A Learning Based Model for Chinese Co-reference Resolution by Mining Contextual Evidence

Feifan Liu and Jun Zhao

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
{ffliu, jzhao}@nlpr.ia.ac.cn

Abstract. This paper presents a learning based model for Chinese co-reference resolution, in which diverse contextual features are explored inspired by related linguistic theory. Our main motivation is to try to boost the co-reference resolution performance only by leveraging multiple shallow syntactic and semantic features, which can escape from tough problems such as deep syntactic and semantic structural analysis. Also, reconstruction of surface features based on contextual semantic similarity is conducted to approximate the syntactic and semantic parallel preferences in resolution linguistic theories. Furthermore, we consider two classifiers in the machine learning framework for the co-reference resolution, and performance comparison and combination between them are conducted and investigated. We experimentally evaluate our approaches on standard ACE (Automatic Content Extraction) corpus with promising results.

1 Introduction

Co-reference resolution refers to the problem of determining whether discourse references in text correspond to the same real world entities [1]. In the context of ACE (Automatic Context Extraction) we address only specified set of entities [2] for co-reference resolution in Chinese texts here. A mention is a referring expression of an object, and a set of mentions referring to the same object within a document constitute an entity, i.e. an equivalence class of mentions. For example, in the following sentence, mentions are nested bracketed:

"[[微软公司/Microsoft Company]总裁/president][比尔盖茨/Bill Gates]表示/stated[微软/Microsoft]与此事件无关/has nothing to do with this issue，[公司/Company]不需要做任何解释/don't need to give any explanations。"

"微软公司/Microsoft Company", "微软/Microsoft" and "公司/Company" constitute a entity since they refer to the same object. Likewise, "微软公司/Microsoft Company总裁/president" and "比尔盖茨/Bill Gates" also constitute a entity.

Recent research in co-reference resolution has exhibited a shift from knowledge-based approaches to data-driven approaches, yielding learning-based co-reference

* This work was supported by the National Natural Sciences Foundation of China (60372016) and the Natural Science Foundation of Beijing (4052027).

L. Wang et al. (Eds.): FSKD 2006, LNAI 4223, pp. 1249 – 1258, 2006.

systems [3][4][6][9]. These approaches recast the problem as a classification task. Specifically, a pair of mentions is classified as co-referring or not based on a statistical model learned from the training data. Then a separate linking algorithm coordinate the co-referring mentions pairs and partition all the mentions in the document into entities. Soon [4] has been commonly used as a baseline system for comparison under this learning based framework, and many extensions have been conducted at different points. Yang [9] and Strube [10] made improvements in string matching strategy and got good results. Ng [3] proposed a different link-best strategy, and Ng [6] presented a novel ranking approach for partitioning mentions in linking stage.

This paper proposes a Chinese co-reference resolution system employing the statistical framework. Unlike existing work, we focus on exploring the contribution of diverse contextual features inspired by linguistic findings. First, incorporating diverse contextual features try to capture the syntactic structural information, which is inspired by the syntactic constrain rules for anaphora resolution. Second, an information reconstruction method based on contextual similarity is proposed to approximate syntactic and semantic parallel preferences, which plays an important role in co-reference resolution according to linguistic findings. We experimentally evaluate our approaches on standard ACE corpus with promising results.

2 Learning-Based Chinese Co-reference Resolution (Baseline)

Our framework for co-reference resolution is a standard combination of classification and clustering as mentioned above. First, we establish a Chinese co-reference resolution system based on [4] as in figure 1. Note that the dashed part is for offline training.

