr*=1
6	Optimized	ME	47.3	70.8	56.7	*t*=0.5	57.1	70.6	63.1*	*t*=0.5	38.1	60.0	46.6†	*t*=0.5
7	Feature-	RIP	53.5	61.3	57.2	*cr*=1	58.7	69.7	63.7*	*cr*=1	54.2	46.8	50.2†	*cr*=1
8	Based,	RIP	58.3	58.3	58.3*	*t*=0.5	63.5	57.0	60.1†	*t*=0.5	63.4	35.3	45.3†	*t*=0.5
9	Locally-	ME	59.6	51.6	55.3†	*cr*=1	65.6	57.9	61.5	*cr*=1	55.1	46.2	50.3	*cr*=1
10	Optimized	ME	59.6	51.6	55.3†	*t*=0.5	66.0	57.7	61.6	*t*=0.5	54.9	46.7	50.4	*t*=0.5
11	Constraint-	RIP	54.5	68.6	**60.8***	*cr*=5	58.4	68.8	63.2*	*cr*=4	50.5	56.7	53.4*	*cr*=3
12	Based,	RIP	54.1	67.1	59.9*	*t*=0.7	56.5	68.1	61.7	*t*=0.65	50.3	53.8	52.0	*t*=0.7
13	Globally-	ME	54.8	62.9	58.5*	*cr*=5	62.4	65.6	**64.0***	*cr*=3	52.2	57.0	**54.5***	*cr*=3
14	Optimized	ME	54.1	60.6	57.2	*t*=0.7	61.7	64.0	62.8*	*t*=0.7	52.0	52.8	52.4*	*t*=0.7
15	Feature-	RIP	60.8	56.1	58.4*	*cr*=8	62.2	61.3	61.7	*cr*=6	54.6	49.4	51.9	*cr*=8
16	Based,	RIP	59.7	57.0	58.3*	*t*=0.6	63.6	59.1	61.3	*t*=0.8	56.7	48.4	52.3	*t*=0.7
17	Globally-	ME	59.9	51.0	55.1†	*cr*=9	66.5	57.1	61.4	*cr*=1	56.3	46.9	51.2*	*cr*=10
18	Optimized	ME	59.6	51.6	55.3†	*t*=0.95	65.9	57.5	61.4	*t*=0.95	56.5	46.7	51.1*	*t*=0.5

Table 2: Results of the coreference systems using different approaches to anaphoricity determination on the three ACE test data sets. Information on which **L**earner (**RIP**PER or **M**ax**E**nt) is used to train the coreference classifier, as well as performance results in terms of **R**ecall, **P**recision, **F**-measure and the corresponding **C**onservativeness parameter are provided whenever appropriate. The strongest result obtained for each data set is boldfaced. In addition, results that represent statistically significant gains and drops with respect to the baseline are marked with an asterisk (*) and a dagger (†), respectively.

hard bypassing constraints, with which the coreference system attempts to resolve only NPs that the anaphoricity classifier determines to be anaphoric. As a result, we hypothesized that precision would increase in comparison to the baseline system. In addition, we expect that recall will drop owing to the anaphoricity classifier's misclassifications of truly anaphoric NPs. Consequently, overall performance is not easily predictable: F-measure will improve only if gains in precision can compensate for the loss in recall.

Results are shown in rows 3-6 of Table 2. Each row corresponds to a different combination of learners employed in training the coreference and anaphoricity classifiers.[3] As mentioned in Section 2.2, locally-optimized approaches are a special case of their globally-optimized counterparts, with the conservativeness parameter set to the default value of one for RIPPER and 0.5 for MaxEnt.

In comparison to the baseline, we see large gains in precision at the expense of recall. Moreover, CBLO does not seem to be very effective in improving the baseline, in part due to the dramatic loss in recall. In particular, although we see improvements in F-measure in five of the 12 experiments in this group, only one of them is statistically significant.[4] Worse still, F-measure drops significantly in three cases.

The Feature-Based, Locally-Optimized (FBLO) Approach. The experimental setting employed here is essentially the same as that in CBLO, except that anaphoricity information is incorporated into the coreference system as a feature rather than as constraints. Specifically, each training/test coreference instance $i_{(NP_i, NP_j)}$ (created from NP$_j$ and a preceding NP NP$_i$) is augmented with a feature whose value is the anaphoricity of NP$_j$ as computed by the anaphoricity classifier.

