5-2009

Specialization Methods and Cataphoricity in Coreference Resolution

Robert Staubs
College of William and Mary

Specialization Methods and Cataphoricity in Coreference Resolution

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelors of Science in Computer Science from
The College of William and Mary in Virginia

by

Robert Staubs

Accepted for: _____
(Honors, High Honors, or Highest Honors)

_____
Xipeng Shen, Director

_____
Ann Reed

_____
Haining Wang

Williamsburg, Virginia
May 1, 2009

Abstract

This paper investigates two areas of coreference resolution—specialization and cataphoricity. In doing so I attempt to build upon an existing state-of-the-art system to achieve greater performance.

Coreference systems utilizing specialization of models and specialization of features have both been previously proposed, but no investigation has been made as to their relative effectiveness or possible interrelationship. In this paper I demonstrate that most readily constructible specialization models are equivalent.

The existence of cataphoric mentions is largely ignored in coreference resolution. In this paper I introduce several proposals for countering the performance losses due to cataphora. In particular, I propose a method of cataphoricity classification that largely counters these losses. I present results for several potential methods of using this classifier to create performance *gains*—particularly joint determination using integer linear programming. These methods are demonstrated to be ineffective, providing guidelines for future work.

Acknowledgments

This work was made possible by a number of generous individuals. First, I would like to thank my advisors, Xipeng Shen and Jason Baldridge. Dr. Shen took me on as a student and helped me find a project suited to me personally. He was always helpful in pushing my work along. Dr. Baldridge led me to some great work without knowing me beforehand and always gave great feedback and ideas.

I would also like to thank the two people who helped instigate this project: Virginia Torczon, who recommended me to Dr. Shen, and Nick Luchsinger, who helped bring me into contact with Dr. Baldridge.

Apart from Dr. Baldridge, two others were greatly helpful in getting me acquainted with CoRTex and pushing me along on my project—Pascal Denis and Ben Wing. Their help was hugely valuable.

I am also grateful for the participation of the other members of my Honors committee, Haining Wang and Ann Reed. This thesis owes an additional debt to Dr. Reed for giving me my undergraduate education in syntax.

Finally, I also thank Michael Liarakos for his help as my proofreader.

Contents

1 Introduction

1.1 Problem

Coreference resolution is the task of identifying the sequences of textual mentions corresponding to the entities in a text (Ng and Cardie, 2002b). That is, given a text, the mentions in the text are partitioned into a series of *coreference chains*, each corresponding to a given entity in the underlying discourse. The task may thus be seen as an expanded version of *pronoun resolution*—determining the antecedents of pronouns. The following example sentence illustrates the ideal process involved in resolution—the actual resolution task is generally undertaken on much larger texts, however:

"I want her report before it gets out of date," John said.

"I_1 want her_2 report_3 before it_3 gets out of date," John_1 said.

1: {I, John}, 2: {her}, 3: {report, it}

In coreferring mentions are of two non-exclusive types: anaphora and cataphora. Anaphora corefer with a preceding mention, called an *antecedent*. In the example above, *it* is an anaphor referring back to its antecedent, *report*. Cataphora, in contrast, corefer with a following mention. This following mention may be called the *postcedent*. Above, *I* is a cataphor referring to *John*, its postcedent. Mentions that do not corefer with any others are called single mentions or *singletons*.

Coreference resolution is a vital or valuable component for a number of natural language processing applications. Any system using dialog understanding will necessarily incorporate *some* solution to the problem in order to maintain coherence in conceptual understanding. Text summarization and similar problems benefit from coreference resolution as these tasks similarly require entities to be "understood" in terms of their persistence across a document. The usefulness of coreference resolution is not limited to these areas, however, including also tasks from information retrieval and other areas.

1.2 System Background

For this work I create modifications of the Co-Reference at Texas (CoRTex)[1] coreference software originally developed as part of the Discourse Structure and Coreference Resolution (DISCOR) project (Denis and Baldridge, 2008[2]. The software is written primarily in the Python programming language. CoRTex is a state-of-the-art system and, as such, changes must be substantial and distinct from the existing work in order to yield significant improvements.

[1] Available at `http://comp.ling.utexas.edu/software/cortex`

[2] `http://comp.ling.utexas.edu/discor/`

1.3 Related Previous Work

1.3.1 Single-Candidate Classifiers

Single-candidate classifiers (SCCs) treat the coreference resolution process as a pairwise decision between the options "coreferring" and "non-coreferring." (Soon et al., 2001; Ng and Cardie, 2002b). In these models, a preceding mention is selected as coreferring with the mention under examination on the basis of a probability estimate. This estimate is conditioned solely upon the two mentions taken together. A selection algorithm is applied to produce the output coreference judgments. This method is widely used despite necessary weaknesses in viewing coreference as a binary, pairwise classification (Denis and Baldridge, 2007a; Denis and Baldridge, 2008). SCCs thus form the foundation for my comparisons presented below.

1.3.2 Ranking

Ranking resolution systems make a coreference decision on the basis of many simultaneous potential mentions, ranking them in terms of their score when viewed as coreferring under a trained model. Thus, rather than comparing pairs of mentions as in an SCC, a ranker compares (for example) all preceding mentions. This schema may perhaps be considered more natural than pairwise comparison when resolution is viewed as a linguistic task and, accordingly, success has been had in implementing it as a solution (Denis and Baldridge, 2008). Ranker results can emphasize effects of certain model changes and thus they may be necessary to further work on these problems. Ranker results are not presented here, however.

1.3.3 Integer Linear Programming

Recent work has shown success in using integer linear programming (ILP) specifications of coreference problems, using typical ILP solving methods (Denis and Baldridge, 2007a; Denis and Baldridge, 2008). In this view, coreference is a *constraint satisfaction* problem. This formalization has the appealing property of allowing disparate types of constraints on the problem to be implemented in a non-serial fashion. For example, whereas anaphoricity is usable only as a "filter" in many systems, with ILP it may used as a *constraint* and thus avoid some issues of error propagation. I develop an ILP formulation for cataphora and present results for using it for joint determination of cataphoricity and coreference.

1.3.4 Rich Features

As an alternative to variation in model type or resolution method, some authors have achieved (perhaps surprising) results by creating equally rich *feature sets* (Bengston and Roth, 2008). Such work is especially relevant to this study because of the specialized features found in some of these rich feature sets.

1.4 Goals

1.4.1 Specialization and Interpolation

Different types of mentions are known from theoretical linguistics to behave differently in their relative permissiveness in forming anaphoric relationships. Such mention type specification is in fact key to theoretical analyses of this type of phenomenon (e.g. Ariel, 1988). Because of this, such information is crucial to the task of coreference resolution. Previous work demonstrates the effectiveness of two types of mention-type specialization in the task—model specialization (Denis and Baldridge, 2007a; Denis and Baldridge, 2008) and feature specialization (Bengston and Roth, 2008). In this work these methods are compared, using an established specialized model system and a new implementation of specialized features. Also considered are the effects of utilizing *both* methods and interpolating a "general" model with a corresponding specialized version. The goal of these comparisons and new implementations is to establish what traits are most important to consider in the construction of such systems—and whether some could be safely ignored.

