Simple Maximum Entropy Models for Multilingual Coreference Resolution

Xinxin Li, Xuan Wang, Xingwei Liao
Computer Application Research Center
Harbin Institute of Technology Shenzhen Graduate School
Shenzhen, China
lixxin2@gmail.com

Abstract

This paper describes our system participating in the CoNLL-2012 shared task: Modeling Multilingual Unrestricted Coreference in Ontonotes. Maximum entropy models are used for our system as classifiers to determine the coreference relationship between every two mentions (usually noun phrases and pronouns) in each document. We exploit rich lexical, syntactic and semantic features for the system, and the final features are selected using a greedy forward and backward strategy from an initial feature set. Our system participated in the closed track for both English and Chinese languages.

1 Introduction

In this paper, we present our system for the CoNLL-2012 shared task which aims to model coreference resolution for multiple languages. The task of coreference resolution is to group different mentions in a document into coreference equivalent classes (Pradhan et al., 2012). Plenty of machine learning algorithms such as Decision tree (Ng and Cardie, 2002), maximum entropy model, logistic regression (Björkelund and Nugues, 2011), Support Vector Machines, have been used to solve this problem. Meanwhile, the CoNLL-2011 shared task on English language show that a well-designed rule-based approach can achieve a comparable performance as a statistical one (Pradhan et al., 2011).

Our system treats coreference resolution problem as classification problem by determining whether every two mentions in a document has a coreference relationship or not. We use maximum entropy (ME) models to train the classifiers. Previous work reveal that features play an important role on coreference resolution problem, and many different kinds of features has been exploited. In this paper, we use many different lexical, syntactic and semantic features as candidate features, and use a greedy forward and backward approach for feature selection for ME models.

2 System Description

The framework of our system is shown in figure 1. It includes four components: candidate mention selection, training example generation, model generation, and decoding algorithm for test data. The details of each component as described below.

2.1 Candidate Mention Selection

In both training and test sets, our system only consider all noun phrases (NP) and pronouns (PRP, PRP$) as candidate mentions for both English and Chinese. The mentions in each sentence are obtained from given syntactic tree by their syntactic label. Other phrases in the syntactic tree are omitted due to their small proportion. For example, in the English training dataset, our candidate mentions includes about 91% of golden mentions.

2.2 Training Example Generation

There are many different training example generation algorithms, e.g., McCarthy and Lehnert's method, Soon et al.s method, Ng and Cardies method (Ng, 2005). For our baseline system, we choose Soon et al.'s method because it is easily understandable, implemented and popularly used. It

Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task, pages 83–87,
Jeju Island, Korea, July 13, 2012.

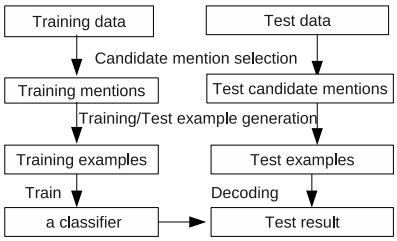

Figure 1: The framework of our coreference resolution system

selects pairs of two coreferent mentions as positive examples, and pairs between mentions among the two mentions and the last mention as negative examples.

2.3 Feature Selection

Rich and meaningful features are important for coreference resolution. Our system starts with Soon's 12 features as baseline features (Soon et al., 2001), and exploits many lexical, syntactic, and semantic features as candidate features. Totally 71 features are considered in our system, and summarized below:

- Distance features: sentence distance, distance in phrases, whether it's a first mention (Strube et al., 2002)
- Lexical features: string match, partial match, apposition, proper name match, head word match, partial head word match, minimum edit distance (Daumé III and Marcu, 2005)
- Grammatical features: pronoun, demonstrative noun phrase, embedded noun, gender agreement, number agreement (Soon et al., 2001)
- Syntactic features: same head, maximal NP, syntactic path (Yang et al., 2006)
- Semantic features: semantic class agreement, governing verb and its grammatical role, predicate (Ponzetto and Strube, 2006)

For English, the number agreement and gender agreement features can be obtained through the gender corpus provided. However, there is no corpus for Chinese. Our system obtains this information by collecting dictionaries for number and gender information from training dataset. For example, the pronoun "他" (he) denotes a male mention, and the noun phrase "女友" (girlfriend) represents a female mention. Similarly for number information, e.g., the mentions containing "和" (and), "群" (group) are plural. We use these words to build number and gender dictionaries, and determine the number and gender information of a new mention by checking whether one of the words in the dictionaries is in the mention.

Algorithm 1 Greedy forward and backward feature selection

```
Initialization: all candidate features in set C
                Choose initial feature set c
                Compute F1 with features c
while forward || backward:
  while forward:
    for each feature f in C-c
      Compute F1 with features c+f
    if best(F1) increases:
      backward = true, c=c+f, continue forward
    else forward = false
  while backward:
    for each feature f in in c
      Compute F1 with features c-f
    if best(F1) increases:
      forward = true, c=c-f continue backward
    else backward = false
```

For semantic class agreement feature in English, the relation between two mentions is extracted from WordNet 3.0 (Ng, 2007),(Miller, 1995). There is no corresponding dictionary for Chinese, so we keep it blank. The head word for each mention is selected by its dependency head, which can be extracted throught the conversion head rules (English [1] and Chinese [2]).

Maximum Entropy modeling is used to train the classifier for our system [3]. We employ a greedy forward and backward procedure for feature selection. The procedure is shown in Algorithm 1.