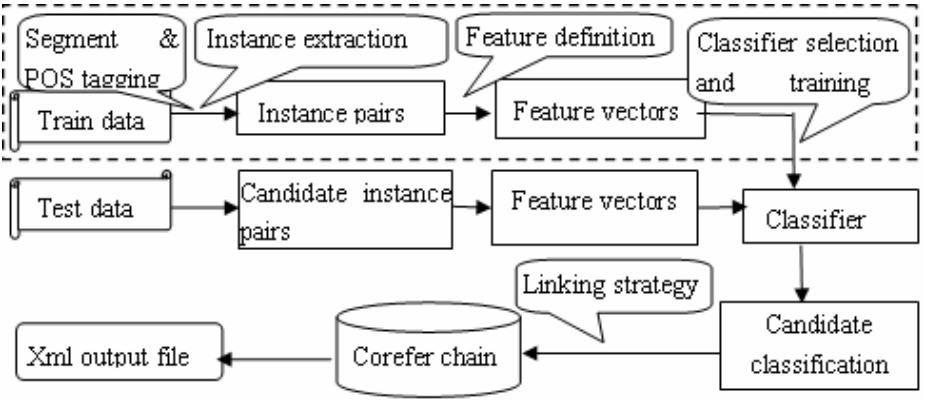

Fig. 1. Statistical Framework for Chinese Co-reference Resolution

2.1 Instance Extraction

Here an instance is a pair of entity mentions (EM) which are either CO-REFERENT or NOT CO-REFERENT. The former is called positive instance, and the latter is

called negative instance. We obtained these instances from the ACE training corpus. And we use the extracting strategy in [4] in our baseline system.

Also, since the named and nominal mentions are typed, and we will consider only instances where two mentions belong to the same type, e.g. we will not extract negative instances with two non-pronominal mentions whose types are inconsistent.

2.2 Feature Definition

Every instance is represented by a feature vector. In our baseline system, we try to simulate the feature set in [4] illustrated in Table 1. We use "I" and "J" to denote EM_i and EM_j in an instance respectively. Note that we make some adaptations or modifications according to Chinese characteristics, marked by star symbol.

(1) StringMatch

Since there is no sufficient information in Chinese, such as capitalized information, to determine alias abbreviation, or shorted form of named and nominal mentions, we modify the string match strategy, replacing the binary match feature and alias feature with matching degree feature. A simple matching function is designed as follows.

$$MatchDegree = \frac{\sum_{w_i \in \{C_m\}} len(w_i)}{\max\{len(EM_i), len(EM_j)\}} \quad (1)$$

where C_m is the matched word set of two mentions, $len(.)$ is measured by characters.

Table 1. Feature Set for the Simulated Soon et al. Baseline System

Feature type	Feature	Feature description
Lexical	StringMatch *	real number value between 0 and 1 by equation(1)
Grammatical	MenType_I	"NOM" if EM_i is nominal mentions; "NAM" if named mentions; "PRO" if pronominal mentions.
	MenType_J	same definition to MenType_I
	Definite_I	"1" if EM_i contains words with definitive and demonstrative sense, such as"□/the,this", "那些/those", else "-1"
	Definite_J	Same definition to Definite_I
	Number	"1" if EM_i and EM_j agree in number; "-1" if they disagree; "0" if they can't be determined
	Gender	"1" if EM_i and EM_j agree in gender; "-1" if they disagree; "0" if they can't be determined
	Appositive*	"1" if mentions are in an appositive relationship;else "-1"
Semantic	EntityType*	"1" if EM_i and EM_j are consistent in entity type; "-1" if they are not; "0" if they can't be determined
Positional	Para*	"1" if EM_i and EM_j are in different paragraphs; else"-1"
	SenNum*	see equation (2)
	SubSenNum*	see equstion (3)

(2) Positional Features

Only the sentence number between two mentions is considered in Soon et al. system. Here we extend this type of feature by adding cross paragraph and cross sub-sentence feature. Sentence is delimited by full stop, question mark, or exclamatory mark, while sub-sentence is delimited by colon, semicolon, or comma. We define SenNum and SubSenNum as follows.

