Specialized models and ranking for coreference resolution

Pascal Denis
ALPAGE Project Team
INRIA Rocquencourt
F-78153 Le Chesnay, France
pascal.denis@inria.fr

Jason Baldridge
Department of Linguistics
University of Texas at Austin
Austin, TX 78712-0198, USA
jbaldrid@mail.utexas.edu

Abstract

This paper investigates two strategies for improving coreference resolution: (1) training separate models that specialize in particular types of mentions (e.g., pronouns versus proper nouns) and (2) using a ranking loss function rather than a classification function. In addition to being conceptually simple, these modifications of the standard single-model, classification-based approach also deliver significant performance improvements. Specifically, we show that on the ACE corpus both strategies produce f-score gains of more than 3% across the three coreference evaluation metrics (MUC, B^3, and CEAF).

1 Introduction

Coreference resolution is the task of partitioning a set of entity mentions in a text, where each partition corresponds to some entity in an underlying discourse model. While early machine learning approaches for the task relied on local, discriminative classifiers (Soon et al., 2001; Ng and Cardie, 2002b; Morton, 2000; Kehler et al., 2004), more recent approaches use joint and/or global models (McCallum and Wellner, 2004; Ng, 2004; Daumé III and Marcu, 2005; Denis and Baldridge, 2007a). This shift improves performance, but the systems are considerably more complex and often less efficient. Here, we explore two simple modifications of the first type of approach that yield performance gains which are comparable, and sometimes better, to those obtained with these more complex systems. These modifications involve: (i) the use of *rankers* instead of classifiers, and (ii) the use of linguistically motivated, *specialized models* for different types of mentions.

Ranking models provide a theoretically more adequate and empirically better alternative approach to pronoun resolution than standard classification-based approaches (Denis and Baldridge, 2007b). In essence, ranking models directly capture during training the competition among potential antecedent candidates, instead of considering them independently. This gives the ranker additional discriminative power and in turn better antecedent selection accuracy. Here, we show that ranking is also effective for the wider task of coreference resolution.

Coreference resolution involves several different types of anaphoric expressions: third-person pronouns, speech pronouns (i.e., first and second person pronouns), proper names, definite descriptions and other types of nominals (e.g., anaphoric uses of indefinite, quantified, and bare noun phrases). Different anaphoric expressions exhibit different patterns of resolution and are sensitive to different factors (Ariel, 1988; van der Sandt, 1992; Gundel et al., 1993), yet most machine learning approaches have ignored these differences and handle these different phenomena with a single, monolithic model. A few exceptions are worth noting. Morton (2000) and Ng (2005b) propose different classifiers models for different NPs for coreference resolution and pronoun resolution, respectively. Other partially capture the differential preferences between different anaphors via different sample selection strategies during training (Ng and Cardie, 2002b; Uryupina, 2004). More recently, Haghighi and Klein (2007) use the distinction between pronouns, nominals and proper nouns

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 660–669,
Honolulu, October 2008. ©2008 Association for Computational Linguistics

in their unsupervised, generative model for coreference resolution; for their model, this is absolutely critical for achieving better accuracy. Here, we show that using specialized models for different types of referential expressions improves performance for supervised models (both classifiers and rankers).

Both these strategies lead to improvements for all three standard coreference metrics: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and CEAF (Luo, 2005). In particular, our specialized ranker system provides absolute f-score improvements against an otherwise identical standard classifier system by 3.2%, 3.1%, and 3.6% for MUC, B^3, and CEAF, respectively.

2 Ranking

Numerous approaches to anaphora and coreference resolution reduce these tasks to a binary classification task, whereby *pairs* of mentions are classified as coreferential or not (McCarthy and Lehnert, 1995; Soon et al., 2001; Ng and Cardie, 2002b). Usually used in combination with a greedy right-to-left clustering, these approaches make very strong independence assumptions. Not only do they model each coreference decision separately, they actually model *each pair* of mentions as a separate event. Recasting these tasks as ranking tasks partly addresses this problem by directly making the comparison between different candidate antecedents for an anaphor part of the training criterion. Each candidate is assigned a conditional probability with respect to the *entire* candidate set. (Re)rankers have been successfully applied to numerous NLP tasks, such as parse selection (Osborne and Baldridge, 2004; Toutanova et al., 2004), parse reranking (Collins and Duffy, 2002; Charniak and Johnson, 2005), question-answering (Ravichandran et al., 2003).

