Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules

Heeyoung Lee*
Stanford University

Angel Chang*
Stanford University

Yves Peirsman**
University of Leuven

Nathanael Chambers†
United States Naval Academy

Mihai Surdeanu‡
University of Arizona

Dan Jurafsky§
Stanford University

We propose a new deterministic approach to coreference resolution that combines the global information and precise features of modern machine-learning models with the transparency and modularity of deterministic, rule-based systems. Our sieve architecture applies a battery of deterministic coreference models one at a time from highest to lowest precision, where each model builds on the previous model's cluster output. The two stages of our sieve-based architecture, a mention detection stage that heavily favors recall, followed by coreference sieves that are precision-oriented, offer a powerful way to achieve both high precision and high recall. Further, our approach makes use of global information through an entity-centric model that encourages the sharing of features across all mentions that point to the same real-world entity. Despite its simplicity, our approach gives state-of-the-art performance on several corpora and genres, and has also been incorporated into hybrid state-of-the-art coreference systems for Chinese and

* Stanford University, 450 Serra Mall, Stanford, CA 94305. E-mail: heeyoung@stanford.edu, angelx@cs.stanford.edu.
** University of Leuven, Blijde-Inkomststraat 21 PO Box 03308, B-3000 Leuven, Belgium. E-mail: yves.peirsman@arts.kuleuven.be.
† United States Naval Academy, 121 Blake Road, Annapolis, MD 21402. E-mail: nchamber@usna.edu.
‡ University of Arizona, PO Box 210077, Tucson, AZ 85721-0077. E-mail: msurdeanu@email.arizona.edu.
§ Stanford University, 450 Serra Mall, Stanford, CA 94305. E-mail: jurafsky@stanford.edu.

Submission received: 27 May 2012; revised submission received: 22 October 2012; accepted for publication: 20 November 2012.

doi:10.1162/COLI_a_00152

Arabic. Our system thus offers a new paradigm for combining knowledge in rule-based systems that has implications throughout computational linguistics.

1. Introduction

Coreference resolution, the task of finding all expressions that refer to the same entity in a discourse, is important for natural language understanding tasks like summarization, question answering, and information extraction.

The long history of coreference resolution has shown that the use of highly precise lexical and syntactic features is crucial to high quality resolution (Ng and Cardie 2002b; Lappin and Leass 1994; Poesio et al. 2004a; Zhou and Su 2004; Bengtson and Roth 2008; Haghighi and Klein 2009). Recent work has also shown the importance of global inference—performing coreference resolution jointly for several or all mentions in a document—rather than greedily disambiguating individual pairs of mentions (Morton 2000; Luo et al. 2004; Yang et al. 2004; Culotta et al. 2007; Yang et al. 2008; Poon and Domingos 2008; Denis and Baldridge 2009; Rahman and Ng 2009; Haghighi and Klein 2010; Cai, Mujdricza-Maydt, and Strube 2011).

Modern systems have met this need for carefully designed features and global or entity-centric inference with machine learning approaches to coreference resolution. But machine learning, although powerful, has limitations. Supervised machine learning systems rely on expensive hand-labeled data sets and generalize poorly to new words or domains. Unsupervised systems are increasingly more complex, making them hard to tune and difficult to apply to new problems and genres as well. Rule-based models like Lappin and Leass (1994) were a popular early solution to the subtask of pronominal anaphora resolution. Rules are easy to create and maintain and error analysis is more transparent. But early rule-based systems relied on hand-tuned weights and were not capable of global inference, two factors that led to poor performance and replacement by machine learning.

We propose a new approach that brings together the insights of these modern supervised and unsupervised models with the advantages of deterministic, rule-based systems. We introduce a model that performs entity-centric coreference, where all mentions that point to the same real-world entity are jointly modeled, in a rich feature space using solely simple, deterministic rules. Our work is inspired both by the seminal early work of Baldwin (1997), who first proposed that a series of high-precision rules could be used to build a high-precision, low-recall system for anaphora resolution, and by more recent work that has suggested that deterministic rules can outperform machine learning models for coreference (Zhou and Su 2004; Haghighi and Klein 2009) and for named entity recognition (Chiticariu et al. 2010).

Figure 1 illustrates the two main stages of our new deterministic model: mention detection and coreference resolution, as well as a smaller post-processing step. In the mention detection stage, nominal and pronominal mentions are identified using a high-recall algorithm that selects all noun phrases (NPs), pronouns, and named entity mentions, and then filters out non-mentions (pleonastic *it*, i-within-i, numeric entities, partitives, etc.).

The coreference resolution stage is based on a succession of ten independent coreference models (or "sieves"), applied from highest to lowest precision. Precision can be informed by linguistic intuition, or empirically determined on a coreference corpus (see Section 4.4.3). For example, the first (highest precision) sieve links first-person pronouns inside a quotation with the speaker of a quotation, and the tenth sieve (i.e., low precision but high recall) implements generic pronominal coreference resolution.

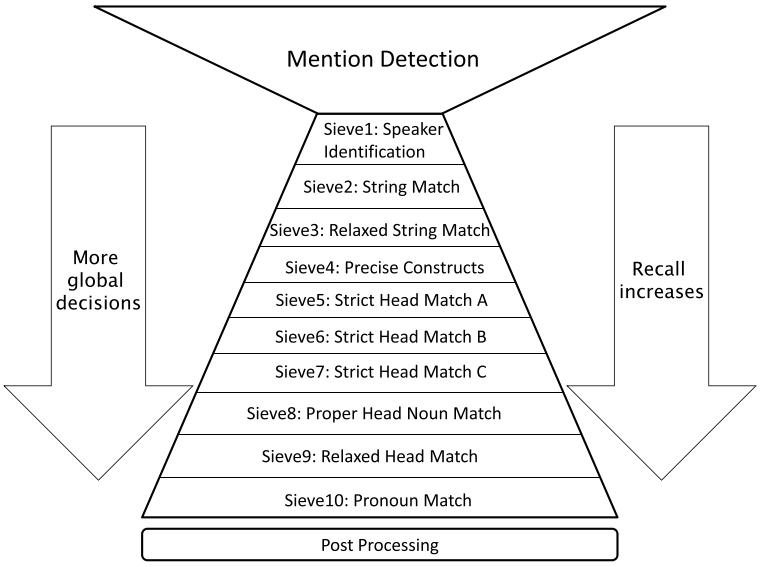

Figure 1
The architecture of our coreference system.

Crucially, our approach is entity-centric—that is, our architecture allows each coreference decision to be globally informed by the previously clustered mentions and their shared attributes. In particular, each deterministic rule is run on the entire discourse, using and extending clusters (i.e., groups of mentions pointing to the same real-world entity, built by models in previous tiers). Thus, for example, in deciding whether two mentions i and j should corefer, our system can consider not just the local features of i and j but also any information (head word, named entity type, gender, or number) about the other mentions already linked to i and j in previous steps.

Finally, the architecture is highly modular, which means that additional coreference resolution models can be easily integrated.

The two stage architecture offers a powerful way to balance both high recall and precision in the system and make use of entity-level information with rule-based architecture. The mention detection stage heavily favors recall, and the following sieves favor precision. Our results here and in our earlier papers (Raghunathan et al. 2010; Lee et al. 2011) show that this design leads to state-of-the-art performance despite the simplicity of the individual components, and that the lack of language-specific lexical features makes the system easy to port to other languages. The intuition is not new; in addition to the prior coreference work mentioned earlier and discussed in Section 6, we draw on classic ideas that have proved to be important again and again in the history of natural language processing. The idea of beginning with the most accurate models or starting with smaller subproblems that allow for high-precision solutions combines the intuitions of "shaping" or "successive approximations" first proposed for learning by Skinner (1938), and widely used in NLP (e.g., the successively trained IBM MT models of Brown et al. [1993]) and the "islands of reliability" approaches to parsing and speech recognition [Borghesi and Favareto 1982; Corazza et al. 1991]). The idea of beginning with a high-recall list of candidates that are followed by a series of high-precision filters dates back to one of the earliest architectures in natural language processing, the part of speech tagging algorithm of the Computational Grammar Coder (Klein and Simmons

1963) and the TAGGIT tagger (Greene and Rubin 1971), which begin with a high-recall list of all possible tags for words, and then used high-precision rules to filter likely tags based on context.

In the next section we walk through an example of our system applied to a simple made-up text. We then describe our model in detail and test its performance on three different corpora widely used in previous work for the evaluation of coreference resolution. We show that our model outperforms the state-of-the-art on each corpus. Furthermore, in these sections we describe analytic and ablative experiments demonstrating that both aspects of our algorithm (the entity-centric aspect that allows the global sharing of features between mentions assigned to the same cluster and the precision-based ordering of sieves) independently offer significant improvements to coreference, perform an error analysis, and discuss the relationship of our work to previous models and to recent hybrid systems that have used our algorithm as a component to resolve coreference in English, Chinese, and Arabic.

2. Walking Through a Sample Coreference Resolution

Before delving into the details of our method, we illustrate the intuition behind our approach with the simple pedagogical example listed in Table 1.

In the mention detection step, the system extracts mentions by inspecting all noun phrases (NP) and other modifier pronouns (PRP) (see Section 3.1 for details). In Table 1, this step identifies 11 different mentions and assigns them initially to distinct entities (Entity id and mention id in each step are marked by superscript and subscript). This component also extracts mention attributes—for example, *John*:{`ne:person`}, and *A girl*:{`gender:female, number:singular`}. These mentions form the input for the following sequence of sieves.

The first coreference resolution sieve (the speaker or quotation sieve) matches pronominal mentions that appear in a quotation block to the corresponding speaker. In general, in all the coreference resolution sieves we traverse mentions left-to-right in a given document (see Section 3.2.1). The first match for this model is my_9^9, which is merged with $John_{10}^{10}$ into the same entity (entity id: 9). This illustrates the advantages of our incremental approach: by assigning a higher priority to the quotation sieve, we avoid linking my_9^9 with $A\ girl_5^5$, a common mistake made by generic coreference models, since anaphoric candidates (especially in subject position) are generally preferred to cataphoric ones (Hobbs 1978).

The next sieve searches for anaphoric antecedents that have the exact same string as the mention under consideration. This component resolves the tenth mention, $John_{10}^9$, by linking it with $John_1^1$. When searching for antecedents, we sort candidates in the same sentential clause from left to right, and we prefer sentences that are closer to the mention under consideration (see Section 3.2.2 for details). Thus, the sorted list of candidates for $John_{10}^9$ is It_7^7, $My\ favorite_8^8$, My_9^9, $A\ girl_5^5$, $the\ song_6^6$, He_3^3, $a\ new\ song_4^4$, $John_1^1$, $a\ musician_2^2$. The algorithm stops as soon as a matching antecedent is encountered. In this case, the algorithm finds $John_1^1$ and does not inspect $a\ musician_2^2$.

The relaxed string match sieve searches for mentions satisfying a looser set of string matching constraints than exact match (details in Section 3.3.3), but makes no change because there are no such mentions. The precise constructs sieve searches for several high-precision syntactic constructs, such as appositive relations and predicate nominatives. In this example, there are two predicate nominative relations in the first and fourth sentences, so this component clusters together $John_1^1$ and $a\ musician_2^2$, and It_7^7 and $my\ favorite_8^8$.

Table 1
A sample run-through of our approach, applied to a made-up sentence. In each step we mark in **bold** the affected mentions; superscript and subscript indicate entity id and mention id.

