Link Type Based Pre-Cluster Pair Model for Coreference Resolution

Yang Song[†], Houfeng Wang[†] and Jing Jiang[‡]
[†]Key Laboratory of Computational Linguistics (Peking University) Ministry of Education,China
[‡]School of Information Systems, Singapore Management University, Singapore
{ysong, wanghf}@pku.edu.cn, jingjiang@smu.edu.sg

Abstract

This paper presents our participation in the CoNLL-2011 shared task, Modeling Unrestricted Coreference in OntoNotes. Coreference resolution, as a difficult and challenging problem in NLP, has attracted a lot of attention in the research community for a long time. Its objective is to determine whether two mentions in a piece of text refer to the same entity. In our system, we implement mention detection and coreference resolution seperately. For mention detection, a simple classification based method combined with several effective features is developed. For coreference resolution, we propose a link type based pre-cluster pair model. In this model, pre-clustering of all the mentions in a single document is first performed. Then for different link types, different classification models are trained to determine wheter two pre-clusters refer to the same entity. The final clustering results are generated by closest-first clustering method. Official test results for closed track reveal that our method gives a MUC F-score of 59.95%, a B-cubed F-score of 63.23%, and a CEAF F-score of 35.96% on development dataset. When using gold standard mention boundaries, we achieve MUC F-score of 55.48%, B-cubed F-score of 61.29%, and CEAF F-score of 32.53%.

1 Introduction

The task of coreference resolution is to recognize all the mentions (also known as noun phrases, including names, nominal mentions and pronouns) in a text and cluster them into equivalence classes where each quivalence class refers to a real-world entity or abstract concept. The CoNLL-2011 shared task[1] uses OntoNotes[2] as the evaluation corpus. The coreference layer in OntoNotes constitutes one part of a multi-layer, integrated annotation of the shallow semantic structures in the text with high inter-annotator agreement. In addition to coreference, this data set is also tagged with syntactic trees, high coverage verb and some noun propositions, partial verb and noun word senses, and 18 named entity types. The main difference between OntoNotes and another wellknown coreference dataset ACE is that the former does not label any singleton entity cluster, which has only one reference in the text. We can delete all the singleton clusters as a postprocessing step for the final results. Alternatively, we can also first train a classifier to separate singleton mentions from the rest and apply this mention detection step before coreference resolution. In this work we adopt the second strategy.

In our paper, we use a traditional learning based pair-wise model for this task. For mention detection, we first extract all the noun phrases in the text and then use a classification model combined with some effective features to determine whether each noun phrase is actually a mention. The features include word features, POS features in the given noun phrase and its context, string matching feature in its context, SRL features, and named entity features among others. More details will be given in Section 3. From our in-house experiments, the final F-scores for coreference resolution can be improved by this mention detection part. For coreference res-

[1]http://conll.bbn.com
[2]http://www.bbn.com/ontonotes/

Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 131–135,
Portland, Oregon, 23-24 June 2011.

Features describing c_i or c_j	
Words	The first and last words of the given NP in c_i (or c_j) , also including the words in the context with a window size 2
POS Tags	The part of speech tags corresponding to the words
Pronoun	Y if mentions in c_i(or c_j) are pronouns; else N
Definite	Y if mentions in c_i(or c_j) are definite NP; else N
Demonstrative	Y if mentions in c_i(or c_j) are demonstrative NP; else N
Number	Singular or Plural, determined using a data file published by Bergsma and Lin (2006)
Gender	Male, Female, Neuter, or Unknown, determined using a data file published by Bergsma and Lin (2006)
Semantic Class	Semantic Classes are given by OntoNotes for named entities
Mentino Type	Common Noun Phrases or Pronouns

Table 1: The feature set describing c_i or c_j.

olution, a traditinal pair-wise model is applied, in which we first use exact string matching to generate some pre-clusters. It should be noted that each pronoun must be treated as a singleton pre-cluster, because they are not like names or nominal mentions, which can be resolved effectively with exact string matching. We then implement a classification based pre-cluster pair model combined with several effective coreference resolution features to determine whether two pre-clusters refer to the same entity. Finally, we use closest-first clustering method to link all the coreferential pre-clusters and generate the final cluster results. As mentioned before, mentions have three types: names, nominal mentions and pronouns. Among them pronouns are very different from names and nominal mentions, because they can only supply limited information literally. So we define three kinds of link types for pre-cluster pairs: NP-NP link, NP-PRP link and PRP-PRP link. (Here NP means Noun Phrases and PRP means Pronominal Phrases.) One link represents one pre-cluster pair. Intuitively, different link types tend to use different features to determine whether this kind of link is coreferential or not. We implement three kinds of pre-cluster pair model based on three link types. Experimental results show that combined with outputs from different link type based pre-cluster pair model can give better results than using an unified classification model for three different kinds of link types. For all the classification models, we use opennlp.maxent[3] package.