The twin-candidate classification approach proposed by (Yang et al., 2003) shares some similarities with the ranker in making the comparison between candidate antecedents part of training. An important difference however is that under the twin-candidate approach, candidates are compared in *pairwise* fashion (and the best overall candidate is the one that has won the most round robin contests), while the ranker considers the entire candidate set at once. Another advantage of the ranking approach is that its complexity is only square in the number of mentions, while that of the twin-candidate model is cubic (see Denis and Baldridge (2007b) for a more detailed comparison in the context of pronoun resolution).

Our ranking models for coreference take the following log-linear form:

$$P_{rk}(\alpha_i|\pi) = \frac{\exp \sum_{j=1}^{m} w_j f_j(\pi, \alpha_i)}{\sum_k \exp \sum_{j=1}^{m} w_j f_j(\pi, \alpha_k)} \qquad (1)$$

where π stands for the anaphoric expression, α_i for an antecedent candidate, f_j the weighted features of the model. The denominator consists of a normalization factor over the k candidate mentions. Model parameters were estimated with the limited memory variable metric algorithm and Gaussian smoothing (σ^2=1000), using TADM (Malouf, 2002).

For the training of the different ranking models, we use the following procedure. For each model, instances are created by pairing each anaphor of the proper type (e.g., definite description) with a set of candidates which contains: (i) a true antecedent, and (ii) a set of non-antecedents. The selection of the true antecedent varies depending on the model we are training: for pronominal forms, the antecedent is selected as the *closest* preceding mention in the chain; for non-pronominal forms, we used the closest preceding *non-pronominal* mention in the chain as the antecedent. For the creation of the non-antecedent set, we collect all the non-antecedents that appear in a window of two sentences around the antecedent.[1] At test time, we consider *all* preceding mentions as potential antecedents.

Not all referential expressions in a given document are anaphors: some expressions introduce a discourse entity, rather than accessing an existing one. Thus, coreference resolvers must have a way of identifying such "discourse-new" expressions. This is easily handled in the standard classification approach: a mention will not be resolved if none of its candidates is classified positively (i.e., as coreferential). The problem is more troublesome for rankers, which always pick an antecedent from the candidate

[1] We suspect that different varying windows might be more appropriate for different types of expressions, but leaves this for further investigations.

set. A natural solution is to use a model that specifically predicts the discourse status (discourse-new vs. discourse-old) of each expression: only expressions that are classified as "discourse-old" by this model are considered by rankers.

Ng and Cardie (Ng and Cardie, 2002a) introduced the use of an "anaphoricity" classifier to act as a filter for coreference resolution in order to correct errors where antecedents are mistakenly identified for non-anaphoric mentions or antecedents are not determined for mentions which are indeed anaphoric. Their approach produced significant improvements in precision, but with consequent larger losses in recall. Ng (2004) improves recall by optimizing the anaphoricity threshold. By using joint inference for anaphoricity and coreference, Denis and Baldridge (2007a) avoid cascade-induced errors without the need to separately optimize the threshold.

We use a similar discourse status classifier to Ng and Cardie's as a filter on mentions for our rankers. We rely on three main types of information sources: (i) the form of mention (e.g., type of linguistic expression, number of tokens), (ii) positional features in the text, (iii) comparisons of the given mention to the mentions that precede it in the text. Evaluated on the ACE datasets, training the model on the `train` texts, and applying the classifier to the `devtest` texts, the model achieves an overall accuracy score of 80.8%, compared to a baseline of 59.7% when predicting the majority class ("discourse-old").

3 Specialized models

Our second strategy is to use different, specialized models for different referential expressions, similarly to Elwell and Baldridge's (2008) use of connective specific models for identifying the arguments of discourse connectives. For this, one must determine along which dimension to split such expressions. For example, Ng (2005b) learns models for each set of anaphors that are lexically identical (e.g., *I*, *he*, *they*, etc.). This option is possible for closed sets like pronouns, but not for other types of anaphors like proper names and definite descriptions. Another option is to rely on the particular *linguistic form* of the different expressions, as signaled by the head word category and the determiner (if any). More concretely, we use separate models for the following types: (i) third person pronouns, (ii) speech pronouns, (iii) proper names, (iv) definite descriptions, and (v) others (i.e., all expressions that don't fall into the previous categories).

