Hybrid Rule-based Algorithm for Coreference Resolution [*]

Heming Shou[1,2]   **Hai Zhao**[1,2†]

[1]Center for Brain-Like Computing and Machine Intelligence,
Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems
Shanghai Jiao Tong University
shouhm@gmail.com, zhaohai@cs.sjtu.edu.cn

Abstract

This paper describes our coreference resolution system for the CoNLL-2012 shared task. Our system is based on the Stanford's *dcoref* deterministic system which applies multiple sieves with the order from high precision to low precision to generate coreference chains. We introduce the newly added constraints and sieves and discuss the improvement on the original system. We evaluate the system using OntoNotes data set and report our results of average F-score 58.25 in the closed track.

1 Introduction

In this paper, our coreference resolution system for CoNLL-2012 shared task (Pradhan et al., 2012) is summarized. Our system is an extension of Stanford's multi-pass sieve system, (Raghunathan et al., 2010) and (Lee et al., 2011), by adding novel constraints and sieves. In the original model , sieves are sorted in decreasing order of precision. Initially each mention is in its own cluster. Mention clusters are combined by satisfying the condition of each sieve in the scan pass. Through empirical studies, we proposed some extensions and algorithms for furthermore enhancing the performance.

Many other existing systems applied supervised or unsupervised (Haghighi and Klein, 2010) learning models. The classical resolution algorithm was proposed by (Soon et al., 2001). Semantic knowledge like word associations was involved by (Kobdani et al., 2011). Most of the supervised learning models in CoNLL-2011 shared task (Chang et al., 2011)(Björkelund and Nugues, 2011) used classifiers (Maximum Entropy or SVM) to train the models for obtaining the pairwise mention scores. However, the training process usually takes much longer time than unsupervised or deterministic systems. In contrast, (Raghunathan et al., 2010) proposed a rule-based model which obtained competitive result with less time.

Two considerable extensions to the Stanford model in this paper are made to guarantee higher precision and recall. First, we recorded error patterns from outputs of the original Stanford system and found that the usual errors are mention boundary mismatches, pronoun mismatches and so on. To avoid the irrational coreference errors, we added some constraints to the mention detection for eliminating some unreasonable mention boundary mismatches. Second, we added some constraints in the coreference sieves based on the errors on the training set and the development set.

We participated in the closed track and received an official F-score (unweighted mean of MUC, BCUBED and CEAF(E) metric) of 58.25 for English. The system with our extensions is briefly introduced in Section 2. We report our evaluation results and discuss in Section 3.

118

2 System Architecture

The original Stanford system consists of three stages: mention detection, coreference resolution and post-processing. The mention detection stage is for extracting mentions with a relative high recall. The coreference resolution stage uses multiple sieves to generate coreference clusters. The post-processing stage makes the output compatible with the shared task and OntoNotes specifications (Pradhan et al., 2007), e.g. removing singletons, appositive, predicate nominatives and relative pronouns.

2.1 Mention Detection

Our system mainly focuses on making extensions for mention detection and coreference resolution. From error analysis, we found that mention boundaries caused many precision and recall errors. For example, for the gold mention *Robert H. Chandross, an economist for Lloyd's Bank in New York*, the original system only extracts *Robert H. Chandross* as the mention and links it with *he* in the following sentence. This mismatch leads to both precision and recall errors since the mention with longer boundary is not detected but the shorter one is used. Another example which omits *today* in the phrase for the predicted mention is mentioned in (Lee et al., 2011) and this boundary mismatch also accounts for precision and recall errors. Some other examples may be like this: *Auto prices had a big effect in the PPI, and at the CPI level they won't*, the gold mentions are *Auto prices*, *the PPI*, *the CPI level* and *they* while the original system only finds out *auto prices*. Considering these boundary mismatches, it is not hard for us to categorize the error types.

By observation, most boundary problems happen in the following cases:

- The predicted mention is embedded in the gold mention.

- The gold mention is embedded in the predicted mention.

- Some gold mentions are totally omitted.

It is very rare for the case that predicted mention overlaps with the gold mention but no one includes the other.

For the first and second cases, some analysis and constraint about prefix and postfix of phrases are applied to get predicted mentions as precise as gold mentions. For the example mentioned above, the clause *,an economist ...* which modifies the person *Robert H. Chandross* is annexed to the person name mention. We also append time and other modifiers to the original mention. As for the third case, we allow more pronouns and proper nouns to be added to the list of mentions.

2.2 Sieve Coreference

Like the constraints on the extension to the mention detection stage, our system also generates error reports for the sieve passes. While our system is rule-based and it also works without training data sets, some statistical information is also helpful to detect and avoid errors.

