Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers,
Mihai Surdeanu, Dan Jurafsky
Stanford NLP Group
Stanford University, Stanford, CA 94305
{heeyoung,peirsman,angelx,natec,mihais,jurafsky}@stanford.edu

Abstract

This paper details the coreference resolution system submitted by Stanford at the CoNLL-2011 shared task. Our system is a collection of deterministic coreference resolution models that incorporate lexical, syntactic, semantic, and discourse information. All these models use global document-level information by sharing mention attributes, such as gender and number, across mentions in the same cluster. We participated in both the open and closed tracks and submitted results using both predicted and gold mentions. Our system was ranked first in both tracks, with a score of 57.8 in the closed track and 58.3 in the open track.

1 Introduction

This paper describes the coreference resolution system used by Stanford at the CoNLL-2011 shared task (Pradhan et al., 2011). Our system extends the multi-pass sieve system of Raghunathan et al. (2010), which applies tiers of deterministic coreference models one at a time from highest to lowest precision. Each tier builds on the entity clusters constructed by previous models in the sieve, guaranteeing that stronger features are given precedence over weaker ones. Furthermore, this model propagates global information by sharing attributes (e.g., gender and number) across mentions in the same cluster.

We made three considerable extensions to the Raghunathan et al. (2010) model. First, we added five additional sieves, the majority of which address the semantic similarity between mentions, e.g., using WordNet distance, and shallow discourse understanding, e.g., linking speakers to compatible pronouns. Second, we incorporated a mention detection sieve at the beginning of the processing flow. This sieve filters our syntactic constituents unlikely to be mentions using a simple set of rules on top of the syntactic analysis of text. And lastly, we added a post-processing step, which guarantees that the output of our system is compatible with the shared task and OntoNotes specifications (Hovy et al., 2006; Pradhan et al., 2007).

Using this system, we participated in both the closed[1] and open[2] tracks, using both predicted and gold mentions. Using predicted mentions, our system had an overall score of 57.8 in the closed track and 58.3 in the open track. These were the top scores in both tracks. Using gold mentions, our system scored 60.7 in the closed track in 61.4 in the open track.

We describe the architecture of our entire system in Section 2. In Section 3 we show the results of several experiments, which compare the impact of the various features in our system, and analyze the performance drop as we switch from gold mentions and annotations (named entity mentions and parse trees) to predicted information. We also report in this section our official results in the testing partition.

[1] Only the provided data can be used, i.e., WordNet and gender gazetteer.

[2] Any external knowledge source can be used. We used additional animacy, gender, demonym, and country and states gazetteers.

Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 28–34,
Portland, Oregon, 23-24 June 2011.

2 System Architecture

Our system consists of three main stages: mention detection, followed by coreference resolution, and finally, post-processing. In the first stage, mentions are extracted and relevant information about mentions, e.g., gender and number, is prepared for the next step. The second stage implements the actual coreference resolution of the identified mentions. Sieves in this stage are sorted from highest to lowest precision. For example, the first sieve (i.e., highest precision) requires an exact string match between a mention and its antecedent, whereas the last one (i.e., lowest precision) implements pronominal coreference resolution. Post-processing is performed to adjust our output to the task specific constraints, e.g., removing singletons.

It is important to note that the first system stage, i.e., the mention detection sieve, favors recall heavily, whereas the second stage, which includes the actual coreference resolution sieves, is precision oriented. Our results show that this design lead to state-of-the-art performance despite the simplicity of the individual components. This strategy has been successfully used before for information extraction, e.g., in the BioNLP 2009 event extraction shared task (Kim et al., 2009), several of the top systems had a first high-recall component to identify event anchors, followed by high-precision classifiers, which identified event arguments and removed unlikely event candidates (Björne et al., 2009). In the coreference resolution space, several works have shown that applying a list of rules from highest to lowest precision is beneficial for coreference resolution (Baldwin, 1997; Raghunathan el al., 2010). However, we believe we are the first to show that this high-recall/high-precision strategy yields competitive results for the complete task of coreference resolution, i.e., including mention detection and both nominal and pronominal coreference.