Input:	John is a musician. He played a new song. A girl was listening to the song. "It is my favorite," John said to her.
Mention Detection:	$[\text{John}]_1^1$ is $[\text{a musician}]_2^2$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$. $[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^6$. "$[\text{It}]_7^7$ is $[[\text{my}]_9^9$ favorite$]_8^8$," $[\text{John}]_{10}^{10}$ said to $[\text{her}]_{11}^{11}$.
Speaker Sieve:	$[\text{John}]_1^1$ is $[\text{a musician}]_2^2$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$. $[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^6$. "$[\text{It}]_7^7$ is $[[\textbf{my}]_9^9$ favorite$]_8^8$," $[\textbf{John}]_{10}^9$ said to $[\text{her}]_{11}^{11}$.
String Match:	$[\textbf{John}]_1^1$ is $[\text{a musician}]_2^2$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$. $[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^6$. "$[\text{It}]_7^7$ is $[[\text{my}]_9^1$ favorite$]_8^8$," $[\textbf{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.
Relaxed String Match:	$[\text{John}]_1^1$ is $[\text{a musician}]_2^2$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$. $[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^6$. "$[\text{It}]_7^7$ is $[[\text{my}]_9^1$ favorite$]_8^8$," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.
Precise Constructs:	$[\textbf{John}]_1^1$ is $[\textbf{a musician}]_2^1$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$. $[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^6$. "$[\textbf{It}]_7^7$ is $[[\textbf{my}]_9^1$ **favorite**$]_8^7$," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.
Strict Head Match A:	$[\text{John}]_1^1$ is $[\text{a musician}]_2^1$. $[\text{He}]_3^3$ played $[\textbf{a new song}]_4^4$. $[\text{A girl}]_5^5$ was listening to $[\textbf{the song}]_6^4$. "$[\text{It}]_7^7$ is $[[\text{my}]_9^1$ favorite$]_8^7$," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.
Strict Head Match B,C:	$[\text{John}]_1^1$ is $[\text{a musician}]_2^1$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$. $[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^4$. "$[\text{It}]_7^7$ is $[[\text{my}]_9^1$ favorite$]_8^7$," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.
Proper Head Noun Match:	$[\text{John}]_1^1$ is $[\text{a musician}]_2^1$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$. $[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^4$. "$[\text{It}]_7^7$ is $[[\text{my}]_9^1$ favorite$]_8^7$," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.
Relaxed Head Match:	$[\text{John}]_1^1$ is $[\text{a musician}]_2^1$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$. $[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^4$. "$[\text{It}]_7^7$ is $[[\text{my}]_9^1$ favorite$]_8^7$," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.
Pronoun Match:	$[\textbf{John}]_1^1$ is $[\text{a musician}]_2^1$. $[\textbf{He}]_3^1$ played $[\text{a new song}]_4^4$. $[\textbf{A girl}]_5^5$ was listening to $[\textbf{the song}]_6^4$. "$[\textbf{It}]_7^4$ is $[[\text{my}]_9^1$ favorite$]_8^4$," $[\text{John}]_{10}^1$ said to $[\textbf{her}]_{11}^5$.
Post Processing:	$[\text{John}]_1^1$ is **a musician**. $[\text{He}]_3^1$ played $[\text{a new song}]_4^4$. $[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^4$. "$[\text{It}]_7^4$ is $[\textbf{my}]_9^1$ **favorite**," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^5$.
Final Output:	$[\text{John}]_1^1$ is a musician. $[\text{He}]_3^1$ played $[\text{a new song}]_4^4$. $[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^4$. "$[\text{It}]_7^4$ is $[\text{my}]_9^1$ favorite," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^5$.

The next four sieves (strict head match A–C, proper head noun match) cluster mentions that have the same head word with various other constraints. *a new song*$_4^4$ and *the song*$_6^6$ are linked in this step.

The last resolution component in this example addresses pronominal coreference resolution. The three pronouns in this text, *He*$_3^3$, *It*$_7^7$, and *her*$_{11}^{11}$ are linked to their

compatible antecedents based on their attributes, such as gender, number, and animacy. In this step we assign He_3^3 and her_{11}^{11} to entities 1 and 5, respectively (same gender), and It_7^7 to entity 4, which represents an inanimate concept.

The system concludes with a post-processing component, which implements corpus-specific rules. For example, to align our output with the OntoNotes annotation standard, we remove mentions assigned to singleton clusters (i.e., entities with a single mention in text) and links obtained through predicate nominative patterns. Note that even though we might remove some coreference links in this step, these links serve an important purpose in the algorithm flow, as they allow new features to be discovered for the corresponding entity and shared between its mentions. See Section 3.2.3 for details on feature extraction.

3. The Algorithm

We first describe our mention detection stage, then introduce the general architecture of the coreference stage, followed by a detailed examination of the coreference sieves. In describing the architecture, we will sometimes find it helpful to discuss the precision of individual components, drawn from our later experiments in Section 4.

3.1 Mention Detection

As we suggested earlier, the recall of our mention detection component is more important than its precision. This is because for the OntoNotes corpus and for many practical applications, any missed mentions are guaranteed to affect the final score by decreasing recall, whereas spurious mentions may not impact the overall score if they are assigned to singleton clusters, because singletons are deleted during post-processing. Our mention detection algorithm implements this intuition via a series of simple yet broad-coverage heuristics that take advantage of syntax, named entity recognition and manually written patterns. Note that those patterns are built based on the OntoNotes annotation guideline because mention detection in general depends heavily on the annotation policy.

We start by marking all NPs, pronouns, and named entity mentions (see the named entity tagset in Appendix A) that were not previously marked (i.e., they appear as modifiers in other NPs) as candidate mentions. From this set of candidates we remove the mentions that match any of the following exclusion rules:

1. We remove a mention if a larger mention with the same head word exists (e.g., we remove *The five insurance companies* in *The five insurance companies approved to be established this time*).
2. We discard numeric entities such as percents, money, cardinals, and quantities (e.g., 9%, \$10,000, *Tens of thousands, 100 miles*).
3. We remove mentions with partitive or quantifier expressions (e.g., *a total of 177 projects, none of them, millions of people*).[1]

1 These are NPs with the word 'of' preceded by one of nine quantifiers or 34 partitives.

4. We remove pleonastic *it* pronouns, detected using a small set of patterns (e.g., *It is possible that . . .*, *It seems that . . .*, *It turns out . . .*). The complete set of patterns, using the tregex[2] notation, is shown in Appendix B.
5. We discard adjectival forms of nations or nationality acronyms (e.g., *American, U.S., U.K.*), following the OntoNotes annotation guidelines.
6. We remove stop words from the following list determined by error analysis on mention detection: *there, ltd., etc, 's, hmm*.

Note that some rules change depending on the corpus we use for evaluation. In particular, adjectival forms of nations are valid mentions in the Automated Content Extraction (ACE) corpus (Doddington et al. 2004), thus they would not be removed when processing this corpus.

3.2 Resolution Architecture

Traditionally, coreference resolution is implemented as a quadratic problem, where potential coreference links between any two mentions in a document are considered. This is not ideal, however, as it increases both the likelihood of errors and the processing time. In this article, we argue that it is better to cautiously construct high-quality mention clusters,[3] and use an entity-centric model that allows the sharing of information across these incrementally constructed clusters. We achieve these goals by: (a) aggressively filtering the search space for which mention to consider for resolution (Section 3.2.1) and which antecedents to consider for a given mention (Section 3.2.2), and (b) constructing features from partially built mention clusters (Section 3.2.3).

3.2.1 Mention Selection in a Given Sieve. Recall that our model is a battery of resolution sieves applied sequentially. Thus, in each given sieve, we have partial mention clusters produced by the previous model. We exploit this information for mention selection, by considering only mentions that are currently first in textual order in their cluster. For example, given the following ordered list of mentions, $\{m_1^1, m_2^2, m_3^2, m_4^3, m_5^1, m_6^2\}$, where the superscript indicates cluster id, our model will attempt to resolve only m_2^2 and m_4^3 (m_1^1 is not resolved because it is the first mention in a text). These two are the only mentions that currently appear first in their respective clusters and have potential antecedents in the document. The motivation behind this heuristic is two-fold. First, early mentions are usually better defined than subsequent ones, which are likely to have fewer modifiers or be pronouns (Fox 1993). Because several of our models use features extracted from NP modifiers, it is important to prioritize mentions that include such information. Second, by definition, first mentions appear closer to the beginning of the document, hence there are fewer antecedent candidates to select from, and thus fewer opportunities to make a mistake.

We further prune the search space using a simple model of *discourse salience*. We disable coreference for mentions appearing first in their corresponding clusters that: (a) are or start with indefinite pronouns (e.g., *some*, *other*), (b) start with indefinite articles

2 `http://nlp.stanford.edu/software/tregex.shtml`.

3 In this article we use the terms *mention cluster* and *entity* interchangeably. We prefer the former when discussing technical aspects of our approach and the latter in a more theoretical context.

(e.g., *a*, *an*), or (c) are bare plurals. One exception to (a) and (b) is the model deployed in the Exact String Match sieve, which only links mentions if their entire extents match exactly (see Section 3.3.2). This model is triggered for all nominal mentions regardless of discourse salience, because it is possible that indefinite mentions are repeated in a document when concepts are discussed but not instantiated, e.g., *a sports bar* in the following:

> *Hanlon, a longtime Broncos fan, thinks it is the perfect place for* **a sports bar** *and has put up a blue-and-orange sign reading, "Wanted Broncos Sports Bar On This Site." . . . In a Nov. 28 letter, Proper states "while we have no objection to your advertising the property as a location for* **a sports bar**, *using the Broncos' name and colors gives the false impression that the bar is or can be affiliated with the Broncos."*

3.2.2 Antecedent Selection for a Given Mention. Given a mention m_i, each model may either decline to propose a solution (in the hope that one of the subsequent models will solve it) or deterministically select a single best antecedent from a list of previous mentions $m_1, \ldots, m_{i-1}$. We sort candidate antecedents using syntactic information provided by the Stanford parser. Candidates are sorted using the following criteria:

- In a given sentential clause (i.e., parser constituents whose label starts with S), candidates are sorted using a left-to-right breadth-first traversal of the corresponding syntactic constituent (Hobbs 1978). Figure 2 shows an example of candidate ordering based on this traversal. The left-to-right ordering favors subjects, which tend to appear closer to the beginning of the sentence and are more probable antecedents. The breadth-first traversal promotes syntactic salience by preferring noun phrases that are closer to the top of the parse tree (Haghighi and Klein 2009).
- If the sentence containing the anaphoric mention contains multiple clauses, we repeat the previous heuristic separately in each S∗ constituent, starting with the one containing the mention.

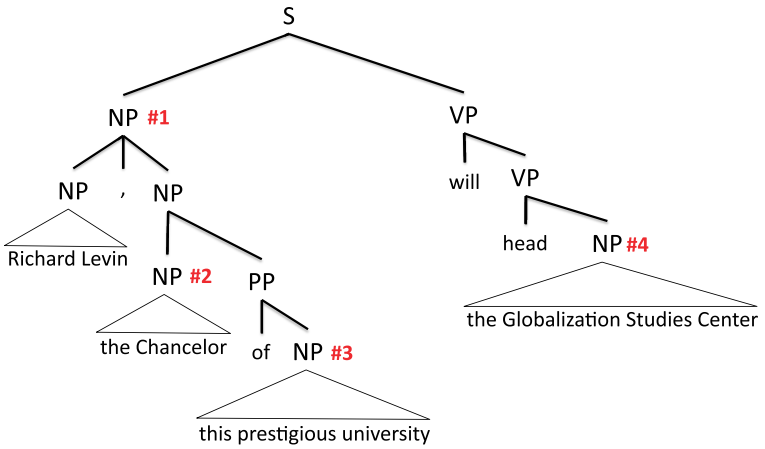

Figure 2
Example of left-to-right breadth-first tree traversal. The numbers indicate the order in which the NPs are visited.

- Clauses in previous sentences are sorted based on their textual proximity to the anaphoric mention.

The sorting of antecedent candidates is important because our algorithm stops at the first match. Thus, low-quality sorting negatively impacts the actual coreference links created.