The first extension we used is a direct way to utilize the training data and the development data. We simply record the erroneous mention pairs in the train and development sets with distance and sieve information. One of the most common errors is that when mentions with particular types appear twice in the same sentence, the original system often puts them into the same cluster. For example, there are often two or more *you* or person names in the dialogue, however, the different occurrences are treated as coreference which produces precision errors. To address this problem, we convert proper nouns to type designator, e.g. *Paul* as *Man Name*. Then we use the formatted error pairs as constraints on the sieve passes since some pairs mostly cause precision errors. If the checking pair matches up some records in the errors with the same sieve information and the error frequency is over a threshold, we must discard this pair in this sieve pass.

Another difference between our system and the Stanford system is the semantic similarity sieve. For each sieve pass, the current clusters are built by stronger sieves (sieves in the earlier passes). The Stanford system selects the most representative mention from a mention cluster to query for semantic information. The preference order is:

1. mentions headed by proper nouns

2. mentions headed by common nouns

3. nominal mentions

4. pronominal mentions

In our system, we not only select the most representative one but compare all the types above, i.e, select the longest string in each type of this cluster. When applying semantic sieves, we also compare representative mention for each type and make synthesized decisions by the number of types which have similar semantic meanings.

We also made some modifications on the sieves and their ordering in the original system. For *Proper Head Word Match* mentioned in (Lee et al., 2011), the Pronoun distance which indicates sentence distance limit between a pronoun and its antecedent. We change the value from 3 to 2.

3   Experiments and Results

Table 1: CoNLL-2012 Shared Task Test Results

Metric	Recall	Precision	F1
MD	75.35	72.08	73.68
MUC	63.46	62.39	62.92
BCUBED	65.31	68.90	67.05
CEAF(M)	55.68	55.68	55.68
CEAF(E)	44.20	45.35	44.77
BLANC	69.43	75.08	71.81
OFFICIAL	-	-	58.25

Table 2: Comparison between original system and our system on the development set

metric	original	our system
MUC F	61.64	62.31
MUC P	58.65	59.58
MUC R	64.95	65.29
BCUBED F	68.61	69.87
BCUBED P	67.23	68.81
BCUBED R	70.04	70.97

Our system enhanced the precision and recall of the original system of (Lee et al., 2011). The table 1. shows the official result for the CoNLL-2012 shared task. The recall of our mention detection approach is 75.35% while the precision is 72.08%. The final official score 58.25 is the unweighed mean of

MUC, BCUBED and CEAF(E). Although the test set is different from that of the previous year, comparing with the original system, our result of MD and MUC shows that our improvement is meaningful. The table 2. indicates the improvement from our system over the original system evaluated by the development set. Since experiments with semantic knowledge like WordNet and Wikipedia cannot give better performance, we omit the semantic function for generating test result. Our explanation is that the predicted mentions are still not precise enough and the fuzziness of the semantic knowledge might cause conflicts with our sieves. If the semantic knowledge tells that two mentions are similar and possibly can be combined while they do not satisfy the sieve constraints, it will be very hard to make a decision since we cannot find an appropriate threshold to let the semantic suggestion pass through.

4   Conclusion

In this paper we made a series of improvements on the existing Stanford system which only uses deterministic rules. Since the rules are high dimensional, i.e., the rules that are adopted in the system may depend on the states of the ongoing clustering process, it is not feasible to apply it in the statistical learning methods since take the intermediate results into consideration will be. The experimental results show that our improvements are effective. For this task, we added constraints on the mention detection stage and the coreference resolution stage. We also added new sieves and conduct a group of empirical studies on semantic knowledge. Our results give a demonstration that the deterministic model for coreference resolution is not only simple and competitive but also has high extendibility.

References

Anders Björkelund and Pierre Nugues. 2011. Exploring lexicalized features for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 45–50, Portland, Oregon, USA, June. Association for Computational Linguistics.

Kai-Wei Chang, Rajhans Samdani, Alla Rozovskaya, Nick Rizzolo, Mark Sammons, and Dan Roth. 2011. Inference protocols for coreference resolution. In *Proceedings of the Fifteenth Conference on Computation-*

al Natural Language Learning: Shared Task, pages 40–44, Portland, Oregon, USA, June. Association for Computational Linguistics.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, Los Angeles, California, June. Association for Computational Linguistics.

Hamidreza Kobdani, Hinrich Schuetze, Michael Schiehlen, and Hans Kamp. 2011. Bootstrapping coreference resolution using word associations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 783–792, Portland, Oregon, USA, June. Association for Computational Linguistics.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA, June. Association for Computational Linguistics.

Sameer S. Pradhan, Lance A. Ramshaw, Ralph M. Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in ontonotes. In *ICSC*, pages 446–453.

Sameer Pradhan, 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 2012)*, Jeju, Korea.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of EMNLP 2010*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.*, 27(4):521–544, December.