2.1 Mention Detection Sieve

In our particular setup, the recall of the mention detection component is more important than its precision, because any missed mentions are guaranteed to affect the final score, but spurious mentions may not impact the overall score if they are left as singletons, which are discarded by our post-processing step. Therefore, our mention detection algorithm focuses on attaining high recall rather than high precision. We achieve our goal based on the list of sieves sorted by recall (from highest to lowest). Each sieve uses syntactic parse trees, identified named entity mentions, and a few manually written patterns based on heuristics and OntoNotes specifications (Hovy et al., 2006; Pradhan et al., 2007). In the first and highest recall sieve, we mark all noun phrase (NP), possessive pronoun, and named entity mentions in each sentence as candidate mentions. In the following sieves, we remove from this set all mentions that match any of the exclusion rules below:

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*.
4. We remove pleonastic *it* pronouns, detected using a set of known expressions, e.g., *It is possible that*.
5. We discard adjectival forms of nations, e.g., *American*.
6. We remove stop words in a predetermined list of 8 words, e.g., *there, ltd., hmm*.

Note that the above rules extract both mentions in appositive and copulative relations, e.g., *[[Yongkang Zhou] , the general manager]* or *[Mr. Savoca] had been [a consultant. . .]*. These relations are not annotated in the OntoNotes corpus, e.g., in the text *[[Yongkang Zhou] , the general manager]*, only the larger mention is annotated. However, appositive and copulative relations provide useful (and highly precise) information to our coreference sieves. For this reason, we keep these mentions as candidates, and remove them later during post-processing.

2.2 Mention Processing

Once mentions are extracted, we sort them by sentence number, and left-to-right breadth-first traversal

order in syntactic trees in the same sentence (Hobbs, 1977). We select for resolution only the first mentions in each cluster,[3] for two reasons: (a) the first mention tends to be better defined (Fox, 1993), which provides a richer environment for feature extraction; and (b) it has fewer antecedent candidates, which means fewer opportunities to make a mistake. 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 subscript indicates textual order and the superscript indicates cluster id, our model will attempt to resolve only m_2^2 and m_4^3. Furthermore, we discard first mentions that start with indefinite pronouns (e.g., *some, other*) or indefinite articles (e.g., *a, an*) if they have no antecedents that have the exact same string extents.

For each selected mention m_i, all previous mentions $m_{i-1}, \ldots, m_1$ become antecedent candidates. All sieves traverse the candidate list until they find a coreferent antecedent according to their criteria or reach the end of the list. Crucially, when comparing two mentions, our approach uses information from the entire clusters that contain these mentions instead of using just information local to the corresponding mentions. Specifically, mentions in a cluster share their attributes (e.g., number, gender, animacy) between them so coreference decision are better informed. For example, if a cluster contains two mentions: *a group of students*, which is `singular`, and *five students*, which is `plural`, the `number` attribute of the entire cluster becomes `singular or plural`, which allows it to match other mentions that are both singular and plural. Please see (Raghunathan et al., 2010) for more details.

2.3 Coreference Resolution Sieves

2.3.1 Core System

The core of our coreference resolution system is an incremental extension of the system described in Raghunathan et al. (2010). Our core model includes two new sieves that address nominal mentions and are inserted based on their precision in a held-out corpus (see Table 1 for the complete list of sieves deployed in our system). Since these two sieves use simple lexical constraints without semantic information, we consider them part of the baseline model.

[3] We initialize the clusters as singletons and grow them progressively in each sieve.

Ordered sieves
1. **Mention Detection Sieve**
2. **Discourse Processing Sieve**
3. Exact String Match Sieve
4. **Relaxed String Match Sieve**
5. Precise Constructs Sieve (e.g., appositives)
6-8. Strict Head Matching Sieves A-C
9. **Proper Head Word Match Sieve**
10. **Alias Sieve**
11. Relaxed Head Matching Sieve
12. **Lexical Chain Sieve**
13. Pronouns Sieve

Table 1: The sieves in our system; sieves new to this paper are in bold.

Relaxed String Match: This sieve considers two nominal mentions as coreferent if the strings obtained by dropping the text following their head words are identical, e.g., *[Clinton]* and *[Clinton, whose term ends in January]*.

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 Raghunathan et al. (2010).

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.

