Easy Victories and Uphill Battles in Coreference Resolution

Greg Durrett and **Dan Klein**
Computer Science Division
University of California, Berkeley
{gdurrett,klein}@cs.berkeley.edu

Abstract

Classical coreference systems encode various syntactic, discourse, and semantic phenomena explicitly, using heterogenous features computed from hand-crafted heuristics. In contrast, we present a state-of-the-art coreference system that captures such phenomena implicitly, with a small number of homogeneous feature templates examining shallow properties of mentions. Surprisingly, our features are actually more effective than the corresponding hand-engineered ones at modeling these key linguistic phenomena, allowing us to win "easy victories" without crafted heuristics. These features are successful on syntax and discourse; however, they do not model semantic compatibility well, nor do we see gains from experiments with shallow semantic features from the literature, suggesting that this approach to semantics is an "uphill battle." Nonetheless, our final system[1] outperforms the Stanford system (Lee et al. (2011), the winner of the CoNLL 2011 shared task) by 3.5% absolute on the CoNLL metric and outperforms the IMS system (Björkelund and Farkas (2012), the best publicly available English coreference system) by 1.9% absolute.

1 Introduction

Coreference resolution is a multi-faceted task: humans resolve references by exploiting contextual and grammatical clues, as well as semantic information and world knowledge, so capturing each of these will be necessary for an automatic system to fully solve the problem. Acknowledging this complexity, coreference systems, either learning-based (Bengtson and Roth, 2008; Stoyanov et al., 2010; Haghighi and Klein, 2010; Rahman and Ng, 2011b) or rule-based (Haghighi and Klein, 2009; Lee et al., 2011), draw on diverse information sources and complex heuristics to resolve pronouns, model discourse, determine anaphoricity, and identify semantically compatible mentions. However, this leads to systems with many heterogenous parts that can be difficult to interpret or modify.

We build a learning-based, mention-synchronous coreference system that aims to use the simplest possible set of features to tackle the various aspects of coreference resolution. Though they arise from a small number of simple templates, our features are numerous, which works to our advantage: we can both implicitly model important linguistic effects and capture other patterns in the data that are not easily teased out by hand. As a result, our data-driven, homogeneous feature set is able to achieve high performance despite only using surface-level document characteristics and shallow syntactic information. We win "easy victories" without designing features and heuristics explicitly targeting particular phenomena.

Though our approach is successful at modeling syntax, we find semantics to be a much more challenging aspect of coreference. Our base system uses only two recall-oriented features on nominal and proper mentions: head match and exact string match. Building on these features, we critically evaluate several classes of semantic features which intu-

[1] The Berkeley Coreference Resolution System is available at http://nlp.cs.berkeley.edu.

itively should prove useful but have had mixed results in the literature, and we observe that they are ineffective for our system. However, these features are beneficial when gold mentions are provided to our system, leading us to conclude that the large number of system mentions extracted by most coreference systems (Lee et al., 2011; Fernandes et al., 2012) means that weak indicators cannot overcome the bias against making coreference links. Capturing semantic information in this shallow way is an "uphill battle" due to this structural property of coreference resolution.

Nevertheless, using a simple architecture and feature set, our final system outperforms the two best publicly available English coreference systems, the Stanford system (Lee et al., 2011) and the IMS system (Björkelund and Farkas, 2012), by wide margins: 3.5% absolute and 1.9% absolute, respectively, on the CoNLL metric.

2 Experimental Setup

Throughout this work, we use the datasets from the CoNLL 2011 shared task[2] (Pradhan et al., 2011), which is derived from the OntoNotes corpus (Hovy et al., 2006). When applicable, we use the standard automatic parses and NER tags for each document. All experiments use system mentions except where otherwise indicated. For each experiment, we report MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and CEAF_e (Luo, 2005), as well as their average, the CoNLL metric. All metrics are computed using version 5 of the official CoNLL scorer.[3]

3 A Mention-Synchronous Framework

We first present the basic architecture of our coreference system, independent of a feature set. Unlike binary classification-based coreference systems where independent binary decisions are made about each pair (Soon et al., 2001; Bengtson and Roth, 2008; Versley et al., 2008; Stoyanov et al., 2010), we use a log-linear model to select at most one antecedent for each mention or determine that it begins a new cluster (Denis and Baldridge, 2008). In this mention-ranking or mention-synchronous framework, features examine single mentions to evaluate whether or not they are anaphoric and pairs of mentions to evaluate whether or not they corefer. While other work has used this framework as a starting point for entity-level systems (Luo et al., 2004; Rahman and Ng, 2009; Haghighi and Klein, 2010; Durrett et al., 2013), we will show that a mention-synchronous approach is sufficient to get state-of-the-art performance on its own.