In general, we hypothesized that FBLO would perform better than the baseline: the addition of an anaphoricity feature to the coreference instance representation might give the learner additional flexibility in creating coreference rules. Similarly, we expect FBLO to outperform its constraint-based counterpart: since anaphoricity information is represented as a feature in FBLO, the coreference learner can incorporate the information selectively rather than as universal hard constraints.

Results using the FBLO approach are shown in rows 7-10 of Table 2. Somewhat unexpectedly, this approach is not effective in improving the baseline: F-measure increases significantly in only two of the 12 cases. Perhaps more surprisingly, we see significant drops in F-measure in five cases. To get a bet-

[3] Bear in mind that different learners employed in training anaphoricity classifiers correspond to different parametric methods. For ease of exposition, however, we will refer to the method simply by the learner it employs.

[4] The Approximate Randomization test described in Noreen (1989) is applied to determine if the difference in the F-measure scores achieved by two coreference systems is statistically significant at the 0.05 level or higher.

	System Variation		BNEWS (dev)				NPAPER (dev)				NWIRE (dev)			
	Experiments	L	R	P	F	C	R	P	F	C	R	P	F	C
1	Constraint-	RIP	62.6	76.3	**68.8**	cr=5	65.5	73.0	**69.1**	cr=4	56.1	58.9	57.4	cr=3
2	Based,	RIP	62.5	75.5	68.4	t=0.7	63.0	71.7	67.1	t=0.65	56.7	54.8	55.7	t=0.7
3	Globally-	ME	63.1	71.3	66.9	cr=5	66.2	71.8	68.9	cr=3	57.9	59.7	**58.8**	cr=3
4	Optimized	ME	62.9	70.8	66.6	t=0.7	61.4	74.3	67.3	t=0.65	58.4	55.3	56.8	t=0.7

Table 3: Results of the coreference systems using a constraint-based, globally-optimized approach to anaphoricity determination on the three ACE held-out development data sets. Information on which **L**earner (**RIP**PER or **M**ax**E**nt) is used to train the coreference classifier as well as performance results in terms of **R**ecall, **P**recision, **F**-measure and the corresponding **C**onservativeness parameter are provided whenever appropriate. The strongest result obtained for each data set is boldfaced.

ter idea of why F-measure decreases, we examine the relevant coreference classifiers induced by RIPPER. We find that the anaphoricity feature is used in a somewhat counter-intuitive manner: some of the induced rules posit a coreference relationship between NP_j and a preceding NP NP_i even though NP_j is classified as non-anaphoric. These results seem to suggest that the anaphoricity feature is an irrelevant feature from a machine learning point of view.

In comparison to CBLO, the results are mixed: there does not appear to be a clear winner in any of the three data sets. Nevertheless, it is worth noticing that the CBLO systems can be characterized as having high precision/low recall, whereas the reverse is true for FBLO systems in general. As a result, even though CBLO and FBLO systems achieve similar performance, the former is the preferred choice in applications where precision is critical.

Finally, we note that there are other ways to encode anaphoricity information in a coreference system. For instance, it is possible to represent anaphoricity as a real-valued feature indicating the probability of an NP being anaphoric rather than as a binary-valued feature. Future work will examine alternative encodings of anaphoricity.

The Constraint-Based, Globally-Optimized (CBGO) Approach. As discussed above, we optimize the anaphoricity model for coreference performance via the conservativeness parameter. In particular, we will use this parameter to maximize the F-measure score for a particular data set and learner combination using held-out development data. To ensure a fair comparison between global and local approaches, we do not rely on additional development data in the former; instead we use $\frac{2}{3}$ of the original training texts for acquiring the anaphoricity and coreference classifiers and the remaining $\frac{1}{3}$ for development for each of the data sets. As far as parameter tuning is concerned, we tested values of 1, 2, ..., 10 as well as their reciprocals for cr and 0.05, 0.1, ..., 1.0 for t.