The correlation between the form of a referential expression and its anaphoric behavior is actually central to various linguistic accounts (Prince, 1981; Ariel, 1988; Gundel et al., 1993). Basically, the idea is that linguistic form is an indicator of the status of the corresponding referent in the discourse model. That is, the use by the speaker of a particular linguistic form corresponds to a particular level of activation (or familiarity or salience or accessibility) in (what she thinks is) the addressee's discourse model. For many authors, the relation takes the form of a continuum and is often represented in the form of a referential hierarchy, such as:

> **Accessibility Hierarchy** (Ariel, 1988)
> Zero pronouns >> Pronouns >> Demonstrative pronouns >> Demonstrative NPs >> Short PNs >> Definite descriptions >> Full PNs >> Full PNs + appositive

The higher up, the more accessible (or salient) the entity is. At the extremes are pronouns (these forms typically require a previous mention in the local context) and proper names (these forms are often used without previous mentions of the entity). This type of hierarchy is validated by corpus studies of the distribution of different types of expressions. For instance, pronouns find their antecedents very locally (in a window of 1-2 sentences), while proper names predominantly find theirs at longer distances (Ariel, 1988).[2] Using discourse structure, Asher et al. (2006) show that while anaphoric pronouns systematically obey the right-frontier constraint (i.e., their antecedents have to appear on the right edge of the discourse graph), this is less so for definites, and even less so for proper names.

From a machine learning perspective, these findings suggest that features encoding some aspect of salience (e.g., distance, syntactic context) are likely to receive different sets of parameters depending on the form of the anaphor. This therefore suggests that better parameters are likely to be learned in the

[2] Haghighi and Klein's (2007) generative coreference model mirrors this in the posterior distribution which it assigns to mention types given their salience (see their Table 1).

Type/Count	train	test
3^{rd} pron.	$4,389$	$1,093$
speech pron.	$2,178$	610
proper names	$7,868$	$1,532$
def. NPs	$3,124$	796
others	$1,763$	568
Total	$19,322$	$4,599$

Table 1: Distribution of the different anaphors in ACE

Linguistic Form	
pn	α is a proper name {1,0}
def_np	α is a definite description {1,0}
indef_np	α is an indefinite description {1,0}
pro	α is a pronoun {1,0}
Context	
left_pos	POS of the token preceding α
right_pos	POS of the token following α
surr_pos	pair of POS for the tokens surrounding α
Distance	
s_dist	Binned values for sentence distance between π and α
np_dist	Binned values for mention distance between π and α
Morphosyntactic Agreement	
gender	pairs of attributes {masc, fem, neut, unk} for π and α
number	pairs of attributes {sg, pl} for π and α
person	pairs of attributes {1, 2, 3, 4, 5, 6} for π and α
Semantic compatibility	
wn_sense	pairs of Wordnet senses for π and α
String similarity	
str_match	π and α have identical strings {1,0}
left_substr	one mention is a left substring of the other {1,0}
right_substr	one mention is a right substring of the other {1,0}
hd_match	π and α have the same head word {1,0}
Apposition	
apposition	π and α are in an appositive structure {1,0}
Acronym	
acronym	π is an acronym of α or vice versa {1,0}

Table 2: Features used by coreference models.

context of different models.[3] While the above studies focus primarily on salience, there are of course other dimensions according to which anaphors differ in their resolution preferences. Thus, the resolution of lexical expressions like definite descriptions and proper names is likely to benefit from the inclusion of features that compare the strings of the anaphor and the candidate antecedent (e.g., string matching) and features that identify particular syntactic configurations like appositive structures. This type of information is however much less likely to help in the resolution of pronominal forms. The problem is that, within a single model, such features are likely to receive strong parameters (due to the fact that they are good predictors for lexical anaphors) in a way that might eventually hurt pronominal resolutions.

Note that our split of referential types only partially cover the referential hierarchies of Ariel (1988) or Gundel et al. (1993). Thus, there is no separate model for demonstrative noun phrases and pronouns: these are very rare in the corpus we used (i.e., the ACE corpus).[4] These expressions were therefore handled through the "others" model. There is however a model for first and second person pronouns (i.e., speech pronouns): this is justified by the fact that these pronouns behave differently from their third person counterparts. These forms indeed often behave like deictics (i.e., they refer to discourse participants) or they appear within a quote.