3.1 Mention Detection

Our system first identifies a set of predicted mentions from text annotated with parses and named entity tags. We extract three distinct types of mentions: proper mentions from all named entity chunks except for those labeled as QUANTITY, CARDINAL, or PERCENT, pronominal mentions from single words tagged with PRP or PRP$, and nominal mentions from all other maximal NP projections. These basic rules are similar to those of Lee et al. (2011), except that their system uses an additional set of filtering rules designed to discard instances of pleonastic *it*, partitives, certain quantified noun phrases, and other spurious mentions. In contrast to this highly engineered approach and to systems which use a trained classifier to compute anaphoricity separately (Rahman and Ng, 2009; Björkelund and Farkas, 2012), we aim for the highest possible recall of gold mentions with a low-complexity method, leaving us with a large number of spurious system mentions that we will have to reject later.

3.2 Coreference Model

Figure 1 shows the mention-ranking architecture that serves as the backbone of our coreference system. Assume we have extracted n mentions from a document x, where x denotes the surface properties of a document and any precomputed information. The ith mention in a document has an associated random variable a_i taking values in the set $\{1, \ldots, i-1, \text{NEW}\}$; this variable specifies mention i's selected antecedent or indicates that it begins a new coreference chain. A setting of the a_i, denoted by $a = (a_1, ..., a_n)$, implies a unique set of coreference chains C that serve as our system output.

[2]This dataset is identical to the English portion of the CoNLL 2012 data, except for the absence of a small pivot text.

[3]Note that this version of the scorer implements a modified version of B^3, described in Cai and Strube (2010), that was used for the CoNLL shared tasks. The implementation of CEAF_e is also not exactly as described in Luo et al. (2004), but for completeness we include this metric as well.

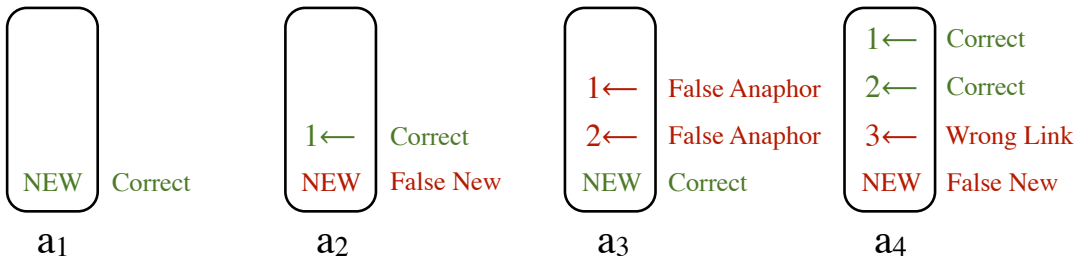

Figure 1: The basic structure of our coreference model. The ith mention in a document has i possible antecedence choices: link to one of the $i-1$ preceding mentions or begin a new cluster. We place a distribution over these choices with a log-linear model. Structurally different kinds of errors are weighted differently to optimize for final coreference loss functions; error types are shown corresponding to the decisions for each mention.

We use a log linear model of the conditional distribution $P(a|x)$ as follows:

$$P(a|x) \propto \exp\left(\sum_{i=1}^{n} \mathbf{w}^\top \mathbf{f}(i, a_i, x)\right)$$

where $\mathbf{f}(i, a_i, x)$ is a feature function that examines the coreference decision a_i for mention i with document context x. When $a_i = \text{NEW}$, the features fired indicate the suitability of the given mention to be anaphoric or not; when $a_i = j$ for some j, the features express aspects of the pairwise linkage, and can examine any relevant attributes of the anaphor i or the antecedent j, since information about each mention is contained in x.

Inference in this model is efficient: because $\log P(a|x)$ decomposes linearly over mentions, we can compute $a_i = \arg\max_{a_i} P(a_i|x)$ separately for each mention and return the set of coreference chains implied by these decisions.

3.3 Learning

During learning, we optimize for conditional log-likelihood augmented with a parameterized loss function (Durrett et al., 2013). The main complicating factor in this process is that the supervision in coreference consists of a gold clustering C^* defined over gold mentions. This is problematic for two reasons: first, because the clustering is defined over gold mentions rather than our system mentions, and second, because a clustering does not specify a full antecedent structure of the sort our model produces. We can address the first of these problems by imputing singleton clusters for mentions that do not appear in the gold standard; our system will then simply learn to put spurious mentions in their own clusters. Singletons are always removed before evaluation because the OntoNotes corpus does not annotate them, so in this way we can neatly dispose of spurious mentions. To address the lack of explicit antecedents in C^*, we simply sum over all possible antecedent structures licensed by the gold clusters.

Formally, we will maximize the conditional log-likelihood of the set $\mathcal{A}(C^*)$ of antecedent vectors a for a document that are *consistent* with the gold annotation.[4] Consistency for an antecedent choice a_i under gold clusters C^* is defined as follows:

1. If $a_i = j$, a_i is consistent iff mentions i and j are present in C^* and are in the same cluster.

2. If $a_i = \text{NEW}$, a_i is consistent off mention i is not present in C^*, or it is present in C^* and has no gold antecedents, or it is present in C^* and none of its gold antecedents are among the set of system predicted mentions.

