

# Illinois-Coref: The UI System in the CoNLL-2012 Shared Task

Kai-Wei Chang Rajhans Samdani Alla Rozovskaya Mark Sammons Dan Roth

University of Illinois at Urbana-Champaign

{kchang10 | rsamdan2 | rozovska | mssammon | danr}@illinois.edu

## Abstract

The CoNLL-2012 shared task is an extension of the last year’s coreference task. We participated in the closed track of the shared tasks in both years. In this paper, we present the improvements of *Illinois-Coref* system from last year. We focus on improving mention detection and pronoun coreference resolution, and present a new learning protocol. These new strategies boost the performance of the system by 5% MUC F1, 0.8% BCUB F1, and 1.7% CEAF F1 on the OntoNotes-5.0 development set.

## 1 Introduction

Coreference resolution has been a popular topic of study in recent years. In the task, a system requires to identify denotative phrases (“mentions”) and to cluster the mentions into equivalence classes, so that the mentions in the same class refer to the same entity in the real world.

Coreference resolution is a central task in the Natural Language Processing research. Both the CoNLL-2011 (Pradhan et al., 2011) and CoNLL-2012 (Pradhan et al., 2012) shared tasks focus on resolving coreference on the OntoNotes corpus. We also participated in the CoNLL-2011 shared task. Our system (Chang et al., 2011) ranked first in two out of four scoring metrics (BCUB and BLANC), and ranked third in the average score. This year, we further improve the system in several respects. In Sec. 2, we describe the Illinois-Coref system for the CoNLL-2011 shared task, which we take as the baseline. Then, we discuss the improvements on mention detection (Sec. 3.1), pronoun resolution (Sec. 3.2), and learning algorithm (Sec. 3.3).

Section 4 shows experimental results and Section 5 offers a brief discussion.

## 2 Baseline System

We use the Illinois-Coref system from CoNLL-2011 as the basis for our current system and refer to it as the *baseline*. We give a brief outline here, but focus on the innovations that we developed; a detailed description of the last year’s system can be found in (Chang et al., 2011).

The *Illinois-Coref* system uses a machine learning approach to coreference, with an inference procedure that supports straightforward inclusion of domain knowledge via constraints.

The system first uses heuristics based on Named Entity recognition, syntactic parsing, and shallow parsing to identify candidate mentions. A pairwise scorer  $\mathbf{w}$  generates compatibility scores  $w_{uv}$  for pairs of candidate mentions  $u$  and  $v$  using extracted features  $\phi(u, v)$  and linguistic constraints  $c$ .

$$w_{uv} = \mathbf{w} \cdot \phi(u, v) + c(u, v) + t, \quad (1)$$

where  $t$  is a threshold parameter (to be tuned). An inference procedure then determines the optimal set of links to retain, incorporating constraints that may override the classifier prediction for a given mention pair. A post-processing step removes mentions in singleton clusters.

Last year, we found that a *Best-Link* decoding strategy outperformed an *All-Link* strategy. The *Best-Link* approach scans candidate mentions in a document from left to right. At each mention, if certain conditions are satisfied, the pairwise scores of all previous mentions are considered, together with any constraints that apply. If one or more viable

links is available, the highest-scoring link is selected and added to the set of coreference links. After the scan is complete, the transitive closure of edges is taken to generate the coreference clusters, each cluster corresponding to a single predicted entity in the document.

The formulation of this best-link solution is as follows. For two mentions  $u$  and  $v$ ,  $u < v$  indicates that the mention  $u$  precedes  $v$  in the document. Let  $y_{uv}$  be a binary variable, such that  $y_{uv} = 1$  only if  $u$  and  $v$  are in the same cluster. For a document  $d$ , *Best-Link* solves the following formulation:

$$\begin{aligned} \arg \max_y \quad & \sum_{u,v:u<v} w_{uv} y_{uv} \\ \text{s.t.} \quad & \sum_{u<v} y_{uv} \leq 1 \quad \forall v, \\ & y_{uv} \in \{0, 1\}. \end{aligned} \quad (2)$$

Eq. (2) generates a set of connected components and the set of mentions in each connected component constitute an entity. Note that we solve the above *Best-Link* inference using an efficient algorithm (Bengtson and Roth, 2008) which runs in time quadratic in the number of mentions.

### 3 Improvements over the Baseline System

Below, we describe improvements introduced to the baseline *Illinois-Coref* system.

#### 3.1 Mention Detection

Mention detection is a crucial component of an end-to-end coreference system, as mention detection errors will propagate to the final coreference chain. *Illinois-Coref* implements a high recall and low precision rule-based system that includes all noun phrases, pronouns and named entities as candidate mentions. The error analysis shows that there are two main types of errors.

