Zhao J, Liu FF. Linguistic theory based contextual evidence mining for statistical Chinese co-reference resolution. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 22(4): 608~617 July 2007

Linguistic Theory Based Contextual Evidence Mining for Statistical Chinese Co-Reference Resolution

Jun Zhao (赵 军) and Fei-Fan Liu (刘非凡)

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China

E-mail: {jzhao,ffliu}@nlpr.ia.ac.cn

Received July 4, 2006; revised March 19, 2007.

Abstract Under statistical learning framework, the paper focuses on how to use traditional linguistic findings on anaphora resolution as a guide for mining and organizing contextual features for Chinese co-reference resolution. The main achievements are as follows. (1) In order to simulate "syntactic and semantic parallelism factor", we extract "bags of word form and POS" feature and "bag of semes" feature from the contexts of the entity mentions and incorporate them into the baseline feature set. (2) Because it is too coarse to use the feature of bags of word form, POS tag and seme to determine the syntactic and semantic parallelism between two entity mentions, we propose a method for contextual feature reconstruction based on semantic similarity computation, in order that the reconstructed contextual features could better approximate the anaphora resolution factor of "Syntactic and Semantic Parallelism Preferences". (3) We use an entity-mention-based contextual feature representation instead of isolated word-based contextual feature representation, and expand the size of the contextual windows in addition, in order to approximately simulate "the selectional restriction factor" for anaphora resolution. The experiments show that the multi-level contextual features are useful for co-reference resolution, and the statistical system incorporated with these features performs well on the standard ACE datasets.

Keywords natural language processing, information extraction, co-reference resolution, anaphora resolution

1 Introduction

Co-reference resolution is a quite challenging problem which becomes more and more important in a lot of natural language processing applications such as machine translation, information retrieval, discourse analysis etc. Co-reference resolution refers to the problem of determining whether discourse references in text correspond to the same real world entity[1]. In the context of ACE (Automatic Context Extraction)[2], all the mentions referring to the same object within a document will be clustered into an equivalent class, called entity class. Here a mention in discourse is a referring expression of an object, and an entity class is in fact a co-reference chain. For example, in Fig.1, the mentions are nestedly bracketed and the co-reference resolution results in two co-reference chains.

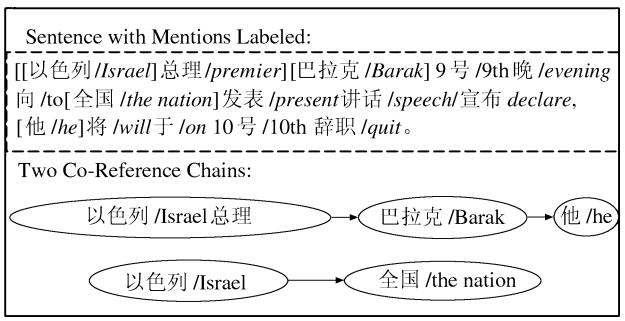

Fig.1. Mentions and co-reference chains in the context of ACE.

Note that we address a specific set of entities defined by ACE[2] for co-reference resolution in Chinese texts here. Under the framework of ACE, to recognize all the entity mentions in the texts are the precondition of co-reference resolution. The task of co-reference resolution is to find the correct co-reference links between mentions. In the next section, we will give a brief introduction to ACE.

Recent research in co-reference resolution has exhibited a shift from knowledge-based approaches to data-driven approaches, yielding learning-based co-reference resolution systems that rival their hand-crafted counterparts in performance[3~6]. Under this learning based framework, the algorithm put forward by Soon *et al.*[3] has been commonly used as a baseline system for comparison, and many extensions have been proposed from different points of view. Strube *et al.*[7] and Yang *et al.*[6] made improvements in string matching strategy and obtain good results. Ng and Cardie[4] proposed a different link-best strategy, and Ng[5] presented a novel ranking approach for partitioning mentions in linking stage.

Currently, for Chinese language processing, research work related to anaphora/co-reference resolution is mainly focused on pronominal resolution using the traditional rule-based methods[8,9], which mainly focus on local resolution for the pronominal mentions.