The total number of anaphors (i.e., of mentions that are not chain heads) in the data is $19,322$ and $4,599$ for training and testing, respectively. The distribution of each anaphoric type is presented in Table 1. Roughly, third person pronouns account for 22-24% of all anaphors in the entire corpus, speech pronouns for 11-13%, proper names for 33-40%, and definite descriptions for 16-17%. The distribution is slightly different from one dataset to another, probably reflecting genre differences. For instance, BNEWS shows a larger proportion of pronouns in general (pronominal forms account for 40-44% of all the anaphoric forms).

We use five broad types of features for all mention types, plus three others used by specific types, summarized in Table 3. Our feature extraction relies on limited linguistic processing: we only made use of a sentence detector, a tokenizer, a POS tagger (as provided by the OpenNLP Toolkit[5]) and the WordNet[6] database. Since we did not use parser, lexical heads for the NP mentions were computed using simple heuristics relying solely on POS sequences. Table 2 describes in detail the entire feature set, and Table 3 shows which features were used for which models.

Linguistic form: the referential form of the antecedent candidate: a proper name, a definite de-

[3] Another possible approach would consist in introducing different salience-based features encoding the form of the anaphor.

[4] There are only 114 demonstrative NPs and 12 demonstrative pronouns in the entire ACE training.

[5] http://opennlp.sf.net.

[6] http://wordnet.princeton.edu/

Features/Types	3P	SP	PN	Def-NP	Oth
Ling. form	√	√	√	√	√
Context	√	√	√	√	√
Distance	√	√	√	√	√
Agreement.	√	√	√	√	√
Sem. compat.	√	√	√	√	√
Str. sim.			√	√	√
Apposition			√	√	
Acronym			√		

Table 3: Features for each type of referential expression.

scription, an indefinite NP, or a pronoun.

Context: the context of the antecedent candidate: these features can be seen as approximations of the grammatical roles, as indicators of the salience of the potential candidate (Grosz et al., 1995). For instance, this includes the part of speech tags surrounding the candidate, as well as a feature that indicates whether the potential antecedent is the first mention in a sentence (approximating subjecthood), and a feature indicating whether the candidate is embedded inside another mention.

Distance: the distance between the anaphor and the candidate, measured by the number of sentences and mentions between them.

Morphosyntactic agreement: indicators of the gender, number, and person of the two mentions. These are determined for non-pronominal NPs with heuristics based on POS tags (e.g., NN vs. NNS for number) and actual mention strings (e.g., whether the mention contains a male/female first name or honorific for gender). These features consist of pairs of attributes, ensuring that not only strict agreement (e.g., *singular-singular*) but also mere compatibility (e.g., *masculine-unknown*) is captured.

Semantic compatibility: features designed to assess whether the two mentions are semantically compatible. For these features, we use the WordNet database: in particular, we collected the synonym set (or synset) as well as the synset of their direct hypernyms associated with each mention. In the case of common nouns, we used the synset associated with the first sense associated with the mention's head word. In the case of proper names, we used the synset associated with the name if available, and the string itself otherwise. For pronouns (which are not part of Wordnet), we simply used the pronominal form.

All these features were used in all five models. While one may question the use of distance for non-pronominal anaphors,[7] their inclusion can be justified in that they might predict some "obviation" effects. Definite descriptions and proper names are sensitive to distance too, although not in the same way as pronouns are: they show a preference for antecedents that appear outside a window of one or two sentences (Ariel, 1988).

Several features are used only for particular mention types:

String similarity: similarity of the anaphor and the candidate strings. Examples are perfect string match, substring matches, and head match (i.e., the two mentions share the same head word).

Appositive: whether the anaphor is an appositive of the antecedent candidate. Since we do not have access to syntactic structure, we use heuristics (e.g., the presence of a comma between the two mentions) to extract this feature.

Acronym: whether the anaphor string is an acronym of the candidate string (or vice versa): e.g., NSF and National Science Foundation.

4 Coreference systems

We evaluate several systems to explore the effect of ranking versus classification and specialized versus monolithic models. The different systems follow a generic architecture. Let $\mathcal{M}$ be the set of mentions present in a document. For all models, each mention $m \in \mathcal{M}$ is associated at test time with a set of antecedent candidates $\mathcal{C}_m$, which includes all the mentions that linearly precede m. The best candidate is determined by the model in use. The final output of each system consists in a list of mention pairs (i.e., the coreference links) which in turn defines (through reflexive, transitive closure) a partition over the set $\mathcal{M}$. Our models are summarized in Table 4.

The use of the discourse status filter is straightforward. For each mention $m \in \mathcal{M}$, the discourse status

[7] In fact, Morton (2000) does not use distance in this case.