This antecedent selection algorithm applies to all the coreference resolution sieves described in this article, with the exception of the speaker identification sieve (Section 3.3.1) and the sieve that applies appositive and predicate nominative patterns (Section 3.3.4).

3.2.3 Feature Sharing in the Entity-Centric Model. In a significant departure from previous work, each model in our framework gets (possibly incomplete) entity information for each mention from the clusters constructed by the earlier coreference models. In other words, each mention m_i may already be assigned to an entity E_j containing a set of mentions: $E_j = \{m_1^j, \ldots, m_k^j\}$; $m_i \in E_j$. Unassigned mentions are unique members of their own cluster. We use this information to share information between same-entity mentions.

This is especially important for pronominal coreference resolution (discussed later in this section), which can be severely affected by missing attributes (which introduce precision errors because incorrect antecedents are selected due to missing information) and incorrect attributes (which introduce recall errors because correct links are not generated due to attribute mismatch between mention and antecedent). To address this issue, we perform a union of all mention attributes (e.g., number, gender, animacy) for a given entity and share the result with all corresponding mentions. If attributes from different mentions contradict each other we maintain all variants. For example, our naive number detection assigns `singular` to the mention *a group of students* and `plural` to *five students*. When these mentions end up in the same cluster, the resulting number attributes becomes the set {`singular`, `plural`}. Thus this cluster can later be merged with both singular and plural pronouns.

3.3 Coreference Resolution Sieves

We describe next the sequence of coreference models proposed in this article. Table 2 lists all these models in the order in which they are applied. We discuss their individual contribution to the overall system later, in Section 4.4.3.

Table 2
Sequence of sieves as they are applied in the overall model.

Sequence	Model Name
Pass 1	Speaker Identification Sieve
Pass 2	Exact String Match Sieve
Pass 3	Relaxed String Match Sieve
Pass 4	Precise Constructs Sieve (e.g., appositives)
Passes 5–7	Strict Head Match Sieves A–C
Pass 8	Proper Head Noun Match Sieve
Pass 9	Relaxed Head Match Sieve
Pass 10	Pronoun Resolution Sieve

3.3.1 Pass 1 – Speaker Identification. This sieve matches speakers to compatible pronouns, using shallow discourse understanding to handle quotations and conversation transcripts, following the early work of Baldwin (1995, 1997). We begin by identifying *speakers* within text. In non-conversational text, we use a simple heuristic that searches for the subjects of reporting verbs (e.g., *say*) in the same sentence or neighboring sentences to a quotation. In conversational text, speaker information is provided in the data set.

The extracted speakers then allow us to implement the following sieve heuristics:

- ⟨I⟩s[4] assigned to the same speaker are coreferent.
- ⟨you⟩s with the same speaker are coreferent.
- The speaker and ⟨I⟩s in her text are coreferent.

Thus for example *I*, *my*, and *she* in the following sentence are coreferent: *"[I] voted for [Nader] because [he] was most aligned with [my] values," [she] said.*

In addition to this sieve, we impose speaker constraints on decisions made by subsequent sieves:

- The speaker and a mention which is not ⟨I⟩ in the speaker's utterance cannot be coreferent.
- Two ⟨I⟩s (or two ⟨you⟩s, or two ⟨we⟩s) assigned to different speakers cannot be coreferent.
- Two different person pronouns by the same speaker cannot be coreferent.
- Nominal mentions cannot be coreferent with ⟨I⟩, ⟨you⟩, or ⟨we⟩ in the same turn or quotation.
- In conversations, ⟨you⟩ can corefer only with the previous speaker.

The constraints result in causing *[my]* and *[he]* to not be coreferent in the earlier example (due to the third constraint).

3.3.2 Pass 2 – Exact Match. This model links two mentions only if they contain exactly the same extent text, including modifiers and determiners (e.g., *[the Shahab 3 ground-ground missile]* and *[the Shahab 3 ground-ground missile]*). As expected, this model is very precise, with a precision over 90% B^3 (see Table 8 in Section 4.4.3).

3.3.3 Pass 3 – Relaxed String Match. This sieve considers two nominal mentions as coreferent if the strings obtained by dropping the text following their head words (such as relative clauses and PP and participial postmodifiers) are identical (e.g., *[Clinton]* and *[Clinton, whose term ends in January]*).

3.3.4 Pass 4 – Precise Constructs. This model links two mentions if any of the following conditions are satisfied:

- **Appositive** – the two nominal mentions are in an appositive construction (e.g., *[Israel's Deputy Defense Minister], [Ephraim Sneh] , said . . .*). We use the standard Haghighi and Klein (2009) definition to detect appositives: third children of a parent NP whose expansion begins with `(NP , NP)`, when there is not a conjunction in the expansion.

4 We define ⟨I⟩ as *I, my, me*, or *mine*, ⟨we⟩ as first person plural pronouns, and ⟨you⟩ as second person pronouns.

- **Predicate nominative** – the two mentions (nominal or pronominal) are in a copulative subject–object relation (e.g., *[The New York-based College Board] is [a nonprofit organization that administers the SATs and promotes higher education]* [Poon and Domingos 2008]).
- **Role appositive** – the candidate antecedent is headed by a noun and appears as a modifier in an NP whose head is the current mention (e.g., *[[actress] Rebecca Schaeffer]*). This feature is inspired by Haghighi and Klein (2009), who triggered it only if the mention is labeled as a person by the Stanford named entity recognizer (NER). We constrain this heuristic more in our work: We allow this feature to match only if: (a) the mention is labeled as a person, (b) the antecedent is animate (we detail animacy detection in Section 3.3.9), and (c) the antecedent's gender is not neutral.
- **Relative pronoun** – the mention is a relative pronoun that modifies the head of the antecedent NP (e.g., *[the finance street [which] has already formed in the Waitan district]*).
- **Acronym** – both mentions are tagged as NNP and one of them is an acronym of the other (e.g., *[Agence France Presse] ... [AFP]*). Our acronym detection algorithm marks a mention as an acronym of another if its text equals the sequence of upper case characters in the other mention. The algorithm is simple, but our error analysis suggests it nonetheless does not lead to errors.
- **Demonym**[5] – one of the mentions is a demonym of the other (e.g., *[Israel] ... [Israeli]*). For demonym detection we use a static list of countries and their gentilic forms from Wikipedia.[6]

All of these constructs are very precise; we show in Section 4.4.3 that the B^3 precision of the overall model after adding this sieve is approximately 90%. In the OntoNotes corpus, this sieve does not enhance recall significantly, mainly because appositions and predicate nominatives are not annotated in this corpus (they are annotated in ACE). Regardless of annotation standard, however, this sieve is important because it grows entities with high quality elements, which has a significant impact on the entity's features (as discussed in Section 3.2.3).

3.3.5 Pass 5 – Strict Head Match. Linking a mention to an antecedent based on the naive matching of their head words generates many spurious links because it completely ignores possibly incompatible modifiers (Elsner and Charniak 2010). For example, *Yale University* and *Harvard University* have similar head words, but they are obviously different entities. To address this issue, this pass implements several constraints that must all be matched in order to yield a link:

- **Entity head match** – the mention head word matches *any* head word of mentions in the antecedent entity. Note that this feature is actually more relaxed than naive head matching in a pair of mentions because here it is satisfied when the mention's head matches the head of any mention in the candidate entity. We constrain this feature by enforcing a conjunction with the following features.

5 Demonym is not annotated in OntoNotes but we keep it in the system.
6 http://en.wikipedia.org/wiki/List_of_adjectival_and_demonymic_forms_of_place_names.

- **Word inclusion** – all the non-stop[7] words in the current entity to be solved are included in the set of non-stop words in the antecedent entity. This heuristic exploits the discourse property that states that it is uncommon to introduce novel information in later mentions (Fox 1993). Typically, mentions of the same entity become shorter and less informative as the narrative progresses. For example, based on this constraint, the model correctly clusters together the two mentions in the following text:

  > *. . . intervene in the [Florida Supreme Court]'s move . . . does look like very dramatic change made by [the Florida court]*

  and avoids clustering the two mentions in the following text:

  > *The pilot had confirmed . . . he had turned onto [the correct runway] but pilots behind him say he turned onto [the wrong runway].*

- **Compatible modifiers only** – the mention's modifiers are all included in the modifiers of the antecedent candidate. This feature models the same discourse property as the previous feature, but it focuses on the two individual mentions to be linked, rather than their corresponding entities. For this feature we only use modifiers that are nouns or adjectives.
- **Not i-within-i** – the two mentions are not in an i-within-i construct, that is, one cannot be a child NP in the other's NP constituent (Chomsky 1981).

This pass continues to maintain high precision (over 86% B^3) while improving recall significantly (approximately 4.5 B^3 points).

3.3.6 Passes 6 and 7 – Variants of Strict Head Match. Sieves 6 and 7 are different relaxations of the feature conjunction introduced in Pass 5, that is, Pass 6 removes the `compatible modifiers only` feature, and Pass 7 removes the `word inclusion` constraint. All in all, these two passes yield an improvement of 0.9 B^3 F1 points, due to recall improvements. Table 8 in Section 4.4.3 shows that the `word inclusion` feature is more precise than `compatible modifiers only`, but the latter has better recall.

3.3.7 Pass 8 – Proper Head Word Match. This sieve marks two mentions headed by proper nouns as coreferent if they have the same head word and satisfy the following constraints:

- **Not i-within-i** - same as in Pass 5.
- **No location mismatches** - the modifiers of two mentions cannot contain different location named entities, other proper nouns, or spatial modifiers. For example, *[Lebanon]* and *[southern Lebanon]* are not coreferent.
- **No numeric mismatches** - the second mention cannot have a number that does not appear in the antecedent, e.g., *[people]* and *[around 200 people]* are not coreferent.

7 Our stopword list includes person titles as well.

3.3.8 Pass 9 – Relaxed Head Match. This pass relaxes the entity head match heuristic by allowing the mention head to match any word in the antecedent entity. For example, this heuristic matches the mention *Sanders* to an entity containing the mentions {*Sauls*, *the judge*, *Circuit Judge N. Sanders Sauls*}. To maintain high precision, this pass requires that both mention and antecedent be labeled as named entities and the types coincide. Furthermore, this pass implements a conjunction of the given features with `word inclusion` and `not i-within-i`. This pass yields less than 0.4 point improvement in most metrics.

3.3.9 Pass 10 – Pronominal Coreference Resolution. With one exception (Pass 1), all the previous coreference models focus on nominal coreference resolution. It would be incorrect to say that our framework ignores pronominal coreference in the previous passes, however. In fact, the previous models prepare the stage for pronominal coreference by constructing precise entities with shared mention attributes. These are crucial factors for pronominal coreference.

We implement pronominal coreference resolution using an approach standard for many decades: enforcing agreement constraints between the coreferent mentions. We use the following attributes for these constraints:

- **Number** – we assign number attributes based on: (a) a static list for pronouns; (b) NER labels: mentions marked as a named entity are considered singular with the exception of organizations, which can be both singular and plural; (c) part of speech tags: `NN*S` tags are plural and all other `NN*` tags are singular; and (d) a static dictionary from Bergsma and Lin (2006).
- **Gender** – we assign gender attributes from static lexicons from Bergsma and Lin (2006), and Ji and Lin (2009).
- **Person** – we assign person attributes only to pronouns. We do not enforce this constraint when linking two pronouns, however, if one appears within quotes. This is a simple heuristic for speaker detection (e.g., *I* and *she* point to the same person in *"[I] voted my conscience," [she] said*).
- **Animacy** – we set animacy attributes using: (a) a static list for pronouns; (b) NER labels (e.g., `PERSON` is animate whereas `LOCATION` is not); and (c) a dictionary bootstrapped from the Web (Ji and Lin 2009).
- **NER label** – from the Stanford NER.
- **Pronoun distance** - sentence distance between a pronoun and its antecedent cannot be larger than 3.