This paper aims to establish a Chinese co-reference resolution system employing the above statistical framework with some adaptations, which can not only realize the global resolution but also enhance the system's robustness. Unlike existing work, we focus on how to marriage linguistic findings used in anaphora resolution to

Regular Paper

Supported by the National Natural Science Foundation of China under Grant Nos. 60372016, 60121302, 60673042, the National High Technology Development 863 Program of China under Grant No. 2006AA01Z144, and the Natural Science Foundation of Beijing under Grant No. 4052027.

statistical approaches for co-reference resolution without deep analysis. 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.

In the following, Section 2 will give a brief introduction to ACE, a framework of content extraction, under which we study the problem of Chinese co-reference resolution; Section 3 describes a baseline system which we developed for statistical learning based Chinese co-reference resolution; Section 4 is an emphasis of the paper, which discusses how to mine multi-level contextual evidence based on linguistic findings on anaphora resolution for Chinese co-reference resolution, and how to incorporate these features into the baseline model; Section 5 will give the experiments to show whether the contextual features are beneficial for Chinese co-reference resolution under the statistical learning framework; finally Section 6 is the conclusion.

2 Brief Introduction to Framework of ACE

In the field of information extraction, there are two important evaluation programs, Message Understanding Conference (MUC)[10,11] and ACE[2]. MUC acted as a critical role in initiating the direction of information extraction and promoting the research in this field. However, the evaluation tasks of MUC have relatively limited coverage. For example, the task of filling the slots of Scenario Templates is conducted on the specific templates in the specific domains. This kind of task definition greatly limits the portability of information extraction technologies developed under the framework. Therefore, how to create adaptive information extraction systems becomes the hotspot of the research of information extraction in the current stage.

Another important evaluation program is ACE sponsored by NIST[2,12]. The evaluation program was initiated in 2002, which aims at developing adaptive content extraction technologies to support automatic processing of human language in three source types, namely newswire, broadcast news (with text derived from ASR), and newspaper (with text derived from OCR). In the following, we will briefly introduce the framework of ACE, that is excerpted from [12].

In the framework of ACE, Entities, Relations and Events are believed to be the most important elements for representing the content of the texts. The ACE research objectives are viewed as the detection and characterization of Entities, Relations and Events[12].

The main evaluation tasks in ACE are Entity Detection and Tracking (EDT), Relation Detection and Characterization (RDC), Event Detection and Characterization (VDC), and Entity Linking (LNK). EDT is the core annotation task, providing the foundation for all remaining tasks. The current ACE task identifies the entities such as Person, Organization, Location, Facility, Weapon, Vehicles and Geo-Political Entity, etc. Annotators tag all mentions of each entity within a document, whether named, nominal or pronominal. During the LNK annotation task, annotators review the entire document to group mentions of the same entity together[12]. Please refer to [12] for the description for RDC and VDC.

The paper is related to the task of LNK.

3 Baseline: Learning Based Model for Chinese Co-Reference Resolution

The focus of the paper is Chinese co-reference resolution. The input of the task is the texts which have been processed by EDT module, i.e., we have recognized all the mentions (including named, nominal and pronominal mentions) of the entities. LNK module will review the entire document to group the mentions of the same entity together. Fig.1 is an example for the LNK task.

The method we adopted is learning-based co-reference resolution, which recasting the task of co-reference resolution as a classification process. First, 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 coordinates the co-referring mention pairs and partitions all the mentions in the document into entity classes.

Our machine learning framework for co-reference resolution is a standard combination of classification and clustering as mentioned above. This learning based framework can be divided into four modules: Instance Extraction, Feature Definition, Classifier Selection, and Linking Strategy. Accordingly, we make some adaptations to the system of Soon *et al.*[3] for establishing our Chinese co-reference resolution framework, illustrated in Fig.2. Note that the dashed part is for offline training.

In the framework, *Instance Extraction* module is used to extract positive instances and negative instances from the ACE manually annotated corpus, in order to support the training of the co-refer classifier which will determine whether two entity mentions are co-referent or not; *Feature Definition* module is used to convert the training instances into feature vectors for the computation in training the co-refer classifier in the training phase, and to convert the testing instances into feature vectors for the computation in classifying the instances into co-refer or not in the testing phase; *Co-Reference Classifier Selection* module is used to select which kind of classification model to be used for co-refer classification, which will determine whether two entity mentions are co-refer or not; *Link Strategy* module is used to link the co-refer pairs into co-refer chains.