**Non-referential Noun Phrases.** Non-referential noun phrases are candidate noun phrases, identified through a syntactic parser, that are unlikely to refer to any entity in the real world (e.g., “the same time”). Note that because singleton mentions are not annotated in the OntoNotes corpus, such phrases are not considered as mentions. Non-referential noun phrases are a problem, since during the coreference stage they may be incorrectly linked to a valid mention, thereby decreasing the precision of the system.

To deal with this problem, we use the training data to count the number of times that a candidate noun phrase happens to be a gold mention. Then, we remove candidate mentions that frequently appear in the training data but never appear as gold mentions. Relaxing this approach, we also take the predicted head word and the words before and after the mention into account. This helps remove noun phrases headed by a preposition (e.g., the noun “fact” in the phrase “in fact”). This strategy will slightly degrade the recall of mention detection, so we tune a threshold learned on the training data for the mention removal.

**Incorrect Mention Boundary.** A lot of errors in mention detection happen when predicting mention boundaries. There are two main reasons for boundary errors: parser mistakes and annotation inconsistencies. A mistake made by the parser may be due to a wrong attachment or adding extra words to a mention. For example, if the parser attaches the relative clause inside of the noun phrase “President Bush, who traveled to China yesterday” to a different noun, the algorithm will predict “President Bush” as a mention instead of “President Bush, who traveled to China yesterday”; thus it will make an error, since the gold mention also includes the relative clause. In this case, we prefer to keep the candidate with a larger span. On the other hand, we may predict “President Bush at Dayton” instead of “President Bush”, if the parser incorrectly attaches the prepositional phrase. Another example is when extra words are added, as in “Today President Bush”.

A correct detection of mention boundaries is crucial to the end-to-end coreference system. The results in (Chang et al., 2011, Section 3) show that the baseline system can be improved from 55.96 avg F1 to 56.62 in avg F1 by using gold mention boundaries generated from a gold annotation of the parsing tree and the name entity tagging. However, fixing mention boundaries in an end-to-end system is difficult and requires additional knowledge. In the current implementation, we focus on a subset of mentions to further improve the mention detection stage of the baseline system. Specifically, we fix mentions starting with a stop word and mentions ending with a punctuation mark. We also use training data to learn patterns of inappropriate mention boundaries. The mention candidates that match the patterns are re-

moved. This strategy is similar to the method used to remove non-referential noun phrases.

As for annotation inconsistency, we find that in a few documents, a punctuation mark or an apostrophe used to mark the possessive form are inconsistently added to the end of a mention. The problem results in an incorrect matching between the gold and predicted mentions and downgrades the performance of the learned model. Moreover, the incorrect mention boundary problem also affects the training phase because our system is trained on a union set of the predicted and gold mentions. To fix this problem, in the training phase, we perform a relaxed matching between predicted mentions and gold mentions and ignore the punctuation marks and mentions that start with one of the following: adverb, verb, determiner, and cardinal number. For example, we successfully match the predicted mention “now the army” to the gold mention “the army” and match the predicted mention “Sony ’s” to the gold mention “Sony.” Note that we cannot fix the inconsistency problem in the test data.

### 3.2 Pronoun Resolution

The baseline system uses an identical model for coreference resolution on both pronouns and non-pronominal mentions. However, in the literature (Bengtson and Roth, 2008; Rahman and Ng, 2011; Denis and Baldrige, 2007) the features for coreference resolution on pronouns and non-pronouns are usually different. For example, lexical features play an important role in non-pronoun coreference resolution, but are less important for pronoun anaphora resolution. On the other hand, gender features are not as important in non-pronoun coreference resolution.

We consider training two separate classifiers with different sets of features for pronoun and non-pronoun coreference resolution. Then, in the decoding stage, pronoun and non-pronominal mentions use different classifiers to find the best antecedent mention to link to. We use the same features for non-pronoun coreference resolution, as the baseline system. For the pronoun anaphora classifier, we use a set of features described in (Denis and Baldrige, 2007), with some additional features. The augmented feature set includes features to identify if a pronoun or an antecedent is a speaker in the sen-

---

### Algorithm 1 Online Latent Structured Learning for Coreference Resolution

---

Loop until convergence:

For each document  $D_t$  and each  $v \in D_t$

1. Let  $u^* = \max_{u \in y(v)} \mathbf{w}^T \phi(u, v)$ , and
  2.  $u' = \max_{u \in \{u < v\} \cup \{\emptyset\}} \mathbf{w}^T \phi(u, v) + \Delta(u, v, y(v))$
  3. Let  $\mathbf{w} \leftarrow \mathbf{w} + \eta \mathbf{w}^T (\phi(u', v) - \phi(u^*, v))$ .
- 

tence. It also includes features to reflect the document type. In Section 4, we will demonstrate the improvement of using separate classifiers for pronoun and non-pronoun coreference resolution.

