Incorporating Rule-based and Statistic-based Techniques for Coreference Resolution

Ruifeng Xu, Jun Xu, Jie Liu, Chengxiang Liu, Chengtian Zou, Lin Gui, Yanzhen Zheng, Peng Qu
Human Language Technology Group, Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, China
{xuruifeng.hitsz;hit.xujun;lyjxcz;matitalk;chsky.zou;monta3pt; zhyz.zheng;viphitqp@gmail.com}

Abstract

This paper describes a coreference resolution system for CONLL 2012 shared task developed by HLT_HITSZ group, which incorporates rule-based and statistic-based techniques. The system performs coreference resolution through the mention pair classification and linking. For each detected mention pairs in the text, a Decision Tree (DT) based binary classifier is applied to determine whether they form a coreference. This classifier incorporates 51 and 61 selected features for English and Chinese, respectively. Meanwhile, a rule-based classifier is applied to recognize some specific types of coreference, especially the ones with long distances. The outputs of these two classifiers are merged. Next, the recognized coreferences are linked to generate the final coreference chain. This system is evaluated on English and Chinese sides (Closed Track), respectively. It achieves 0.5861 and 0.6003 F1 score on the development data of English and Chinese, respectively. As for the test dataset, the achieved F_1 scores are 0.5749 and 0.6508, respectively. This encouraging performance shows the effectiveness of our proposed coreference resolution system.

1 Introduction

Coreference resolution aims to find out the different mentions in a document which refer to the same entity in reality (Sundheim and Beth, 1995; Lang et al. 1997; Chinchor and Nancy, 1998;). It is a core component in natural language processing and information extraction. Both rule-based approach (Lee et al. 2011) and statistic-based approach (Soon et al., 2001; Ng and Cardie, 2002; Bengtson and Roth, 2008; Stoyanov et al., 2009; Chen et al. 2011) are proposed in coreference resolution study. Besides the frequently used syntactic and semantic features, the more linguistic features are exploited in recent works (Versley, 2007; Kong et al. 2010).

CoNLL-2012 proposes a shared task, "Modeling multilingual unrestricted coreference in the OntoNotes" (Pradhan et al. 2012). This is an extension of the CoNLL-2011 shared task. The task involves automatic anaphoric mention detection and coreference resolution across three languages including English, Chinese and Arabic. HLT_HITSZ group participated in the Closed Track evaluation on English and Chinese side. This paper presents the framework and techniques of HLT_HITSZ system which incorporates both rule-based and statistic-based techniques. In this system, the mentions are firstly identified based on the provided syntactic information. The mention pairs in the document are fed to a Decision Tree based classifier to determine whether they form a coreference or not. The rule-based classifiers are then applied to recognize some specific types of coreference, in particular, the long distance ones. Finally, the recognized coreference are linked to obtain the final coreference resolution results. This system incorporates lexical, syntactical and semantic features. Especially for English, WordNet is used to provide semantic information of the mentions, such as semantic distance and the

Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task, pages 107–112,
Jeju Island, Korea, July 13, 2012.

category of the mentions and so on. Other than the officially provided number and gender data, we generated some lexicons from the training dataset to obtain the values of some features. This system achieves 0.5861 and 0.6003 F_1 scores on English and Chinese development data, respectively, and 0.5749 and 0.6508 F_1 scores on English and Chinese testing data, respectively. The achieved encouraging performances show that the proposed incorporation of rule-based and statistic-based techniques is effective.

The rest of this report is organized as below. Section 2 presents the mention detection. Section 3 presents the coreference determination and Section 4 presents the coreference linking. The experimental results are given in Section 5 in detail. Finally, Section 6 concludes this report.

2 Mention Detection

In this stage, the system detects the mentions from the text. The pairs of these mentions in one document are regarded as the coreference candidates. Thus, the high recall is a more important target than higher precision for this stage. Corresponding to English and Chinese, we adopted different detection methods, respectively.