3.1 Instance Extraction

Here an instance is a pair of entity mentions (EM) which are either CO-REFERENT or NOT CO-REFE-

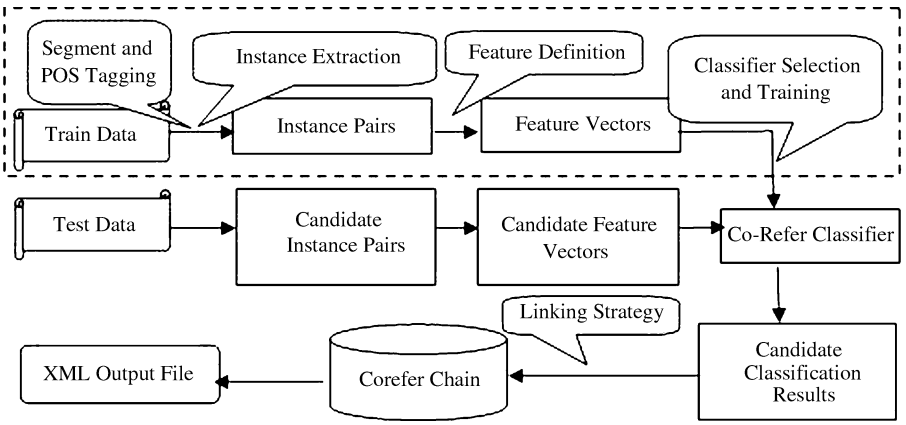

Fig.2. Learning based framework for Chinese co-reference resolution.

RENT. The former is called positive instances, and the latter is called negative instances. We obtained these instances from the ACE training corpus. A positive instance is extracted for each mention that is involved in a co-reference chain but is not the head of the chain. A negative instance is extracted for each of the remaining mentions.

More formally, let EM_k be the k-th EM in a document. An instance created from EM_i and EM_j $(i < j)$ is denoted by $Ins(EM_i, EM_j)$. Following previous work, a valid instance $Ins(EM_i, EM_j)$ should satisfy the constrains that EM_i precedes EM_j. So we can rely on co-reference chains from ACE training data to create 1) a positive instance for each mention in the co-reference chain, EM_j, and its closest preceding antecedent, EM_i; and 2) a negative instance for EM_j paired with each of the intervening EMs, i.e., EM_{i+1}, $EM_{i+2}, \ldots, EM_{j-1}$.

To make this more clearly, we give an example here.

ACE Data: [[以色列 1/Israel] 总理 2/premier] [巴拉克 2/Barak] 9 号/9th 晚/evening 通过/via [电视台 3/television station] 向/to [全国 1/the nation] 发表/present 讲话/speech 宣布/declare, [他 2/he] 将 .....

In the example, there are 6 bracketed mentions and 3 co-reference chains indexed by 1, 2 and 3 (which have been manually annotated), i.e.,

1) 以色列/Israel → 全国/the nation;

2) 以色列/Israel 总理/premier → 巴拉克/Barak → 他/he;

3) 电视台/television station.

The instance extraction module will get (以色列/Israel, 全国/the nation), (以色列总理/Israel premier, 巴拉克/Barak), (巴拉克/Barak, 他/he) as three positive instances, and (电视台/television station, 全国/the nation), (全国/the nation, 他/he), (电视台/television station, 他/he) as negative instances.

3.2 Feature Definition

Every instance, whether positive or negative, is represented using a feature vector. In our baseline system, we try to simulate the feature set proposed by Soon *et al.*[3] 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 morphological variance information in Chinese text which can be used to detect alias abbreviation or shorted form of named and nominal mentions, we modify the string matching strategy, replacing the binary matching feature and alias feature with matching degree feature. A simple matching function is designed as follows.

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

where $MatchDegree(EM_i, EM_j)$ is a function which describes the matching degree of two mentions EM_i and EM_j; C_m indicates the matched word set of the two mentions; $len(.)$ is measured by characters.