### 3.3 Learning Protocol for Best-Link Inference

The baseline system applies the strategy in (Bengtson and Roth, 2008, Section 2.2) to learn the pairwise scoring function  $\mathbf{w}$  using the Averaged Perceptron algorithm. The algorithm is trained on mention pairs generated on a per-mention basis. The examples are generated for a mention  $v$  as

- Positive examples:  $(u, v)$  is used as a positive example where  $u < v$  is the closest mention to  $v$  in  $v$ ’s cluster
- Negative examples: for all  $w$  with  $u < w < v$ ,  $(w, v)$  forms a negative example.

Although this approach is simple, it suffers from a severe label imbalance problem. Moreover, it does not relate well to the best-link inference, as the decision of picking the closest preceding mention seems rather ad-hoc. For example, consider three mentions belonging to the same cluster:  $\{m_1$ : “President Bush”,  $m_2$ : “he”,  $m_3$ : “George Bush”}. The baseline system always chooses the pair  $(m_2, m_3)$  as a positive example because  $m_2$  is the closest mention of  $m_3$ . However, it is more proper to learn the model on the positive pair  $(m_1, m_3)$ , as it provides more information. Since the *best links* are not given but are latent in our learning problem, we use an online latent structured learning algorithm (Connor et al., 2011) to address this problem.

We consider a structured problem that takes mention  $v$  and its preceding mentions  $\{u \mid u < v\}$  as inputs. The output variables  $y(v)$  is the set of antecedent mentions that co-refer with  $v$ . We define a latent structure  $\mathbf{h}(v)$  to be the bestlink decision of  $v$ . It takes the value  $\emptyset$  if  $v$  is the first mention

| Method                              | Without Separating Pronouns |       |       |       |       | With Separating Pronouns |       |       |       |       |
|-------------------------------------|-----------------------------|-------|-------|-------|-------|--------------------------|-------|-------|-------|-------|
|                                     | MD                          | MUC   | BCUB  | CEAF  | AVG   | MD                       | MUC   | BCUB  | CEAF  | AVG   |
| <i>Binary Classifier (baseline)</i> | 70.53                       | 61.63 | 69.26 | 43.03 | 57.97 | 73.24                    | 64.57 | 69.78 | 44.95 | 59.76 |
| <i>Latent-Structured Learning</i>   | 73.02                       | 64.98 | 70.00 | 44.48 | 59.82 | 73.95                    | 65.75 | 70.25 | 45.30 | 60.43 |

Table 1: The performance of different learning strategies for best-link decoding algorithm. We show the results with/without using separate pronoun anaphora resolver. The systems are trained on the TRAIN set and evaluated on the **CoNLL-2012 DEV** set. We report the F1 scores (%) on mention detection (MD) and coreference metrics (MUC, BCUB, CEAF). The column AVG shows the averaged scores of the three coreference metrics.

| System   | MD    | MUC   | BCUB  | CEAF  | AVG   |
|----------|-------|-------|-------|-------|-------|
| Baseline | 64.58 | 55.49 | 69.15 | 43.72 | 56.12 |
| New Sys. | 70.03 | 60.65 | 69.95 | 45.39 | 58.66 |

Table 2: The improvement of *Illinois-Coref*. We report the F1 scores (%) on the DEV set from **CoNLL-2011** shared task. Note that the CoNLL-2011 data set does not include corpora of bible and of telephone conversation.

in the equivalence class, otherwise it takes values from  $\{u \mid u < v\}$ . We define a loss function  $\Delta(\mathbf{h}(v), v, y(v))$  as

$$\Delta(\mathbf{h}(v), v, y(v)) = \begin{cases} 0 & \mathbf{h}(v) \in y(v), \\ 1 & \mathbf{h}(v) \notin y(v). \end{cases}$$

We further define the feature vector  $\phi(\emptyset, v)$  to be a zero vector and  $\eta$  to be the learning rate in Perceptron algorithm. Then, the weight vector  $\mathbf{w}$  in (1) can be learned from Algorithm 1. At each step, Alg. 1 picks a mention  $v$  and finds the Best-Link decision  $u^*$  that is consistent with the gold cluster. Then, it solves a loss-augmented inference problem to find the best link decision  $u'$  with current model ( $u' = \emptyset$  if the classifier decides that  $v$  does not have coreferent antecedent mention). Finally, the model  $\mathbf{w}$  is updated by the difference between the feature vectors  $\phi(u', v)$  and  $\phi(u^*, v)$ .

Alg. 1 makes learning more coherent with inference. Furthermore, it naturally solves the data imbalance problem. Lastly, this algorithm is fast and converges very quickly.