Model Name	Model Type	Specialized?	Disc. Status
CLASS	class	No	No
CLASS+DS	class	No	Yes
CLASS+SP	class	Yes	No
CLASS+DS+SP	class	Yes	Yes
RANK+DS+SP	rank	Yes	Yes

Table 4: Model names and their properties.

model is first applied to determine whether m introduces a new discourse entity (i.e., it is classified as "new") or refers back to an existing entity (i.e., it is classified as "old"). If m is classified as "new", the process terminates and goes to the next mention. If m is classified as "old", m along with its set of antecedent candidates $\mathcal{C}_m$ is sent to the model.

For classifiers, we replicate the procedures of Ng and Cardie (2002b). During training, instances are formed by pairing each anaphor with each of its preceding candidates, until the antecedent is reached: the closest preceding antecedent in the case of a pronominal anaphor, or the closest non-pronominal antecedent for other anaphor types. For classifiers, the use of a discourse status filter at test time is optional. When a filter is not used, then a mention is left unresolved if none of the pairs created for a given mention is classified positively. If several pairs for a given mention are classified positively, then the pair with the highest score is selected (i.e., "Best-First" link selection). If a filter is used, then the candidate with the highest score is selected, even if the probability of coreference is less than one-half.[8]

The use of specialized models is simple, for both classifiers and rankers. Specialized models are created for: (i) third person pronouns, (ii) speech pronouns, (iii) proper names, (iv) definite descriptions, (v) other types of phrases. The mention type is determined and the best candidate is chosen by the appropriate model Following Elwell and Baldridge (2008), these models could be interpolated with a monolithic model, or even word specific models, but we have not explored that option here.

System	Accuracy
3^{rd} pron.	82.2
speech pron.	66.9
proper names	83.5
def. NPs	66.5
others	63.6

Table 5: Accuracy of the different ranker models.

The feature sets for the classifiers in the baseline systems includes all the features that were used for the described in Section 3. For the classifiers that do not use specialized models (**CLASS** and **CLASS+DS**), we have also added extra features describing the linguistic form of the potential anaphor (whether it is a pronoun, a proper name, and so on). This is in accordance with standard feature sets in the pairwise approach. It gives these models a chance to learn weights more appropriately for the different types within a single, monolithic model.

5 Experiments

We use the ACE corpus (Phase 2). The corpus has three parts, each corresponding to a different genre: newspaper texts (NPAPER), newswire texts (NWIRE), and broadcast news (BNEWS). Each set is split into a `train` part and a `devtest` part. In our experiments, we consider only *true* ACE mentions.

5.1 Antecedent selection results

We first evaluate the specialized ranker models individually on the task of anaphora resolution: their ability to select a correct antecedent for each anaphor. Following common practice in this task, we report results in terms of *accuracy*, which is simply the ratio of correctly resolved anaphors. The candidate set during testing was formed by taking *all* the mentions that appear before the anaphor. Also, we assume that correctly resolving an anaphor amounts to selecting any of the previous mentions in the entity as the antecedent. The accuracy scores for the different models are presented in Table 5.

[8] This is very similar to the approach of Ng and Cardie (2002a). An important difference is that their system does not necessarily yield an antecedent for each of the anaphors proposed by the discourse status model. In their system, if the coreference classifier finds that none of the candidates for a "new" mention are coreferential, it leaves it unresolved. In this case, the coreference model acts as an additional filter. Not surprisingly, these authors report gains in precision but comparatively larger losses in recall. Our development experiments revealed that forcing a decision on items identified as new provided performed better across all metrics.

The best accuracy results on the entire ACE corpus are found first for the proper name resolver with a score of 83.5%, then for the third person pronoun resolver with 82.2%, then for the definite description and speech pronoun resolvers with 66.9% and 66.5% respectively. The worst scores are obtained for the "others" category. The high scores for the third person pronoun and the proper name rankers most likely follow from the fact that the resolution of these expressions relies on simple, reliable predictors, such as distance and morphosyntactic agreement for pronouns, and string similarity features for proper names. The resolution of definite descriptions and other types of lexical NPs (which are handled through the "others" model) are much more challenging: they rely on lexical semantic and world knowledge, which is only partially encoded via our WordNet-based features. Finally, note that the resolution of speech pronouns is also much harder than that of the other pronominal forms: these expressions are much less (if at all) constrained by recency and agreement. Furthermore, these expressions show a lot of cataphoric uses, which are not considered by our models. The low scores for the "others" category is likely due to the fact that it encompasses very different referential expressions.