In addition to the above, a few more rules are added to get better performance for predicted mentions.

Pronoun distance - sentence distance between a pronoun and its antecedent cannot be larger than 3.

Bare plurals - bare plurals are generic and cannot have a coreferent antecedent.

2.3.2 Semantic-Similarity Sieves

We first extend the above system with two new sieves that exploit semantics from WordNet, Wikipedia infoboxes, and Freebase records, drawing on previous coreference work using these databases (Ng & Cardie, 2002; Daumé & Marcu, 2005; Ponzetto & Strube, 2006; Ng, 2007; Yang & Su,

2007; Bengston & Roth, 2008; Huang et al., 2009; inter alia). Since the input to a sieve is a collection of mention clusters built by the previous (more precise) sieves, we need to link mention clusters (rather than individual mentions) to records in these three knowledge bases. The following steps generate a query for these resources from a mention cluster.

First, we select the most representative mention in a cluster by preferring mentions headed by proper nouns to mentions headed by common nouns, and nominal mentions to pronominal ones. In case of ties, we select the longer string. For example, the mention selected from the cluster {*President George W. Bush, president, he*} is *President George W. Bush*. Second, if this mention returns nothing from the knowledge bases, we implement the following query relaxation algorithm: (a) remove the text following the mention head word; (b) select the lowest noun phrase (NP) in the parse tree that includes the mention head word; (c) use the longest proper noun (NNP*) sequence that ends with the head word; (d) select the head word. For example, the query *president Bill Clinton, whose term ends in January* is successively changed to *president Bill Clinton*, then *Bill Clinton*, and finally *Clinton*. If multiple records are returned, we keep the top two for Wikipedia and Freebase, and all synsets for WordNet.

Alias Sieve

This sieve addresses name aliases, which are detected as follows. Two mentions headed by proper nouns are marked as aliases (and stored in the same entity cluster) if they appear in the same Wikipedia infobox or Freebase record in either the 'name' or 'alias' field, or they appear in the same synset in WordNet. As an example, this sieve correctly detects *America Online* and *AOL* as aliases. We also tested the utility of Wikipedia categories, but found little gain over morpho-syntactic features.

Lexical Chain Sieve

This sieve marks two nominal mentions as coreferent if they are linked by a WordNet lexical chain that traverses hypernymy or synonymy relations. We use all synsets for each mention, but restrict it to mentions that are at most three sentences apart, and lexical chains of length at most four. This sieve correctly links *Britain* with *country*, and *plane* with *aircraft*.

To increase the precision of the above two sieves, we use additional constraints before two mentions can match: attribute agreement (number, gender, animacy, named entity labels), no i-within-i, no location or numeric mismatches (as in Section 2.3.1), and we do not use the abstract entity synset in WordNet, except in chains that include 'organization'.

2.3.3 Discourse Processing Sieve

This sieve matches speakers to compatible pronouns, using shallow discourse understanding to handle quotations and conversation transcripts. Although more complex discourse constraints have been proposed, it has been difficult to show improvements (Tetreault & Allen, 2003; 2004).

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 dataset.

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.

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 the above 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.

For example, *[my]* and *[he]* are not coreferent in the above example (third constraint).

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

Annotations	Coref	R	P	F1
Gold	Before	92.8	37.7	53.6
Gold	After	75.1	70.1	72.6
Not gold	Before	87.9	35.6	50.7
Not gold	After	71.7	68.4	70.0

Table 2: Performance of the mention detection component, before and after coreference resolution, with both gold and actual linguistic annotations.

2.4 Post Processing

To guarantee that the output of our system matches the shared task requirements and the OntoNotes annotation specification, we implement two post-processing steps:

- We discard singleton clusters.
- We discard the mention that appears later in text in appositive and 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.