In general, we hypothesized that CBGO would outperform both the baseline and the locally-optimized approaches, since coreference performance is being explicitly maximized. Results using CBGO, which are shown in rows 11-14 of Table 2, are largely consistent with our hypothesis. The best results on all of the three data sets are achieved using this approach. In comparison to the baseline, we see statistically significant gains in F-measure in nine of the 12 experiments in this group. Improvements stem primarily from large gains in precision accompanied by smaller drops in recall. Perhaps more importantly, CBGO never produces results that are significantly worse than those of the baseline systems on these data sets, unlike CBLO and FBLO. Overall, these results suggest that CBGO is more robust than the locally-optimized approaches in improving the baseline system.

As can be seen, CBGO fails to produce statistically significant improvements over the baseline in three cases. The relatively poorer performance in these cases can potentially be attributed to the underlying learner combination. Fortunately, we can use the development data not only for parameter tuning but also in predicting the best learner combination. Table 3 shows the performance of the coreference system using CBGO on the development data, along with the value of the conservativeness parameter used to achieve the results in each case. Using the notation $Learner_1/Learner_2$ to denote the fact that $Learner_1$ and $Learner_2$ are used to train the underlying coreference classifier and anaphoricity classifier respectively, we can see that the RIPPER/RIPPER combination achieves the best performance on the BNEWS development set, whereas MaxEnt/RIPPER works best for the other two. Hence, if we rely on the development data to pick the best learner combination for use in testing, the resulting coreference system will outperform the baseline in all three data sets and yield the best-performing system on all but the NPAPER data sets, achieving an F-measure of 60.8 (row 11), 63.2 (row 11), and 54.5 (row 13) for the BNEWS, NPAPER,

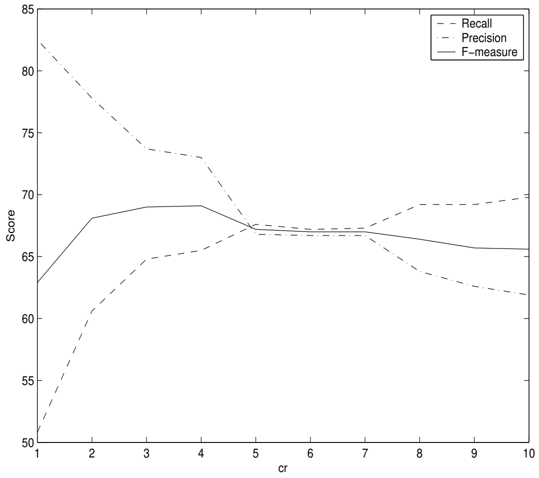

Figure 1: Effect of cr on the performance of the coreference system for the NPAPER development data using RIPPER/RIPPER

and NWIRE data sets, respectively. Moreover, the high correlation between the relative coreference performance achieved by different learner combinations on the development data and that on the test data also reflects the stability of CBGO.

In comparison to the locally-optimized approaches, CBGO achieves better F-measure scores in almost all cases. Moreover, the learned conservativeness parameter in CBGO always has a larger value than the default value employed by CBLO. This provides empirical evidence that the CBLO anaphoricity classifiers are too liberal in classifying NPs as non-anaphoric.

To examine the effect of the conservativeness parameter on the performance of the coreference system, we obtain in Figure 1 the recall, precision, F-measure curves by varying cr for the NPAPER development data using the RIPPER/RIPPER learner combination. As cr increases, recall rises and precision drops. This should not be surprising, since (1) increasing cr causes fewer anaphoric NPs to be misclassified and allows the coreference system to find a correct antecedent for some of them, and (2) decreasing cr causes more truly non-anaphoric NPs to be correctly classified and prevents the coreference system from attempting to resolve them. The best F-measure in this case is achieved when cr=4.

The Feature-Based, Globally-Optimized (FBGO) Approach. The experimental setting employed here is essentially the same as that in the CBGO setting, except that anaphoricity information is incorporated into the coreference system as a feature rather than as constraints. Specifically, each training/test instance $i_{(NP_i,NP_j)}$ is augmented with a feature whose value is the computed anaphoricity of NP_j. The development data is used to select the anaphoricity model (and hence the parameter value) that yields the best-performing coreference system. This model is then used to compute the anaphoricity value for the test instances. As mentioned before, we use the same parametric anaphoricity model as in CBGO for achieving global optimization.

Since the parametric model is designed with a constraint-based representation in mind, we hypothesized that global optimization in this case would not be as effective as in CBGO. Nevertheless, we expect that this approach is still more effective in improving the baseline than the locally-optimized approaches.

