Chinese Coreference Resolution via Ordered Filtering*

Xiaotian Zhang[1,2] **Chunyang Wu**[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, Shanghai, China, 200240
xtian.zh@gmail.com, chunyang506@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

We in this paper present the model for our participation (BCMI) in the CoNLL-2012 Shared Task. Following the work of (Lee et al., 2011), we extend their English deterministic coreference resolution model to Chinese. This paper describes a pure rule-based method, which assembles different filters in a proper order. Different filters handle different situations and the filtering strategies are designed manually. These filters are assigned to different ordered tiers from general to special cases. We participated in the Chinese and English closed tracks, scored 54.21 and 59.24 respectively.

1 Introduction

This paper presents the approaches that we utilized for our participation in the CoNLL-2012 Shared Task. This year's shared task targets at modeling coreference resolution for multiple languages. Following (Lee et al., 2011), we extend the methodology of deterministic coreference model, using manually designed rules to recognize expressions with corresponding entities. The deterministic coreference model (Raghunathan et al., 2010) has shown quite good performance in the shared task of CoNLL-2011. This kind of model focuses on filtering with ordered tiers: One filter is applied at every turn, from highest to lowest precision. However, compared with statistical machine learning approaches (Soon et al., 2001), since effective rules are quite heterogeneous in different languages, several filtering methods should be redesigned when different languages are considered. We thus modified the original Stanford English coreference system[1] to let it adapt to the Chinese scenario. For the English participation, we implemented the semantic-based filters which can not be obtained from the open source toolkit.

The rest of this paper is organized as follows: In Section 2, the related works are reviewed; In Section 3, the details of our model of handling coreference resolution in Chinese are given; Experimental results are reported in Section 4 and the conclusion is in Section 5.

2 Related Works

Many existing works have been on learning relation extractors via supervised (Soon et al., 2001) or unsupervised (Haghighi and Klein, 2010; Poon and Domingos, 2008) approaches. For involved semantics, (Rahman and Ng, 2011) proposed a coreference resolution model with world knowledge; By using word associations, (Kobdani et al., 2011) showed its effectiveness to coreference resolution. Compared with machine learning methods, (Raghunathan et al., 2010) proposed a rule-base model which had been witnessed good performance.

* This work was partially supported by the National Natural Science Foundation of China (Grant No. 60903119 and Grant No. 61170114), the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No. 20110073120022, the National Basic Research Program of China (Grant No. 2009CB320901) and the European Union Seventh Framework Programme (Grant No. 247619).

† Corresponding author.

[1]http://nlp.stanford.edu/software/dcoref.shtml

Researchers began to work on Chinese coreference resolution at a comparatively late date and most of them adopt a machine learning approach. (Guochen and Yunfei, 2005) reported their Chinese personal pronoun coreference resolution system based on decision trees and (Naiquan et al., 2009) realized a Chinese coreference resolution system based on maximum entropy model. (Weixuan et al., 2009) proposed a SVM-based approach to anaphora resolution of noun phrases in Chinese and achieved the F-measure of 63.3% in the evaluation on ACE 2005. (Guozhi et al., 2011) presented a model for personal pronouns anaphora resolution based on corpus,which used rule pretreatment combined with maximum entropy.

3 Model for Chinese

In general, we adapt Stanford English coreference system to Chinese by making a series of necessary changes. The sketch of this deterministic model is to extract mentions and relevant information firstly, then several manually designed rules, or filtering sieves are applied to identify the coreference. Moreover, these sieves are utilized in a pre-designed order as same as (Lee et al., 2011), which are sorted from highest to lowest precision. The ordered filtering sieves are listed in Table 1.

Ordered Sieves
1. **Mention Detection Sieve**
2. **Discourse Processing Sieve**
3. Exact String Match Sieve
4. **Relaxed String Match Sieve**
5. Precise Constructs Sieve
6. **Head Matching Sieves**
7. **Proper Head Word Match Sieve**
8. **Pronouns Sieve**
9. Post-Processing Sieve

Table 1: Ordered filtering sieves for Chinese. Modified sieves are bold.

We remove the semantic-based sieves due to the resource constraints for Chinese closed track. The simplified version consists of nine filtering sieves. The bold ones in Table 1 are the modified sieves for Chinese. First of all, we adopt the head finding rules for Chinese used in (Levy and Manning, 2003), and this affects sieve 4, 6 and 7 which all take advantage of the head words. And our changes to other sieves are described as follows.