1.4.2 Cataphora Resolution

Cataphoric relationships—those in which a resolvable coreferring mention *follows* the mention otherwise considered as "anaphoric"—are largely ignored in coreference resolution systems. Such systems simply force a strict ordering of the resolved elements, ignoring a property of the data being used. In this work I examine concerns for cataphora resolution including features, specialization, and the possibility of a cataphoricity classifier.

2 Models

Model parameters were estimated throughout by maximum entropy (Berger, et al., 1996) using the limited memory variable metric algorithm implemented in the Toolkit for Advanced Discriminative

Modeling[3] (Malouf, 2002). Throughout a Gaussian prior of variance 1000 was used.

2.1 Single-Candidate Classifier

In a single-candidate classifier (SCC), coreference resolution is viewed as a *classification* task. Roughly stated, for every pair of mentions $\langle i, j \rangle$, a decision is made as to whether or not the two corefer (Soon, et al., 2001; Ng and Cardie, 2002b; Denis and Baldridge, 2007a). In greater detail, the decision is divided into a probability estimation step and a selection step. In the first, an estimation is made as to the probability $P_C\left(COREF|\langle i, j \rangle\right)$ that a given pair $\langle i, j \rangle$ corefers. In the second, a single mention i is chosen out of all the possible mentions for a set j according to a given selection algorithm (e.g. choosing i to maximize $P_C\left(COREF|\langle i, j \rangle\right)$). This pair is chosen as coreferring if a threshold probability is attained (often 0.5).

I use a maximum entropy model for the SCCs used here. Such models are equipped to deal with the potential featural dependencies in coreference resolution and are well-established in the literature (e.g. Kehler, 1997; Morton, 1999; Soon, et al., 2001). The model is defined as:

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

where $Z\left(\langle i, j \rangle\right)$ is a normalization factor, f_k are the weighted features, and λ_k is an additional trained parameter.

2.2 Feature Set

Feature extraction for the models presented here uses only limited processing. Data was processed with a tokenizer, sentence detector, part-of-speech (POS) tagger[4], and the WordNet semantic database[5]. Lexical heads were computed using heuristics based on POS information.

2.2.1 General Features

When a model is not specialized for a particular mention type, all features were used. These are presented here.

[3] Available from `http://tadm.sf.net`

[4] POS tagger from the OpenNLP Toolkit, available at `http://opennlp.sf.net/`

[5] Available at `http://wordnet.princeton.edu/`

- Linguistic Form
  - Third-Person Pronoun
  - Speech Pronoun
  - Reflexive Pronoun
  - Possessive Pronoun
  - Proper Noun
  - Definite Noun Phrase
  - Indefinite Noun Phrase
- Salience
  - First Mention in Sentence
  - Embedded Noun Phrase
- Context
  - Preceding POS
  - Following POS
  - Surrounding POS
- Morphosyntactic Agreement
  - Gender
  - Number
  - Person
- Distance
  - Sentence Distance
  - Mention Distance
- String Similarity
  - Exact Match
  - Left Substring Match
  - Right Substring Match
  - Head Word Match
- Acronym
- Appositive

- **Linguistic Form**

  These features describe the linguistic form of a mention. There is a binary feature for each type of form considered. These forms are: third-person pronoun, speech pronoun, reflexive pronoun, possessive pronoun, proper noun, definite noun phrase (NP), and indefinite noun phrase. Features were created for both the linguistic form of the anaphor and that of the antecedent.

- **Salience**

  These features attempt to capture the relative salience of an antecedent. The first is true if an antecedent is the first mention within a sentence. The second is true if the antecedent is an embedded NP.

- **Context**

  Three features describe the part-of-speech of the mentions near the antecedent. One lists the POS of the preceding mention, one the following mention, and the final one describes the pair of surrounding mentions taken together.

- **Morphosyntactic Agreement**

  These three features describe pairs of attributes held by the anaphor and antecedent. The three attributes used are: gender, number, and person.

- **Distance**

  These features are binned values for distances between the anaphor and antecedent. One distance value is counted in sentences, the other in mentions.

- **String Similarity**

  Four features describe the similarity of the anaphor and antecedent in terms of their string properties. The first feature is true if an exact match exists between the two. The second is true if one is a left substring of the other, the third if a substring is found on the right instead. Finally, if they share the same head word, a fourth feature is true.

- **Apposition**

  True if anaphor and antecedent are in an appositive structure.

- **Acronym**

  True if the anaphor is an acronym of the antecedent or *vice versa*.

2.2.2 Specialized Models

For specialized models, described in Section 3.2, subsets of the above feature set are used. These are listed in Table 1.

Feature	*3P*	*SP*	*PN*	*Def-NP*	*Other*
Linguistic Form	X	X	X	X	X
Salience	X	X	X	X	X
Context	X	X	X	X	X
Agreement	X	X	X	X	X
Distance	X	X	X	X	X
String Similarity			X	X	X
Apposition			X	X	
Acronym			X		

Table 1: *Features used for types of mentions in specialized models.*

3 Specialization and Interpolation

3.1 Problem

The problem of coreference is necessarily deeply concerned with distinctions of mention type. Different antecedent-anaphor relations are seen depending on the type of anaphor, meaning that such variation is likely critical to fully-developed resolution systems. In theoretical linguistics, the behaviors of different noun phrase types are analyzed in accordance with their information status,

salience, and (thus) their "accessibility," evaluated throughout sentences and the discourse (Ariel, 1988). The mention types with which we are concerned in coreference are: speech pronouns (those used in dialog situations—first and second person pronouns), third-person pronouns, proper names, definite descriptions, and other (e.g. certain indefinite, quantified, and bare noun phrases). Despite the theoretical importance of mention type to anaphoric phenomena, coreference resolution systems have typically not given them more than basic featural importance, using fully monolithic systems (Denis and Baldridge, 2008). In this work I compare two ways of going beyond this understanding of mention type towards heterogeneous models and present potential ways of integrating *both*.

3.2 Approaches

The first adaptation to mention-type specialization is specialization of *models*. Denis and Baldridge (2008) develop a set of models for coreference resolution by SCCs or rankers, each model trained and operating on a particular subset of mentions (according to the division outlined above). This simple idea allows very distinct training results for parameters, reflecting the distinct behavior of these types as observed.

present an SCC with an expanded featural inventory and demonstrate that it competes well with seemingly more sophisticated models. One of their significant introductions is *specialized features*. In their approach, the feature set is supplemented with features formed by the conjunction of each pair of other features. Given the (quite typical) features in their set for mention type, this generates a subset of features indexed by mention type. This setup forms an interesting contrast with specialized models—in the model specialization view, learning proceeds with entirely divided feature sets. In the feature specialization view, however, the set used for learning is unified (but not homogenous).

3.3 Combined Approaches

Specialized model and specialized feature solutions may not capture the same properties of the data due to their different architectures. It is thus worth considering combining the two into one system. As a first option one might immediately think of using specialized features *within* specialized models—this approach, however, is not as different from specialized models as might first be thought. Specialized models represent a partition of mentions—separating their processing in accordance with their type. This separation is concerned with anaphora, and as such all anaphor-type conjoined features simply duplicate the theoretical motivation implied by the division of models. A distinction *may* be present, however, in the manner in which the weights are learned given each structure. I

will therefore include comparative results which include specialized resolution utilizing specialized features.

