A Mixed Deterministic Model for Coreference Resolution

Bo Yuan[1], Qingcai Chen, Yang Xiang, Xiaolong Wang[2]
Liping Ge, Zengjian Liu, Meng Liao, Xianbo Si
Intelligent Computing Research Center, Key Laboratory of Network Oriented Intelligent
Computation, Computer Science and technology Department, Harbin Institute of Technology
Shenzhen graduate School, Shenzhen, Guangdong, 518055, China
{yuanbo.hitsz[1], windseedxy, qingcai.chen, geliping123,
autobotsonearth, dream2009gd, sixianbo}@gmail.com
wangxl@insun.hit.edu.cn[2]

Abstract

This paper presents a mixed deterministic model for coreference resolution in the CoNLL-2012 shared task. We separate the two main stages of our model, mention detection and coreference resolution, into several sub-tasks which are solved by machine learning method and deterministic rules based on multi-filters, such as lexical, syntactic, semantic, gender and number information. We participate in the closed track for English and Chinese, and also submit an open result for Chinese using tools to generate the required features. Finally, we reach the average F1 scores 58.68, 60.69 and 61.02 on the English closed task, Chinese closed and open tasks.

1 Introduction

The coreference resolution task is a complicated and challenging issue of natural language processing. Although many sub-problems, such as noun phrase to noun phrase and pronouns to noun phrase, are contained in this issue, it is interesting that humans do not get too confused when they determine whether two mentions refer to the same entity. We also believe that automatic systems should copy the human behavior (Kai-Wei et al., 2011). In our understanding, the basis for human making judgment on different sub-problems is different and limited. Although there are some complicated and ambiguous cases in this task, and we are not able to cover all the prior knowledge of human mind, which plays a vital role in his solution, the mixed deterministic model we constructed can solve a big part of this task. We present a mixed deterministic model for coreference resolution in the CoNLL-2012 shared task (Sameer et al., 2011).

Different methods such as Relaxation labeling (Emili et al., 2011), Best-Link (Kai-Wei et al., 2011), Entropy Guided Transformation Learning (Cicero et al., 2011) and deterministic models (Heeyoung et al., 2011), were attempted in the CoNLL-2011 shared task (Sameer et al., 2011). The system performance reported by the task shows that a big part of this task has been solved but some sub-problems need more exploration.

We also participate in the Chinese closed and open tracks. However, the lack of linguistic annotations makes it more difficult to build a deterministic model. Basic solutions such as Hobbs Algorithm and Center Theory have been listed in (Wang et al., 2002; Jun et al., 2007). The recent research on Chinese contains non-anaphors detection using a composite kernel (Kong Fang, et al., 2012(a)) and a tree kernel method to anaphora resolution of pronouns (Kong Fang et al., 2012(b)).

We accept the thought of Stanford (Karthik et al., 2010; Heeyoung et al., 2011). In Stanford system the coreference resolution task is divided into several problems and each problem is solved by rule based methods. For English we did some research on mention detection which uses Decision Tree to decide whether the mention ‘it’ should refer to some other mention. For Chinese we submit closed and open result. The lack of gender, number and name entities make it more difficult for the Chinese closed task and we try to extract information from the training data to help enhance the performance. For the open task, we use some dictionaries such as appellation dictionary, gender dictionary, geographical name dictionary and temporal word dictionary (Bo et al., 2009), and some tools such as conversion of pinyin-to-character and LTP which is a Chinese parser that can generate the features such as Part-of-Speech, Parse bit, Named Entities (Liu et al., 2011) to generate the similar information.

We describe the system architecture in section 2. Section 3 illustrates the mention detection process. Section 4 describes the core process of coreference resolution. In section 5, we show the results and discussion of several experiments. Finally, we give the conclusion of our work in section 6.

2 System Architecture

Our system mainly contains mention detection and coreference resolution. Recall is the determining factor in mention detection stage. The reason is that if some mention is missed in this stage, the coreference resolution part will miss the chains which contain this mention. Yet some mentions still need to be distinguished because in some cases they refer to no entity. For example 'it', in the sentence 'it + be + weather/ time', 'it' should refer to no entity. But the 'it' in the phrase 'give it to me' might refer to some entity. The coreference resolution module of our system follows the idea of Stanford. In the English task we did some more exploration on mention detection, pronoun coreference and partial match of noun phrases. The Chinese task is more complicated and because gender, number and name entities are not provided, the feature generation from the training data has to be added before the coreference resolution process. Some Chinese idiomatic usages are also considered in this stage.