- **Mention Detection Sieve:** We in this sieve first extract all the noun phrases, pronouns (the words with part-of-speech (POS) tag **PN**), proper nouns (the words with POS tag **NR**) and named entities. Thus a mention candidate set is produced. We then refine this set by removing several types of candidates listed as follows:

  1. The *measure word*s, a special word pattern in Chinese such as “一年” (a year of), “一吨” (a ton of).
  2. Cardinals, percents and money.
  3. A mention if it is nested in a larger mention with the same head word.

- **Discourse Processing & Pronouns Sieve:** In these two sieves, we adapt the common pronouns to Chinese. They include “你” (you), “我” (I or me),“他” (he or him),“她” (she or her),“它” (it),“你们” (plural of “you”), “我们” (we or us),“他们” (they, gender: male),“她们” (they, gender: female),“它们” (plural of “it”) and “自己” (self). In addition, we enrich the pronouns set by adding “咱”, “咱们”, “俺” and “俺们” which are more often to appear in spoken dialogs as first person pronouns, and “您” which is the polite expression in Chinese for the second person pronoun “you”, and the third person pronoun “其”.

For mention processing of the original system (Lee et al., 2011), whether a mention is *singular* or *plural* should be given. However, different from English POS tags, in Chinese plural nouns cannot be distinguished from single nouns in terms of the POS. Therefore, we add two rules to judge whether a noun is plural or not.

- A noun that ends with “们” (plural marker for pronouns and a few of nouns that represent living things), and “等” (etc, and so on) is plural.

- A noun phrase that includes the coordinating conjunction words such as “和” (and) is plural.

4 Experiments

4.1 Modification for the English system

We implement the semantic-similarity sieves proposed in (Lee et al., 2011) with the WordNet (Stark and Riesenfeld, 1998). These modifications consider the alias sieve and lexical chain sieve. For the alias sieve, two mentions are marked as aliases if they appear in the same synset which contains a unordered set of words that denote the same concept and are interchangeable in many contexts in WordNet. For the lexical chain sieve, two mentions are marked as coreference if they are linked by a WordNet lexical chain that traverses hypernymy or synonymy relations.

4.2 Numerical Results

Lang.	Coref	Anno.	R	P	F
Ch	Before	gold	87.78	40.63	55.55
		auto	80.37	38.95	52.47
	After	gold	69.56	62.77	65.99
		auto	65.02	59.76	62.28
En	Before	gold	93.65	42.32	58.30
		auto	88.84	40.17	55.32
	After	gold	77.49	74.59	76.01
		auto	72.88	74.53	73.69

Table 2: Performance of the mention detection component, before and after coreference resolution, with both gold and auto linguistic annotations on development set.

Lang.	R	P	F
Ch	67.52	65.33	66.41%
En	75.23	72.24	73.71%

Table 3: Performance of the mention detection component, after coreference resolution, with auto linguistic annotations on test set.

Table 2 shows the performance of mention detection both before and after the coreference resolution with gold and predicted linguistic annotations on development set. The performance of mention detection on test set is presented in Table 3. The recall is much higher than the precision so as to make sure less mentions are missed, and because spurious mentions are left as singletons and removed at last, a low precision does not affect the final result.

	Metric	R	P	F1	avg F1
Ch	MUC	56.11	52.81	54.41	54.21
	BCUBED	68.27	64.49	66.33	
	CEAF (M)	51.92	51.92	51.92	
	CEAF (E)	40.39	43.47	41.88	
	BLANC	69.16	65.03	66.80	
En	MUC	64.08	63.57	63.82	59.24
	BCUBED	66.45	70.71	68.51	
	CEAF (M)	57.24	57.24	57.24	
	CEAF (E)	45.13	45.67	45.40	
	BLANC	71.12	77.92	73.95	

Table 5: Results on the official test set (closed track).

Our results on the development set for both languages are listed in Table 4 and the official test results are in Table 5. Avg F1 is the arithmetic mean of MUC, B3, and CEAFE.

We further examine the performance by testing on different data types (broadcast conversations, broadcast news, magazine articles, newswire, pivot text, conversational speech, and web data) of the development set, and the results are shown in Table 6. The system does better on *bn, mz, tc* than *bc, nw, wb* for both Chinese and English. And it performs the worst on wb due to a relative lower recall in mention detection. For Chinese, we also compare the performance when handling the three different mention types, proper nominal, pronominal, and other nominal. Table 7 shows the scores output by the official scorer when only each kind of mentions are provided in the keys file and response file each time and both the quality of the coreference links among the nominal of each mention type and the corresponding performance of mention detection are presented. The performance of coreference resolution among proper nominal and pronominal is significant higher than that of other nominal which highly coincides with the results in Table 6.