A perhaps less obvious approach to combining the two types of specialization is *interpolation*. In this approach, building from previous work on discourse connectives (Elwell and Baldridge, 2008), the estimated probability for a given evaluation is changed to reflect a weighted average of the probabilities given by each model. In this way, more emphasis may be given to the estimates originating from specialized models as compared to specialized features (or vice versa). An estimate representing such an interpolation is constructed as follows (using an SCC probability statement for simplicity):

$$P_{ISC}\left(COREF|\langle i,j\rangle\right) = \lambda_{t_k} P_{t_k}\left(COREF|\langle i,j\rangle\right) + \left(1-\lambda_{t_k}\right) P_{SC}\left(COREF|\langle i,j\rangle\right)$$

where P_{ISC} is the interpolated probability, P_{t_k} is the probability given by the specialized model for type k, P_{SC} is the probability given by the specialized feature model, and λ_{t_k} is the relative weight given to the specialized model for type k ($\lambda_{t_k} \in [0,1]$).

The form of the interpolation—as well as comments by Denis and Baldridge (2008)—suggest an interpolation between a general model and the specialized models, removing the specialized features:

$$P_{IC}\left(COREF|\langle i,j\rangle\right) = \lambda_{t_k} P_{t_k}\left(COREF|\langle i,j\rangle\right) + \left(1-\lambda_{t_k}\right) P_{C}\left(COREF|\langle i,j\rangle\right)$$

where P_{IC} is the interpolated probability and P_C is the probability given by the general model. Such an interpolated model is useful to compare with the one above—it may give results in allowing wider generalizations without being pinned down by specialized features.

The definition of λ_{t_k} must be addressed. One approach would be to set it to a fixed value λ_0 for all mention types. This value would be chosen to maximize performance on held-out data. An alternative, however, is to weight λ_{t_k} by the frequency of encountered mentions of type k as:

$$\lambda_{t_k} = \frac{n\left(k\right)}{n\left(k\right)+C}$$

where $n\left(k\right)$ is the frequency of mentions of type k in the training data and C is a constant chosen for performance on held-out data. This approach serves to increase the relative importance of a specialized model when more data is acquired for it. As $n\left(k\right)$ grows, the fraction approaches 1. That is, as the amount of data on a particular type increases, the reliance on a specialized model for that type also increases. C here then serves to bias this importance, establishing the degree

to which the "general" model is relied upon. In the interpolation between two different types of specialization this trait is perhaps not vital—it is potentially of more use with the interpolation with a fully general model. Both definitions of λ_{t_k} are considered below. However, the mention count-weighted definition is *a priori* considered more valuable.

4 Cataphora Resolution

4.1 Problem

Cataphora is a phenomenon which is in essence the inverse of anaphora. A mention is "cataphoric" if it corefers with a *following* mention—the *postcedent*. This is opposed to anaphora, in which the anaphoric mention corefers with a *preceding* mention. Because of the symmetry of these two phenomena, many anaphora may be viewed as postcedents from the perspective of their antecedents—taking these, in turn, as cataphora. The distinction is of particular use in cases where a more general mention precedes a specific one. Perhaps the most salient example is of pronouns preceding their coreferents in quotations, for example:

"I_i am going to the store," John_i said.

Such expressions do occur in text and speech, but they are largely ignored in coreference resolution. They are, in fact, often categorically excluded from acting as candidates for links. The reasoning behind this prohibition is the duality of anaphoric and cataphoric roles previously mentioned. If a mention can be taken as either anaphoric or cataphoric (which is generally the case), any true chain of mentions of length three or greater is potentially disruptable. That is, the second mention may be linked with the third, treating the second as cataphoric. If this is done the potential to link the first mention with the rest of the chain may be lost. The essential issue of a system that allows cataphora, then, is to not overgenerate cataphoric pairings. Efforts in this direction are discussed below.

4.2 Features

The feature set described in Section 2.2 is insufficient for cataphora resolution in two respects. The first is ordering and the second is quotation.

No account is actually present in the current feature set as to relative order of anaphor and

antecedent[6]. This is because no distinction needs to be made—the antecedent is assumed to always precede the anaphor. In cataphora resolution this is not necessarily the case—the "antecedent" may actually follow the "anaphor" as the postcedent of a cataphor. Because of this, an additional feature **Antecedent Precedes Anaphor** must be added. This feature is true if and only if the "antecedent" considered occurs before the mention for which features are assembled.

Critically, also, no mention is made in the features above of whether a given mention is quoted or not. This is clearly an important feature for cataphora and is therefore included as **Quoted**, a feature true of a given mention if and only if that mention occurs within quotes. An alternate formulation would split quotation into three features: **Anaphor Quoted** (true if the anaphor is within a quotation), **Antecedent Quoted** (true if the antecedent is), and **Intervening Quote** (true if anaphor is within a quotation but the antecedent is not). This distinction creates only negligibly different results (e.g. identical to within ± 0.1 on a given score) and therefore the simpler single-feature system is presented.

These two features may be most appropriately related to particular types of mentions. Quotation, in particular, appears most closely related to speech pronouns. This distinction is not, however, made in the compilation of features. They are both assigned to all mention types regardless of the specialization of a model.

4.3 Cataphoricity Classifier

Due to the risks outlined above for overgenerating cataphoric links, it is perhaps desirable to make a decision for each mention as to whether to resolve it cataphorically. We would therefore like to create a "cataphoricity" classifier in line with anaphoricity classifiers (Ng and Cardie, 2002a; Denis and Baldridge, 2007a). Such a classifier would be used to determine whether to resolve a particular mention anaphorically or cataphorically.

Because of the very nature of the problem we would like to solve, however, definitional problems develop at this point. It is not possible to tag a given mention as being "cataphoric" as *opposed* to being anaphoric for a given chain. There is no hard boundary between the two—unless a mention occurs at an end of a chain it may only be viewed as more or less typical for a given stance. It is thus useful to relax the goals of the classifier.

Instead of creating a classifier to overtly make the anaphoric versus cataphoric decision, it is more tractable to use a classifier which decides whether or not to *permit* a cataphoric link. That

[6]When discussing features and the general principles of coreference resolution, *anaphor* and *antecedent* are taken to stand also for *cataphor* and *postcedent* when applicable.

is, instead of permitting a cataphoric link and *prohibiting* an anaphoric one, it would instead be used only to remove a general prohibition on cataphoric links. In this model, therefore, anaphora is taken as default—the unmarked case. It is *deviations* which we would like to be able to detect and accommodate.

Formally, then, the cataphoricity classifier decides for every mention i whether to permit a cataphoric link involving i. That is, it performs a binary classification into the classes $\textsc{Cata}_{base}$ and $\neg CATA$. $P_{Cata}(CATA|i)$ is estimated and a decision is made for $\textsc{Cata}_{base}$ if this probability exceeds a threshold P_{thresh} ($= 0.5$ without qualification). $\neg CATA$ is predicted otherwise. The form of $P_{Cata}(CATA|i)$ is similar to $P_C(COREF|\langle i,j\rangle)$ except that it uses the single mention i rather than the pair $\langle i,j\rangle$:

$$P_{Cata}(CATA|i) = \frac{exp\left(\sum_{k=1}^{n}(\lambda_k f_k(i, CATA))\right)}{Z(i)}$$

This "permission" based schema must be trained on an approximation or heuristic. Here, a mention is marked as "allowing a cataphoric interpretation" if any of the following is true:

1. The next following coreferring mention is closer than the preceding one.
2. The mention is quoted.
3. There are no preceding coreferring mentions.