When we cannot extract an attribute, we set the corresponding value to `unknown` and treat it as a wildcard—that is, it can match any other value. As expected, pronominal coreference resolution has a big impact on the overall score (e.g., 5 B^3 F1 points in the development partition of OntoNotes).

3.4 Post Processing

This step implements several transformations required to guarantee that our output matches the annotation specification in the corresponding corpus. Currently this

step is deployed only for the OntoNotes corpus and it contains the following two operations:

- We discard singleton clusters.
- We discard the shorter mentions in appositive patterns and the mentions that appear later in text in copulative relations. For example, in the text *[[Yongkang Zhou] , the general manager]* or *[Mr. Savoca] had been [a consultant. . .]*, the mentions *Yongkang Zhou* and *a consultant. . .* are removed in this stage.

4. Experimental Results

We start this section with overall results on three corpora widely used for the evaluation of coreference resolution systems. We continue with a series of ablative experiments that analyze the contribution of each aspect of our approach and conclude with error analysis, which highlights cases currently not solved by our approach.

4.1 Corpora

We used the following corpora for development and formal evaluation:

- **OntoNotes-Dev** – development partition of OntoNotes v4.0 provided in the CoNLL2011 shared task (Pradhan et al. 2011).
- **OntoNotes-Test** – test partition of OntoNotes v4.0 provided in the CoNLL-2011 shared task.
- **ACE2004-Culotta-Test** – partition of the ACE 2004 corpus reserved for testing by several previous studies (Culotta et al. 2007; Bengtson and Roth 2008; Haghighi and Klein 2009).
- **ACE2004-nwire** – newswire subset of the ACE 2004 corpus, utilized by Poon and Domingos (2008) and Haghighi and Klein (2009) for testing.
- **MUC6-Test** – test corpus from the sixth Message Understanding Conference (MUC-6) evaluation.

The corpora statistics are shown in Table 3. We used the first corpus (OntoNotes-Dev) for development and all others for the formal evaluation. We parsed all documents in the ACE and MUC corpora using the Stanford parser (Klein and Manning 2003) and the Stanford NER (Finkel, Grenager, and Manning 2005). We used the provided parse

Table 3
Corpora statistics.

Corpora	# Documents	# Sentences	# Words	# Entities	# Mentions
OntoNotes-Dev	303	6,894	136K	3,752	14,291
OntoNotes-Test	322	8,262	142K	3,926	16,291
ACE2004-Culotta-Test	107	1,993	33K	2,576	5,455
ACE2004-nwire	128	3,594	74K	4,762	11,398
MUC6-Test	30	576	13K	496	2,136

trees and named entity labels (not gold) in the OntoNotes corpora to facilitate the comparison with other systems.

4.2 Evaluation Metrics

We use five evaluation metrics widely used in the literature. B^3 and CEAF have implementation variations in how to take system mentions into account. We followed the same implementation as used in CoNLL-2011 shared task.

- **MUC** (Vilain et al. 1995) – link-based metric which measures how many predicted and gold mention clusters need to be merged to cover the gold and predicted clusters, respectively.
  $R = \frac{\sum(|G_i| - |p(G_i)|)}{\sum(|G_i|-1)}$ (G_i: a gold mention cluster, $p(G_i)$: partitions of G_i)
  $P = \frac{\sum(|S_i| - |p(S_i)|)}{\sum(|S_i|-1)}$ (S_i: a system mention cluster, $p(S_i)$: partitions of S_i)
  $F1 = \frac{2PR}{P+R}$
- **B^3** (Bagga and Baldwin 1998) – mention-based metric which measures the proportion of overlap between predicted and gold mention clusters for a given mention. When G_{m_i} is the gold cluster of mention m_i and S_{m_i} is the system cluster of mention m_i,
  $R = \sum_i \frac{|G_{m_i} \cap S_{m_i}|}{|G_{m_i}|}$, $P = \sum_i \frac{|G_{m_i} \cap S_{m_i}|}{|S_{m_i}|}$, $F1 = \frac{2PR}{P+R}$
- **CEAF** (Constrained Entity Aligned F-measure) (Luo 2005) – metric based on entity alignment.
  For best alignment $g^* = argmax_{g \in G_m} \Phi(g)$ ($\Phi(g)$: total similarity of g, a one-to-one mapping from G: gold mention clusters to S: system mention clusters),
  $R = \frac{\Phi(g^*)}{\sum_i \Phi(G_i,G_i)}$, $P = \frac{\Phi(g^*)}{\sum_i \Phi(S_i,S_i)}$, $F1 = \frac{2PR}{P+R}$
  If we use $\phi(G,S) = |G \cap S|$, it is called mention-based CEAF (CEAF-ϕ_3), if we use $\phi(G,S) = \frac{2|R \cap S|}{|R|+|S|}$, it is called entity-based CEAF (CEAF-ϕ_4).
- **BLANC** (BiLateral Assessment of NounPhrase Coreference) (Recasens and Hovy 2011) – metric applying the Rand index (Rand 1971) to coreference to deal with imbalance between singletons and coreferent mentions by considering coreference and non-coreference links.
  $P_c = \frac{rc}{rc+wc}$, $P_n = \frac{rn}{rn+wn}$, $R_c = \frac{rc}{rc+wn}$, $R_n = \frac{rn}{rn+wc}$,
  $F_c = \frac{2P_cR_c}{P_c+R_c}$, $F_n = \frac{2P_nR_n}{P_n+R_n}$, $BLANC = \frac{F_c+F_n}{2}$
  (rc: the number of correct coreference links, wc: the number of incorrect coreference links, rn: the number of correct non-coreference links, wn: the number of incorrect non-coreference links)
- **CoNLL F1** Average of MUC, B^3, and CEAF-ϕ_4 F1. This was the official metric in the CoNLL-2011 shared task (Pradhan et al. 2011).

4.3 Experimental Results

Tables 4 and 5 compare the performance of our system with other state-of-the-art systems in the CoNLL-2011 shared task and the ACE and MUC corpora, respectively. For the CoNLL-2011 shared task we report results in the closed track, which did not allow the use of external resources, and the open track, which allowed any other

Table 4
Performance of the top systems in the CoNLL-2011 shared task. All these systems use automatically detected mentions. We report results for both the closed and the open tracks, which allowed the use of resources not provided by the task organizers. MD indicates mention detection, and gold boundaries indicate that mention boundary information is given.

System	MD			MUC			B^3			CEAF-ϕ_4			BLANC			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Closed Track																
This paper	75.1	66.8	70.7	61.8	57.5	59.6	68.4	68.2	68.3	43.4	47.8	45.5	70.6	76.2	73.0	57.8
Sapena	92.4	28.2	43.2	56.3	63.2	59.6	62.8	72.1	67.1	44.8	38.4	41.3	69.5	73.1	71.1	56.0
Chang	68.1	62.0	64.9	57.2	57.2	57.2	67.1	70.5	68.8	41.9	41.9	41.9	71.2	77.1	73.7	56.0
Nugues	69.9	68.1	69.0	60.2	57.1	58.6	66.7	64.2	65.5	38.1	41.1	39.5	72.0	70.3	71.1	54.5
Santos	67.8	63.3	65.5	59.2	54.3	56.7	68.8	62.8	65.7	35.9	40.2	37.9	73.4	66.9	69.5	53.4
Song	57.8	80.4	67.3	53.7	67.8	60.0	60.7	66.1	63.2	43.4	30.7	36.0	69.5	59.7	61.5	53.1
Stoyanov	70.8	65.0	67.8	63.6	54.0	58.4	72.6	53.3	61.4	32.0	40.8	35.9	73.2	58.9	60.9	51.9
Sobha	67.8	62.1	64.8	51.1	49.9	50.5	62.6	65.4	64.0	40.7	41.8	41.2	61.4	68.4	63.9	51.9
Kobdani	62.1	60.0	61.0	55.6	51.5	53.5	69.7	62.4	65.9	32.3	35.4	33.8	61.9	63.5	62.6	51.0
Zhou	61.1	63.6	62.3	45.7	52.8	49.0	57.1	72.9	64.1	43.2	36.8	39.7	61.1	73.9	64.7	50.9
Charton	65.9	62.8	64.3	55.1	50.1	52.5	66.3	58.4	62.1	34.3	39.1	36.5	69.9	62.2	64.8	50.4
Yang	71.9	57.5	63.9	59.9	46.4	52.3	71.6	55.1	62.3	30.3	42.4	35.3	71.1	61.8	64.6	50.0
Hao	64.5	64.1	64.3	57.9	51.4	54.5	67.8	55.4	61.0	30.1	35.8	32.7	72.6	62.4	65.4	49.4
Xinxin	65.5	58.7	61.9	48.5	44.9	46.6	61.6	62.3	61.9	35.2	38.6	36.8	63.0	65.8	64.3	48.5
Zhang	55.4	68.3	61.1	42.0	55.6	47.9	52.6	73.1	61.1	42.0	30.3	35.2	62.8	69.2	65.2	48.1
Kummerfeld	69.8	57.0	62.7	46.4	39.6	42.7	63.6	57.3	60.3	35.1	42.3	38.3	58.7	61.6	59.9	47.1
Zhekova	67.5	37.6	48.3	28.9	20.7	24.1	67.1	56.7	61.5	31.6	41.2	35.8	52.8	57.1	53.8	40.4
Irwin	17.1	61.1	26.7	12.5	50.6	20.0	35.1	89.9	50.5	45.8	17.4	25.2	51.5	56.8	51.1	31.9
Open Track																
This paper	74.3	67.9	70.9	62.8	59.3	61.0	68.9	69.0	68.9	43.3	46.8	45.0	71.9	76.6	74.0	58.3
Cai	67.2	67.6	67.4	56.7	58.9	57.8	64.6	71.0	67.7	42.7	40.7	41.7	69.8	74.0	71.6	55.7
Uryupina	70.6	66.3	68.4	59.7	55.7	57.6	66.3	64.1	65.2	38.3	42.2	40.2	69.2	68.5	68.9	54.3
Klenner	64.4	60.3	62.3	49.0	50.7	49.9	61.7	68.6	65.0	41.3	39.7	40.5	66.1	73.9	69.1	51.8
Irwin	24.6	62.3	35.3	18.6	51.0	27.2	39.0	85.6	53.6	43.3	19.4	26.8	51.6	52.9	51.8	35.8
Closed Track - gold boundaries																
This paper	79.5	71.3	75.2	65.9	62.1	63.9	69.5	70.6	70.0	46.3	50.5	48.3	72.0	78.6	74.8	60.7
Nugues	74.2	70.7	72.4	64.3	60.1	62.1	68.3	65.2	66.7	39.9	44.2	41.9	72.5	71.0	71.8	56.9
Chang	63.4	73.2	67.9	55.0	65.5	59.8	62.2	76.7	68.7	46.8	37.2	41.4	71.0	79.3	74.3	56.6
Santos	65.8	69.9	67.8	57.8	61.4	59.5	64.5	70.3	67.3	41.4	38.2	39.7	72.7	72.0	72.3	55.5
Kobdani	67.1	65.1	66.1	62.6	56.8	59.6	73.2	62.2	67.3	32.9	37.3	34.9	64.1	64.1	64.1	53.9
Stoyanov	76.9	64.7	70.3	69.8	55.0	61.5	77.1	52.5	62.5	31.0	44.8	36.6	76.6	60.3	63.0	53.6
Zhang	59.6	71.2	64.9	46.1	58.8	51.6	53.9	73.4	62.2	43.5	32.1	37.0	64.1	70.5	66.5	50.3
Song	58.4	77.6	66.7	46.7	68.4	55.5	54.4	70.2	61.3	43.8	25.9	32.5	66.3	58.8	60.2	49.8
Zhekova	69.2	57.3	62.7	33.5	37.2	35.2	55.5	68.2	61.2	38.3	34.7	36.4	53.5	63.3	54.8	44.3

resources. For the closed track, the organizers provided dictionaries for gender and number information, in addition to parse trees and named entity labels (Pradhan et al. 2011). For the open track, we used the following additional resources: (a) a hand-built list of genders of first names that we created, incorporating frequent names from census lists and other sources (Vogel and Jurafsky 2012) (b) an animacy list (Ji and Lin 2009), (c) a country and state gazetteer, and (d) a demonym list. These resources were also used for the results reported in Table 5.