$$SenNum = \begin{cases} SenNum & if\ \ SenNum <= 2 \\ "SenNum > 2" & if\ \ 2 < SenNum <= 5 \\ "SenNum > 5" & if\ \ 5 < SenNum <= 10 \\ "SenNum > 10" & if\ \ SenNum > 10 \end{cases} \tag{2}$$

$$SubSenNum = \begin{cases} SubSenNum & if\ \ SubSenNum <= 2 \\ "SenNum > 2" & if\ \ 2 < SubSenNum <= 5 \\ "SenNum > 5" & if\ \ SubSenNum > 5 \end{cases} \tag{3}$$

2.3 Co-reference Classifier Selection

Diverse machine learning methods have been used for co-reference resolution, such as decision tree(DT) model C4.5[3][4][6], maximum entropy(ME) model[6][11], support vector machine(SVM) model[7][12], and etc. Bryant's work [12] proved experimentally that SVM model (F-value: 72.4) outperform the traditional DT model (F-value: 70.7) in the machine learning framework for co-reference resolution. We consider two learning models in our baseline system: SVM and ME. Our motivation is to compare the two models' performance in the context of co-reference resolution and try some combining strategy on them.

2.4 Linking Strategy

Linking strategy is used to apply the classifier predictions to create co-reference chains. The most popular linking strategy is the link-first strategy [4], which links a mention, EM_j, to the first preceding mention, EM_i, predicated as co-referent. An alternative linking strategy, which can be called link-best strategy [3], links a mention, EM_j, to the most probable preceding mention, EM_i, where the probability is measured by the confidence of the co-reference classifier prediction.

3 Incorporating Multi-level Contextual Evidence

Co-reference is a discourse-level problem, which depends on not only the two candidate mentions themselves but also diverse contextual evidence. Here are two examples.

(1) 小明/Xiaoming 说/said 在学校/in school 数学/math 老师/ teacher 常/often 责怪/blame他/he。

(2) 李刚/Ligang 常/often 找/go with 刘辉/Liuhui 打篮球/play basketball，徐庆/Xuqing 常/often 找/go with 他/he 去游泳/go swimming。

In the two above sentence, it is very hard for a classifier to predict correct co-referent relations between underlined mentions only using features in the baseline system. But this can be well explained and resolved by linguistic findings on constrains and preferences involved in traditional anaphora resolution theory [13]. Syntactic constrains (他/he≠数学老师/math teacher) should be considered in sentence (1), and syntactic constrains and semantic parallelism preferences (他/he＝刘辉/Liuhui) should be used in sentence (2).

Deep syntactic and semantic knowledge, however, is quite difficult for current Chinese processing technology. So we try to mine multi-level contextual features and investigate their contribution to system performance accordingly. We hope to capture the properties implied in deep syntactic and semantic analysis by incorporating multi-level surface features and reconstruct them using some strategy.

3.1 Word Form and POS (Part of Speech) Evidence

Word forms and POS of them are the most fundamental contextual features at lexical and shallow syntactic level. For each of the two mentions in question, we consider a 5-width window to extract those contextual cues to enrich the feature vector of a training instance, which is expected to be helpful in co-reference resolution.

3.2 Bag of Sememes (BS) Feature

Although deep semantic analysis is not available, we can resort to a Chinese-English knowledge base called HowNet (http://www.keenage.com) to acquire shallow semantic features. HowNet is a bilingual common-sense knowledge base, which uses a set of non-decomposable sememes to define a sense of a word. A total of over 1,600 sememes are involved and they are organized hierarchically. For example, the sense of "研究所/research institute" is defined as "InstitutePlace|场所,*research|研究,#knowledge|知识". Here there are three sememes split by commas, and symbols such as "*" represent specific relations.

So we can acquire a set of sememes for each word in the contextual window. Bag of sememes (BS) is used for modeling the semantic context of each mention, including preceding context BS and post context BS.

3.3 Feature Reconstruction Based on Contextual Semantic Similarity

There are two problems in using BS features in Section 3.2. First, we don't use any disambiguation strategy in extracting word sense from HowNet, which can introduce harmful noises. Second, we just use bag of sememes to model the context, losing the useful associated information between sememes from the same word. We will make it manifest by looking at two different contexts: "校园/schoolyard 歌手/singer" and "学生/student 歌厅/singing hall". Sense for every word and BS features of context are shown in figure 2.