3 Mention detection

All the NPs, pronouns and the phrases which are indexed as named entities are selected as candidates. NPs are extracted from the parse tree. Yet some mentions do not refer to any entity in some cases. In our system we attempt to distinguish these mentions in this stage. The reason is that the deterministic rules in coreference resolution part are not complete to distinguish these mentions. The methods below can also be added to the coreference resolution part as a pre-processing. For the conveniences of system design, we finish this work in this stage.

For English, the pronoun 'it' and NPs 'this, that, those and these' need to be distinguished. We take 'it' as an example to illustrate the process. First we use regular expressions to select 'it', which refers to no entity, such as 'it + be + weather/ time', 'it happened that' and 'it makes (made) sense that'.
Second we use Decision Tree (C4.5) to classify the two kinds of 'it' based on the training data. The features contain the Part-of-Speech, Parse bit, Predicate Arguments of 'it', the word before and after 'it'. The number of total 'it' is 9697 and 4043 of them have an entity to refer to in the training data.

Category	Precision	Recall	F
no entity refered	0.576	0.596	0.586
entity refered	0.747	0.731	0.739
total	0.682	0.679	0.68

Table 1: Results of 'it' classification using C4.5

Table 1 shows the classification result of 'it' in the development data v4. The number of total 'it' is 1401 and 809 of them have an entity to refer to. The result is not perfect but can help enhance the performance of coreference resolution. However, the results of 'this, that, those and these' are not acceptable and we skip over these words. We did not do any process on 'verb' mention detection and coreference resolution.

In addition, we divide mentions into groups in which they are nested in position. And for mentions which have the same head word in one group, only the mentions with the longest span should be left (for the English task and a set of Chinese articles). For some Chinese articles of which names contain 'chtb', both in the training data and the development data, the nest is permitted based on the statistic results.

For Chinese we also attempt to train a model for pronouns '你'(you) and '那'(that). However, the results are not acceptable either since the features we select are not enough for the classifier.

After the mentions have been extracted, the related features of each mention are also extracted. We transform the 'conll' document into mention document. Each mention has basic features such as position, part-of-speech, parse tree, head word, speaker, Arguments, and the gender and number of head word. The head word feature is very important and regular expression can almost accomplish the process but not perfectly. Firstly, we extract the key NPs of a mention based on parse feature. Then the regular expressions are to extract the head word. For example, the mention:

(NP (DNP (LCP (NP (NP (NR 中国)) (NP (NN 大地))) (LC 上)) (DEG 的)) (NP (NR 二战)) (NP (NN 标志))) (NP (DNP (LCP (NP (NP (NR 中国)) (NP (NN 大地))) (LC 上)) (DEG 的)) (NP (NR 二战)) (NP (NN 标志)))

The key NPs of this mention is:

(NP (NR 二战)) (NP (NN 标志)) .The head word of this mention is: ***NN 标志***

However, there are still some cases that need to be discussed. For example, the head word of ‘the leader of people’ should be ‘leader’, while the head word of ‘the city of Beijing’ should be ‘city’ and ‘Beijing’ for the mentions of ‘the city’ and ‘Beijing’ both have the same meaning with ‘the city of Beijing’. Finally, we only found the words of ‘city’ and ‘country’ should be processed.

4 Coreference resolution

The deterministic rules are the core methods to solve the coreference resolution task. All the mentions in the same part can be seen as a list. The mentions which refer to the same entity will be clustered based on the deterministic rules. After all the clusters have generated, the merge program will merge the clusters into chains based on the position information. The mentions in one chain cannot be reduplicative in position. Basically the nested mentions are not allowed.

The process contains two parts NP-NP and NP-pronoun. Each part has several sub-problems to be discussed. First, the same process of English task and Chinese task will be illustrated. Then the different parts will be discussed separately.

4.1 NP-NP

Exact match: the condition of exact match is the two NP mentions which have no other larger parent mentions in position are coreferential if they are exactly the same. The stop words such as ‘a’, ‘the’, ‘this’ and ‘that’ have been removed.

Partial match: there are two conditions for partial match which are the two mentions have the same head word and one of them is a part of the other in form simultaneously.

Alias and abbreviation: some mentions have alias or abbreviation. For example the mentions ‘USA’ and ‘America’ should refer to the mention ‘the United States’.

Similar match: there are three forms of this match. The first one is all the modifiers of two NPs are same and the head words are similar based on WordNet[1] which is provided for the English closed task. We only use the English synonym sets of the WordNet to solve the first form. The second one is the head words are same and the modifiers are not conflicted. The third form is that the head words and modifiers are all different. The result of similar match may be reduplicative with that of exact match and partial match. This would be eliminated by the merge process.