Results using FBGO are shown in rows 15-18 of Table 2. As expected, FBGO is less effective than CBGO in improving the baseline, underperforming its constraint-based counterpart in 11 of the 12 cases. In fact, FBGO is able to significantly improve the corresponding baseline in only four cases. Somewhat surprisingly, FBGO is by no means superior to the locally-optimized approaches with respect to improving the baseline. These results seem to suggest that global optimization is effective only if we have a "good" parameterization that is able to take into account how anaphoricity information will be exploited by the coreference system. Nevertheless, effective global optimization with a feature-based representation may not easy to accomplish, since we have no control over whether or how the anaphoricity feature will be used by the coreference learner.

6 Analyzing Anaphoricity Features

So far we have focused on computing and using anaphoricity information to improve the performance of a coreference system. In an attempt to gain linguistic insights into the problem of anaphoricity determination, we examine which anaphoricity features are informative in this section.

Specifically, we measure the informativeness of a feature by computing its **information gain** (see p.22 of Quinlan (1993) for details) on our three data sets for training anaphoricity classifiers. Overall, the most informative features are HEAD_MATCH (whether the NP under consideration has the same head as one of its preceding NPs), STR_MATCH (whether the NP under consideration is the same string as one of its preceding NPs), and PRONOUN (whether the NP under consideration is a pronoun). The high discriminating power of HEAD_MATCH and STR_MATCH is a probable consequence of the fact that an NP is likely to be anaphoric if there is a lexically similar noun phrase preceding it in the text. The informativeness of PRONOUN can also be expected: most pronominal NPs are anaphoric.

Features that determine whether the NP under consideration is a PROPER_NOUN, whether it is a BARE_SINGULAR or a BARE_PLURAL, and whether it begins with an "a" or a "the" (ARTICLE) are also highly informative. This is consistent with our intuition that the (in)definiteness of an NP plays an important role in determining its anaphoricity.

7 Conclusions

We have examined two largely unexplored issues in computing and using anaphoricity information for improving learning-based coreference systems: representation and optimization. In particular, we have systematically evaluated all four combinations of local vs. global optimization and constraint-based vs. feature-based representation of anaphoricity information in terms of their effectiveness in improving a learning-based coreference system.

Extensive experiments on the three ACE coreference data sets using a symbolic learner (RIPPER) and a statistical learner (MaxEnt) for training coreference classifiers demonstrate the effectiveness of the constraint-based, globally-optimized approach to anaphoricity determination, which employs our conservativeness-based anaphoricity model. Not only does this approach improve a "no anaphoricity" baseline coreference system, it is more effective than the commonly-adopted locally-optimized approach without relying on additional labeled data.

Acknowledgments

We thank Regina Barzilay, Claire Cardie, Bo Pang, and the anonymous reviewers for their invaluable comments on earlier drafts of the paper. This work was supported in part by NSF Grant IIS–0208028.

References

David Bean and Ellen Riloff. 1999. Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the ACL*, pages 373–380.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Stanley Chen and Ronald Rosenfeld. 2000. A survey of smoothing techniques for ME models. *IEEE Transactions on Speech on Audio Processing*, 8(1):37–50.

William Cohen. 1995. Fast effective rule induction. In *Proceedings of ICML*.

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393.

Michel Denber. 1998. Automatic resolution of anaphora in English. Technical report, Eastman Kodak Co.

Richard Evans. 2001. Applying machine learning toward an automatic classification of *it*. *Literary and Linguistic Computing*, 16(1):45–57.

Christopher Kennedy and Branimir Boguraev. 1996. Anaphor for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of COLING*, pages 113–118.

Shalom Lappin and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–562.

Ruslan Mitkov, Richard Evans, and Constantin Orasan. 2002. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In Al. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 169–187.

MUC-6. 1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*.

MUC-7. 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Vincent Ng and Claire Cardie. 2002a. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of COLING*, pages 730–736.

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

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypothesis: An Introduction*. John Wiley & Sons.

Chris Paice and Gareth Husk. 1987. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun 'it'. *Computer Speech and Language*, 2.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

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.

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the ACL*, pages 168–175.

Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competitive learning approach. In *Proceedings of the ACL*, pages 176–183.