4.3.1 Features

- Linguistic Form
  - Pronoun
  - Proper Noun
  - Common Noun
  - Speech Pronoun
  - Reflexive Pronoun
  - Short Pronoun
  - Definite NP
  - Short Definite NP
  - Indefinite NP
  - Quantified NP
  - Demonstrative NP
  - Possessive NP
  - Bare NP
- Comparison with Previous Mention
  - String Match with Following
  - Head Match with Following
  - Appositive with Following
  - Embedded with Following
- Comparison with Following Mention
  - String Match with Previous
  - Head Match with Previous
  - Appositive with Previous
  - Embedded with Previous
- Length

The features used in the cataphoricity classifier are more simplistic than those presented in Section 2.2 and there are relatively few of them. They largely constitute the features used in Denis and Baldridge (2007a) for their anaphoricity classifier.

Features corresponding to linguistic form are similar to those in the general model, with different specifications. A feature is also introduced which encodes mention length. Features relating the mention with its surrounding mentions replace previous features between the anaphor and its antecedent. The added features comparing the mention with its following mention are where I differ from Denis and Baldridge's (2007a)—they have no need of these features as they create an anaphoricity classifier, not a cataphoricity one.

4.4 Joint Determination of Cataphoricity

One risk of using a cataphoricity classifier as developed here is a problem of *error cascade*. When a classifier is used as a kind of filter in sequence before coreference, interrelationships between these decisions are lost. If cataphoricity and coreference judgments are made simultaneously the system might better be able to balance competing constraints. Denis and Baldridge (2007a; 2009) use this insight to develop joint determination formulations of anaphoricity classification, named entity classification, and coreference resolution. They do this by formulating these tasks as an integer linear programming (ILP) problem. They use probabilities for the models to set up *costs* used in an *objective function*. The properties of the models' mutual restrictions are then set up as ILP *constraints*. Classification and resolution then is the problem of minimizing the objective function within the constraints. I will present their model for joint determination of anaphoricity and coreference, as it is most relevant to the formulation I develop.

Given coreference classifier probabilities $p_C = P_C\left(COREF|\langle i,j\rangle\right)$, the associated cost of establishing a coreference link is defined as $c^C_{\langle i,j\rangle} = -log\left(p_C\right)$. An associated complement cost of *not* establishing a link is also created: $\bar{c}^C_{\langle i,j\rangle} = -log\left(1-p_C\right)$. Analogously, anaphoricity classifier probabilities $p_A = P_A\left(ANAPH|j\right)$ are used to create the costs of treating anaphorically and non-anaphorically as $c^A_j = -log\left(p_A\right)$ and $\bar{c}^A_j = -log\left(1-p_A\right)$, respectively.

$\mathcal{M}$ is defined as the set of all mentions in a document and $\mathcal{P}$ as the set of all possible links between them: $\mathcal{P} = \{(i,j)|(i,j) \in \mathcal{M}\times\mathcal{M}, i<j\}$. An indicator variable $x_{\langle i,j\rangle}$ is 1 if i and j corefer, 0 otherwise. Similarly, y_j is 1 if j is anaphoric, 0 otherwise.

The objective function is thus:

$$\min\left(\sum_{(i,j)\in\mathcal{P}}\left(c^C_{\langle i,j\rangle}x_{\langle i,j\rangle}+\bar{c}^C_{\langle i,j\rangle}\left(1-x_{\langle i,j\rangle}\right)\right)+\sum_{j\in\mathcal{M}}\left(c^A_j y_j+\bar{c}^A_j\left(1-y_j\right)\right)\right)$$

such that $\forall(i,j)\in\mathcal{P}$, $x_{\langle i,j\rangle}\in\{0,1\}$ and $\forall j\in\mathcal{M}$, $y_j\in\{0,1\}$. That is, the predicted cost of coreference decisions over all mention pairs and the predicted cost of anaphoricity decisions over all mentions is minimized.

This minimization is subject to two constraints: "resolve only anaphora" and "resolve all anaphora." The definitions of these require the additional set $\mathcal{M}^-_j$, the set of all mentions preceding j. The constraints are expressed as restrictions on the corresponding indicator variables.

Resolve only anaphora requires that only mentions marked as anaphoric may serve as the anaphoric mention in a coreference link:

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

Resolve all anaphora requires that if a mention is marked as anaphoric it must be an anaphor in *some* coreference link:

$$y_j\leq\sum_{i\in\mathcal{M}^-_j}x_{\langle i,j\rangle}\quad\forall j\in\mathcal{M}$$

This formulation allows anaphoricity and coreference resolution decisions to influence one another. Thus, they are jointly determined. This schema is potentially valuable to adopt for cataphoricity. I therefore propose an ILP-based joint determination approach to cataphoricity classification and coreference resolution.

I define additional cataphoricity costs. These are based on the probabilities from the cataphoricity classifier, $p_{Cata}=P_{Cata}\left(CATA|j\right)$, as $c^{Cata}_j=-log\left(p_{Cata}\right)$ and $\bar{c}^{Cata}_j=-log\left(1-p_{Cata}\right)$.

As the previous definition of the set $\mathcal{P}$ required that mention i precede j, it must be redefined to allow cataphoric resolution. Thus the set $\mathcal{P}'$ is defined as $\mathcal{P}'=\{(i,j)|(i,j)\in\mathcal{M}\times\mathcal{M},i\neq j\}$. Otherwise the objective function is the same as the preceding one, replacing anaphoric variables and costs with cataphoric ones. The indicator variable q_j is defined to be 1 if j is cataphoric, 0 otherwise. This gives the following objective function:

$$\min\left(\sum_{(i,j)\in\mathcal{P}'}\left(c^C_{\langle i,j\rangle}x_{\langle i,j\rangle}+\bar{c}^C_{\langle i,j\rangle}\left(1-x_{\langle i,j\rangle}\right)\right)+\sum_{j\in\mathcal{M}}\left(c^{Cata}_j q_j+\bar{c}^{Cata}_j\left(1-q_j\right)\right)\right)$$

such that $\forall(i,j)\in\mathcal{P}'$, $x_{\langle i,j\rangle}\in\{0,1\}$ and $\forall j\in\mathcal{M}$, $q_j\in\{0,1\}$.

Two parallel constraints to the anaphoric ones are possible, "cataphorically resolve only cataphora" and "cataphorically resolve all cataphora." For these I define M_j^+, the set of all mentions following j:

Cataphorically resolve only cataphora requires that if a mention is decided to be cataphoric, it must be resolved as the cataphor in a coreference link. That is, it must not only be resolved but also must be resolved *cataphorically*:

$$x_{\langle i,j \rangle} \leq q_j \quad \forall (i,j) \in \mathcal{P}', j > i$$

Cataphorically resolve all cataphora requires, in turn, that if a mention is marked as cataphoric it must be resolved as a cataphor in some coreference link:

$$q_j \leq \sum_{i \in \mathcal{M}_j^+} x_{\langle i,j \rangle} \quad \forall j \in \mathcal{M}$$

The objective function, taken together with these two constraints, forms a joint determination approach to coreference resolution and cataphoricity classification. A potential question is whether "cataphorically resolve all cataphora" is a tenable constraint for the definition of the cataphoricity classifier used here. The classifier as constructed *cannot* know that a cataphoric link is the best one for a given mention. It therefore can only decide between allowing the *possibility* of a cataphoric mention or not. *Requiring* these decisions to be fulfilled as cataphoric resolutions, therefore, seems likely to cause error. I will therefore present results for ILP formulations with and without "cataphorically resolve all cataphora."