Given t training examples of the form (x_k, C_k^*), we write the following likelihood function:

$$\ell(\mathbf{w}) = \sum_{k=1}^{t} \log\left(\sum_{a \in \mathcal{A}(C_k^*)} P'(a|x_k)\right) + \lambda\|\mathbf{w}\|_1$$

where $P'(a|x_k) \propto P(a|x_k)\exp(l(a, C_k^*))$ with $l(a, C^*)$ being a real-valued loss function. The loss

[4] Because of this marginalization over latent antecedent choices, our objective is non-convex.

here plays an analogous role to the loss in structured max-margin objectives; incorporating it into a conditional likelihood objective is a technique called softmax-margin (Gimpel and Smith, 2010).

Our loss function $l(a, C^*)$ is a weighted linear combination of three error types, examples of which are shown in Figure 1. A false anaphor (FA) error occurs when a_i is chosen to be anaphoric when it should start a new cluster. A false new (FN) error occurs in the opposite case, when a_i wrongly indicates a new cluster when it should be anaphoric. Finally, a wrong link (WL) error occurs when the antecedent chosen for a_i is the wrong antecedent (but a_i is indeed anaphoric). Our final parameterized loss function is a weighted sum of the counts of these three error types:

$$l(a, C^*) = \alpha_{\text{FA}}\text{FA}(a, C^*) + \alpha_{\text{FN}}\text{FN}(a, C^*) + \alpha_{\text{WL}}\text{WL}(a, C^*)$$

where $\text{FA}(a, C^*)$ gives the number of false anaphor errors in prediction a with gold chains C^* (FN and WL are analogous). By setting α_{FA} low and α_{FN} high relative to α_{WL}, we can counterbalance the high number of singleton mentions and bias the system towards making more coreference linkages. We set $(\alpha_{\text{FA}}, \alpha_{\text{FN}}, \alpha_{\text{WL}}) = (0.1, 3.0, 1.0)$ and $\lambda = 0.001$ and optimize the objective using AdaGrad (Duchi et al., 2011).

4 Easy Victories from Surface Features

Our primary goal with this work is to show that a high-performance coreference system is attainable with a small number of feature templates that use only surface-level information sources. These features will be general-purpose and capture linguistic effects to the point where standard heuristic-driven features are no longer needed in our system.

4.1 SURFACE Features and Conjunctions

Our SURFACE feature set only considers the following properties of mentions and mention pairs:

- Mention type (nominal, proper, or pronominal)
- The complete string of a mention
- The semantic head of a mention
- The first word and last word of each mention
- The word immediately preceding and the word immediately following a mention
- Mention length, in words
- Two distance measures between mentions (number of sentences and number of mentions)

Feature name	Count
Features on the current mention	
[ANAPHORIC] + [HEAD WORD]	41371
[ANAPHORIC] + [FIRST WORD]	18991
[ANAPHORIC] + [LAST WORD]	19184
[ANAPHORIC] + [PRECEDING WORD]	54605
[ANAPHORIC] + [FOLLOWING WORD]	57239
[ANAPHORIC] + [LENGTH]	4304
Features on the antecedent	
[ANTECEDENT HEAD WORD]	57383
[ANTECEDENT FIRST WORD]	24239
[ANTECEDENT LAST WORD]	23819
[ANTECEDENT PRECEDING WORD]	53421
[ANTECEDENT FOLLOWING WORD]	55718
[ANTECEDENT LENGTH]	4620
Features on the pair	
[EXACT STRING MATCH (T/F)]	47
[HEAD MATCH (T/F)]	46
[SENTENCE DISTANCE, CAPPED AT 10]	2037
[MENTION DISTANCE, CAPPED AT 10]	1680

Table 1: Our SURFACE feature set, which exploits a small number of surface-level mention properties. Feature counts for each template are computed over the training set, and include features generated by our conjunction scheme (not explicitly shown in the table; see Figure 2), which yields large numbers of features at varying levels of expressivity.

Table 1 shows the SURFACE feature set. Features that look only at the current mention fire on all decisions ($a_i = j$ or $a_i = \text{NEW}$), whereas features that look at the antecedent in any way (the latter two groups of features) only fire on pairwise linkages ($a_i \neq \text{NEW}$).

Two conjunctions of each feature are also included: first with the "type" of the mention being resolved (either NOMINAL, PROPER, or, if it is pronominal, the citation form of the pronoun), and then additionally with the antecedent type (only if the feature is over a pairwise link). This conjunction process is shown in Figure 2. Note that features that just examine the antecedent will end up with

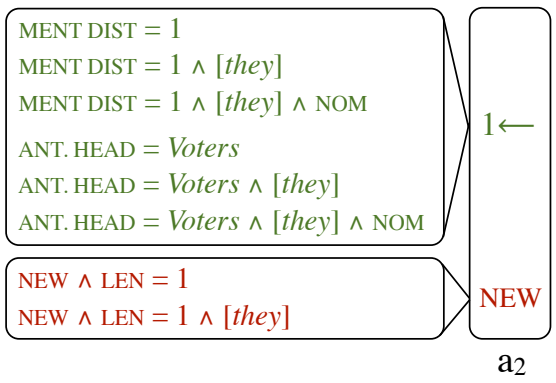

Figure 2: Demonstration of the conjunction scheme we use. Each feature on anaphoricity is conjoined with the type (NOMINAL, PROPER, or the citation form if it is a pronoun) of the mention being resolved. Each feature on a mention pair is additionally conjoined with the types of the current and antecedent mentions.

conjunctions that examine properties of the current mention as well, as shown with the ANT. HEAD feature in the figure.