A significant difference between Tables 4 and 5 is that in the former (other than its last block) we used predicted mentions (detected with the algorithm described in Section 3.1), whereas in the latter we used gold mentions. The only reason for this distinction is to facilitate comparison with previous work (all systems listed in Table 5 used gold mention boundaries).

The two tables show that, regardless of evaluation corpus and methodology, our system generally outperforms the previous state of the art. In the CoNLL shared task,

our system scores 1.8 CoNLL F1 points higher than the next system in the closed track and 2.6 points higher than the second-ranked system in the open track. The Chang et al. (2011) system has marginally higher B^3 and BLANC F1 scores, but does not outperform our model on the other two metrics and the average F1 score. Table 5 shows that our model has higher B^3 F1 scores than all the other models in the two ACE corpora. The model of Haghighi and Klein (2009) minimally outperforms ours by 0.6 B^3 F1 points in the MUC corpus. All in all, these results prove that our approach compares favorably with a wide range of models, which include most aspects deemed important for coreference resolution, among other things, supervised learning using rich feature sets (Sapena, Padró, and Turmo 2011; Chang et al. 2011), joint inference using spectral clustering (Cai, Mujdricza-Maydt, and Strube 2011), and deterministic rule-based models (Haghighi and Klein 2009). We discuss in more detail the similarities and differences between our approach and previous work in Section 6.

Table 4 shows that using additional resources yields minimal improvement: There is a difference of only 0.5 CoNLL F1 points between our open-track and closed-track systems. We show in Section 5 that the explanation of this modest improvement is that most of the remaining errors require complex, context-sensitive semantics to be solved. Such semantic models cannot be built with our shallow feature set that relies on simple semantic dictionaries (e.g., animacy or even hyponymy).

It is not trivial to compare the mention detection system alone because its score is affected by the performance of the coreference resolution model. For example, even if we start with a perfect set of gold mentions, if we miss all coreference relations in a text, every mention will remain as a singleton and will be removed by the OntoNotes post

Table 5
Comparison of our system with the other reported results on the ACE and MUC corpora. All these systems use gold mention boundaries.

System	MUC			B^3		
	R	P	F1	R	P	F1
ACE2004-Culotta-Test						
This paper	70.2	82.7	75.9	74.5	88.7	81.0
Haghighi and Klein (2009)	77.7	74.8	79.6	78.5	79.6	79.0
Culotta et al. (2007)	–	–	–	73.2	86.7	79.3
Bengston and Roth (2008)	69.9	82.7	75.8	74.5	88.3	80.8
ACE2004-nwire						
This paper	75.1	84.6	79.6	74.1	87.3	80.2
Haghighi and Klein (2009)	75.9	77.0	76.5	74.5	79.4	76.9
Poon and Domingos (2008)	70.5	71.3	70.9	–	–	–
Finkel and Manning (2008)	58.5	78.7	67.1	65.2	86.8	74.5
MUC6-Test						
This paper	69.1	90.6	78.4	63.1	90.6	74.4
Haghighi and Klein (2009)	77.3	87.2	81.9	67.3	84.7	75.0
Poon and Domingos (2008)	75.8	83.0	79.2	–	–	–
Finkel and Manning (2008)	55.1	89.7	68.3	49.7	90.9	64.3

processing, resulting in zero mentions in the final output. Therefore, we included the score using gold mention boundaries in the last part of Table 4 ("Closed Track – gold boundaries") to isolate the performance of the coreference resolution component. This experiment shows that our system outperforms the others with a considerable margin, demonstrating that our coreference resolution model, rather than the mention detection component, is the one responsible for the overall performance.

4.4 Analysis

In this section, we present a series of analytic and ablative experiments that demonstrate that both aspects of our algorithm (the entity-centric approach and the multi-pass model with precision-ordered sieves) independently offer significant improvements to coreference. We also analyze the contribution of each proposed sieve and of the features deployed in our model. We conclude with an experiment that measures the performance drop as we move from an oracle system that uses gold information for mention boundaries, syntactic analysis, and named entity labels, to the actual system where all this information is predicted. For all the experiments reported here we used the OntoNotes-Dev corpus.

4.4.1 Contribution of the Entity-Centric Model. Table 6 shows the impact of our entity-centric approach, which enables the sharing of features between mentions assigned to the same cluster (detailed in Section 3.2.3). As a baseline, we use a typical mention-pair model where this sharing is disabled. That is, when two mentions are compared, this model uses only the features that were extracted from the corresponding textual extents. The table shows that feature sharing has a considerable impact on all evaluation metrics, with an overall contribution of approximately 3.4 CoNLL F1 points. This is further proof that an entity-centric approach is beneficial for coreference resolution.

As an illustration, the following text shows an example where the incorrect decision is taken if feature sharing is disabled:

> ***This*** *was the best result of a Chinese gymnast in 4 days of competition. . . .* ***It*** *was the best result for Greek gymnasts since they began taking part in gymnastic internationals.*

In the example text, the mention-pair model incorrectly links *This* and *It*, because all the features that can be extracted locally are compatible (e.g., number is singular for both pronouns). On the other hand, the entity-centric model avoids this decision because, in a previous sieve driven by predicate nominative relations, these pronouns are each

Table 6
Comparison of our entity-centric model against a baseline that handles mention pairs independently. The former model shares mention features across entities as they are constructed. The latter model does not.

	MUC			B^3			CEAF-ϕ_4			BLANC			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Entity-centric	60.0	60.9	60.3	68.6	73.3	70.9	47.5	46.2	46.9	73.5	79.3	76.0	59.3
Mention-pair	61.4	51.1	55.8	73.2	64.3	68.5	39.1	48.8	43.4	74.6	74.1	74.3	55.9

Table 7
Impact of the multi-pass model. The single-pass baseline uses the same sequence of sieves as the multi-pass model (i.e., all the sieves introduced in Section 3 with the exception of the optional ones) but it applies all of them at the same time.

	MUC			B^3			CEAF-ϕ_4			BLANC			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Multi-pass	59.6	60.9	60.3	68.6	73.3	70.9	47.5	46.2	46.9	73.5	79.3	76.0	59.3
Single-pass	44.7	63.1	52.3	55.1	80.1	65.3	51.2	34.8	41.5	64.2	78.4	68.5	53.0

linked to incompatible noun phrases, i.e., *the best result of a Chinese gymnast* and *the best result for Greek gymnasts*.

4.4.2 Impact of the Multi-Pass Model. Table 7 shows the contribution of our multi-pass model. We compare this model with a single-pass baseline, which uses the same sieves as the multi-pass system but applies all of them at the same time. That is, for each mention under consideration, we select the first antecedent that matches any of the available sieves. This experiment shows that our multi-pass model, which sorts and deploys sieves using precision-based ordering, yields improvements across the board, with more than 6 CoNLL F1 points overall improvement.

This multi-pass model goes hand-in-hand with the entity-centric approach. That is, the higher the quality of mention clusters built in the previous sieves, the better the features extracted from these clusters will be in the current sieve—and, of course, better features drive better clustering decisions in the next sieve, and so on. This incremental process is highlighted in the given example: Because the sieve based on predicate nominative patterns runs before pronominal coreference resolution, the two pronouns under consideration have additional, high-quality features that stops the incorrect clustering decision.

4.4.3 Contribution of Individual Sieves. Table 8 lists the performance of our system as ten sieves are incrementally added. This table illustrates our tuning process, which allowed us to deploy the sieves in descending order of their precision. With respect to individual

Table 8
Cumulative performance as sieves are added to the system.

	MUC			B^3			CEAF-ϕ_4			BLANC			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Sieve 1	8.7	72.7	15.5	32.4	96.4	48.5	50.6	15.4	23.7	57.2	80.3	60.2	29.2
+ Sieve 2	29.5	71.8	41.9	46.4	90.4	61.4	51.8	23.8	32.6	63.0	82.2	67.8	45.3
+ Sieve 3	29.7	71.2	41.9	46.7	90.1	61.5	51.6	24.0	32.7	63.0	82.0	67.8	45.4
+ Sieve 4	30.2	71.0	42.3	47.1	89.9	61.8	51.5	24.1	32.9	63.2	81.7	68.0	45.7
+ Sieve 5	34.4	66.1	45.2	51.5	86.6	64.6	50.8	27.6	35.8	64.1	80.8	68.8	48.5
+ Sieve 6	34.9	65.8	45.6	51.9	86.1	64.8	50.4	27.8	35.9	64.2	80.6	68.9	48.8
+ Sieve 7	35.8	64.0	45.9	53.3	85.0	65.5	49.8	28.9	36.6	64.4	80.3	69.1	49.3
+ Sieve 8	36.2	63.5	46.1	53.7	84.5	65.7	49.4	29.1	36.6	64.6	79.9	69.2	49.5
+ Sieve 9	36.7	63.2	46.5	54.2	84.0	65.9	49.2	29.4	36.8	64.7	79.5	69.2	49.7
+ Sieve 10	59.6	60.9	60.3	68.6	73.3	70.9	47.5	46.2	46.9	73.5	79.3	76.0	59.3

contributions, this analysis highlights three significant performance increases. The first is caused by Sieve 2, exact string match. This sieve accounts for approximately 16 CoNLL F1 points improvement, which proves that a significant percentage of mentions in text are indeed repetitions of previously seen concepts. The second big jump in performance, almost 3 CoNLL F1 points, is caused by Sieve 5, strict head match, which is the first pass that compares individual headwords. These results are consistent with error analyses from earlier work which have shown the importance of string match in general (Zhou and Su 2004; Bengtson and Roth 2008; and Recasens, Can, and Jurafsky 2013) and the high precision of strict head match (Recasens and Hovy 2010).

Lastly, pronominal coreference resolution (Sieve 10) is responsible for approximately 9.5 CoNLL F1 points improvement. Thus it would be possible to build an even simpler system, with just three sieves, that achieves 97% of the performance of our best model (based on the CoNLL score). This suggests that what is most important for coreference resolution, at least relative to today's state of the art, is not necessarily the clustering decision mechanism, but rather the entire architecture behind it, and in particular the use of cautious decision-making based on high precision information, entity-centric modeling, and so forth.

4.4.4 Contribution of Feature Groups. Table 9 lists the results of an ablative experiment where each feature group was individually removed from the complete model. When a feature is eliminated, two mentions under consideration are always considered compatible with respect to that feature. For example, singular and plural mentions are number compatible when the number feature is removed.

As the table shows, the most significant feature in our model is the number feature. This feature alone is responsible for 2.6 CoNLL F1 points. Removing this feature has a considerable negative impact on the pronoun resolution sieve, which makes a considerable number of errors without it (e.g., linking *our* and *Jiaju Hou*). The second most relevant feature is animacy, with an overall contribution of 1 CoNLL F1 point. Animacy helps disambiguate clustering decisions where the two mentions under consideration are otherwise number and gender compatible. For example, animacy enables the linking of *firms from Taiwan* and *they*, and avoids the linking of *17 year* and *she*. Lastly, the NE and gender features contribute 0.5 and 0.4 F1 points, respectively. This relatively minor contribution is caused by the overlap with the other features (e.g., many errors corrected by using NE information are corrected also by a combination of animacy and number). Nevertheless, these features are still useful. For example, the NE feature covers many mentions that do not exist in our animacy dictionaries, which helps in several decisions, e.g., avoiding linking *it* and *Saddam Hussein*.