4.5 Anaphoricity Biasing

A simple alternative method of allowing but restricting cataphoric links is to implement a filter in the link selection algorithm. As default in my tests I used the "Best-First" link selection method (Ng and Cardie, 2002b). That is, the highest scoring pair is the one chosen for a link. As an alternative for cataphora resolution I propose an "Anaphora-Biased Best-First" selection algorithm. This method is identical to Best-First except that if *any* potential anaphoric interpretation exists for a mention (i.e. some candidate antecedent exceeds the threshold probability of 0.5) this interpretation is chosen over a cataphoric interpretation. Thus cataphoric links are only created when no strong anaphoric links can be. This severely limits the generation of cataphoric links—perhaps to a fault. It should, however, form a useful comparison with other methods.

5 Experiments, Results, and Discussion

All experiments were performed using modifications of the CoRTex[7] coreference system.

5.1 Metrics

In reporting results I use the following three metrics: *MUC* (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and *CEAF* (Luo, 2005). For each I report recall, precision, and F-score.

F-score is computed in the same way for each as the harmonic mean of precision and recall:

$$F = \frac{2 \times P \times R}{P + R}$$

Additionally I report the mean F-score over the three metrics, $\bar{F}$.

Details on these metrics follow below. Each makes use of the set $\mathcal{S}$ of coreference chains created by the system as well as the set $\mathcal{T}$ of true chains.

5.1.1 *MUC*

The *MUC* (Message Understanding Conference) metric is a *link-based* metric. That is, it evaluates resolution on the basis of *pairs* of mentions. Recall is defined as the percentage of true links (links in $\mathcal{T}$) contained in $\mathcal{S}$. Precision is defined as the percentage of predicted links (links in $\mathcal{S}$) contained in $\mathcal{T}$. *MUC* therefore is biased towards systems that create long chains and does not evaluate singleton entities.

5.1.2 B^3

The B^3 metric, in contrast, is *mention-based*. In this metric, precision and recall are calculated for every mention individually. For every mention m, recall is calculated as the percentage of mentions in the true chain of m within $\mathcal{T}$ that are also contained within the chain of m within $\mathcal{S}$. Similarly, precision is defined as the percentage of mentions in the predicted chain of m in $\mathcal{S}$ that are also contained within its true chain in $\mathcal{T}$. The overall B^3 metric reported is the mean over all m. The problems of *MUC* do not arise for this metric.

[7] v0.1. Available at `http://comp.ling.utexas.edu/software/cortex`

5.1.3 *CEAF*

Finally, the *CEAF* metric is based on *entities*. This metric finds the best mapping between chains in $\mathcal{S}$ and $\mathcal{T}$—that is, it creates a set of pairs $(\mathcal{S}_i, \mathcal{T}_i)$ which maximizes the overall number of mentions shared between the pairs. Recall is defined as the percentage of mentions in $\mathcal{T}$ which are shared between these pairs thus created. Precision is the percentage of mentions in $\mathcal{S}$ which are shared in such a way. Note that for *CEAF* precision is equal to recall for true mentions. Thus I present only a single value for this metric below.

5.2 Data

I use Phase 2 of the ACE corpus. This corpus is split into three genres: newspaper texts (NPAPER), newswire texts (NWIRE), and broadcast news texts (BNEWS). Each genre is in turn divided into `train` and `devtest` sections. For development I made use of the `train` sections of NPAPER primarily, with reliance on the train section of all genres for some parameters. No use was made of `devtest` prior to final experiments. All resulted presented here are for training and testing on *all* sections unless otherwise noted.

5.3 Specialization and Interpolation

Results for specialization method comparison and interpolation were obtained using the models described above. For both training and testing, all sections of the ACE corpus were used.

General classifier training and application methodology was that of Ng and Cardie (2002). During training instances were created by pairing each anaphor in the training set with all preceding candidate mentions up to and including its antecedent. If the closest preceding antecedent was pronominal the antecedent used was the closest overall. Otherwise, the closest non-pronominal antecedent was used instead. During classification the candidate pair with the highest score was selected—that is, Best-First link selection was used.

5.3.1 Specialization of Models and Features

Experiments were first conducted to compare the performance of a baseline general model lacking specialized features (BASE), a set of specialized models (SPEC$_{models}$), and a model with specialized features (SPEC$_{features}$). These results are presented in Table 2.

As anticipated, both SPEC$_{models}$ and SPEC$_{features}$ outperform BASE on all metrics. This dif-

Test	$MUC:R$	$MUC:P$	$MUC:F$	$B^3:R$	$B^3:P$	$B^3:F$	$CEAF$	$\bar{F}$
BASE	60.6	72.3	65.9	62.2	77.2	68.9	62.2	65.667
SPEC$_{models}$	64.4	74.2	68.9	64.7	78.5	70.9	64.5	68.100
SPEC$_{features}$	63.2	73.6	68.0	64.2	78.2	70.5	64.2	67.567

Table 2: *Baseline model and two types of specialization.*

ference is significant for all measures.[8] It might perhaps also be expected that the specialized models perform better than a model with specialized features. However, the difference observed does not reach the level of significance for any measure. This is an interesting observation because the overhead costs associated with the two types of system may differ. Significance is approached for differences in precision but reached only if SPEC$_{models}$ is assumed to outperform SPEC$_{features}$.

5.3.2 Combined Models

Next, experiments were carried out to compare differing techniques for combining specialized models and specialized features—a set of specialized models with specialized features (SPEC$_{models,features}$), a set of specialized models interpolated with a general model (I(SPEC$_{models}$, BASE)), and a set of specialized models interpolated with a specialized feature model (I(SPEC$_{mod}$, SPEC$_{feat}$)). Experiments for specialized models using specialized features interpolated with a specialized feature model were also performed (I(SPEC$_{mod,feat}$, SPEC$_{feat}$)). The results of these experiments are shown in Table 3.