5.2 Coreference Results

For evaluating the coreference performance, we rely on three primary metrics: (i) the **link** based MUC metric (Vilain et al., 1995), the **mention** based B^3 metric (Bagga and Baldwin, 1998), and the **entity** based CEAF metric (Luo, 2005). Common to these metrics is: (i) they operate by comparing the set of chains $\mathcal{S}$ produced by the system against the true chains $\mathcal{T}$, and (ii) they report performance in terms of *recall* and *precision*. There are however important differences in how each metric computes these scores, each producing a different bias.

MUC scores are based on the number of *links* (pairs of mentions) common to $\mathcal{S}$ and $\mathcal{T}$. Recall is the number of common links divided by the total number of links in $\mathcal{T}$; precision is the number of common links divided by the total number of links in $\mathcal{S}$. This focus gives MUC two main biases. First, it favors systems that create large chains (and thus fewer entities). For instance, a system that produces a single chain achieves 100% recall without severe degradation in precision. Second, it ignores single mention entities, which are involved in no links.[9]

The B^3 metric was designed to address the MUC metric's shortcomings. It is *mention-based*: it computes both recall and precision scores for each mention i. Let S be the system chain containing m, T be the true chain containing m. The set of correct elements in S is thus $|S \cap T|$. The recall score for a mention i is $\frac{|S \cap T|}{|T|}$, while the precision score for i is $\frac{|S \cap T|}{|S|}$. Overall recall/precision is obtained by averaging over the individual mention scores. The fact that this metric is mention-based by definition solves the problem of single mention entities. Also solved is the bias favoring larger chains, since this will be penalized in the precision score of *each* mention.

The Constrained Entity Aligned F-Measure (CEAF) (Luo, 2005). aligns each system chain S with *at most one* true chain T. It finds the best one-to-one mapping between the set of chains $\mathcal{S}$ and $\mathcal{T}$, which is equivalent to finding the optimal alignment in a bipartite graph. The best mapping maximizes the similarity over pairs of chains (S_i, T_i), where the similarity between two chains is the number of common mentions to the two chains. With CEAF, recall is computed as the total similarity divided by the number of mentions in all the $\mathcal{T}$ (i.e., the self-similarity), while precision is the total similarity divided by the number of mentions in $\mathcal{S}$.

Table 6 gives scores for all three metrics for the different models on the entire ACE corpus. Two main patterns emerge: significant improvements are obtained by using specialized models (**CLASS** vs **CLASS+SP** and **CLASS+DS** vs **CLASS+DS+SP**) and by using a ranker (**CLASS+DS+SP** vs **RANK+DS+SP**). Overall, the **RANK+DS+SP** system significantly outperforms the other systems on the three different metrics.[10]

The f-scores for **RANK+DS+SP** are 71.6% with the MUC metric, 72.7% with the B^3, and 67.0% with the CEAF metric. These scores place the **RANK+DS+SP** among the best coreference resolution systems, since most existing systems are typically under the bar of the 70% in f-score with the

[9] It is worth noting that the MUC corpus does not annotate single mention entities.

[10] Statistical significance was determined with t-tests for both recall and precision scores, with $p < 0.05$.

System	MUC			B^3			CEAF
	R	P	F	R	P	F	F
CLASS	60.8	72.6	66.2	62.4	77.7	69.2	62.3
CLASS+DS	64.9	72.3	68.4	65.6	74.1	69.6	63.4
CLASS+SP	64.8	74.5	69.3	65.3	79.1	71.5	65.0
CLASS+DS+SP	66.8	74.4	70.4	66.4	77.0	71.3	65.3
RANK+DS+SP	67.9	75.7	71.6	66.8	79.8	72.7	67.0

Table 6: Recall (R), Precision (P), and f-score (F) results on the entire ACE corpus using the MUC, B^3, and CEAF metrics. Note that R=P=F for CEAF when using true mentions, as we do here.