5 Conclusion

This paper presents the rule-base approach for the BCMI's participation in the shared task of CoNLL-2012. We extend the work by (Lee et al., 2011) and modified several tiers to adapt to Chinese. Numerical results show the effectiveness in the evaluation for Chinese and English. For the Chinese scenario, we firstly show it is possible to consider special POS-tags and common pronouns as indicators for

Lang.	Setting	MUC			BCUBED			CEAF (E)			avg F1
		R	P	F1	R	P	F1	R	P	F1	
Ch	AUTO	52.38	47.44	49.79	68.25	62.36	65.17	37.43	41.89	39.54	51.50
	GOLD	58.16	53.55	55.76	70.66	68.65	69.64	41.44	45.60	43.42	56.27
	GMB	63.60	87.63	73.70	62.71	88.32	73.34	74.08	42.83	54.28	67.11
En	AUTO	64.24	64.95	64.59	68.22	73.16	70.60	47.03	46.29	46.66	60.61
	GOLD	67.45	66.94	67.20	69.76	73.62	71.64	47.86	48.42	48.14	62.33
	GMB	71.78	90.55	80.08	65.45	88.95	75.41	77.42	46.47	58.08	71.19

Table 4: Results on the official development set (closed track). GMB stands for Gold Mention Boundaries

Lang.	Anno.	bc	bn	mz	nw	pt	tc	wb
Ch	AUTO	50.31	53.87	52.80	47.82	-	55.10	47.54
	GOLD	53.19	63.63	58.23	50.65	-	58.96	50.15
En	AUTO	59.26	62.40	63.17	57.57	65.24	60.91	56.88
	GOLD	60.34	64.51	64.36	59.71	67.07	62.44	58.47

Table 6: Results (Avg F1) on different data types of the development set (closed track).

Data Type	Proper nominal		Pronominal		Other nominal	
	MD (Recall)	avg F1	MD (Recall)	avg F1	MD (Recall)	avg F1
bc	94.5 (550/582)	68.06	94.5 (1372/1452)	66.40	80.5 (1252/1555)	47.74
bn	96.7 (1213/1254)	67.46	97.8 (264/270)	77.39	83.7 (1494/1786)	53.51
mz	92.0 (526/572)	67.05	94.8 (91/96)	56.89	76.1 (834/1096)	53.68
nw	91.4 (402/440)	67.44	90.6 (29/32)	83.54	51.0 (1305/2559)	44.86
tc	100 (23/23)	95.68	84.5 (572/677)	61.96	71.2 (272/382)	53.88
wb	93.2 (218/234)	72.23	95.9 (397/414)	72.55	77.1 (585/759)	43.37
all	94.4 (2932/3105)	68.30	92.7 (2725/2941)	68.10	70.6 (5742/8137)	49.56

Table 7: Results (Recall of mention detection and Avg F1) on different data types and different mention types of the development set with linguistic annotations (closed track).

improving the performance. This work could be extended by introducing more feasible filtering tiers or utilizing some automatic rule generating methods.

References

Li Guochen and Luo Yunfei. 2005. 采用优先选择策略的中文人称代词的指代消解. (Personal pronoun coreference resolution in Chinese using a priority selection strategy). *Journal of Chinese Information Processings*, 19:24–30.

Dong Guozhi, Zhu Yuquan, and Cheng Xianyi. 2011. Research on personal pronoun anaphora resolution in Chinese. *Application Research of Computers*, 28:1774–1776.

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 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.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 439–446, Sapporo, Japan, July. Association for Computational Linguistics.

Hu Naiquan, Kong Fang, Wang Haidong, and Zhou Guodong. 2009. Realization on Chinese coreference resolution system based on maximum entropy model. *Application Research of Computers*, 26:2948–2951.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 650–659, Honolulu, Hawaii, October. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nate Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Cambridge, MA, October. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 814–824, Portland, Oregon, USA, June. Association for Computational Linguistics.

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

Michael M. Stark and Richard F. Riesenfeld. 1998. Wordnet: An electronic lexical database. In *Proceedings of 11th Eurographics Workshop on Rendering*. MIT Press.

Tan Weixuan, Kong Fang, Wang Haidong, and Zhou Guodong. 2009. Svm-based approach to Chinese anaphora resolution. In *Proceedings of China National Computer Congress*, Tianjin, China, October.