For these experiments, λ for interpolation was determined using a C value of 2000. This value was chosen on an 80%/20% (training and testing, respectively) partition of ACE `train` data—no `devtest` data was used. The interpolation used in determining this value was between a set of specialized models and a general-features model (i.e. similar to I(SPEC$_{models}$, BASE)). C was considered to maximize performance if it maximized $\bar{F}$. Surprisingly, there is no significant difference

Test	$MUC:R$	$MUC:P$	$MUC:F$	$B^3:R$	$B^3:P$	$B^3:F$	$CEAF$	$\bar{F}$
SPEC$_{models}$	64.4	74.2	68.9	64.7	78.5	70.9	64.5	68.100
SPEC$_{features}$	63.2	73.6	68.0	64.2	78.2	70.5	64.2	67.567
SPEC$_{models,features}$	65.2	73.7	69.2	65.3	77.5	70.9	64.5	68.200
I(SPEC$_{models}$, BASE)	64.4	74.2	68.9	64.6	78.5	70.9	64.5	68.100
I(SPEC$_{models}$, SPEC$_{feat}$)	65.2	73.7	69.2	65.3	77.6	70.9	64.5	68.200
I(SPEC$_{mod,feat}$, SPEC$_{feat}$)	65.1	73.6	69.1	65.3	77.6	70.9	64.6	68.200

Table 3: *Results comparing models with specialization and three methods of combination. ($C = 2000$)*

[8] All significance judgments presented here are on the basis of a two-tailed t-test with $p = 0.05$. If significance is approached but not reached, a one-tailed test result may be given as additional information—not in lieu of the two-tailed result. Performance is said to differ significantly if F-scores differ significantly.

between any of the systems in Table 3. Again, differences in precision between models with only specialized features and systems with only specialized models approach significance and reach it if assumptions are made about directionality.

The interpolated models are then disappointing in this respect. Each does not improve upon its closest corresponding non-interpolated system. This is possibly related to the chosen value of $C = 2000$. This value is greatly surpassed by the number of instances created for most mention types. As such, the value of λ approaches 1—reliance is almost entirely on the set of specialized models. This would not explain, however, the lack of significant difference between specialized feature models and specialized model systems.

5.3.3 Varying λ

Given that the data on which the decision of $C = 2000$ is different from the data used in experiments, it is worth investigating whether the more simplistic method of assigning a constant λ is effective. Towards this end a series of other experiments were carried out using values of $\lambda \in [0.0, 1.0]$, $\lambda = 0.1k$. These λ values are not mention count-sensitive. Results for interpolation between specialized models and a general model are presented in Table 4. (Note that $\lambda = 0.0$ and $\lambda = 1.0$ correspond to a general unspecialized model and a specialized model system alone, respectively.)

λ	$MUC:R$	$MUC:P$	$MUC:F$	$B^3:R$	$B^3:P$	$B^3:F$	$CEAF$	$\bar{F}$
$C = 2000$	64.4	74.2	68.9	64.6	78.5	70.9	64.5	68.100
0.0	60.6	72.3	65.9	62.2	77.2	68.9	62.2	65.667
0.1	63.9	73.6	68.4	64.1	78.0	70.4	63.6	67.467
0.2	64.0	73.7	68.5	64.3	78.1	70.5	63.9	67.633
0.3	64.3	74.1	68.8	64.7	78.5	70.9	64.4	68.033
0.4	64.3	74.1	68.8	64.6	78.5	70.9	64.4	68.033
0.5	64.3	74.1	68.9	64.6	78.5	70.9	64.4	68.067
0.6	64.4	74.2	69.0	64.6	78.6	70.9	64.5	68.133
0.7	64.5	74.3	69.0	64.7	78.5	71.0	64.5	68.167
0.8	64.4	74.2	69.0	64.7	78.6	71.0	64.6	68.200
0.9	64.4	74.2	69.0	64.7	78.6	71.0	64.6	68.200
1.0	64.4	74.2	68.9	64.7	78.5	70.9	64.5	68.100

Table 4: *Specialized models interpolated with a general model for mention count-weighted interpolation weight λ and 10 values for a generic λ.*

As can be seen, the value of λ which maximizes performance is high, as would be expected with the value of C found previously. It appears that $\lambda_{C=2000}$ is outperformed by a large range of constant λ values—however $\lambda_{C=2000}$ is not significantly different from the other values for $\lambda < 0.3$. This suggests that either the initial assumption about relative effectiveness was incorrect, or that the value of C needs to be altered in some fashion.

Results were also obtained for a similar set of tests on a I(SPEC$_{models}$, SPEC$_{feat}$) system, shown in Table 5.

λ	$MUC:R$	$MUC:P$	$MUC:F$	$B^3:R$	$B^3:P$	$B^3:F$	$CEAF$	$\bar{F}$
$C = 2000$	65.2	73.7	69.2	65.3	77.6	70.9	64.5	68.200
0.0	63.2	73.6	68.0	64.2	78.2	70.5	64.2	67.567
0.1	64.6	73.1	68.6	64.7	77.2	70.4	63.7	67.567
0.2	64.8	73.3	68.8	65.2	77.4	70.8	64.2	67.933
0.3	65.0	73.5	69.0	65.2	77.5	70.8	64.3	68.033
0.4	65.1	73.6	69.1	65.3	77.6	70.9	64.4	68.133
0.5	65.1	73.6	69.1	65.3	77.6	70.9	64.5	68.167
0.6	65.2	73.7	69.2	65.4	77.7	71.0	64.6	68.267
0.7	65.2	73.7	69.2	65.3	77.7	70.9	64.5	68.200
0.8	65.3	73.8	69.3	65.3	77.7	71.0	64.5	68.267
0.9	65.2	73.7	69.2	65.3	77.5	70.9	64.5	68.200
1.0	65.2	73.7	69.2	65.3	77.5	70.9	64.5	68.200

Table 5: *Specialized models interpolated with a model with specialized features for mention count-weighted interpolation weight λ and 10 values for a generic λ.*

A similar pattern holds for the results of tests for interpolation in these systems as well. Here, however, the value of λ obtained by the use of $C = 2000$ is even closer to the overall highest result for a constant λ. This result is only statistically different from $\lambda = 0.1$ and $\lambda = 0.3$. The apparent suggestion of this data, however, is that a constant λ in the range of $0.8 - 0.9$ is close to optimum for interpolation results. Potential gains for a more sophisticated weighting method may be lost in the noise created by distinct training set sizes. Taking Section 5.3.2 into consideration, however, the distinction in λ values may be moot—those results suggest that a λ of 1.0 may be equivalent.

5.4 Cataphora

Results for cataphora resolution methods are reported below.

In training, instances were created by pairing each anaphor in the training set with the members of its candidate set. The sample selection method consisted of the following:

1. The antecedent.
2. All preceding candidate mentions up to the antecedent.
3. The postcedent.
4. All following candidate mentions up to the postcedent *up to three sentences away.*

The fourth component of the candidate set for training was added to control for the very larger possible size of the set intervening between cataphor and postcedent. It is, additionally, primarily close mentions with which we are concerned for cataphora.

For cataphoric classification all preceding and following mentions were considered as candidates. The candidate pair with the highest score was selected—that is, Best-First link selection was used—unless otherwise noted.

5.4.1 Basic Methods and Cataphoricity Classification

BASE is, as before, the typical SCC unspecialized system which prohibits cataphoric links. CATA$_{base}$ is a system relaxing that prohibition, using the sampling selection method and classification candidates described above. CATA$_{feat}$ then introduces the two features relevant to cataphora. CATA$_{bias}$ adds the "Anaphora-Biased Best-First" link selection algorithm. CATA$_{class}$ lacks this algorithm but uses the cataphoricity classifier to make a decision as to whether to permit all mentions for classification or merely preceding ones. CATA$_{class,bias}$ is the Anaphora-Biased Best-First version of CATA$_{class}$, while CATA$_{class,spec}$ is a version using specialized models. CATA$_{class,spec,bias}$ combines the two. Table 6 lists findings on these tests.