Finally, we found it beneficial for our lexical indicator features to only fire on words occurring at least 20 times in the training set; for rare words, we use the part of speech of the word instead.

The performance of our system is shown in Table 2. We contrast our performance with that of the Stanford system (Lee et al. (2011), the winner of the CoNLL 2011 shared task) and the IMS system (Björkelund and Farkas (2012), the best publicly available English coreference system). Despite its simplicity, our SURFACE system is sufficient to outperform these sophisticated systems: the Stanford system uses a cascade of ten rule-based sieves each of which has customized heuristics, and the IMS system uses a similarly long pipeline consisting of a learned referentiality classifier followed by multiple resolvers, which are run in sequence and rely on the outputs of the previous resolvers as features.

4.2 Data-Driven versus Heuristic-Driven Features

Why are the SURFACE features sufficient to give high coreference performance, when they do not make apparent reference to important linguistic phenomena? The main reason is that they actually do capture the same phenomena as standard coreference features, just implicitly. For example, rather than having rules targeting person, number, gender, or animacy of mentions, we use conjunctions with pronoun identity, which contains this information. Rather than explicitly writing a feature targeting definiteness, our indicators on the first word of a mention will capture this and other effects. And finally, rather than targeting centering theory (Grosz et al., 1995) with rule-based features identifying syntactic positions (Stoyanov et al., 2010; Haghighi and Klein, 2010), our features on word context can identify configurational clues like whether a mention is preceded or followed by a verb, and therefore whether it is likely in subject or object position.[5]

	MUC	B^3	CEAF_e	Avg.
STANFORD	60.46	65.48	47.07	57.67
IMS	62.15	65.57	46.66	58.13
SURFACE	**64.39**	**66.78**	**49.00**	**60.06**

Table 2: Results for our SURFACE system, the STANFORD system, and the IMS system on the CoNLL 2011 development set. Complete results are shown in Table 7. Despite using limited information sources, our system is able to substantially outperform the other two, the two best publicly-available English coreference systems. Bolded values are significant with $p < 0.05$ according to a bootstrap resampling test.

Not only are data-driven features able to capture the same phenomena as heuristic-driven features, but they do so at a finer level of granularity, and can therefore model more patterns in the data. To contrast these two types of features, we experiment with three ablated versions of our system, where we replace data-driven features with their heuristic-driven counterparts:

1. Instead of using an indicator on the first word of a mention (1STWORD), we instead fire a feature based on that mention's manually-computed definiteness (DEF).

2. Instead of conjoining features on pronominal-pronominal linkages with the citation form of

[5]Heuristic-driven approaches were historically more appropriate, since past coreference corpora such as MUC and ACE were smaller and therefore more prone to overfitting feature-rich models. However, the OntoNotes dataset contains thousands of documents, so having support for features is less of a concern.

	MUC	B^3	CEAF_e	Avg.
SURFACE	64.39	66.78	49.00	60.06
−1STWORD	63.32	66.22	47.89	59.14
+DEF−1STWORD	63.79	66.46	48.35	59.53
−PRONCONJ	59.97	63.46	47.94	57.12
+AGR−PRONCONJ	63.54	66.10	48.72	59.45
−CONTEXT	60.88	64.66	47.60	57.71
+POSN−CONTEXT	62.45	65.44	48.08	58.65
+DEF+AGR+POSN	64.55	66.93	48.94	60.14

Table 3: CoNLL metric scores on the development set, for the three different ablations and replacement features described in Section 4.2. Feature types are described in the text; + indicates inclusion of that feature class, − indicates exclusion. Each individual shallow indicator appears to do as well at capturing its target phenomenon as the hand-engineered features, while capturing other information as well. Moreover, the hand-engineered features give no benefit over the SURFACE system.

each pronoun (PRONCONJ), we only conjoin with a PRONOUN indicator and add features targeting the person, number, gender, and animacy of the two pronouns (AGR).

3. Instead of using our context features on the preceding and following word (CONTEXT), we use manual determinations of when mentions are in subject, direct object, indirect objection, or oblique position (POSN).

All rules for computing person, number, gender, animacy, definiteness, and syntactic position are taken from the system of Lee et al. (2011).