The rest of this paper is organized as follows. Section 2 describes our mention detection method. We discuss our link type based pre-cluster pair model for coreference resolution in Section 3, evaluate it in Section 4, and conclude in Section 5.

2 Mention Detection

We select all the noun phrases tagged by the OntoNotes corpus as mention candidates and implement a classification-based model combined with several commonly used features to determine whether a given noun phrase is a mention. The features are given below:

- Word Features - They include the first word and the last word in each given noun phrase. We also use words in the context of the noun phrase within a window size of 2.
- POS Features - We use the part of speech tags of each word in the word features.
- Position Features - These features indicate where the given noun phrase appears in its sentence: begining, middle, or end.
- SRL Features - The Semantic Role of the given noun phrase in its sentence.
- Verb Features - The verb related to the Semantic Role of the given noun phrase.

[3]http://incubator.apache.org/opennlp/

Features describing the relationship between c_i and c_j	
Distance	The minimum distance between mentions in c_i and c_j
String Match	Y if mentions are the same string; else N
Substring Match	Y if one mention is a substring of another; else N
Levenshtein Distance	Levenshtein Distance between the mentions
Number Agreement	Y if the mentions agree in number; else N
Gender Agreement	Y if the mentions agree in gender; else N
N & G Agreement	Y if mentions agree in both number and gender; else N
Both Pronouns	Y if the mentions are both pronouns; else N
Verb Agreement	Y if the mentions have the same verb.
SRL Agreement	Y if the mentions have the same semantic role
Position Agreement	Y if the mentions have the same position (Beginning, Middle or End) in sentences

Table 2: The feature set describing the relationship between c_i and c_j.

- Entity Type Features - The named entity type for the given noun phrase.
- String Matching Features - True if there is another noun phrase wich has the same string as the given noun phrase in the context.
- Definite NP Features - True if the given noun phrase is a definite noun phrase.
- Demonstrative NP Features - True if the given noun phrase is a demonstrative noun phrase.
- Pronoun Features - True if the given noun phrase is a pronoun.

Intutively, common noun phrases and pronouns might have different feature preferences. So we train classification models for them respectively and use the respective model to predicate for common noun phrases or pronouns. Our mention detection model can give 52.9% recall, 80.77% precision and 63.93% F-score without gold standard mention boundaries on the development dataset. When gold standard mention boundaries are used, the results are 53.41% recall, 80.8% precision and 64.31% F-score. (By using the gold standard mention boundaries, we mean we use the gold standard noun phrase boundaries.)

3 Coreference Resolution

After getting the predicated mentions, we use some heuristic rules to cluster them with the purpose of generating highly precise pre-clusters. For this task

Metric	**Recall**	**Precision**	**F-score**
MUC	49.64%	67.18%	57.09%
BCUBED	59.42%	70.99%	64.69%
CEAF	45.68%	30.56%	36.63%
AVERAGE	51.58%	56.24%	52.80%

Table 3: Evaluation results on development dataset without gold mention boundaries

Metric	**Recall**	**Precision**	**F-score**
MUC	48.94%	67.72%	56.82%
BCUBED	58.52%	72.61%	64.81%
CEAF	46.49%	30.45%	36.8%
AVERAGE	51.32%	56.93%	52.81%

Table 4: Evaluation results on development dataset with gold mention boundaries

only identity coreference is considered while attributive NP and appositive construction are excluded. That means we cannot use these two important heuristic rules to generate pre-clusters. In our system, we just put all the mentions (names and nominal mentions, except pronouns) which have the same string into the identical pre-clusters. With these pre-clusters and their coreferential results, we implement a classification based pre-cluster pair model to determine whether a given pair of pre-clusters refer to the same entity. We follow Rahman and Ng (2009) to generate most of our features. We also include some other features which intuitively seem effective for coreference resolution. These features