2.1 Mention Detection - English

HLT_HITSZ system chooses the marked noun phrase (NP), pronouns (PRP) and PRP$ in English data as the mentions. The system selects most named entities (NE) as the mentions but filter out some specific types. Firstly, the NEs which cannot be labeled either as NP or NML are filter out because there are too cases that the pairs of these NEs does not corefer even they are in the same form as shown in the training dataset. Second, the NEs of ORDINAL, PERCENT and MONEY types are filtered because they have very low coreference ratio (less than 2%). Furthermore, for the cases that NPs overlapping a shorter NP, normally, only the longer one are choose. An exception is that if the shorter NPs are in parallel structures with the same level to construct a longer NP. For example, for a NP "A and B", "A", "B" and "A and B" as regarded ed as three different mentions.

2.2 Mention Detection – Chinese

HLT_HITSZ system extracts all NPs and PNs as the mention candidates. For the NPs have the overlaps, we handle them in three ways: 1. For the cases that two NPs share the same tail, the longer NP is kept and the rest discarded; 2. For cases that the longer NP has a NR as its tail, the NPs which share the same tail are discarded; 3. In MZ and NW folders, they are many mentions nested marked as the nested co-referent mentions. The system selects the longest NP as mention in this stage while the other mention candidates in the longest NP will be recalled in the post processing stage.

3 Coreference Determination

Any pair of two detected mentions in one document becomes one coreference candidate. In this stage, the classifiers are developed to determine whether this pair be a coreference or not. During the generation of mention pairs, it is observed that linking any two mentions in one document as candidates leads to much noises. The statistical observation on the Chinese training dataset show that 90% corefered mention pairs are in the distance of 10 sentences. Similar results are found in the English training dataset while the context window is set to 5 sentences. Therefore, in this stage, the context windows for generating mention pairs as coreference candidates for English and Chinese are limited to 5 and 10 sentences, respectively.

3.1 The Statistic-based Coreference Determination

The same framework is adopted in the statistical-based coreference determination for English and Chinese, respectively, which is based on a machine learning-based statistical classifier and selected language-dependent features. Through transfer the examples in the training test into feature-valued space, the classifier is trained. This binary classifier will be applied to determine whether the input mention pair be a coreference or not. Here, we evaluated three machine learning based classifiers including Decision Tree, Support Vector Machines and Maximum Entropy on the training data while Decision Tree perform the best. Thus, DT classifier is selected. Since the annotations on the training data from different directory show some inconsistence, multiple classifiers corresponding to each directory are trained individually.

3.1.1 Features - English

51 features are selected for English coreference determination. The features are camped to six categories. Some typical features are listed below:

1. Basic features:
   (1) Syntactic type of the two mentions, includes NP, NE, PRP, PRP$. Here, only the NPs which do not contain any named entities or its head word isn't a named entity are considered as an NP while the others are discarded.
   (2) If one mention is a PRP or PRP$, use an ID to specify which one it is.
   (3) The sentence distance between two mentions.
   (4) Whether one mention is contained by another one.
2. Parsing features:
   (1) Whether two mentions belong to one NP.
   (2) The phrase distance between the two mentions.
   (3) The predicted arguments which the two mentions belong to.
3. Named entity related features:
   (1) If both of the two mentions may be considered as named entities, whether they have the same type.
   (2) If one mention is a common NP or PRP and another one can be considered as named entity, whether the words of the common NP or PRP can be used to refer this type of named entity. This knowledge is extracted from the training dataset.
   (3) Whether the core words of the two named entity type NP match each other.
4. Features for PRP:
   (1) If both mentions are PRP or PRP$, use an ID to show what they are. The PRP$ with the same type will be assigned the same ID, for example, *he*, *him* and *his*.
   (2) Whether the two mentions has the same PRP ID.
5. Semantic Features:
   (1) Whether the two mentions have the same headword.
   (2) Whether the two mentions belong to the same type. Here, we use WordNet to get three most common sense of each NP and compare the type they belong to.
   (3) The semantic distance between two mentions. WordNet is used here.
   (4) The natures of the two mentions, including number, gender, is human or not, and match each other or not. We use WordNet and a lexicon extracted from the gender and number file here.
6. Document features:
   (1) How many speakers in this document.
   (2) Whether the mention is the first or the last sentence of the document.
   (3) Whether the two mentions are from the same speaker.