4.2 Pronoun - NP

There are seven categories of pronoun to NP in our system. For English second person, it is difficult to distinguish the plural form from singular form and we put them in one deterministic rule. For each kind of pronouns shown below, the first cluster is the English form and the second cluster is the Chinese form.

First Person (singular) = {'I', 'my', 'me', 'mine', 'myself'}{‘我’}

Second Person= {'you', 'your', 'yours', 'yourself', 'yourselves'}{‘你’，‘你们’}

Third Person (male) = {'he', 'him', 'his', 'himself'}{‘他’}

Third Person (female) = {'she', 'her', 'hers', 'herself'}{‘她’}

Third Person (object) = {'it', 'its', 'itself'}{‘它’}

First Person (plural) = {'we', 'us', 'our', 'ours', 'ourselves'}{‘我们’}

Third Person (plural) = {'they', 'them', 'their', 'theirs', 'themselves'}{‘他们’，‘她们’，‘它们’}

In the Chinese task the possessive form of pronoun is not considered. For example, the mention ‘我们 的’(our) is a DNP in the parse feature and it contains two words ‘我们’ and ‘的’. We only selected the NP ‘我们’as a mention. The reflexive pronouns are composed by two words which are the pronoun itself and the word ‘自己’.

[1] http://wordnet.princeton.edu/

For example, the mention ‘我 自己’(myself) is processed as ‘我’(I or me).

Gender, number and distance between pronoun and NP are the most important features for this part (Shane et al., 2006). We only allow pronoun to find NPs at first. We find out the first mention of which all the features are satisfied ahead of the pronoun. If there is no matching mention, search backward from the pronoun. For the first person and second person, we merged all the pronouns with the same form and the same speaker. If the context is a conversation of two speakers, the second person of a speaker should refer to the first person of the other speaker. The scene of multi-speakers conversation is too difficult to be solved.

In the Chinese task there are some other pronouns. The pronoun ‘双方’(both sides) should refer to a plural mention which contain ‘和’(and) in the middle. The pronoun ‘其’ has similar meaning of third person and refers to the largest NP mention before it. The pronouns ‘这’(this), ‘那’(that), ‘这里’(here), ‘那里’(there) are not processed for we did not find a good solution.

However in some cases the provided gender and number are not correct or missing and we had to label these mentions based on the appellation words of the training data. For example, if the appellation word of a person is ‘Mr.’ or ‘sir’, the gender should be male.

4.3 Chinese closed task

For the Chinese closed task NE, the gender and number are not provided. We used regular patterns to generate these features from the training data.

In the NE (named entities) feature ‘PERSON’ is a very important category because most pronouns will refer to the person entity. To extract ‘PERSON’, we build a PERSON dictionary which contains all the PERSON mentions in training data, such as ‘先生’(Mr.) and ‘教授’(Professor). If the same mention appears in the test data, we believe it is a person entity. However, the PERSON dictionary cannot cover all the PERSON mentions. The appellation words are extracted before or after the person entity. When some appellation word appears in the test data, the NP mention before or after the appellation word should be a person entity, if they compose a larger NP mention.

The Gender feature was generated at the same time of the ‘PERSON’ generation. We separate the ‘PERSON’ dictionary and appellation dictionary into male cluster and female cluster by the pronouns in the same chain.

The generation of number feature is a little complicated. Since the Chinese word does not have plural form, the numerals and the quantifiers of the mention are the main basis to extract the number feature. We extract the numerals and the quantifiers from the training data and built regular expressions for determine the number feature of a mention in test data. Other determinative rules for number feature extraction are shown below:

If the word ‘们’ appears in a mention tail, this mention is plural. For example ‘同学’(student) is singular and ‘同学们’(students) is plural.

If the word ‘和’(and) appears in the middle of a mention A, and the two parts separated by ‘和’ are sub-mentions of A, mention A should be plural. Other words which have the similar meaning of ‘和’, such as ‘同’, ‘与’ and ‘跟’, are considered.

The time and date coreference resolution is also considered. The NP mentions which contain temporal words are processed separately since these categories of name entity are not provided. These temporal words are also extracted from training data. Since the head words of these mentions are themselves, the two time or date mentions are coreferential if they are the same or one must be a part of the other’s tail. For example ‘今年九月’(this September) and ‘九月’(September) which are not nested should be coreferential.

4.4 Chinese open task

For the Chinese open task we use several tools to generate features we need.