Table 9
Contribution of each feature group. This is an ablative experiment, that is, each feature group is analyzed by removing it from the complete system listed in the first row.

	MUC			B^3			CEAF-ϕ_4			BLANC			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Complete system	59.6	60.9	60.3	68.6	73.3	70.9	47.5	46.2	46.9	73.5	79.3	76.0	59.3
− Number	57.0	56.4	56.7	66.2	68.6	67.4	45.6	46.2	45.9	67.6	72.6	69.7	56.7
− Gender	59.3	60.2	59.7	68.2	72.3	70.2	47.2	46.3	46.7	72.6	77.8	74.9	58.9
− Animacy	58.2	58.6	58.4	67.8	71.6	69.6	47.1	46.8	47.0	71.6	77.3	74.0	58.3
− NE	58.5	60.4	59.5	67.5	73.3	70.3	47.6	45.7	46.6	72.3	78.8	75.1	58.8

Table 10
The relevance of gold information. The "no gold" system is our final system used in the formal evaluation. The system with "gold annotations" uses gold part-of-speech tags, syntactic analysis, and named entity labels.

	MUC			B^3			CEAF-ϕ_4			BLANC			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
No gold	59.6	60.9	60.3	68.6	73.3	70.9	47.5	46.2	46.9	73.5	79.3	76.0	59.3
Gold NE	60.3	61.1	60.7	69.0	73.3	71.1	47.5	46.7	47.1	74.0	79.5	76.4	59.6
Gold syntax	62.3	62.5	62.4	69.9	73.5	71.7	47.8	47.6	47.7	74.8	80.0	77.1	60.6
Gold annotations	62.8	62.6	62.7	70.3	73.5	71.9	47.9	48.1	48.0	75.1	80.1	77.4	60.9
Gold mentions	73.0	90.3	80.7	69.1	89.5	78.0	79.2	51.4	62.4	78.8	89.4	83.1	73.7

4.4.5 Gold versus Predicted Information. We conclude this section with an analysis of the performance penalty suffered when using predicted information as input in our system (a realistic scenario) versus using gold information. We consider both linguistic information (i.e., part of speech tags, named entity labels, and syntax) and mention boundaries. Table 10 shows the results when various inputs were replaced with gold information.

The table shows that, out of the linguistic resources, syntax is the most important. This is to be expected, because we use a constituent parser for mention identification, mention traversal, and for some of the sieves (e.g., the precise constructs model). All in all, if all linguistic information is replaced with gold annotations, the performance of the system increases by 1.6 CoNLL F1 points, or 2.7% relative improvement. We consider this relatively small difference a success story for the quality of natural language processors, especially considering our heavy reliance on such tools throughout the entire system. On the other hand, the difference between our actual system and the oracle system with gold mentions is 14.4 F1 points. This is because the gold mentions include the anaphoricity information, detection of which is already a hard task by itself.

4.4.6 Automatic Ordering. The ordering of our sieves was determined using linguistic intuition about how precise each sieve is (for example exact match is clearly more precise than partial match). We also supplemented this intuition, early on in our design process, by measuring the actual precision of some of the sieves on a development set from ACE.

But because this development set, not to mention our intuition, may not match the circumstances in the OntoNotes corpus, we performed a study to see if an automatically learned ordering for sieves could result in superior performance.

We used greedy search to find an ordering, choosing the best precision sieve at each pass. We tuned the ordering on OntoNotes-Train data, and evaluated this comparison on the OntoNotes-Dev set.

Our optimization resulted in 0.1 CoNLL F1 improvement, and gave a very similar ordering to our hand-built order:

Hand Ordered:

Speaker Match, String Match, Relaxed String Match, Precise Constructs, Strict Head MatchA-C, Proper Head Noun Match, Relaxed Head Match, Pronoun Match

Learned Ordering:

String Match, Relaxed String Match, Speaker Match, Proper Head Noun Match, Strict Head MatchA-C, Relaxed Head Match, Pronoun Match, Precise Constructs

The main change is that the learned ordering downplays the importance of the precise constructs sieves, which is easily explained by the fact that OntoNotes does not annotate appositive or predicate nominative relations.

This experiment confirms that hand ordering sieves by linguistic intuition of how precise they are does remarkably well at choosing an ordering, despite the fact that the ordering was originally designed for ACE, a completely different corpus.

5. Error Analysis

To understand the errors in the system, we analyzed and categorized them into five distinct groups. The distribution of the errors is given in Table 11, with specific examples for each category given in Table 12. For this analysis, we inspected 115 precision and recall errors.

Semantics, discourse. Whereas simple examples can be solved by using shallow semantics such as knowledge about the semantic compatibility of headwords (e.g., *McCain – senator*), most of the errors in this class require context-dependent semantics or discourse. For example, to know that *the thrift* and *his property* are coreferent, we need to understand the context and that both *the thrift* and *his property* are being seized, involving relations not only between the coreferent words, but also between other parts of the sentence as well.

Pronominal resolution errors. Our pronominal resolution algorithm includes several strong heuristics that model the matching of attributes (e.g., gender, number, animacy), the position of mentions in discourse (e.g., we model only the first mention in text for a given entity), or the distance between pronouns and antecedents. This is still far from language understanding, however. Table 12 shows that our approach often generates incorrect links when it finds other compatible antecedents that appear closer, according to our antecedent ordering, to the pronoun under consideration. In the example shown in the table, *the land* is selected as the antecedent for the pronoun *its*, because *the land* appears earlier than the correct antecedent, *the ANC*, in the sentence. Implementing a richer model of pronominal anaphora using syntactic and discourse information is an important next step.

Non-referential mentions. The third significant cause of errors is due to non-referential mentions such as pleonastic *it* or generic mentions. Our mention detection model removes some of these non-referential mentions, but there are still many left, which generate precision errors. For example, in Table 12, the pronoun *you* is generic, but our system incorrectly links them. The large number of these errors suggests the need to

Table 11
Distribution of errors.

Error type	Percentage
Semantics, discourse	41.7
Pronominal resolution errors	28.7
Non-referential mentions	14.8
Event mentions	6.1
Miscellaneous	8.7

Table 12
Examples of errors in each class. The mention to be resolved is in **boldface**, its correct antecedent is in *italics*, and we underlined the incorrect antecedent from our system result.

Error type	Example
Semantics, discourse	• Lincoln's parent company, American Continental Corp., entered bankruptcy - law proceedings this April 13, and regulators seized *the thrift* the next day. . . . Mr. Keating has filed his own suit, alleging that **his property** was taken illegally. • *New pictures* reveal the sheer power of that terrorist bomb . . . In **these photos obtained by NBC News**, the damage much larger than first imagined . . . • Of all the one-time expenses incurred by a corporation or professional firm, few are larger or longer term than *the purchase of real estate or the signing of a commercial lease* . . . To take full advantage of the financial opportunities in **this commitment**, . . .
Pronominal resolution errors	Under the laws of the land, *the ANC* remains an illegal organization , and **its** headquarters are still in Lusaka, Zambia.
Non-referential mentions	When you become a federal judge, all of a sudden **you** are relegated to a paltry sum.
Event mentions	"*Support* the troops, not the regime" **That** 's a noble idea until you're supporting the weight of an armoured vehicle on your chest.
Miscellaneous (inconsistent annotations, parser or NER errors, enumerations)	• Inconsistent annotation - Inclusion of *'s*: . . . that's without adding in [*Business Week 's*] charge . . . Small wonder that [*Britain*] 's Labor Party wants credit controls. • Parser or NER error: Um alright uh *Mister Zalisko* do **you** know anything from your personal experience of having been on the cruise as to what happened? – *Mister Zalisko* is not recognized as a PERSON • Enumerations: This year, the economies of the five large special economic zones, namely, Shenzhen, Zhuhai, Shantou, Xiamen and Hainan, have maintained strong growth momentum. . . . A three dimensional traffic frame in **Zhuhai** has preliminarily taken shape and the investment environment improves daily.

add more sophisticated anaphoricity detection to our system (Vieira and Poesio 2000; Ng and Cardie 2002a; Poesio et al. 2004b; Boyd, Gegg-Harrison, and Byron 2005; Gupta, Purver, and Jurafsky 2007; Bergsma, Lin, and Goebel 2008; Ng 2009).

Event mentions. Our system was tailored for the resolution of entity coreference and does not have any event-specific features, such as, for example, matching event participants. Furthermore, our model considers only noun phrases as antecedent candidates, thus missing all mentions that are verbal phrases. Therefore, our system misses most coreference links between event mentions. For example, in Table 12 the pronoun *That*

is coreferent with the event mention *Support*. Our system fails to detect the latter event mention and, as a consequence, incorrectly links *That* to *the regime*.

Miscellaneous. There are several other reasons for errors, including inconsistent annotations, parse or NER errors, and incorrect processing of enumerations. For example, the possessive (*'s*) is annotated inconsistently in several cases: sometimes it is included in the possessor mention in the gold mention annotation, but sometimes it is not. This will penalize the final score twice (once for recall due to the missed mention and once for precision due to the incorrectly detected mention).

Another considerable source of errors is caused by incorrect NER labels or parse trees. NER errors can result in incorrect pronoun resolution due to incorrect attributes. Parser errors are responsible for many additional coreference resolution errors. First, incorrect syntactic attachments lead to incorrect mention boundaries, which are penalized by our strict scorer. Second, parser errors often lead to the selection of an incorrect head word for a given constituent, which influences many of our sieves. Thirdly, because our parser does not always distinguish between coordinated nominal phrases and appositions, our system sometimes takes an entire coordinated phrase as a single mention, leading to a series of mention errors. For example, the last example in the table shows a compounded syntactic error: first, the parser failed to identify the entire construct (*Shenzhen, Zhuhai, Shantou, Xiamen, and Hainan*) as a single enumeration. Second, our system believed that *Zhuhai, Shantou, Xiamen* is an appositive phrase and kept it as a single mention, rather than separate it into three distinct mentions.

Lastly, our processing of enumerations needs to be improved. Because we prefer to assign content words as head words of syntactic constituents, we take the head word of the first noun phrase in the enumeration to be the head word of the coordinated nominal phrase (Kuebler, McDonald, and Nivre 2009; de Marneffe and Manning 2008). Because of this, the coordinated phrase is often linked to another mention of the first element in the enumeration. For example, our system marks *Zhuhai, Shantou, Xiamen* as a unique mention and incorrectly links it to *Zhuhai*, because they have the same headword.

6. Comparison with Previous Work

Algorithms for coreference (or just pronominal anaphora) include rule-based systems (Hobbs 1978; Brennan, Friedman, and Pollard 1987; Lappin and Leass 1994; Baldwin 1995; Zhou and Su 2004; Haghighi and Klein 2009, inter alia), supervised systems (Connolly, Burger, and Day 1994; McCarthy and Lehnert 1995; Kehler 1997; Soon, Ng, and Lim 2001; Ng and Cardie 2002b; Rahman and Ng 2009, inter alia), and unsupervised approaches (Cardie and Wagstaff 1999; Haghighi and Klein 2007; Ng 2008; Kobdani et al. 2011a). Our deterministic system draws from all of these, but specifically from three strands in the literature that cross-cut this classification.