3 Results and Discussion

Table 2 shows the performance of our mention detection algorithm. We show results before and after coreference resolution and post-processing (when singleton mentions are removed). We also list results with gold and predicted linguistic annotations (i.e., syntactic parses and named entity recognition). The table shows that the recall of our approach is 92.8% (if gold annotations are used) or 87.9% (with predicted annotations). In both cases, precision is low because our algorithm generates many spurious mentions due to its local nature. However, as the table indicates, many of these mentions are removed during post-processing, because they are assigned to singleton clusters during coreference resolution. The two main causes for our recall errors are lack of recognition of event mentions (e.g., verbal mentions such as *growing*) and parsing errors. Parsing errors often introduce incorrect mention boundaries, which yield both recall and precision errors. For example, our system generates the predicted mention, *the working meeting of the "863 Program" today*, for the gold mention *the working meeting of the "863 Program"*. Due to this boundary mismatch, all mentions found to be coreferent with this predicted mention are counted as precision errors, and all mentions in the same coreference cluster with the gold mention are counted as recall errors.

Table 3 lists the results of our end-to-end system on the development partition. "External Resources", which were used only in the open track, includes: (a) a hand-built list of genders of first names that we created, incorporating frequent names from census lists and other sources, (b) an animacy list (Ji and Lin, 2009), (c) a country and state gazetteer, and (d) a demonym list. "Discourse" stands for the sieve introduced in Section 2.3.3. "Semantics" stands for the sieves presented in Section 2.3.2. The table shows that the discourse sieve yields an improvement of almost 2 points to the overall score (row 1 versus 3), and external resources contribute 0.5 points. On the other hand, the semantic sieves do not help (row 3 versus 4). The latter result contradicts our initial experiments, where we measured a minor improvement when these sieves were enabled and gold mentions were used. Our hypothesis is that, when predicted mentions are used, the semantic sieves are more likely to link spurious mentions to existing clusters, thus introducing precision errors. This suggests that a different tuning of the sieve parameters is required for the predicted mention scenario. For this reason, we did not use the semantic sieves for our submission. Hence, rows 2 and 3 in the table show the performance of our official submission in the development set, in the closed and open tracks respectively.

The last three rows in Table 3 give insight on the impact of gold information. This analysis indicates that using gold linguistic annotation yields an improvement of only 2 points. This implies that the quality of current linguistic processors is sufficient for the task of coreference resolution. On the other hand, using gold mentions raises the overall score by 15 points. This clearly indicates that pipeline architectures where mentions are identified first are inadequate for this task, and that coreference resolution might benefit from the joint modeling of mentions and coreference chains.

Finally, Table 4 lists our results on the held-out testing partition. Note that in this dataset, the gold mentions included singletons and generic mentions

Components					MUC			B^3			CEAFE			BLANC			
ER	D	S	GA	GM	R	P	F1	R	P	F1	R	P	F1	R	P	F1	avg F1
√					58.8	56.5	57.6	68.0	68.7	68.4	44.8	47.1	45.9	68.8	73.5	70.9	57.3
	√				59.1	57.5	58.3	69.2	71.0	70.1	46.5	48.1	47.3	72.2	78.1	74.8	58.6
√	√				60.1	59.5	59.8	69.5	71.9	70.7	46.5	47.1	46.8	73.8	78.6	76.0	59.1
√	√	√			60.3	58.5	59.4	69.9	71.1	70.5	45.6	47.3	46.4	73.9	78.2	75.8	58.8
√	√		√		63.8	61.5	62.7	71.4	72.3	71.9	47.1	49.5	48.3	75.6	79.6	77.5	61.0
√	√			√	73.6	90.0	81.0	69.8	89.2	78.3	79.4	52.5	63.2	79.1	89.2	83.2	74.2
√	√		√	√	74.0	90.1	81.3	70.2	89.3	78.6	79.7	53.1	63.7	79.5	89.6	83.6	74.5

Table 3: Comparison between various configurations of our system. ER, D, S stand for External Resources, Discourse, and Semantics sieves. GA and GM stand for Gold Annotations, and Gold Mentions. The top part of the table shows results using only predicted annotations and mentions, whereas the bottom part shows results of experiments with gold information. Avg F1 is the arithmetic mean of MUC, B^3, and CEAFE. We used the development partition for these experiments.

		MUC			B^3			CEAFE			BLANC			
Track	Gold Mention Boundaries	R	P	F1	R	P	F1	R	P	F1	R	P	F1	avg F1
Close	Not Gold	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
Open	Not Gold	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
Close	Gold	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
Open	Gold	66.9	63.9	65.4	70.1	71.5	70.8	46.3	49.6	47.9	73.4	79.0	75.8	61.4

Table 4: Results on the official test set.

as well, whereas in development (lines 6 and 7 in Table 3), gold mentions included only mentions part of an actual coreference chain. This explains the large difference between, say, line 6 in Table 3 and line 4 in Table 4.