Table 3 shows each of the target ablations, as well as the SURFACE system with the DEF, AGR, and POSN features added. While the heuristic-driven feature always help over the corresponding ablated system, they cannot do the work of the fine-grained data-driven features. Most tellingly, though, none of the heuristic-driven features give statistically significant improvements on top of the data-driven features we have already included, indicating that we are at the point of diminishing returns on modeling those specific phenomena. While this does not preclude further engineering to take better advantage of other syntactic constraints, our simple features represent an "easy victory" on this subtask.

5 Uphill Battles on Semantics

In Section 4, we gave a simple set of features that yielded a high-performance coreference system; this high performance is possible because features targeting only superficial properties in a fine-grained way can actually model complex linguistic constraints. However, while our existing features capture syntactic and discourse-level phenomena surprisingly well, they are not effective at capturing semantic phenomena like type compatibility. We will show that due to structural aspects of the coreference resolution problem, even a combination of several shallow semantic features from the literature fails to adequately model semantics.

5.1 Analysis of the SURFACE System

What can the SURFACE system resolve correctly, and what errors does it still make? To answer this question, we will split mentions into several categories based on their observable properties and the gold standard coreference information, and examine our system's accuracy on each mention subclass in order to more thoroughly characterize its performance.[6] These categories represent important distinctions in terms of the difficulty of mention resolution for our system.

We first split mentions into three categories by their *status* in the gold standard: singleton (unannotated in the OntoNotes corpus), starting a new entity with at least two mentions, or anaphoric. It is important to note that while singletons and mentions starting new entities are outwardly similar in that they have no antecedents, and the prediction should be the same in either case (NEW), we treat them as distinct because the factors that impact the coreference decision differ in the two cases. Mentions that start new clusters are semantically similar to anaphoric mentions, but may be marked by heaviness or by a tendency to be named entities, whereas singletons may be generic or temporal NPs which might be thought of as coreferent in a loose sense, but are not

[6]This method of analysis is similar to that undertaken in Stoyanov et al. (2009) and Rahman and Ng (2011b), though we split our mentions along different axes, and can simply evaluate on accuracy because our decisions do not directly imply multiple links, as they do in binary classification-based systems (Stoyanov et al., 2009) or in entity-mention models (Rahman and Ng, 2011b).

	Nominal/Proper				Pronominal	
	1st w/head		2nd+ w/head			
Singleton	99.7%	18.1K	85.5%	7.3K	66.5%	1.7K
Starts Entity	98.7%	2.1K	78.9%	0.7K	48.5%	0.3K
Anaphoric	7.9%	0.9K	75.5%	3.9K	72.0%	4.4K

Table 4: Analysis of our SURFACE system on the development set. We characterize each predicted mention by its status in the gold standard (singleton, starting a new entity, or anaphoric), its type (pronominal or nominal/proper), and by whether its head has appeared as the head of a previous mention. Each cell shows our system's accuracy on that mention class as well as the size of the class. The biggest weakness of our system appears to be its inability to resolve anaphoric mentions with new heads (bottom-left cell).

included in the OntoNotes dataset due to choices in the annotation standard.

Second, we divide mentions by their type, pronominal versus nominal/proper; we then further subdivide nominals and propers based on whether or not the head word of the mention has appeared as the head of a previous mention in the document.

Table 4 shows the results of our analysis. In each cell, we show the fraction of mentions that we correctly resolve (i.e., for which we make an antecedence decision consistent with the gold standard), as well as the total number of mentions falling into that cell. First, we observe that there are a surprisingly large number of singleton mentions with misleading head matches to previous mentions (often recurring temporal nouns phrases, like *July*). The features in our system targeting anaphoricity are useful for exactly this reason: the more bad head matches we can rule out based on other criteria, the more strongly we can rely on head match to make correct linkages.

Our system is most noticeably poor at resolving anaphoric mentions whose heads have not appeared before. The fact that exact match and head match are our only recall-oriented features on nominals and propers is starkly apparent here: when we cannot rely on head match, as is true for this mention class, we only resolve 7.9% of anaphoric mentions correctly.[7] Many of the mentions in this category can only be correctly resolved by exploiting world knowledge, so we will need to include features that capture this knowledge in some fashion.

5.2 Incorporating Shallow Semantics

As we were able to incorporate syntax with shallow features, so too might we hope to incorporate semantics. However, the semantic information contained even in a coreference corpus of thousands of documents is insufficient to generalize to unseen data,[8] so system designers have turned to external resources such as semantic classes derived from WordNet (Soon et al., 2001), WordNet hypernymy or synonymy (Stoyanov et al., 2010), semantic similarity computed from online resources (Ponzetto and Strube, 2006), named entity type features, gender and number match using the dataset of Bergsma and Lin (2006), and features from unsupervised clusters (Hendrickx and Daelemans, 2007; Durrett et al., 2013). In this section, we consider the following subset of these information sources:

- WordNet hypernymy and synonymy
- Number and gender data for nominals and propers from Bergsma and Lin (2006)
- Named entity types
- Latent clusters computed from English Gigaword (Graff et al., 2007), where a latent cluster label generates each nominal head (excluding pronouns) and a conjunction of its verbal governor and semantic role, if any (Durrett et al., 2013). We use twenty clusters, which include clusters like *president* and *leader* (things which *announce*).