Test	$MUC:R$	$MUC:P$	$MUC:F$	$B^3:R$	$B^3:P$	$B^3:F$	$CEAF$	$\bar{F}$
BASE	60.6	72.3	65.9	62.2	77.2	68.9	62.2	65.667
CATA$_{base}$	48.7	69.1	57.1	52.8	84.1	64.9	55.0	59.000
CATA$_{feat}$	51.2	68.4	58.5	53.9	82.4	65.2	55.9	59.867
CATA$_{bias}$	53.8	69.2	60.5	55.5	81.0	65.9	57.7	61.367
CATA$_{class}$	51.6	71.8	60.1	55.3	84.2	66.8	58.5	61.800
CATA$_{class,bias}$	53.4	72.3	61.4	56.5	83.4	67.43	59.5	62.777
CATA$_{class,spec}$	59.4	74.1	66.0	59.8	81.5	68.9	61.9	65.600
CATA$_{class,spec,bias}$	60.1	73.6	66.1	60.9	80.2	69.2	62.6	65.977

Table 6: *Results comparing baseline with various cataphoric models.*

First it is obvious that performance suffers by allowing cataphoric links naively. BASE is significantly different from CATA$_{base}$, CATA$_{feat}$, and CATA$_{bias}$ in both F-score and recall—that is, any cataphoric system without a cataphoricity classifier performs worse than a system that ignores cataphora. CATA$_{base}$ and CATA$_{feat}$ are thus outperformed by all systems with cataphoricity classifiers. CATA$_{bias}$ is more marginal—it is outperformed (among cataphoric systems) only by the three systems which also incorporate specialized models or the bias (CATA$_{class,bias}$, CATA$_{class,spec}$, and CATA$_{class,spec,bias}$). CATA$_{class}$ is outperformed by CATA$_{class,bias}$ and CATA$_{class,spec,bias}$, nearly so by CATA$_{class,spec}$. The other system with an anaphoricity classifier alone, CATA$_{class,bias}$ is nearly outperformed by those with specialized models (CATA$_{class,spec}$ and CATA$_{class,spec,bias}$). This difference does not reach significance but does under a stronger assumption.

Allowing cataphora appears to cause great detriments in recall and precision. With the addition of the sophistications developed above it appears that progress can be made to reattain near equivalence with the baseline but not to make significant advances. The only system which out-

performs the baseline is CATA$_{class,spec,bias}$, which fails to outperform the non-cataphoric equivalent, SPEC$_{models}$. The additional features do not have a significant impact on scores. Similarly, systems using Anaphora-Biased Best-First link selection perform only slightly differently from ones using Best-First. The introduction of the cataphoricity classifier, however, does have an effect. Within the classifiers, specialization appears (potentially) important.

Of note is the fact that the cataphoric systems losses are mostly in recall, especially in the more complex systems. In fact, CATA$_{class,spec,bias}$ achieves higher precision values on all metrics—although this difference does not turn out to be significant. There is, however, an interesting pair of significant differences in precision. CATA$_{feat}$ differs from CATA$_{class}$, as does CATA$_{bias}$ from CATA$_{class,bias}$. Thus there is some evidence for an impact of the classifier on precision—as well as a distinction between Anaphora-Biased Best-First link selection and Best-First not seen elsewhere.

5.4.2 Threshold Optimization

Systems which use classify anaphoricity and resolve coreference in a cascade are known to suffer from issues of performance degradation, particularly in recall (Ng and Cardie, 2002a; Denis and Baldridge, 2007b). Ng (2004) addresses this problem by optimizing the probability threshold at which an anaphoric judgment is made. That is, rather than merely accepting a mention as anaphoric if a probability greater than 0.5 is given by the classifier, some other value is used chosen on the basis of performance on development data. This suggests that a similar optimization process might also be advantageous for the cataphoricity classifier.

The cataphoricity threshold was optimized from the options $P_{thresh} \in [0.5, 1.0), P_{thresh} = 0.1k$. That is, the threshold must be at least the default of 0.5 and is limited to increments of 0.1. The result which maximized $\bar{F}$ on development data was $P_{thresh} = 0.9$—that is, the value which most strictly prohibits cataphoric links. This suggests that such a threshold optimization method will not help to engender better performance than the baseline. Results for using this value for CATA$_{class}$ are presented in Table 7.

Test	$MUC:R$	$MUC:P$	$MUC:F$	$B^3:R$	$B^3:P$	$B^3:F$	$CEAF$	$\bar{F}$
BASE	60.6	72.3	65.9	62.2	77.2	68.9	62.2	65.667
CATA$_{class}$	53.9	73.3	62.1	57.8	83.1	68.2	60.7	63.667

Table 7: *Results comparing specialized model system with its cataphoricity classifier-based equivalent using* $P_{thresh} = 0.9$.

The difference between these models is not significant, but approaches so for a laxer test. Thus this "optimization" appears ineffective even at rendering CATA$_{class}$ *as effective* as BASE, let alone

bringing it above this bar. Such optimization seems unlikely to be a solution to the problem of cataphora, therefore.

5.4.3 Joint Determination

Joint-determination is the other method used for dealing with cascade error (Denis and Baldridge, 2007b; Denis and Baldridge, 2009). I developed a similar ILP formulation for cataphoricity in Section 4.4. Results for experiments on these ILP systems are listed in Table 8. The GNU Linear Programming Kit was used to solve ILP problems.[9].

BASE and CATA$_{class,spec,bias}$ are as before and are included for comparison (the model I seek to improve and the best-performing system from the previous set). ILP$_{base}$ is an ILP coreference resolver without cataphoricity costs or constraints—but using cataphora features. It does not use the cataphoric sample selection method or allow cataphora for resolution. ILP$_{cata}$, in contrast, additionally uses the method and allows such resolution. ILP$_{class}$ is the ILP system with cataphoric costs but only "cataphorically resolve only cataphora" as a constraint. ILP$_{class,resolveall}$ adds "cataphorically resolve all cataphora."

Test	$MUC:R$	$MUC:P$	$MUC:F$	$B^3:R$	$B^3:P$	$B^3:F$	$CEAF$	$\bar{F}$
BASE	60.6	72.3	65.9	62.2	77.2	68.9	62.2	65.667
CATA$_{class,spec,bias}$	60.1	73.6	66.1	60.9	80.2	69.2	62.6	65.977
ILP$_{base}$	70.1	72.1	71.1	73.2	63.2	67.8	58.5	65.800
ILP$_{cata}$	65.5	71.7	68.4	67.6	70.6	69.1	61.4	66.300
ILP$_{class}$	63.1	73.2	67.8	65.9	74.1	69.8	62.3	66.633
ILP$_{class,resolveall}$	66.6	69.0	67.8	71.7	57.6	63.8	53.2	61.600

Table 8: *Results comparing ILP cataphora systems.*

The F-scores of the results show high variance. Only a single significance judgment can be made—ILP$_{class,resolveall}$ performs significantly *worse* than ILP$_{base}$. The same pattern holds for precision—only these exhibit a significant difference.

Joint determination thus does not improve performance over the baseline. In fact, it hurts it if both constraints are used. As noted previously, "cataphorically resolve all cataphora" is not particularly suited to the type of classifier used here. Cataphora are *forced* when they "should" merely be *allowed*, considering the nature of the task carried out by the classifier. The classifier cannot state unequivocally that a cataphoric link *must* be made in most cases. Including this constraint thus proves to be fatal for the performance of ILP$_{class,resolveall}$. ILP$_{class}$ does not suffer the extreme performance losses of ILP$_{class,resolveall}$, but does not offer much in the way of improvement. This is likely owing to the very fact that it lacks "cataphorically resolve all cataphora."