NE generation: LTP is a Chinese parser that can generate the features such as Part-of-Speech, Parse bit, Named Entities (Liu et al., 2011). We only use LTP for the NE generation. However, the NE labels of LTP are different with that provided by the gold training data and need to be transformed. The difference of word segmentation between LTP and the provided data also made some errors. At last we find the NE feature from LTP does not perform well and it will be discussed in section 5.

The conversion of pinyin-to-character is also used in the Chinese open task. The speaker provided in the training data is given in pinyin form. The speaker might be the ‘PERSON’ mention in the context. When we determine the pronoun coreference, we need to know whether the speaker and the 'PERSON' mention are same.

Other tools used in open task contain appellation dictionary, gender dictionary, geographical name dictionary and temporal word dictionary (Bo et al., 2009). These dictionaries are more complete than those used in the closed task, although the enhancements are also limited.

5 Results and Discussion

Table 2 to table 4 show the results of English coreference resolution on the gold and auto development and the test data. The results of the auto development data and the test data are close and lower than that of the gold data. Since the deterministic rules can not cover all the cases, there is still an improvement if we could make the deterministic rules more complete.

Measure	R	P	F1
Mention detection	77.7	71.8	74.6
MUC	65.1	62.9	64
B^3	69.2	70.9	70.1
CEAF(E)	46.4	48.9	47.6
(CEAF(E)+MUC+B^3)/3			60.6

Table 2: Results of the English gold development data

Measure	R	P	F1
Mention detection	72.4	71.5	72
MUC	62.3	62.8	62
B^3	66.7	71.8	69.1
CEAF(E)	46.4	44.9	45.6
(CEAF(E)+MUC+B^3)/3			58.9

Table 3: Results of the English auto development data

Measure	R	P	F1
Mention detection	73.2	71.9	72.53
MUC	62.1	63	63
B^3	66.2	70.5	68.3
CEAF(E)	45.7	44.7	45.2
CEAF(M)	57.3	57.3	57.3
BLANC	72.1	76.9	74.2
(CEAF(E)+MUC+B^3)/3			58.68

Table 4: Results of English test data

The results of the closed Chinese performance on the gold and auto development and the test data are shown in table 5 to table 7. The performance of the auto development data and the test data has about 4% decline to that of the gold development on F1 of coreference resolution. It means the Chinese results are also partly affected by the parse feature. In fact we attempted to revise the parse feature of the auto development data using regular expressions. Yet the complicacy and unacceptable results made us abandon that.

Measure	R	P	F1
Mention detection	82.3	69.8	75.5
MUC	71.6	64.3	67.7
B^3	76.7	74.2	75.4
CEAF(E)	49	56.5	52.5
(CEAF(E)+MUC+B^3)/3			65.2

Table 5: Closed results of the Chinese gold development data

Measure	R	P	F1
Mention detection	74.2	66	70
MUC	63.6	60	61.7
B^3	73.1	73.5	73.3
CEAF(E)	47.3	50.6	48.9
(CEAF(E)+MUC+B^3)/3			61.3

Table 6: Closed results of the Chinese auto development data

Measure	R	P	F1
Mention detection	72.8	64.1	68.15
MUC	62.4	58.4	60.3
B^3	73.1	72.7	72.9
CEAF(E)	47.1	50.7	48.8
CEAF(M)	59.6	59.6	59.6
BLANC	73.7	78.2	75.8
(CEAF(E)+MUC+B^3)/3			60.69

Table 7: Closed results of the Chinese test data

The results of the open Chinese performance on the gold and auto development and the test data are shown in table 8 to table 10. The performance is similar with that of the closed task. However, the improvement between F1 of the open task and F1 of the closed task is limited. We also get the F1 of the closed and open test results using gold parser which are 66.46 and 66.38. The open result is even lower. This can be explained. The performance enhanced by the dictionaries we used for the open task are limited because the open dictionaries information which appears in the test data is not much more than that of the closed dictionaries which generated from the training data, although the total information of the former is much larger. The named entities generated by LTP have some errors such as person identification errors and will caused coreferential errors in Pronoun-NP stage. For the time we did not use LTP well and some other open tools such as Wikipedia and Baidu Baike should be applied in the open task.