Together, we call these the SEM features. We show results from this expansion of the feature set in Table 5. When using system mentions, the improvements are not statistically significant on every metric, and are quite marginal given that these features add information that is intuitively central to coreference and otherwise unavailable to the system. We explore the reasons behind this in the next section.

[7] There are an additional 346 anaphoric nominal/proper mentions in the 2nd+ category whose heads only appeared previously as part of a different cluster; we only resolve 1.7% of these extremely tricky cases correctly.

[8] We experimented with bilexical features on head pairs, but they did not give statistically significant improvements over the SURFACE features.

	MUC	B^3	CEAF_e	Avg.
SURFACE	64.39	66.78	49.00	60.06
SURFACE+SEM	64.70	**67.27**	**49.28**	**60.42**
SURFACE (G)	82.80	74.10	68.33	75.08
SURFACE+SEM (G)	**84.49**	**75.65**	**69.89**	**76.68**

Table 5: CoNLL metric scores on the development set for our SEM features when added on top of our SURFACE features. We experiment on both system mentions and gold mentions. Surprisingly, despite the fact that absolute performance numbers are much higher on gold mentions and there is less room for improvement, the semantic features help much more than they do on system mentions.

5.3 Analysis of Semantic Features

The main reason that weak semantic cues are not more effective is the small fraction of positive coreference links present in the training data. From Table 4, the number of annotated coreferent spans in the OntoNotes data is about a factor of five smaller than the number of system mentions.[9] This both means that most NPs are not coreferent, and for those that are, choosing the correct links is much more difficult because of the large number of possible antecedents. Even head match, which is generally considered a high-precision indicator (Lee et al., 2011), would introduce many spurious coreference arcs if applied too liberally (see Table 4).

In light of this fact, a system needs very strong evidence to overcome the default hypothesis that a mention is not coreferent, and a weak indicator will have such a high "false positive" rate that it cannot be relied on (given high weight, this feature would do more harm than good, by introducing many false linkages).

To confirm this intuition, we show in the bottom part of Table 5 results when we apply these semantic features on top of our SURFACE system on *gold mentions*, where there are no singletons. In the gold mention setting, we see that the semantic features give a consistent improvement on every metric. Moreover, if we look at a breakdown of errors, the main improvement the semantic features give us is on resolution of anaphoric nominals with no head match: accuracy on the 1601 mentions that fall into this category improves from 28.0% to 37.9%. On predicted mentions, by contrast, this category only improves from 7.9% to 12.2%, a much smaller absolute improvement and one that comes at the expense of performance on most other resolution class. The one class that does not get worse, singleton pronouns, actually improves by a similar 4% margin, indicating that roughly half of the gains we observe are not even necessarily a result of our features doing what they were designed to do.

Our weak cues do yield some small gains, so there is hope that better weak indicators of semantic compatibility could prove more useful. However, while extremely high-precision approaches with carefully engineered features have been shown to be successful (Rahman and Ng, 2011a; Bansal and Klein, 2012; Recasens et al., 2013a), we conclude that capturing semantics in a data-driven, shallow manner remains an uphill battle.

6 FINAL System and Results

While semantic features ended up giving only marginal benefit, we have demonstrated that nevertheless our SURFACE system is a state-of-the-art English coreference system. However, there remain a few natural features that we omitted in order to keep the system as simple as possible, since they were orthogonal to the discussion of data-driven versus heuristic-driven features and do not target world knowledge. Before giving final results, we will present a small set of additional features that consider four additional mention properties beyond those in Section 4.1:

- Whether two mentions are nested
- Ancestry of each mention head: the dependency parent and grandparent POS tags and arc directions (shown in Figure 3)
- The speaker of each mention
- Number and gender of each mention as determined by Bergsma and Lin (2006)

The specific additional features we use are shown in Table 6. Note that unlike in Section 5, we use the number and gender information only on the antecedent. Due to our conjunction scheme, both this

[9] This observation is more general than just our system: the majority of coreference systems, including the winners of the CoNLL shared tasks (Lee et al., 2011; Fernandes et al., 2012), opt for high mention recall and resolve a relatively large number of system mentions.

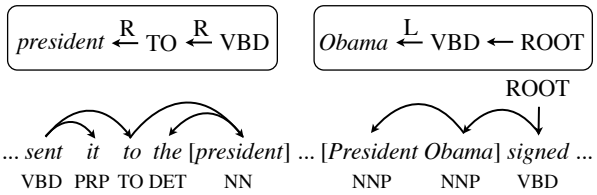

Figure 3: Demonstration of the ancestry extraction process. These features capture more sophisticated configurational information than our context word features do: in this example, *president* is in a characteristic indirect object position based on its dependency parents, and *Obama* is the subject of the main verb of the sentence.