Sense representation of each word in HowNet:

1.校园/schoolyard: {part|部件,%InstitutePlace|场所,education|教育}

2.歌手/singer: {human|人,entertainment|艺,*sing|唱,#music|音乐}

3.学生/student: { human|人,*study|学,education|教育}

4.歌厅/singing hall: {InstitutePlace|场所,@recreation|娱乐,@sing|唱}

Bag of sememes feature of two contexts:

1. 校园/schoolyard 歌手/singe: {part|部件,**InstitutePlace|**场所,**education|**教育,**human|**人,entertainment|艺,**sing|**唱,music|音乐}

2. 学生/student 歌厅/singing hall: {**human|**人,study|学,**education|**教育,**InstitutePlace|**场所,recreation|娱乐,**sing|**唱}

Fig. 2. Bag of Sememes of the Two Example Contexts

From Figure 2 we can see that, although the two contexts are semantically discrepant, their BS features are similar to a large extent.

We now have no effective way to resolve the first problem. Motivated by linguistic findings, we can make improvements on modeling the context related to the second question. Undoubtedly, BS features can express the semantic property of the context, which can help us to approximate the semantic parallelism in anaphora theory. But the information in BS is unordered without any relational or structural information. Intuitively, we can use some strategy to reconstruct those unordered BS features to give a better representation of context.

A word similarity computing approach [8] based HowNet is used for reference in our information reconstruction method. We regard context as a set of word and contextual similarity can be acquired by computing the similarity between two word sets.

In our case of computing word similarity, we didn't consider the relation sememe description and the relation symbol description as [8] does. We only consider two aspects: one is the similarity of first basic sememe between two words, denoted as $Sim_1(S_1, S_2)$; one is the similarity of all the other basic sememes in word sense representation from HowNet, denoted as $Sim_2(S_1, S_2)$. Here S_1 and S_2 are sense representations for W_1 and W_2 in consideration. So we compute the similarity of two words using the following equation.

$$Sim(W_1, W_2) = \beta_1 Sim_1(S_1, S_2) + \beta_2 Sim_2(S_1, S_2) \tag{4}$$

where $\beta_1 + \beta_2 = 1$. Since the first basic sememe indicates the most important semantic feature, we set $\beta_1 = 0.7$. Similarity between two sememes is computed according to the distance in the sememe hierarchy of HowNet [8].

Now we turn to the computation of contextual similarity based on word similarity. As mentioned above, contextual similarity can be formulated as a problem of computing similarity between two word sets. We should first find the possible corresponding word pairs between two sets and compute the arithmetical average of the similarity

```
ContextSim ← 0.0
ArrayWordSim ←∅
for each w_i∈C_1
    for each w_j∈C_2
        WordSim->value ← Sim(w_i,w_j), WordSim->index_1←i, WordSim->index_2 ←j
        ArrayWordSim ← ArrayWordSim ∪ {WordSim}
    end for
end for
DeSort(ArrayWordSim)    // sort in decreasing order of similarity value
ProcessedIndex_1←∅, ProcessedIndex_2←∅
for each k = 0,1,2,…,min{ size(C_1),size(C_2)}
    if(WordSim_k->index_1∉ProcessedIndex_1 && WordSim_k->index_2∉ProcessedIndex_2)
        ContextSim←ContextSim+WordSim_k->value
        ProcessedIndex_1 ← ProcessedIndex_1 ∪ {WordSim_k->index_1}
        ProcessedIndex_2 ← ProcessedIndex_2 ∪ {WordSim_k->index_2}
    end if
end for
ContextSim ← ContextSim / min{ size(C_1),size(C_2)}
```

Fig. 3. Description of Computing Algorithm for Contextual Similarity

values of all the corresponding word pairs. Let C_1 and C_2 denote the word set containing the words in the context of EM_i and EM_j respectively. The algorithm description for contextual similarity computation is illustrated in figure 3.