MUC and B^3 metrics (Ng, 2005a). An interesting point of comparison is provided by Ng (2007), who also relies on true mentions and reports MUC f-scores only slightly superior to ours (73.8%) while relying on perfect semantic class information. His best results otherwise are 64.6%. The fact that our improvements are consistent across the different evaluation metrics is remarkable, especially given that these three metrics are quite different in the way they compute their scores. The gains in f-score range from 1.2 to 5.4% on the MUC metric (i.e., error reductions of 4 to 15.9%), from 1.4 to 3.5% on the B^3 metric (i.e., error reductions of 4.8 to 11.4%), and from 1.7 to 4.7% on the CEAF metric (i.e., error reductions of 6.9 to 17%). The larger improvements come from recall, with improvements ranging from 1.9 to 7.1% with MUC, from 2.4 to 5.6% with B^3.[11] This suggests that **RANK+DS+SP** predicts many more valid coreference links than the other systems. Smaller but still significant gains are made in precision: **RANK+DS+SP** is also able to reduce the proportion of invalid links.

The overall improvements found with **RANK+DS+SP** suggest that it is able to capitalize on the better antecedent selection capabilities offered by the ranking approach. This is supported by the error analysis on the development data. Errors made by a coreference system can be conceptualized as falling into three main classes: (i) "missed anaphors" (i.e., an anaphoric mention that fails to be linked to a previous mention), (ii) "spurious anaphors" (i.e., an non-anaphoric mention that is linked to a previous mention), and (iii) "invalid resolutions" (i.e., a true anaphor that is linked to a incorrect antecedent). The two first types of error pertain to the determination of the discourse status of the mention, while the third regards the selection of an antecedent (i.e., anaphora resolution). Considering the systems' invalid resolutions, we found that the **RANK+DS+SP** had a much lower error rate: only 17.9% of all true anaphors were incorrectly resolved by this system, against 23.1% for **CLASS**, 24.9% for **CLASS+DS**, 20.4% for **CLASS+SP**, and 22.1% for **CLASS+DS+SP**.

Denis (2007) provides multi-metric scores for the **JOINT-ILP** model of Denis and Baldridge (2007a), which uses integer linear programming for joint inference over coreference resolution and discourse status: f-scores of 73.3%, 68.0%, and 58.9% for MUC, B^3, and CEAF, respectively. Despite the fact that this MUC score beats **RANK+DS+SP**'s, it is actually *worse* than even the basic model **CLASS** for B^3 and CEAF. This difference fact that MUC gives more recall credit for large chains without a consequent precision reduction, and shows the importance of using B^3 and CEAF scores in addition to MUC.

Denis (2007) also extends the **JOINT-ILP** system by adding named entity resolution and constraints on transitivity with respect to coreference links. The best model reported there (**JOINT-DS-NE-AE-ILP**) obtains f-scores of 70.1%, 72.7%, and 66.2% for MUC, B^3, and CEAF, respectively. Interestingly, **RANK+DS+SP** actually performs better across all metrics despite being a simpler model with fewer sources of information.

5.3 Oracle results

Using specialized rankers with a discourse status classifier yields coreference performance superior to that given by various classification-based baseline systems. Crucially, these improvements have been

[11] Recall that recall and precision scores are identical with CEAF, due to the fact that we are using true mention boundaries.

System	MUC			B^3			CEAF
	R	P	F	R	P	F	F
RANK+DS+SP	67.9	75.7	71.6	66.8	79.8	72.7	67.0
RANK+DS-ORACLE+SP	79.1	79.1	79.1	75.4	76.0	75.7	76.9
LINK-ORACLE	78.8	100.0	88.1	74.3	100.0	85.2	79.7

Table 7: Recall (R), Precision (P), and f-score (F) results for **RANK+DS-ORACLE+SP** and **LINK-ORACLE** on the entire ACE corpus.

possible using a discourse status model that has an accuracy of just 80.8%. Clearly, the performance of the discourse status module has a direct impact on the performance of the entire coreference system. On the one hand, misclassified anaphors are simply not resolved by the rankers: this limits the recall of the coreference system. On the other hand, misclassified non-anaphors are linked to a previous mention: this limits precision.

In order to further assess the impact of the errors made by the discourse status classifier, we build two different oracle systems. The first oracle system, **RANK+DS-ORACLE+SP**, uses the specialized rankers in combination with a perfect discourse status classifier. That is, this system knows for each mention whether it is anaphoric or not: the only errors made by such a system are invalid resolutions. **RANK+DS-ORACLE+SP** thus provides an upperbound for the **RANK+DS+SP** model. The results for this oracle are given in Table 7: they show substantial improvements over **RANK+DS+SP**, which suggests that the **RANK+DS+SP** has also the potential to be further improved if used in combination with a more accurate discourse status classifier.