The algorithm will iterate forward and backward procedures until the performance does not improve. We use two initial feature sets: a blank set and Soon's baseline feature set. Both feature sets start

[1] http://w3.msi.vxu.se/ nivre/research/headrules.txt
[2] http://w3.msi.vxu.se/ nivre/research/chn_headrules.txt
[3] http://homepages.inf.ed.ac.uk/lzhang10/maxent.html

with a forward procedure.

2.4 Decoding

For every candidate mention pair, to determine their coreference relationship is simple because the probability whether they are coreferent can be obtained by our maximum entropy model. We can just set a threshold $\theta = 0.5$ and select the pairs with probability larger than θ. But usually it is hard for multiple mentions. Suppose there are three mentions A, B, C where the probability between A and B, A and C is larger than θ, but B and C is small. Thus choosing an appropriate decoding algorithm is necessary.

We use best-first clustering method for our system which for each candidate mention in a document, chooses the mention before it with best probability larger than threshold θ. The difference between English and Chinese is that we consider the coreference relationship of two mentions nested in Chinese, but not in English.

3 Experiments

3.1 Setting

Our system participates in the English and Chinese closed tracks with auto mentions. For both the English and Chinese datasets, we use gold annotated training data for training, and a portion of auto annotated development data for feature selection. Only part of development data is chosen because the evaluation procedure takes lot of time. To simplify, We only select one or two file in each directory as our development data.

The performance of the system is evaluated on MUC, B-CUBED, CEAF(M), CEAF(E), BLANC metrics. The official metric is calculated as $(MUC+B^3+CEAF)/3$.

3.2 Development set

Figures 2 and 3 show the performance on the English and Chinese development datasets using feature selection starting from a empty feature set and Soon's baseline feature set. The x-axis means the number of iterations with either forward or backward selection. The performance on Soon's baseline feature set for both languages are shown on 1st iteration. The performance from empty feature set starts on 2nd iteration. From these figures, we can see that

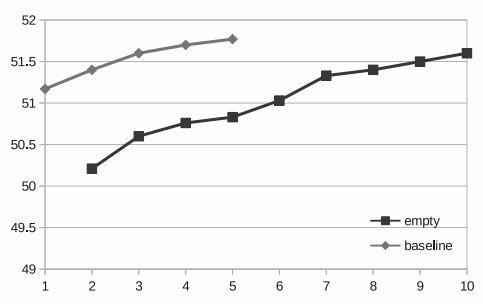

Figure 2: Performance of English development data with Feature selection

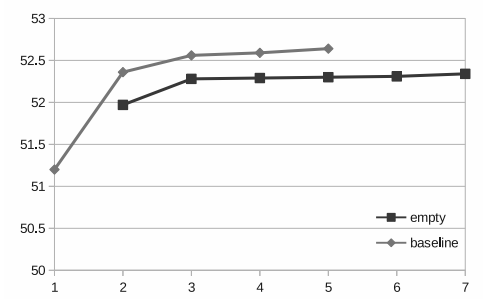

Figure 3: Performance of Chinese development data with Feature selection

using feature selection in both initial feature sets, the performance improves.

However the performance of our system is improved only on a few iteration. The best system for English stops at the 4th iteration with total 10 features left, which starts from Soon's baseline feature set. Similarly, the system for Chinese achieves its best performance at the 4th iteration with only 8 features. The phenomenon reveals that most of the features left for our system are still from Soon's baseline features, and our newly exploited lexical, syntactic, and semantic features are not well utilized.

Then we evaluate our model on the entire development data. The results are shown on Table 1. Comparing Figures 2, 3 and Table 1, we can observe that the performance on entire development data is lower than part one, about 1% decrease.

3.3 Test

For test data, we retrain our model on both gold training data and development data using the selected features. The final results for English and Chinese are shown in Table 2.

Model	English	Chinese
MUC	49.28	48.31
B^3	62.79	67.97
CEAF(M)	46.77	49.49
CEAF(E)	38.19	38.9
BLANC	66.31	68.91
Average	50.09	51.73

Table 1: Results on entire development data

Model	English	Chinese
MUC	48.27	48.09
B^3	61.37	68.31
CEAF(M)	44.83	49.92
CEAF(E)	36.68	38.89
BLANC	65.42	71.44
Official	48.77	51.76

Table 2: Results on test data

Comparing tables 2 and 1, we can observe that the performance for the Chinese test data is similar as the development data. The result seems reasonable because the model for testing use additional development data which is much smaller than training data. However, the result on English test data seem a little odd. The performance is about 1.4% less than that on the development data. The result needs further analysis.

4 Conclusion

In this paper, we presented our coreference resolution system which uses maximum entropy model to determine the coreference relationship between two mentions. Our system exploits many lexical, syntactic and semantic features. However, using greedy forward and backward feature selection strategy for ME model, these rich features are not well utilized. In future work we will analyze the reason for this phenomenon and extend these features to other machine learning algorithms.

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.

Hal Daumé III and Daniel Marcu. 2005. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 97–104, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38:39–41, November.

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

Vincent Ng. 2005. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 157–164, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Vincent Ng. 2007. Semantic class induction and coreference resolution. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 536–543, Prague, Czech Republic, June. Association for Computational Linguistics.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 192–199, New York City, USA, June. Association for Computational Linguistics.

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: Shared Task*, pages 1–27, Portland, Oregon, USA, June. Association for Computational Linguistics.

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.

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

Michael Strube, Stefan Rapp, and Christoph Müller. 2002. The influence of minimum edit distance on

reference resolution. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 312–319, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2006. Kernel-based pronoun resolution with structured syntactic knowledge. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 41–48, Sydney, Australia, July. Association for Computational Linguistics.