The idea of doing accurate reference resolution by starting with a set of very high-precision constraints was first proposed for pronominal anaphora in Baldwin's (1995) important but undercited dissertation. Baldwin suggested using seven high-precision rules as filters, combining them so as to achieve reasonable recall. One of his rules, for example, resolved pronouns whose antecedents were unique in the discourse, and another resolved pronouns in quoted speech. Baldwin's idea of starting with high-precision knowledge was adopted by later researchers, such as Ng and Cardie (2002b), who trained to the highest-confidence rather than nearest antecedent, or Haghighi and Klein (2009), who began with syntactic constraints (which tend to be higher-precision) before applying semantic constraints. This general idea is known by different names in

many NLP applications: Brown et al. (1993) used simple models as "stepping stones" for more complex word alignment models; Collins and Singer (1999) used "cautious" decision list learning for named entity classification; Borghesi and Favareto (1982) and Corazza et al. (1991) used "islands of reliability" approaches to parsing and speech recognition, and Spitkovsky et al. (2010) used "baby steps" for unsupervised dependency parsing, and so forth. Our work extends the intuition of Baldwin and others to the full coreference task (i.e., including mention detection and both nominal and pronominal coreference) and shows that it can result in extremely high-performing resolution when combined with global inference.

Our second inspiration comes from two works: Zhou and Su (2004) and Haghighi and Klein (2009), both of which extended Baldwin's approach to generic nominal coreference. Zhou and Su proposed a multi-agent model that triggers a different agent with a specific set of deterministic constraints for each anaphor depending on its type and context (e.g., there are different constraints for noun phrases in appositive constructs, definite noun phrases, or bare noun phrases). Some of the constraints' parameters (e.g., size of candidate search space for a given anaphor type) are learned from training data. The authors showed that this model outperforms the state of the art on the MUC-6 and MUC-7 domains. To our knowledge, Zhou and Su's approach is the first work to demonstrate that a deterministic approach obtains state-of-the-art results for both nominal and pronominal coreference resolution. Our approach extends Zhou and Su's model in two significant ways. First, Zhou and Su solve the coreference task in a single pass over the text. We show that a multi-pass approach, which applies a series of sieves incrementally from highest to lowest precision, performs considerably better (see Table 7). Second, Zhou and Su's model follows a mention-pair approach, where coreference decisions are taken based only on information extracted from the two mentions under consideration. We demonstrate that an entity-centric approach, which allows features to be shared between mentions of the same entity, outperforms the mention-pair model (see Table 6).

Haghighi and Klein's (2009) two-pass system based on deterministic rules further proved that deterministic rules could achieve state-of-the-art performance. Haghighi and Klein's first, purely syntactic pass, uses high-precision syntactic information to assign possible coreference. The second, transductive pass identifies Wikipedia articles relevant to the entity mentions in the test set, and then bootstraps a database of hyponyms and other semantically related head pairs from known syntactic patterns for apposition and predicate-nominatives. Haghighi and Klein found that this transductive learning was essential for semantic knowledge to be useful (Aria Haghighi, personal communication); other researchers have found that semantic knowledge derived from Web resources can be quite noisy (Uryupina et al. 2011a). But although transductive learning (learning using test set mentions) thus offers advantages in precision, running a Web-based bootstrapping learner whenever a new data set is encountered is not practical and, ultimately, reduces the usability of this NLP component. Our system thus offers the deterministic simplicity and high performance of the Haghighi and Klein (2009) system without the need for gold mention labels or test-time learning. Furthermore, our work extends the multi-pass model to ten passes and shows that this approach can be naturally combined with an entity-centric model for better results.

Finally, recent work has shown the importance of performing coreference resolution jointly for all mentions in a document (McCallum and Wellner 2004; Daumé III and Marcu 2005; Denis and Baldridge 2007; Haghighi and Klein 2007; Culotta et al. 2007; Poon and Domingos 2008; Haghighi and Klein 2010; Cai, Mujdricza-Maydt, and Strube 2011) rather than the classic method of simply aggregrating local decisions about pairs of mentions. Like these systems, our model adopts the *entity-mention model* (Morton

2000; Luo et al. 2004; Yang et al. 2008; Ng 2010)[8] in which features can be extracted over not just pairs of mentions but over entire clusters of mentions defining an entity. Previous systems do this by encoding constraints using rich probabilistic models and complex global inference algorithms. By contrast, global reasoning is implemented in our system just by allowing the rules in each stage to reason about any features of a cluster from a previous stage, including attributes like gender and number as well as headword information derived from the first (most informative) mention. Because our system begins with high-precision clusters, accurate information naturally propagates to later stages.

7. Other Systems Incorporating this Algorithm

A number of recent systems have incorporated our algorithm as an important component in resolving coreference. For example, the CoNLL-2012 shared task focused on coreference resolution in a multi-lingual setting: English, Chinese, and Arabic (Pradhan et al. 2012). Forty percent of the systems in the shared task (6 of the 15 systems) made use of our sieve architecture (Chen and Ng 2012; Fernandes, dos Santos, and Milidiu 2012; Shou and Zhao 2012; Xiong and Liu 2012; Yuan et al. 2012; Zhang, Wu, and Zhao 2012), including the systems that were the highest scoring for each of the three languages (Fernandes, dos Santos, and Milidiu 2012; Chen and Ng 2012).

The system of Fernandes, dos Santos, and Milidiu (2012) had the highest average score over all languages, and the best score for English and Arabic, by implementing a stacking of two models. Our sieve-based approach was first used to generate mention-link candidates, which are then reranked by a supervised model inspired from dependency parsing. This result demonstrates that our deterministic approach can be naturally combined with more-complex supervised models for further performance gains.

The system of Chen and Ng (2012) performed the best for Chinese by making the observation that most sieves in our model are minimally lexicalized so they can be easily adapted to other languages. Their coreference model for Chinese incorporated our English sieves with only four modifications, only two of which were related to the differences between Chinese and English: The precise constructs sieve was extended to add patterns for Chinese name abbreviations, and the relaxed head-match sieve was removed, because Chinese tends not to have post-nominal modifiers.[9] Chen and Ng (2012) then added a second component which first linked mentions with high string-pair or head-pair probabilities before running the sieve architecture. The strong performance of our English sieve system on Chinese with only this small number of changes speaks to the multi-lingual strength of our approach.

The intuition of our system can be further extended to the task of event coreference resolution. Our recent work (Lee et al. 2012) showed that an iterative method that cautiously constructs clusters of entity and event mentions, using linear regression to model cluster merge operations, allows information flow between entity and event coreference.

8 In this article, we call this approach *entity-centric* to avoid confusion with individual mentions of entities.

9 Two changes were related to differences between the English and Chinese shared task in the supplied annotations and data: The pronoun sieve was extended to determine gender for Chinese NPs, because the gender gazeteer used for the shared task and for our system only provides gender for English, and a new head-match sieve was added to deal with embedded heads, because the Chinese annotation marked embedded heads differently than the English annotation.

A similar easy-first machine learning based approach to entity coreference by Stoyanov and Eisner (2012) also adopts this intuition. Their system greedily merges clusters with the highest score (the current easiest decision), using higher precision classifications ('easier decisions') to guide harder decisions later.

In summary, recent systems have used the sieve architecture as a component in hybrid machine learning systems, either as a first pass in generating candidate links which are then incorporated in a probabilistic system, or as a second pass for generating links after high-probability mention-pairs have already been linked. These hybrid systems are the state-of-the-art in English, Chinese, and Arabic coreference resolution. Further, our algorithm can be extended to other tasks, for example, event coreference resolution.

8. Conclusion

We have presented a simple deterministic approach to coreference resolution that incorporates document-level information, which is typically exploited only by more complex, joint learning models. Our approach exploits document-level information through an entity-centric model, which allows features to be shared across mentions that point to the same real-world entity. The sieve architecture applies a battery of deterministic coreference models one at a time from highest to lowest precision, where each model builds on the previous model's entity output. Despite its simplicity, our approach outperforms or performs comparably to the state of the art on several corpora.

An additional benefit of the sieve framework is its modularity: New features or models can be inserted in the system with limited understanding of the other features already deployed. Our code is publicly released[10] and can be used both as a stand-alone coreference system and as a platform for the development of future systems.

The state-of-the-art performance of our system in coreference, either directly or as a component in hybrid systems, and that of other recent rule-based systems in named entity recognition (Chiticariu et al. 2010) suggests that rule-based systems are still an important tool for modern natural language processing. Our results further suggest that precision-ordered sieves may be an important way to structure rule based systems, and suggests the use of sieves in other NLP tasks for which a variety of very high-precision features can be designed and non-local features can be shared. Likely candidates include relation and event extraction, template slot filling, and author name deduplication.

Our error analysis points to a number of places where our system could be improved, including better performance on pronouns. More sophisticated anaphoricity detection, drawing on the extensive literature in this area, could also help (Vieira and Poesio 2000; Ng and Cardie 2002a; Poesio et al. 2004b; Boyd, Gegg-Harrison, and Byron 2005; Gupta, Purver, and Jurafsky 2007; Bergsma, Lin, and Goebel 2008; Ng 2009).

The main conclusion of our error analysis, however, is that the plurality of our errors are due to shallow knowledge of semantics and discourse. This result points to the crucial need for more sophisticated methods of incorporating semantic and discourse knowledge. Unsupervised or semi-supervised approaches to semantics such as Yang and Su (2007), Kobdani et al. (2011b), Uryupina et al. (2011b), Bansal and Klein (2012), or Recasens, Can, and Jurafsky (2013) may point the way forward. Although sieve-based architectures are at the modern state of the art, it is only by incorporating these more powerful models of meaning that we can eventually deal with the full complexity and richness of coreference.

10 `http://nlp.stanford.edu/software/dcoref.shtml`.

Appendix A: The OntoNotes Named Entity Tag Set

PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FACILITY	Buildings, airports, highways, bridges, etc.
ORGANIZATION	Companies, agencies, institutions, etc.
GPE	Countries, cities, states
LOCATION	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Vehicles, weapons, foods, etc. (Not services)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK OF ART	Titles of books, songs, etc.
LAW	Named documents made into laws
LANGUAGE	Any named language
DATE	Absolute or relative dates or periods
TIME	Times smaller than a day
PERCENT	Percentage (including "%")
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
ORDINAL	"first", "second"
CARDINAL	Numerals that do not fall under another type

Appendix B: Set of Patterns for Detecting Pleonastic *it*

```
NP < (PRP=m1) $.. (VP < ((/^V.*/ < /^(?:is|was|become|became)/) $.. (VP < (VBN $.. /S|SBAR/))))
NP < (PRP=m1) $.. (VP < ((/^V.*/ < /^(?:is|was|become|became)/) $.. (ADJP $.. (/S|SBAR/))))
NP < (PRP=m1) $.. (VP < ((/^V.*/ < /^(?:is|was|become|became)/) $.. (ADJP < (/S|SBAR/))))
NP < (PRP=m1) $.. (VP < ((/^V.*/ < /^(?:is|was|become|became)/) $.. (NP < /S|SBAR/)))
NP < (PRP=m1) $.. (VP < ((/^V.*/ < /^(?:is|was|become|became)/) $.. (NP $.. ADVP $.. /S|SBAR/)))
NP < (PRP=m1) $.. (VP < (MD $ .. (VP < ((/^V.*/ < /^(?:be|become)/) $.. (VP < (VBN $.. /S|SBAR/))))))
NP < (PRP=m1) $.. (VP < (MD $ .. (VP < ((/^V.*/ < /^(?:be|become)/) $.. (ADJP $.. (/S|SBAR/))))))
NP < (PRP=m1) $.. (VP < (MD $ .. (VP < ((/^V.*/ < /^(?:be|become)/) $.. (ADJP < (/S|SBAR/))))))
NP < (PRP=m1) $.. (VP < (MD $ .. (VP < ((/^V.*/ < /^(?:be|become)/) $.. (NP < /S|SBAR/)))))
NP < (PRP=m1) $.. (VP < (MD $ .. (VP < ((/^V.*/ < /^(?:be|become)/) $.. (NP $.. ADVP $.. /S|SBAR/)))))
NP < (PRP=m1) $.. (VP < ((/^V.*/ < /^(?:seems|appears|means|follows)/) $.. /S|SBAR/))
NP < (PRP=m1) $.. (VP < ((/^V.*/ < /^(?:turns|turned)/) $.. PRT $.. /S|SBAR/)
```

Acknowledgments

We gratefully acknowledge the support of the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA, AFRL, or the U.S. government. We gratefully thank Aria Haghighi, Marta Recasens, Karthik Raghunathan, and Chris Manning for useful suggestions; Sameer Pradhan for help with the CoNLL infrastructure; the Stanford NLP Group for help throughout; and the four anonymous reviewers for extremely helpful feedback.