Metric	Recall	Precision	F-score
MUC	42.66%	53.7%	47.54%
BCUBED	61.05%	74.32%	67.04%
CEAF	40.54%	32.35%	35.99%
AVERAGE	48.08%	53.46%	50.19%

Table 5: Evaluation results on development dataset with gold mention boundaries using unified classification model

Metric	Recall	Precision	F-score
MUC	53.73%	67.79%	59.95%
BCUBED	60.65%	66.05%	63.23%
CEAF	43.37%	30.71%	35.96%
AVERAGE	52.58%	54.85%	53.05%

Table 6: Evaluation results on test dataset without gold mention boundaries

are shown in Table 1 and Table 2. For simplicity, we use c_i and c_j to represent pre-clusters i and j. Each pre-cluster pair can be seen as a link. We have three kinds of link types: NP-NP link, NP-PRP link and PRP-PRP link. Different link types may have different feature preferences. So we train the classification based pre-cluster pair model for each link type separately and use different models to predicate the results. With the predicating results for pre-cluster pairs, we use closest-first clustering to link them and form the final cluster results.

4 Experimental Results

We present our evaluation results on development dataset for CoNLL-2011 shared Task in Table 3, Table 4 and Table 5. Official test results are given in Table 6 and Table 7. Three different evaluation metrics were used: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) and CEAF (Luo, 2005). Finally, the average scores of these three metrics are used to rank the participating systems. The difference between Table 3 and Table 4 is whether gold standard mention boundaries are given. Here "mention boundaries" means a more broad concept than the mention definition we gave earlier. We should also detect real mentions from them. From the tables, we can see that the scores can be improved littte by using gold standard mention boundaries. Also the results from Table 5 tell us that combining different link-type based classification models performed

Metric	Recall	Precision	F-score
MUC	46.66%	68.40%	55.48%
BCUBED	54.40%	70.19%	61.29%
CEAF	43.77%	25.88%	32.53%
AVERAGE	48.28%	54.82%	49.77%

Table 7: Evaluation results on test dataset with gold mention boundaries

better than using an unified classification model. For official test results, our system did not perform as well as we had expected. Some possible reasons are as follows. First, verbs that are coreferential with a noun phrase are also tagged in OntoNotes. For example, "grew " and "the strong growth" should be linked in the following case: "Sales of passenger cars grew 22%. The strong growth followed year-to-year increases." But we cannot solve this kind of problem in our system. Second, we should perform feature selection to avoid some useless features harming the scores. Meanwhile, we did not make full use of the WordNet, PropBank and other background knowledge sources as features to represent pre-cluster pairs.

5 Conclusion

In this paper, we present our system for CoNLL-2011 shared Task, Modeling Unrestricted Coreference in OntoNotes. First some heuristic rules are performed to pre-cluster all the mentions. And then we use a classification based pre-cluster pair model combined with several cluster level features. We hypothesize that the main reason why we did not achieve good results is that we did not carefully examine the features and dropped the feature selection procedure. Specially, we did not make full use of background knowledge like WordNet, PropBank, etc. In our future work, we will make up for the weakness and design a more reasonable model to effectively combine all kinds of features.

Acknowledgments

This research is supported by National Natural Science Foundation of Chinese (No.60973053, No.91024009) and Research Fund for the Doctoral Program of Higher Education of China (No.20090001110047).

References

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), Portland, Oregon.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. *A Model-Theoretic Coreference Scoring Scheme.* In Proceedings of the Sixth Message Understanding Conference (MUC-6), pages 4552, San Francisco, CA. Morgan Kaufmann.

Amit Bagga and Breck Baldwin. 1998. *Algorithms for Scoring Coreference Chains.* In Proceedings of the 1st International Conference on Language Resources and Evaluation, Granada, Spain, pp. 563566.

Xiaoqiang Luo. 2005. *On Coreference Resolution Performance Metrics.* In Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing, Vancouver, B.C., Canada, pp. 2532.

Vincent Ng. 2008. *Unsupervised Models for Coreference Resolution.* In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 640–649.

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

Vincent Ng. 2010. *Supervised Noun Phrase Coreference Research: The First Fifteen Years.* In Proceedings of the 48th Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, pages 1396-1411.

Shane Bergsma and Dekang Lin. 2006. *Bootstrapping Path-Based Pronoun Resolution.* In COLING–ACL 2006, pages 33–40.