3.1.2 Features - Chinese

There are 61 features adopted in Chinese side. Because of the restriction of closed crack, most of features use the position and POS information. It is mentionable that the ways for calculating the features values. For instance, the sentence distance is not the real sentence distance in the document. For instead, the value is the number of sentences in which there are at least one mention between the mention pair. This ignores the sentences of only modal particles.

The 61 features are camped into five groups. Some example features are listed below.

1. Basic information:
   (1) The matching degree of two mentions
   (2) The word distance of two mentions
   (3) The sentence distance of two mentions
2. Parsing information:
   (1) Predicted arguments which the two mentions belong to and corresponding layers.
3. POS features
   (1) Whether the mention is NR
   (2) Whether the two mentions are both NR and are matched
4. Semantic features:
   (1) Whether the two mention is related
   (2) Whether the two mentions corefer in the history. Since the restriction of closed track, we did not use any additional semantic resources. Here, we extract the co-reference history from the training set to obtain some semantic information, such as "NN 歹徒" and "NN 绑匪" corefered in the training data, and they are regarded as coreference in the testing data.
5. Document Features:

(1) Whether the two mentions have the same speaker.
(2) Whether the mention is a human.
(3) Whether the mention is the first mention in the sentence.
(4) Whether the sentence to which the mention belongs to is the first sentence.
(5) Whether the sentence to which the mention belongs to is the second sentence
(6) Whether the sentence to which the mention belongs to is the last sentence
(7) The number of the speakers in the document.

3.2 The Rule-based Coreference Determination

The rule-based classifier is developed to recognize some specific types of coreference and especially, the long distance ones.

3.2.1 Rule-based Classifier - English

To achieve a high precision, only the mention pairs of NE-NE (include NPs those can be considered as NE) or NP-NP types with the same string are classified here.

For the NE-NE pair, the classifier identifies their NE part from the whole NP, if their strings are the same, they are considered as coreference.

For the NP-NP pair, the pairs satisfy the following rules are regarded as coreference.

(1) The POS of the first word isn't "JJR" or "JJ".
(2) If NP has only one word, its POS isn't "NNS" or "NNPS".
(3) The NP have no word like "*every*", "*every-*", "*none*", "*no*", "*any*", "*some*", "*each*".
(4) If the two NP has article, they can't be both "*a*" or "*an*".

Additionally, for the PRP mention pairs, only "*I*", "*me*", "*my*" with the same speaker can be regarded as coreference.

3.2.2 Rule-based Classifier - Chinese

A rule-based classifier is developed to determine whether the mention pairs between PNs and mentions not PN corefer or not. For instance, the mention pairs between the PN "他" which is after a comma and the mention which is marked as ARG0 in the same sentence. In the sentence "埃斯特拉达表示，他希望上帝能够赐给他智慧", because the mention pair between "埃斯特拉达" and the first "他" match the mentioned above rule, it is classified as a positive one. The result on the development set shows that the rule-based classifier brings good improvement.

4 Coreference Chain Construction

4.1 Coreference Chain Construction-English

The evaluation on development data shows that the achieved precision of our system is better than recall. Thus, in this stage, we simply link every pair of mentions together if there is any links can link them together to generate the initial coreference chain. After that, the mentions have the distance longer than 5 sentences are observed. The NE-NE or NP-NP mention pairs between one known coreference and an observing mention with long distance are classified to determine they are corefered or not by using a set of rules. The new detected conference will be linked to the initial coreference chain.