References

Bagga, Amit and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, volume 1, pages 563–566, Granada.

Baldwin, Breck. 1995. *CogNIAC: A Discourse Processing Engine*. University of Pennsylvania Department of Computer and Information Sciences. Ph.D. thesis.

Baldwin, Breck. 1997. Cogniac: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45, Madrid.

Bansal, Mohit and Dan Klein. 2012. Coreference semantics from web features. In *Proceedings of ACL 2012*, pages 389–398, Jeju Island.

Bengtson, Eric and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of EMNLP 2008*, pages 294–303, Honolulu, HI.

Bergsma, Shane and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of COLING-ACL*, pages 33–40, Stroudsburg, PA.

Bergsma, Shane, Dekang Lin, and Randy Goebel. 2008. Distributional identification of non-referential pronouns. In *Proceedings of ACL-HLT 2008*, pages 10–18, Columbus, OH.

Borghesi, Luigi and Chiara Favareto. 1982. Flexible parsing of discretely uttered sentences. In *Proceedings of the 9th Conference on Computational Linguistics-Volume 1*, pages 37–42, Prague.

Boyd, Adriane, Whitney Gegg-Harrison, and Donna Byron. 2005. Identifying non-referential *it*: A machine learning approach incorporating linguistically motivated features. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP*, pages 40–47, Ann Arbor, MI.

Brennan, Susan E., Marilyn W. Friedman, and Carl Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*, pages 155–162, Stanford, CA.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Cai, Jie, Eva Mujdricza-Maydt, and Michael Strube. 2011. Unrestricted coreference resolution via global hypergraph partitioning. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 56–60, Portland, OR.

Cardie, Claire and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89, College Park, MD.

Chang, Kai-Wei, Rajhans Samdani, Alla Rozovskaya, Nick Rizzolo, Mark Sammons, and Dan Roth. 2011. Inference protocols for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 40–44, Portland, OR.

Chen, Chen and Vincent Ng. 2012. Combining the best of two worlds: A hybrid approach to multilingual coreference resolution. In *Proceedings of the CoNLL-2012 Shared Task*, pages 56–63, Jeju Island.

Chiticariu, Laura, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1,002–1,012, Cambridge, MA.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Mouton de Gruyter, Berlin.

Collins, Michael and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110, College Park, MD.

Connolly, Dennis, John D. Burger, and David S. Day. 1994. A machine learning approach to anaphoric reference. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP-1)*, pages 255–261, Manchester.

Corazza, A., R. De Mori, R. Gretter, and G. Satta. 1991. Stochastic context-free grammars for island-driven probabilistic parsing. In *Proceedings of Second International Workshop on Parsing Technologies (IWPT 91)*, pages 210–217, Cancun.

Culotta, Aron, Michael Wick, Robert Hall, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *Proceedings of HLT-NAACL 2007*, pages 81–88, Rochester, NY.

Daumé III, Hal and Daniel Marcu. 2005. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *HLT-EMNLP 2005*, pages 97–104, Vancouver.

de Marneffe, Marie-Catherine and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Proceedings of COLING Workshop on Cross-framework and*

Cross-domain Parser Evaluation, pages 1–8, Manchester.
Denis, Pascal and Jason Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of NAACL-HLT 2007*, pages 236–243, Rochester, NY.
Denis, Pascal and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42:87–96.
Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) program—Tasks, data, and evaluation. In *Proceedings of LREC 2004*, pages 837–840, Lisbon.
Elsner, Micha and Eugene Charniak. 2010. The same-head heuristic for coreference. In *Proceedings of ACL 2010 Short Papers*, pages 33–37, Uppsala.
Fernandes, Eraldo, Cicero dos Santos, and Ruy Milidiu. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Proceedings of the CoNLL-2012 Shared Task*, pages 41–48, Jeju Island.
Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA.
Finkel, Jenny Rose and Christopher D. Manning. 2008. Enforcing transitivity in coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 45–48, Columbus, OH.
Fox, Barbara A. 1993. *Discourse Structure and Anaphora: Written and Conversational English*. Cambridge University Press.
Greene, Barbara B. and Gerald M. Rubin. 1971. *Automatic Grammatical Tagging of English*. Brown University Press.
Gupta, Surabhi, Matthew Purver, and Dan Jurafsky. 2007. Disambiguating between generic and referential "you" in dialog. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 105–108, Prague.
Haghighi, Aria and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of ACL 2007*, pages 848–855, Prague.
Haghighi, Aria and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of EMNLP 2009*, pages 1,152–1,161, Suntec.
Haghighi, Aria and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Proceedings of HLT-NAACL 2010*, pages 385–393, Los Angeles, CA.
Hobbs, Jerry R. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.
Ji, Heng and Dekang Lin. 2009. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation*, pages 220–229, Hong Kong.
Kehler, Andrew. 1997. Probabilistic coreference in information extraction. In *Proceedings of EMNLP 1997*, pages 163–173, Providence, RI.
Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA.
Klein, Sheldon and Robert F. Simmons. 1963. A computational approach to grammatical coding of English words. *Journal of the Association for Computing Machinery*, 10(3):334–347.
Kobdani, Hamidreza, Hinrich Schuetze, Michael Schiehlen, and Hans Kamp. 2011a. Bootstrapping coreference resolution using word associations. In *Proceedings of ACL HLT 2011*, pages 783–792, Portland, OR.
Kobdani, Hamidreza, Hinrich Schütze, Michael Schiehlen, and Hans Kamp. 2011b. Bootstrapping coreference resolution using word associations. In *Proceedings of ACL*, pages 783–792, Portland, OR.
Kuebler, Sandra, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Morgan and Claypool Publishers.
Lappin, Shalom and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.
Lee, Heeyoung, 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 CoNLL 2011: Shared Task*, pages 28–34, Portland, OR.

Lee, Heeyoung, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 489–500, Jeju Island.

Luo, Xiaoqiang. 2005. On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP 2005*, pages 25–32, Vancouver.

Luo, Xiaoqiang, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell tree. In *Proceedings of ACL 2004*, pages 21–26, Barcelona.

McCallum, Andrew and Ben Wellner. 2004. Conditional models of identity uncertainty with application to noun coreference. In *Proceedings of NIPS 2004*, pages 905–912, Vancouver.

McCarthy, Joseph F. and Wendy G. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of IJCAI 1995*, pages 1,050–1,055, Montréal.

Morton, Thomas S. 2000. Coreference for NLP applications. In *Proceedings of ACL 2000*, pages 173–180, Hong Kong.

Ng, Vincent. 2008. Unsupervised models for coreference resolution. In *Proceedings of EMNLP 2008*, pages 640–649, Honolulu, HI.

Ng, Vincent. 2009. Graph-cut-based anaphoricity determination for coreference resolution. In *Proceedings of NAACL-HLT 2009*, pages 575–583, Boulder, CO.

Ng, Vincent. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of ACL*, pages 1,396–1,411, Uppsala.

Ng, Vincent and Claire Cardie. 2002a. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of COLING*, pages 1–7, Taipei.

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

Poesio, Massimo, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004a. Learning to resolve bridging references. In *Proceedings of ACL*, pages 143–150, Barcelona.

Poesio, Massimo, Olga Uryupina, Renata Vieira, Mijail Alexandrov-Kabadjov, and Rodrigo Goulart. 2004b. Discourse-new detectors for definite description resolution: A survey and a preliminary proposal. In *ACL 2004: Workshop on Reference Resolution and its Applications*, pages 47–54, Barcelona.

Poon, Hoifung and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov logic. In *Proceedings of EMNLP 2008*, pages 650–659, Honolulu, HI.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL)*, page 1, Jeju Island.

Pradhan, Sameer, 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 Fifteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 1–27, Portland, OR.

Raghunathan, Karthik, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Chris Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of EMNLP 2010*, pages 492–501, Cambridge, MA.

Rahman, Altaf and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–977, Suntec.

Rand, William M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Recasens, Marta and Eduard Hovy. 2010. Coreference resolution across corpora: Languages, coding schemes, and preprocessing information. In *Proceedings of ACL 2010*, pages 1,423–1,432, Uppsala.

Recasens, Marta, Matthew Can, and Dan Jurafsky. 2013. Same referent, different words: Unsupervised mining of opaque coreferent mentions. In *Proceedings of NAACL 2013*, pages 897–906, Atlanta.

Recasens, Marta and Eduard Hovy. 2011. BLANC: Implementing the Rand index for

coreference evaluation. *Natural Language Engineering*, 17(4):485–510.
Sapena, Emili, Lluís Padró, and Jordi Turmo. 2011. Relaxcor participation in CoNLL-shared task on coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 35–39, Portland, OR.
Shou, Heming and Hai Zhao. 2012. System paper for CoNLL-2012 shared task: Hybrid rule-based algorithm for coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 118–121, Jeju Island.
Skinner, B. F. 1938. *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century-Crofts.
Soon, Wee M., Hwee T. Ng, and Daniel C. Y. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.
Spitkovsky, Valentin I., Hiyan Alshawi, and Daniel Jurafsky. 2010. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 751–759, Stroudsburg, PA.
Stoyanov, Veselin and Jason Eisner. 2012. Easy-first coreference resolution. In *Proceedings of COLING 2012*, pages 2,519–2,534, Mumbai.
Uryupina, Olga, Massimo Poesio, Claudio Giuliano, and Kateryna Tymoshenko. 2011a. Disambiguation and filtering methods in using web knowledge for coreference resolution. In *FLAIRS Conference*, pages 317–322, Palm Beach, FL.
Uryupina, Olga, Massimo Poesio, Claudio Giuliano, and Kateryna Tymoshenko. 2011b. Disambiguation and filtering methods in using web knowledge for coreference resolution. In *Proceedings of FLAIRS*, pages 317–322, Palm Beach, FL.
Vieira, Renata and Massimo Poesio. 2000. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.
Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of MUC-6*, pages 45–52, Columbia, MD.
Vogel, Adam and Dan Jurafsky. 2012. He Said, She Said: Gender in the ACL Anthology. In *ACL Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island.
Xiong, Hao and Qun Liu. 2012. Ict: System description for CoNLL-2012. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 71–75, Jeju Island.
Yang, Xiaofeng and Jian Su. 2007. Coreference resolution using semantic relatedness information from automatically discovered patterns. In *Proceedings of ACL 2007*, pages 525–535, Prague.
Yang, Xiaofeng, Jian Su, Jun Lang, Chew L. Tan, Ting Liu, and Sheng Li. 2008. An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of ACL-HLT 2008*, pages 843–851, Columbus, OH.
Yang, Xiaofeng, Guodong Zhou, Jian Su, and Chew L. Tan. 2004. An NP-cluster approach to coreference resolution. In *Proceedings of COLING 2004*, pages 219–225, Geneva.
Yuan, Bo, Qingcai Chen, Yang Xiang, Xiaolong Wang, Liping Ge, Zengjian Liu, Meng Liao, and Xianbo Si. 2012. A mixed deterministic model for coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 76–82, Jeju Island.
Zhang, Xiaotian, Chunyang Wu, and Hai Zhao. 2012. Chinese coreference resolution via ordered filtering. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 95–99, Jeju Island.
Zhou, Guodong and Jian Su. 2004. A high-performance coreference resolution system using a constraint-based multi-agent strategy. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, page 522, Geneva.