[9]Available at `http://www.gnu.org/software/glpk/`

There is no constraint that forces cataphoric judgments to be used. This means that the coreference classifier is constrained by the cataphoricity classifier without being able to benefit much from it.

There is, however, evidence that ILP yields benefits. ILP_{class} has an $\bar{F}$ of 66.633, compared to 61.800 for CATA_{class}. The F-scores are actually near-significantly different, and are so under a constrained test. Thus, joint determination of cataphoricity with coreference actually does yield benefits—the performance losses from introducing cataphora are somewhat combatted by this addition.

In the previous set of cataphoric systems, precision benefited at the expense of recall. For the joint determination systems, the problems with recall are somewhat alleviated but precision is lost. This suggests that an ILP solution with a better cataphoricity classifier might be able to acquire gains in both scores.

5.4.4 Elaborated Joint Determination

In this work I jointly determine only cataphoricity and coreference. With a more precise cataphoricity classifier, joint determination of anaphoricity in addition to these two would likely be fruitful.

Such a system would be very similar to the ones discussed here. The objective function would take the form:

$$\min \left[\sum_{(i,j)\in\mathcal{P}'} \left(c^C_{\langle i,j\rangle} x_{\langle i,j\rangle} + \bar{c}^C_{\langle i,j\rangle} \left(1 - x_{\langle i,j\rangle}\right)\right) + \sum_{j\in\mathcal{M}} \left(c^A_j y_j + \bar{c}^A_j \left(1 - y_j\right)\right) + \sum_{j\in\mathcal{M}} \left(c^{Cata}_j q_j + \bar{c}^{Cata}_j \left(1 - q_j\right)\right) \right]$$

such that $\forall (i,j) \in \mathcal{P}'$, $x_{\langle i,j\rangle} \in \{0,1\}$, $\forall j \in \mathcal{M}$, $y_j \in \{0,1\}$, and $\forall j \in \mathcal{M}$, $q_j \in \{0,1\}$. That is, the cost of coreference, anaphoricity, and cataphoricity assignments is minimized.

Constraints, too, would be similar to those previously explicated. Two would be identical—"anaphorically resolve all anaphora" (equivalent to "resolve all anaphora") and "cataphorically resolve all cataphora":

Anaphorically resolve all anaphora. An anaphoric mention must be an anaphor in *some* coreference link:

$$y_j \leq \sum_{i\in\mathcal{M}_j^-} x_{\langle i,j\rangle} \quad \forall j \in \mathcal{M}$$

Cataphorically resolve all cataphora. A cataphoric mention must be a cataphor in some coreference link:

$$q_j \leq \sum_{i \in \mathcal{M}_j^+} x_{\langle i,j \rangle} \quad \forall j \in \mathcal{M}$$

A third constraint is similar to the previously developed "resolve only anaphora," but differs in that it permits resolution of either anaphora or cataphora:

Resolve all resolvable mentions. A cataphoric or anaphoric mention must be resolved as a cataphor or anaphor.

$$x_{\langle i,j \rangle} \leq y_j + q_i \quad \forall (i,j) \in \mathcal{P}'$$

This set constitutes a formulation of joint determination of coreference, anaphoricity, and cataphoricity. It is not viable with the current cataphoricity classifier but may be useful in future work.

6 Discussion and Conclusion

In this work I modify a state-of-the-art coreference system. For such systems, it is difficult for modifications to yield significant improvements. Indeed, the modifications given here would likely have had more impact on performance in more simplistic setups. However, the intention of this project was to investigate possibilities for *improving* on a state-of-the-art coreference system, not merely demonstrating that a given property has some value for coreference resolution. Additionally, I attempt to maximize three disparate metrics (MUC, B^3, and $CEAF$). $CEAF$ and B^3 are more difficult to raise, and raising all three together is more difficult still.

I have tested a variety of specialization techniques against one another within the same basic system. I have found that once specialization of one type is introduced into a coreference system, adding a second type does not increase performance in a significant way. This is true regardless of the method used to integrate the two types of specialization—interpolated systems as well as directly combined ones perform similarly. There is, however, some slight evidence that specialized models might be preferable in some cases.

The conclusion that can be reached from this is that in comparing coreference systems any type of specialization covered here may be considered equivalent to another for large data sets. This can only not be so if there are compelling interactions elsewhere in the methods of the system—necessarily interactions not investigated or conceived of here. When adding specialization to future large models it would be best to concentrate on the relative efficiency of the technique used—its

relative effect is likely not important. However, this may vary depending on the amount of training data or its type. Here tests on the full ACE corpus or its individual sections failed to demonstrate a significant difference between any of the methods.

I have also advanced and compared a number of systems which incorporate cataphoric resolution of coreference. This problem cannot yet be said to be resolved. However, I have demonstrated that systems using the type of simple "anaphoricity classifier" advanced here significantly outperform cataphoric systems that do without it. As such, it appears that such a classifier is likely an useful step towards productive systems to be built in the future. Performance problems that arise with the cataphoricity classifier cannot be dealt with by simple thresholding as is partially the case for anaphoricity classifiers.

I proposed an ILP formulation for cataphoricity. However, an integration of the cataphoricity classifier into a joint determination approach proved to be somewhat fruitless. If the type of cataphoricity classifier advanced here is used, it appears that the ILP problem is *overly* constrained in ways that cause performance losses while being not constrained *enough* in ways that would yield gains.

If performance improvements are to be had using a cataphoricity classifier they may likely require a different view of such a classifier. It may need to be taken in explicit *contrast* to an anaphoricity classifier. If possible, a way of identifying "true" cataphora for training would similarly aid in addressing the problem. The recall gains taken by the ILP systems here, taken with the precision gains of the classifier alone, suggest that such a classifier would be beneficial to *overall* performance. In either case, the ILP problem specification presented here is likely to be useful in its integration into a full coreference system—especially the unimplemented formulation which integrates anaphoricity. How an improved cataphoricity classifier can be constructed is an open question. It may necessitate the use of deeper parsing than is used here—or merely different heuristics.

7 References

1. M. Ariel. 1988. Referring and accessibility. *Journal of Linguistics*, pages 65-87.

2. A. Bagga and B. Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of LREC 1998*, pages 563-566.

3. Eric Bengston and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of EMNLP-2008*. Honolulu, Hawaii.

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

5. Pascal Denis and Jason Baldridge. 2007a. A ranking approach to pronoun resolution. In *Proceedings of IJCAI-2007*. Hyderabad, India.

6. Pascal Denis and Jason Baldridge. 2007b. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of NAACL-2007*. Rochester, NY.

7. Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of EMNLP-2008*. Honolulu, Hawaii.

8. Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. Procesamiento del Lenguaje Natural 42. ISSN:1135-5948.

9. Robert Elwell and Jason Baldridge. 2008. Discourse connective argument identification with connective specific rankers. In *Proceedings of ICSC- 2008*. Santa Clara, CA.

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

11. 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*.

12. 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.

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

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

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

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

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

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

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