Measure	R	P	F1
Mention detection	82.4	69.3	75.3
MUC	72.3	63.8	67.8
B^3	77.7	73.3	75.4
CEAF(E)	48.3	56.8	52.2
(CEAF(E)+MUC+B^3)/3			65.1

Table 8: Open results of the Chinese gold development data

Measure	R	P	F1
Mention detection	75.1	65.7	70.1
MUC	64.9	59.9	62.3
B^3	74.2	72.6	73.4
CEAF(E)	46.7	51.5	49
(CEAF(E)+MUC+B^3)/3			61.6

Table 9: Open results of the Chinese auto development data

Measure	R	P	F1
Mention detection	73.7	64	68.49
MUC	63.7	58.5	61
B^3	74	72.2	73.1
CEAF(E)	60.1	60.1	60.1
CEAF(M)	46.8	51.5	49
BLANC	74.3	78	76
(CEAF(E)+MUC+B^3)/3			61.02

Table 10: Open results of the Chinese test data

The results of the gold-mention-boundaries and gold-mentions data of the English and Chinese closed task are shown in table 11 and 12. Although the mention detection stage is optimized by the gold-mention-boundaries and gold-mentions data and the final performance is enhanced, there is still space to enhance in the coreference resolution stage. The recall of mention detection of gold-mentions is 99.8. This problem will be explored in our future work.

Data	R	P	F1
Mention detection(A)	75.7	70.8	73.2
gold-mention-boundaries			59.50
Mention detection(B)	80	100	88.91
gold-mentions			69.88

Table 11: Results of the English closed gold-mention-boundaries and gold-mentions data, (A) is the mention detection score of the gold-mention-boundaries and (B) is the score of the gold-mentions.

Data	R	P	F1
Mention detection(A)	82.9	66.9	74.02
gold-mention-boundaries			64.42
Mention detection(B)	81.7	99.8	89.85
gold-mentions			76.05

Table 12: Results of the Chinese closed gold-mention-boundaries and gold-mentions data

6 Conclusion

In this paper we described a mixed deterministic model for coreference resolution of English and Chinese. We start the mention detection from extracting candidates based on the parse feature. The pre-processing which contains static rules and decision tree is applied to remove the defective candidates. In the coreference resolution stage the task is divided into several sub-problems and for each sub-problem the deterministic rules are constructed based on limited features. For the Chinese closed task we use regular patterns to generate named entities, gender and number from the training data. Several tools and dictionaries are applied for the Chinese open task. The result is not as good as we supposed since the feature errors caused by these tools also made the coreferential errors.

However, a deeper error analysis is needed in the construction of deterministic rules. The feature of the predicate arguments is not used well. Although the open performance of the Chinese task is not good, we still believe that complete and accurate prior knowledge can help solve the task.

Acknowledgement

This work is supported in part by the National Natural Science Foundation of China (No. 61173075 and 60973076), ZTE Foundation and Science and Technology Program of Shenzhen.

References

Bo Yuan, Qingcai Chen, Xiaolong Wang, Liwei Han. 2009. Extracting Event Temporal Information based on Web. 2009 Second International Symposium on Knowledge Acquisition and Modeling, pages.346-350

Cicero Nogueira dos Santos, Davi Lopes Carvalho. 2011. Rule and Tree Ensembles for Unrestricted Coreference Resolution. Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 51–55.

Emili Sapena, Llu´ıs Padr´o and Jordi Turmo. 2011. RelaxCor Participation in CoNLL Shared Task on Coreference Resolution. Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 35–39.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky. 2011. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 28–34, Portland, Oregon.

Liu Ting, Che Wanxiang, Li Zhenghua. 2011. Language Technology Platform. Journal of Chinese Information Processing. 25(6): 53-62

Jun Lang, Bing Qin, Ting Liu, Sheng Li. 2007. Intra-document Coreference Resolution: The state of the art. Journal of Chinese Language and Computing. 17 (4):227-253

Kai-Wei Chang Rajhans Samdani. 2011. Inference Protocols for Coreference Resolution. Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 40–44, Portland, Oregon.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, Christopher Manning. 2010. A Multi-Pass Sieve for Coreference Resolution. In EMNLP.

Kong Fang, Zhu Qiaoming and Zhou Guodong. 2012(a). Anaphoricity determination for coreference resolution in English and Chinese languages. Computer Research and Development (Chinese).

Kong Fang and Zhou Guodong. 2012(b). Tree kernel-based pronoun resolution in English and Chinese languages. Journal of Software (Chinese). Accepted: 23(8).

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel and Nianwen Xue.2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011). Portland, OR.

Sameer Pradhan and Alessandro Moschitti and Nianwen Xue and Olga Uryupina and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012). Jeju, Korea.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping Path-Based Pronoun Resolution Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 33–40, Sydney

Wang Houfeng. 2002. Survey: Computational Models and Technologies in Anaphora Resolution. Journal of Chinese Information Processing. 16(6): 9-17.