Feature name	Count
Features of the SURFACE system	418704
Features on the current mention	
[ANAPHORIC] + [CURRENT ANCESTRY]	46047
Features on the antecedent	
[ANTECEDENT ANCESTRY]	53874
[ANTECEDENT GENDER]	338
[ANTECEDENT NUMBER]	290
Features on the pair	
[HEAD CONTAINED (T/F)]	136
[EXACT STRING CONTAINED (T/F)]	133
[NESTED (T/F)]	355
[DOC TYPE] + [SAME SPEAKER (T/F)]	437
[CURRENT ANCESTRY] + [ANT. ANCESTRY]	2555359

Table 6: FINAL feature set; note that this includes the SURFACE feature set. As with the features of the SURFACE system, two conjoined variants of each feature are included: first with the type of the current mention (NOMINAL, PROPER, or the citation form of the pronoun), then with the types of both mentions in the pair. These conjunctions allow antecedent features on gender and number to impact pronoun resolution, and they allow speaker match to capture effects like *I* and *you* being coreferent when the speakers differ.

semantic information and the speaker information can apply in a fine-grained way to different pronouns, and can therefore improve pronoun resolution substantially; however, these features generally only improve pronoun resolution.

Full results for our SURFACE and FINAL feature sets are shown in Table 7. Again, we compare to Lee et al. (2011) and Björkelund and Farkas (2012).[10] Despite our system's emphasis on one-pass resolution with as simple a feature set as possible, we are able to outperform even these sophisticated systems by a wide margin.

7 Related Work

Many of the individual features we employ in the FINAL feature set have appeared in other coreference systems (Björkelund and Nugues, 2011; Rahman and Ng, 2011b; Fernandes et al., 2012). However, other authors have often emphasized bilexical features on head pairs, whereas our features are heavily monolexical. For feature conjunctions, other authors have exploited three classes (Lee et al., 2011) or automatically learned conjunction schemes (Fernandes et al., 2012; Lassalle and Denis, 2013), but to our knowledge we are the first to do fine-grained modeling of every pronoun. Inclusion of a hierarchy of features with regularization also means that we organically get distinctions among different mention types without having to choose a level of granularity a priori, unlike the distinct classifiers employed by Denis and Baldridge (2008).

In terms of architecture, many coreference systems operate in a pipelined fashion, making partial decisions about coreference or pruning arcs before full resolution. Some systems use separate rule-based and learning-based passes (Chen and Ng, 2012; Fernandes et al., 2012), a series of learning-based passes (Björkelund and Farkas, 2012), or referentiality classifiers that prune the set of mentions before resolution (Rahman and Ng, 2009; Björkelund and Farkas, 2012; Recasens et al., 2013b). By contrast, our system resolves all mentions in one pass and does not need pruning: the SURFACE system can train in less than two hours without any subsampling of coreference arcs, and rule-based pruning of coreference arcs actually causes our system to perform less well, since our features can learn valuable information from these negative examples.

[10] Discrepancies between scores here and those printed in Pradhan et al. (2012) arise from two sources: improvements to the system of Lee et al. (2011) since the first CoNLL shared task, and a fix to the scoring of B^3 in the official scorer since results of the two CoNLL shared tasks were released. Unfortunately, because of this bug in the scoring program, direct comparison to the printed results of the other highest-scoring English systems, Fernandes et al. (2012) and Martschat et al. (2012), is impossible.

	MUC			B^3			CEAF_e			Avg.
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1
	CoNLL 2011 Development Set									
STANFORD	61.62	59.34	60.46	74.05	58.70	65.48	45.98	48.22	47.07	57.67
IMS	66.67	58.20	62.15	77.60	56.77	65.57	42.92	51.11	46.66	58.13
SURFACE*	68.42	60.80	64.39	76.57	59.21	66.78	45.30	53.36	49.00	60.06
FINAL*	68.97	63.47	**66.10**	76.58	62.06	**68.56**	47.32	53.19	**50.09**	**61.58**
	CoNLL 2011 Test Set									
STANFORD	60.91	62.13	61.51	70.61	57.31	63.27	45.79	44.56	45.17	56.65
IMS	68.15	61.60	64.71	75.97	56.39	64.73	42.30	48.88	45.35	58.26
FINAL*	66.81	66.04	**66.43**	71.07	61.89	**66.16**	47.37	48.22	**47.79**	**60.13**

Table 7: CoNLL metric scores for our systems on the CoNLL development and blind test sets, compared to the results of Lee et al. (2011) (STANFORD) and Björkelund and Farkas (2012) (IMS). Starred systems are contributions of this work. Bolded F_1 values represent statistically significant improvements over other systems with $p < 0.05$ using a bootstrap resampling test. Metric values reflect version 5 of the CoNLL scorer.