Our scores are comparable to previously reported state-of-the-art results for coreference resolution with predicted mentions. For example, Haghighi and Klein (2010) compare four state-of-the-art systems on three different corpora and report B^3 scores between 63 and 77 points. While the corpora used in (Haghighi and Klein, 2010) are different from the one in this shared task, our result of 68 B^3 suggests that our system's performance is competitive. In this task, our submissions in both the open and the closed track obtained the highest scores.

4 Conclusion

In this work we showed how a competitive end-to-end coreference resolution system can be built using only deterministic models (or sieves). Our approach starts with a high-recall mention detection component, which identifies mentions using only syntactic information and named entity boundaries, followed by a battery of high-precision deterministic coreference sieves, applied one at a time from highest to lowest precision. These models incorporate lexical, syntactic, semantic, and discourse information, and have access to document-level information (i.e., we share mention attributes across clusters as they are built). For this shared task, we extended our existing system with new sieves that model shallow discourse (i.e., speaker identification) and semantics (lexical chains and alias detection). Our results demonstrate that, despite their simplicity, deterministic models for coreference resolution obtain competitive results, e.g., we obtained the highest scores in both the closed and open tracks (57.8 and 58.3 respectively). The code used for this shared task is publicly released.[5]

Acknowledgments

We thank the shared task organizers for their effort.

This material is based upon work supported by the Air Force Research Laboratory (AFRL) under prime contract no. FA8750-09-C-0181. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the Air Force Research Laboratory (AFRL).

[5]See `http://nlp.stanford.edu/software/dcoref.shtml` for the standalone coreference resolution system and `http://nlp.stanford.edu/software/corenlp.shtml` for Stanford's suite of natural language processing tools, which includes this coreference resolution system.

References

B. Baldwin. 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*.

E. Bengston & D. Roth. 2008. Understanding the value of features for coreference resolution. In *EMNLP*.

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. *Extracting Complex Biological Events with Rich Graph-Based Feature Sets*. Proceedings of the Workshop on BioNLP: Shared Task.

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

B. A. Fox 1993. *Discourse structure and anaphora: written and conversational English*. Cambridge University Press.

A. Haghighi and D. Klein. 2010. Coreference resolution in a modular, entity-centered model. In Proc. of *HLT-NAACL*.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel 2006. OntoNotes: The 90% Solution. In *HLT/NAACL*.

Z. Huang, G. Zeng, W. Xu, and A. Celikyilmaz 2009. Accurate semantic class classifier for coreference resolution. In *EMNLP*.

J.R. Hobbs. 1977. Resolving pronoun references. *Lingua*.

H. Ji and D. Lin. 2009. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *PACLIC*.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. *Overview of the BioNLP'09 Shared Task on Event Extraction*. Proceedings of the NAACL-HLT 2009 Workshop on Natural Language Processing in Biomedicine (BioNLP'09).

V. Ng 2007. Semantic Class Induction and Coreference Resolution. In *ACL*.

V. Ng and C. Cardie. 2002. Improving Machine Learning Approaches to Coreference Resolution. in *ACL 2002*

S. Ponzetto and M. Strube. 2006. Exploiting semantic role labeling, Wordnet and Wikipedia for coreference resolution. *Proceedings of NAACL*.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted Coreference: Indentifying Entities and Events in OntoNotes. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*.

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

K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning 2010. A Multi-Pass Sieve for Coreference Resolution. In *EMNLP*.

J. Tetreault and J. Allen. 2003. An Empirical Evaluation of Pronoun Resolution and Clausal Structure. In *Proceedings of the 2003 International Symposium on Reference Resolution*.

J. Tetreault and J. Allen. 2004. Dialogue Structure and Pronoun Resolution. In *DAARC*.

X. Yang and J. Su. 2007. Coreference Resolution Using Semantic Relatedness Information from Automatically Discovered Patterns. In *ACL*.