By calculating contextual similarity, the unordered BS features are reconstructed. The computation is word-based, and the first sememe similarity is given a larger weight, so the BS shortcomings discussed in figure 2 can be overcome to some extent.

4 Experiments and Analysis

In our experiments, two standard classification toolkits are used, namely Maximum Entropy Toolkit (MaxEnt)[1] and Support Vector Machine Toolkit (libSvm)[2]. Parameters in the models are selected by 5-fold cross validation.

4.1 Experiment Data and Evaluation Metric

We now focus on the empirical performance analysis of an implementation of the statistical co-reference model described above. We evaluate the co-reference system on the standard ACE-05 co-reference data. The co-reference classifier is trained on 80% of the data set, and other 20% is used for testing the co-reference resolution systems. Statistics of train data and test data is shown in table 2.

[1] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html
[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Table 2. Statistics on Experiment Data

	#Doc	#Entity Mention				#Co-reference Chain
		Named	Nominal	Pronominal	Total	
Train data	511	11649	12952	2763	27364	12258
Test data	122	3048	3326	583	6957	3156

Performance of our co-reference system is reported in terms of recall, precision, and F-measure using the model-theoretic MUC scoring program [14].

4.2 Impact of Multi-level Contextual Features on System Performance

This experiment reports the performance of our baseline system and explores the contribution of multi-level contextual features, which is shown in table 3. In table 3, "CR" denotes Co-reference Resolution, "WP" denotes the word and POS features, "BS" denotes the semantic feature represented by Bag of Sememes, and also classification accuracy of classifiers is given for reference.

Table 3. Improved Performances by Incorporating Multi-level Contextual Features (SVM)

	Classification Accuracy	CR recall	CR precision	CR F-value
Baseline System	85.49%	61.06%	93.4%	73.84%
Baseline+WP	85.85%	74.82%	85.37%	79.74%
Baseline+WP+BS	86.27%	74.56%	86.06%	**79.89%**

Table 3 shows that incorporating the contextual lexical and shallow syntactic feature (WP) acquires significant increase in recall, but some drops in precision. The resulting F-value, however, increase non-trivially from 73.84% to 79.74%. The introduction of BS features based on HowNet can further boost the system's performance.

The experimental results are largely consistent with our hypothesis. System performance improves dramatically by applying diverse contextual features. This can be explained that combining multiple contextual features can capture the syntactic constrain information which is definitely helpful for co-reference resolution according to traditional linguistic findings.

4.3 Performance Comparison Between Different Classifiers

We consider two classifiers, ME(Maximum Entropy) and SVM(Support Vector Machine), in our machine learning co-reference resolution framework. Comparison results under three different configurations are demonstrated in table 4.

The results reveal that SVM performs better than ME under all the three configurations. From the principle point of view, ME is a probability model based on log-linear

Table 4. Performance Comparison between ME and SVM

	Baseline			Baseline+WP			Baseline+WP+BS		
	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)
ME	59.46	89.72	71.52	65.43	88.25	75.14	66.27	86.71	75.13
SVM	61.06	93.40	73.84	74.82	85.37	**79.74**	74.56	86.06	**79.89**
SVM +ME	62.12	91.87	**74.12**	70.46	89.24	78.74	72.12	86.97	78.85

regression while SVM is classification model based on large-margin principle. For this reason, SVM can outperform ME when the training data is not much sufficient.

For three systems, only baseline system can benefit from the combination. When adding contextual features, combination doesn't help at all according to the results. We don't know exactly what the reason is now, but we guess it may have something to do with the learning mechanism and confidence measurement of the classifiers.

4.4 Performance Improvement by Similarity Based Information Reconstruction

In this section we try to investigate whether our motivation of modeling context by similarity based information reconstruction is valid or not.