4.2 Coreference Chain Construction-Chinese

The coreference chain construction for Chinese is similar to English. Furthermore, as mentioned above, in MZ and NW folders, there are many mentions nested marked as the nested co-referenced mentions. In this stage, HLT_HITSZ system generates the nested co-reference mentions for improving the analysis for these two folders. Additionally, the system uses some rules to improve the coreference chain construction. We find that the trained classifier performs poor in co-reference resolution related to Pronoun. So, most rules adopted here are related to these Pronouns: "自己", "我", "你", "他", "她", "两国", "双方", "其". We use these rules to bridge the chain of pronouns and the chain of other type.

Although high precision for NT co-reference cases are achieved through string matching, the recall is not satisfactory. It partially attributes to the fact that the flexible use of Chinese. For example, to express the year of 1980, we found "一九八零年", "一九八零", " 一九八○", "八零年", "1980 年". Similar situation happens for month (月, 月份) and day (日,号), we conclude most situations to several templates to improve the rule-based conference resolution.

5 Evaluation Results

5.1 Dataset

The status of training dataset, development dataset and testing dataset in CoNLL 2012 for English and Chinese are given in Table 1 and Table 2, respectively.

	Files	Sentence	Cluster	Coreference
Train	1,940	74,852	35,101	155,292
Development	222	9,603	4,546	19,156
Test	222	9,479	n/a	n/a

Table 1. Status of CoNLL 2012 dataset - English

	Files	Sentence	Cluster	Coreference
Train	1,391	36,487	28,257	102,854
Develop	172	6,083	3,875	14,383
Test	166	4,472	n/a	n/a

Table 2. Status of CoNLL 2012 dataset - Chinese

5.2 Evaluation on Mention Detection

Firstly, the mention detection performance is evaluated. The performance achieved on the development dataset (Gold/Auto) and test data on English and Chinese are given in Table 3 and Table 4, respectively. In which, Gold means the development dataset with gold manually annotation and Auto means the automatically generated annotations.

	Precision	Recall	F_1
Develop-Gold	0.8499	0.6716	0.7503
Develop-Auto	0.8456	0.6256	0.7192
Test	0.8455	0.6264	0.7196

Table 3. Performance on Mention Detection - English

	Precision	Recall	F_1
Develop-Gold	0.7402	0.7360	0.7381
Develop-Auto	0.6987	0.6429	0.6697
Test	0.7307	0.7502	0.7403

Table 4. Performance on Mention Detection - Chinese

Generally speaking, our system achieves acceptable mention detection performance, but further improvements are desired.

5.3 Evaluation on Coreference Resolution

The performance on coreference resolution is next evaluated. The achieved performances on the development data (Gold/Auto) and test dataset on English and Chinese are given in Table 5 and Table 6, respectively. It is shown that the OF performance drops 0.0309(Gold) and 0.0112(Auto) from development dataset to test dataset on English, respectively. On the contrary, the OF performance increases 0.0096(Gold) and 0.0505(Auto) from development dataset to test dataset on Chinese, respectively. Compared with the performance reported in CoNLL2012 shared task, our system achieves a good result, ranked 3rd, on Chinese. The results show the effectiveness of our proposed system.

	Precision	Recall	F_1
MUC	0.7632	0.6455	0.6994
BCUB	0.7272	0.6797	0.7027
CEAFE	0.3637	0.4840	0.4154
OF-Develop-Gold			0.6058
MUC	0.7571	0.5993	0.6691
BCUB	0.7483	0.6441	0.6923
CEAFE	0.3350	0.4865	0.3968
OF-Develop-Auto			0.5861
MUC	0.7518	0.5911	0.6618
BCUB	0.7329	0.6228	0.6734
CEAFE	0.3264	0.4829	0.3895
OF-Test			0.5749