8 Conclusion

We have presented a coreference system that uses a simple, homogeneous set of features in a discriminative learning framework to achieve high performance. Large numbers of lexicalized, data-driven features implicitly model linguistic phenomena such as definiteness and centering, obviating the need for heuristic-driven rules explicitly targeting these same phenomena. Additional semantic features give only slight benefit beyond head match because they do not provide strong enough signals of coreference to improve performance in the system mention setting; modeling semantic similarity still requires complex outside information and deep heuristics.

Our system, the Berkeley Coreference Resolution System, is publicly available at `http://nlp.cs.berkeley.edu`.

Acknowledgments

This work was partially supported by BBN under DARPA contract HR0011-12-C-0014 and by an NSF fellowship for the first author. Thanks to Sameer Pradhan for helpful discussions regarding the CoNLL scoring program, and thanks to Leila Zilles and the anonymous reviewers for their insightful comments.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for Scoring Coreference Chains. In *Proceedings of the Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*.

Mohit Bansal and Dan Klein. 2012. Coreference Semantics from Web Features. In *Proceedings of the Association for Computational Linguistics*.

Eric Bengtson and Dan Roth. 2008. Understanding the Value of Features for Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping Path-Based Pronoun Resolution. In *Proceedings of the Conference on Computational Linguistics and the Association for Computational Linguistics*.

Anders Björkelund and Richárd Farkas. 2012. Data-driven Multilingual Coreference Resolution using Resolver Stacking. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning - Shared Task*.

Anders Björkelund and Pierre Nugues. 2011. Exploring Lexicalized Features for Coreference Resolution. In *Proceedings of the Conference on Computational Natural Language Learning: Shared Task*.

Jie Cai and Michael Strube. 2010. Evaluation Metrics for End-to-End Coreference Resolution Systems. In *Proceedings of the Special Interest Group on Discourse and Dialogue*.

Chen Chen and Vincent Ng. 2012. Combining the Best of Two Worlds: A Hybrid Approach to Multilingual Coreference Resolution. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning - Shared Task*.

Pascal Denis and Jason Baldridge. 2008. Specialized Models and Ranking for Coreference Resolution. In

Proceedings of the Conference on Empirical Methods in Natural Language Processing.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159, July.

Greg Durrett, David Hall, and Dan Klein. 2013. Decentralized Entity-Level Modeling for Coreference Resolution. In *Proceedings of the Association for Computational Linguistics*.

Eraldo Rezende Fernandes, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú. 2012. Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Proceedings and Conference on Computational Natural Language Learning - Shared Task*.

Kevin Gimpel and Noah A. Smith. 2010. Softmax-Margin CRFs: Training Log-Linear Models with Cost Functions. In *Proceedings of the North American Chapter for the Association for Computational Linguistics*.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2007. English Gigaword Third Edition. Linguistic Data Consortium, Catalog Number LDC2007T07.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225, June.

Aria Haghighi and Dan Klein. 2009. Simple Coreference Resolution with Rich Syntactic and Semantic Features. In *Proceedings of Empirical Methods in Natural Language Processing*.

Aria Haghighi and Dan Klein. 2010. Coreference Resolution in a Modular, Entity-Centered Model. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Iris Hendrickx and Walter Daelemans, 2007. *Adding Semantic Information: Unsupervised Clusters for Coreference Resolution*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Short Papers*.

Emmanuel Lassalle and Pascal Denis. 2013. Improving Pairwise Coreference Models Through Feature Space Hierarchy Learning. In *Proceedings of the Association for Computational Linguistics*.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the Conference on Computational Natural Language Learning: Shared Task*.

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 Association for Computational Linguistics*.

Xiaoqiang Luo. 2005. On Coreference Resolution Performance Metrics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Sebastian Martschat, Jie Cai, Samuel Broscheit, Éva Mújdricza-Maydt, and Michael Strube. 2012. A Multigraph Model for Coreference Resolution. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Proceedings and Conference on Computational Natural Language Learning - Shared Task*.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution. In *Proceedings of the North American Chapter of the Association of Computational Linguistics*.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the Conference on Computational Natural Language Learning: Shared Task*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*.

Altaf Rahman and Vincent Ng. 2009. Supervised Models for Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Altaf Rahman and Vincent Ng. 2011a. Coreference Resolution with World Knowledge. In *Proceedings of the Association for Computational Linguistics: Human Language Technologies*.

Altaf Rahman and Vincent Ng. 2011b. Narrowing the Modeling Gap: A Cluster-Ranking Approach to Coreference Resolution. *Journal of Artificial Intelligence Research*, 40(1):469–521, January.

Marta Recasens, Matthew Can, and Daniel Jurafsky. 2013a. Same Referent, Different Words: Unsupervised Mining of Opaque Coreferent Mentions. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013b. The Life and Death of Discourse Entities: Identifying Singleton Mentions. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544, December.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art. In *Proceedings of the Association for Computational Linguistics*.

Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. 2010. Coreference Resolution with Reconcile. In *Proceedings of the Association for Computational Linguistics: Short Papers*.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A Modular Toolkit for Coreference Resolution. In *Proceedings of the Association for Computational Linguistics: Demo Session*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the Conference on Message Understanding*.