From figure 4, we can see that RC (feature reconstruction) outperforms BS feature and get increase both in recall and precision. The resulting F-value increases from 79.89% to 80.28%. This commendably verifies our analysis in section 3. By contextual similarity, the semantic information can be leveraged in a more reasonable way.

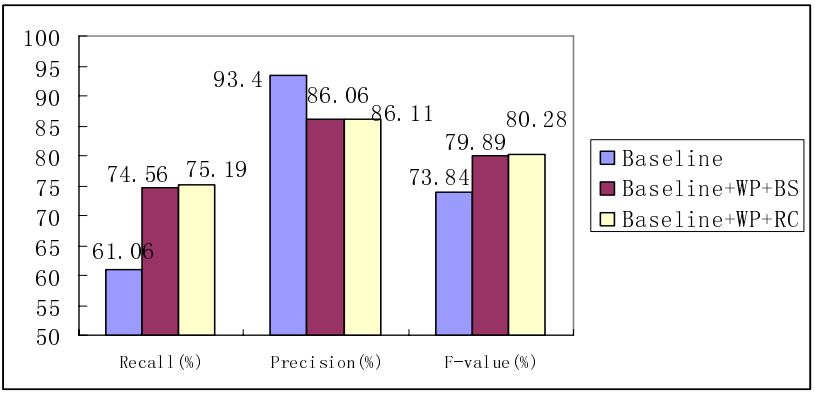

Fig. 4. Performance Improvement by Similarity Based Information Reconstruction

5 Conclusions and Future Work

We propose a learning based model for Chinese co-reference resolution and investigate multiple contextual features to improve the system performance based on related linguistic theory. Experimental results prove that our approach performs very well on the standard ACE data sets without deep syntactic and semantic analysis.

Our future work will focus on the following.

- How to reduce the noise of introducing semantic features more efficiently;
- How to find global features useful for Chinese co-reference resolution.

References

1. R. Mitkov. 2002. Anaphora Resolution. Long-man.
2. NIST. 2005. The Official Evaluation Plan for the ACE 2005 Evaluation.
3. V. Ng and C. Cardie. 2002. Improving Machine Learning Approaches to coreference resolution. In Proc. of the ACL 2002, pages: 104-111.
4. W.M. Soon, H. T. Ng, and D. Lim. 2001. A Machine Learning Approach to Co-reference Resolution of Noun Phrases. Computational Linguistics, 27(4):521-544.
5. M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A Model-Theoretic Coreference Soring Scheme. In Proc. of MUC-6, pages: 45-52.
6. Vincent Ng. 2005. Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. In Proceedings of ACL 2005, Ann Arbor, MI, June 2005, pp. 157-164.
7. R. Iida, K. Inui, H. Takamura, and Y. Matsumoto. 2003. Incorporating contextual cues in trainable models for coreference resolution. In Proc. of the EACL'03 Workshop on the Computational Treatment of Anaphora, pp: 23–30.
8. Qun Liu, Sujian Li. 2002. Word Similarity Computing Based on How-net. Computational Linguistics and Chinese Language Processing. Vol. 7, No. 2, pp: 59-76.
9. X. Yang, G. Zhou, J. Su & C. L. Tan. 2004. Improving noun phrase co-reference resolution by matching strings, In Proceedings of IJCNLP04, Lecture Notes in Computer Science, Volume 3248, pages 22 - 31. Hainan, China.
10. M. Strube, S. Rapp, and C. Muller. 2002. The Influence of Minimum Edit Distance on Reference Resolution. In Proc. of EMNLP-02, pages:312-319.
11. R Florian, H Hassan, A Ittycheriah, H Jing, N Kambhatla, X Luo, N Nicolov, and S Roukos. 2004. A statistical model for multilingual entity detection and tracking. In Proc. of HLT/NAACL-04, pp: 1-8, Boston Massachusetts, USA.
12. John Bryant. 2004. Combining Feature Based and Semantic Information for Co-reference Resolution. Research report at U.C. Berkeley and ICSI.
13. Daniel Jurafsky, James H. Martin. 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice-Hall.