Table 5. Performance on Coreference Resolution – English

	Precision	Recall	F_1
MUC	0.6892	0.6655	0.6771
BCUB	0.7547	0.7410	0.7478
CEAFE	0.4876	0.5105	0.4988
OF-Develop-Gold			0.6412
MUC	0.6535	0.5643	0.6056
BCUB	0.7812	0.6809	0.7276
CEAFE	0.4322	0.5101	0.4679
OF-Develop-Auto			0.6003
MUC	0.6928	0.6595	0.6758
BCUB	0.7765	0.7328	0.7540
CEAFE	0.5072	0.5390	0.6253
OF-Test(Gold parses)			0.6508
MUC	0.5502	0.6147	0.5807
BCUB	0.6839	0.7638	0.7216
CEAFE	0.5040	0.4481	0.4744
OF-Test-Predicted-mentions (Auto parses)			0.5922
MUC	0.6354	0.6873	0.6603
BCUB	0.7136	0.7870	0.7485
CEAFE	0.5390	0.4907	0.5137
OF-Test-Gold-mention-boundaries(Auto parses)			0.6408
MUC	0.6563	0.9407	0.7732
BCUB	0.6505	0.9123	0.7595
CEAFE	0.7813	0.4377	0.5611
OF-Test-Gold-mentions (Auto parses)			0.6979

Table 6. Performance on Coreference Resolution – Chinese

6 Conclusions

This paper presents the HLT_HITSZ system for CoNLL2012 shared task. Generally speaking, this system uses a statistic-based classifier to handle short distance coreference resolution and uses a rule-based classifier to handle long distance cases. The incorporation of rule-based and statistic-based techniques is shown effective to improve the performance of coreference resolution. In our future work, more semantic and knowledge bases will be incorporated to improve coreference resolution in open track.

Acknowledgement

This research is supported by HIT.NSFIR.201012 from Harbin Institute of Technology, China and China Postdoctoral Science Foundation No. 2011M500670.

References

B. Baldwin. 1997. CogNIAC: High Precision Coreference with Limited Knowledge and Linguistic Resources. Proceedings of Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts.

E. Bengtson, D. Roth. 2008. Understanding the Value of Features for Coreference Resolution. Proceedings of EMNLP 2008, 294-303.

M. S. Beth M. 1995. Overview of Results of the MUC-6 Evaluation. Proceedings of the Sixth Message Understanding Conference (MUC-6)

W. P. Chen, M. Y. Zhang, B. Qin, 2011. Coreference Resolution System using Maximum Entropy Classifier. Proceedings of CoNLL-2011.

N. A. Chinchor. 1998. Overview of MUC-7/MET-2. Proceedings of the Seventh Message Understanding Conference (MUC-7).

F. Kong, G. D. Zhou, L. H. Qian, Q. M. Zhu. 2010. Dependency-driven Anaphoricity Determination for Coreference Resolution. Proceedings of COLING 2010, 599-607

J. Lang, B. Qin, T. Liu. 2007. Intra-document Coreference Resolution: The State of the Art. Journal of Chinese Language and Computing , 2007, 17(4) : 227-253.

H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, D. Jurafsky. 2011. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. Proceedings of CoNLL-2011.

V. Ng and C. Cardie. 2002. Improving Machine Learning Approaches to Coreference Resolution. Proceedings of ACL 2002.

W. M. Soon, H. T. Ng, and D. C. Y. Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics, 27(4):521-544

S. Pradhan and A. Moschitti et al. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. Proceedings of CoNLL 2012

V. Stoyanov, N. Gilbert, C. Cardie, E. Riloff. 2009. Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art. Proceeding ACL 2009

Y. Versley. 2007. Antecedent Selection Techniques for High-recall Coreference Resolution. Proceedings of EMNLP/CoNLL 2007.

Y. Yang, N. W. Xue, P. Anick. 2011. A Machine Learning-Based Coreference Detection System For OntoNotes. Proceedings of CoNLL-2011.