## 4 Experiments and Results

In this section, we demonstrate the performance of *Illinois-Coref* on the OntoNotes-5.0 data set. A previous experiment using an earlier version of this data

can be found in (Pradhan et al., 2007). We first show the improvement of the mention detection system. Then, we compare different learning protocols for coreference resolution. Finally, we show the overall performance improvement of *Illinois-Coref* system.

First, we analyze the performance of mention detection before the coreference stage. Note that singleton mentions are included since it is not possible to identify singleton mentions before running coreference. They are removed in the post-processing stage. The mention detection performance of the end-to-end system will be discussed later in this section. With the strategy described in Section 3.1, we improve the F1 score for mention detection from 55.92% to 57.89%. Moreover, we improve the detection performance on short named entity mentions (name entity with less than 5 words) from 61.36 to 64.00 in F1 scores. Such mentions are more important because they are easier to resolve in the coreference layer.

Regarding the learning algorithm, Table 1 shows the performance of the two learning protocols with/without separating pronoun anaphora resolver. The results show that both strategies of using a pronoun classifier and training a latent structured model with an online algorithm improve the system performance. Combining the two strategies, the avg F1 score is improved by 2.45%.

Finally, we compare the final system with the baseline system. We evaluate both systems on the CoNLL-11 DEV data set, as the baseline system is tuned on it. The results show that *Illinois-Coref* achieves better scores on all the metrics. The mention detection performance after coreference resolution is also significantly improved.

| Task                              | MD     | MUC   | BCUB  | CEAF  | AVG   |
|-----------------------------------|--------|-------|-------|-------|-------|
| English (Pred. Mentions)          | 74.32  | 66.38 | 69.34 | 44.81 | 60.18 |
| English (Gold Mention Boundaries) | 75.72  | 67.80 | 69.75 | 45.12 | 60.89 |
| English (Gold Mentions)           | 100.00 | 85.74 | 77.46 | 68.46 | 77.22 |
| Chinese (Pred Mentions)           | 47.58  | 37.93 | 63.23 | 35.97 | 45.71 |

Table 3: The results of our submitted system on the TEST set. The systems are trained on a collection of TRAIN and DEV sets.

#### 4.1 Chinese Coreference Resolution

We apply the same system to Chinese coreference resolution. However, because the pronoun properties in Chinese are different from those in English, we do not train separate classifiers for pronoun and non-pronoun coreference resolution. Our Chinese coreference resolution on Dev set achieves 37.88% MUC, 63.37% BCUB, and 35.78% CEAF in F1 score. The performance for Chinese coreference is not as good as the performance of the coreference system for English. One reason for that is that we use the same feature set for both Chinese and English systems, and the feature set is developed for the English corpus. Studying the value of strong features for Chinese coreference resolution system is a potential topic for future research.

#### 4.2 Test Results

Table 3 shows the results obtained on TEST, using the best system configurations found on DEV. We report results on both English and Chinese coreference resolution on predicted mentions with predicted boundaries. For English coreference resolution, we also report the results when using gold mentions and when using gold mention boundaries<sup>1</sup>.

### 5 Conclusion

We described strategies for improving mention detection and proposed an online latent structure algorithm for coreference resolution. We also proposed using separate classifiers for making Best-Link decisions on pronoun and non-pronoun mentions. These strategies significantly improve the *Illinois-Coref* system.

<sup>1</sup>Note that, in Ontonotes annotation, specifying gold mentions requires coreference resolution to exclude singleton mentions. Gold mention boundaries are provided by the task organizers and include singleton mentions.

**Acknowledgments** This research is supported by the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181 and the Army Research Laboratory (ARL) under agreement W911NF-09-2-0053. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA, AFRL, ARL or the US government.

#### References

- E. Bengtson and D. Roth. 2008. Understanding the value of features for coreference resolution. In *EMNLP*.
- K. Chang, R. Samdani, A. Rozovskaya, N. Rizzolo, M. Sammons, and D. Roth. 2011. Inference protocols for coreference resolution. In *CoNLL*.
- M. Connor, C. Fisher, and D. Roth. 2011. Online latent structure training for language acquisition. In *IJCAI*.
- P. Denis and J. Baldridge. 2007. A ranking approach to pronoun resolution. In *IJCAI*.
- S. Pradhan, L. Ramshaw, R. Weischedel, J. MacBride, and L. Micciulla. 2007. Unrestricted Coreference: Identifying Entities and Events in OntoNotes. In *ICSC*.
- S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *CoNLL*.
- S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *CoNLL*.
- A. Rahman and V. Ng. 2011. Narrowing the modeling gap: a cluster-ranking approach to coreference resolution. *Journal of AI Research*, 40(1):469–521.