The second oracle system, **LINK-ORACLE**, uses the discourse status classifier with a perfect coreference resolver. That is, this system has perfect knowledge regarding the antecedents of anaphors: the errors made by such a system are only errors in the discourse status of mentions. The results for **LINK-ORACLE** are also reported in Table 7. These figures show that however accurate our models are at picking a correct antecedent for a true anaphor, the best they can achieve in terms of f-scores is 88.1% with MUC, 85.2% with B^3, and 79.7% with CEAF.

6 Conclusion

We present and evaluate two straight-forward tactics for improving coreference resolution: (i) ranking models, and (ii) separate, specialized models for different types of referring expressions. The specialized rankers are used in combination with a discourse status classifier which determines the mentions that are sent to the rankers. This simple pipeline architecture produces significant improvements over various implementations of the standard, classifier-based coreference system. In turn, these strategies could be integrated with the joint inference models we have explored elsewhere (Denis and Baldridge, 2007a; Denis, 2007) and which have obtained performance improvements that are orthogonal to those obtained here.

This paper's improvements are consistent across the three main coreference evaluation metrics: MUC, B^3, and CEAF.[12] We attribute improvements to: (i) the better antecedent selection capabilities offered by the ranking approach, and (ii) the division of labor between specialized models, allowing each one to better model the corresponding distribution.

Acknowledgments

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

References

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

N. Asher, P. Denis, and B. Reese. 2006. Names and pops and discourse structure. In *Workshop on Constraints in Discourse*, Maynooth, Ireland.

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

[12] We *strongly* advocate that coreference results should ***never*** be presented in terms of MUC scores alone.

E. Charniak and M. Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of ACL 2005*, Ann Arbor, Michigan.

M. Collins and N. Duffy. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures and the voted perceptron. In *Proceedings of ACL 2002*, pages 263–270, Philadelphia, PA.

H. Daumé III and D. Marcu. 2005. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of HLT-EMNLP 2005*, Vancouver, Canada.

P. Denis and J. Baldridge. 2007a. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of HLT-NAACL 2007*, Rochester, NY.

P. Denis and J. Baldridge. 2007b. A ranking approach to pronoun resolution. In *Proceedings of IJCAI 2007*, Hyderabad, India.

Pascal Denis. 2007. *New Learning Models for Robust Reference Resolution*. Ph.D. thesis, The University of Texas at Austin.

R. Elwell and J. Baldridge. 2008. Discourse connective argument identification with connective specific rankers. In *Proceedings of the International Conference on Semantic Computing*, Santa Clara, CA.

B. Grosz, A. Joshi, and S. Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 2(21).

J. K. Gundel, N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.

A. Haghighi and D. Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings ACL 2007*, pages 848–855, Prague, Czech Republic.

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

X. Luo. 2005. On coreference resolution performance metrics. In *Proceedings of HLT-NAACL 2005*, pages 25–32.

R. 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.

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

J. F. McCarthy and W. G. Lehnert. 1995. Using decision trees for coreference resolution. In *IJCAI*, pages 1050–1055.

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

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

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

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

V. Ng. 2005a. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of ACL 2005*, pages 157–164, Ann Arbor, MI.

V. Ng. 2005b. Supervised ranking for pronoun resolution: Some recent improvements. In *Proceedings of AAAI 2005*.

V. Ng. 2007. Semantic class induction and coreference resolution. In *Proceedings of ACL 2007*.

M. Osborne and J. Baldridge. 2004. Ensemble-based active learning for parse selection. In *Proceedings of HLT-NAACL 2004*, pages 89–96, Boston, MA.

E. F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York.

D. Ravichandran, E. Hovy, and F. J. Och. 2003. Statistical QA - classifier vs re-ranker: What's the difference? In *Proceedings of the ACL Workshop on Multilingual Summarization and Question Answering–Machine Learning and Beyond*.

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

K. Toutanova, P. Markova, and C. Manning. 2004. The leaf projection path view of parse trees: Exploring string kernels for HPSG parse selection. In *Proceedings of EMNLP 2004*, pages 166–173, Barcelona.

O. Uryupina. 2004. Linguistically motivated sample selection for coreference resolution. In *Proceedings of DAARC 2004*, Furnas.

R. van der Sandt. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics*, 9:333–377.

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

X. Yang, G. Zhou, J. Su, and C.L. Tan. 2003. Coreference resolution using competitive learning approach. In *Proceedings of ACL 2003*, pages 176–183.