2) *Positional Features*

Soon *et al.*'s system only considered sentence number between two mentions[3]. Here we extend this type of feature by adding features indicative of cross paragraph and cross sub-sentence. Sentence is delimited by full stop, question mark, or exclamatory mark, while subsentence is delimited by colon, semicolon, or comma. We define SenNum and SubSenNum as follows.

$$SenNum = \begin{cases} SenNum, & if\ SenNum \leqslant 2; \\ \text{“}SenNum > 2\text{”}, & if\ 2 < SenNum \leqslant 5; \\ \text{“}SenNum > 5\text{”}, & if\ 5 < SenNum \leqslant 10; \\ \text{“}SenNum > 10\text{”}, & if\ SenNum > 10; \end{cases} \quad (2)$$

$$SubSenNum = \begin{cases} SubSenNum, & if\ SubSenNum \leqslant 2; \\ \text{“}SenNum > 2\text{”}, & if\ 2 < SubSenNum \leqslant 5; \\ \text{“}SenNum > 5\text{”}, & if\ SubSenNum > 5. \end{cases} \quad (3)$$

3) *Appositive Feature*

Without deep analysis and disambiguation, it is not trivial to determine whether the two mentions are appositive or not. Here we approximate this property by capturing the relation of adjacency between the two mentions. If the two mentions with the same entity type are adjacent, we set appositive feature to "1".

Table 1. Feature Set for the Baseline System (derived from Soon *et al.*, the features which has been adjusted according to Chinese characteristics are marked by $*$)

Feature	Feature Description
StringMatch*	Real number value between 0 and 1 computed by (1).
MenType_I	"NOM" if EM_i is nominal mentions; "NAM" if named mentions; "PRO" if pronominal mentions.
MenType_J	Similar to MenType_I
Definite_I	"1" if EM_i contains the word with definitive and demonstrative sense, such as "这/the, this", "那些/those", else "−1".
Definite_J	Similar to Definite_I
Number	"1" if EM_i and EM_j agree in number; "−1" if they disagree; "0" if number information for one or both mentions cannot be determined.
Gender	"1" if EM_i and EM_j agree in gender; "−1" if they disagree; "0" if gender information for one or both mentions cannot be determined.
Appositive*	"1" if the two mentions are in an appositive relationship; else "−1".
EntityType*	"1" if EM_i and EM_j are consistent in entity type; "−1" if they are not; "0" if entity type information for one or both mentions cannot be determined.
CrossPara*	"1" if EM_i and EM_j are in different paragraphs; else "−1".
CrossSenNum*	See (2).
CrossSubSenNum*	See (3).

3.3 Co-Reference Classifier Selection

Diverse machine learning methods have been used for co-reference resolution, such as decision tree (DT) model C4.5[3~5], maximum entropy (ME) model[5,13], support vector machine (SVM) model[14,15] and etc. Bryant 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[15].

We consider two learning models in our baseline system: SVM and ME, which are increasingly employed for co-reference resolution in recent years. Our motivation is to compare the two models' performance in the context of co-reference resolution and try some combining strategy on them.

3.4 Linking Strategy

Linking strategy is used to combine the co-refer entity mention pairs into co-reference chains. The most popular linking strategy is the link-first strategy[3], 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[2], 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. Our baseline system uses the link-first strategy.

Section 4 will focus on "feature definition" module.

4 Mining Multi-Level Contextual Evidence Based on Linguistic Findings for Co-Reference Resolution

In Section 3, we give a baseline system for co-reference resolution. The system is under the statistical learning framework based on a series of surface features. However, these features are far from enough for co-reference resolution. Moreover, some of these features are difficult to be obtained in Chinese. For example, in some cases, it is difficult to get gender information and number information for entity mentions in Chinese texts. The following are some instances.

Example 1. 今年的上市公司/The companies that come into the market this year 中/among/...

In the example, we cannot determine that the entity mention "今年的上市公司/The companies that come into the market this year" is a plural structure if we do not take into account the contextual information around it.

Example 2. 这家戏剧团的演员/The players in the troupe 都/all 有/have 一技之长/professional skills.

In the example, it is difficult for us to determine that the entity mention "这家戏剧团的演员/The players in the troupe" is a plural structure if we do not take into account the contextual information around it.

Example 3. 经营着三家公司的李嘉凡/LI Jiafan, who manages three companies 嫁给了/married 一名大学教师/a university teacher.

In the example, it is difficult for us to determine that the entity mention "经营着三家公司的李嘉凡/LI Jiafan, who manages three companies" is a female only depending on the information inside the mention.

From the above 3 examples, we can see that , besides the features inside the entity mentions, the contextual information around them are also important for co-reference resolution, such as "中/among" and "都 /all" imply plural information, "嫁给/married" imply female information.

In order to get more features for co-reference resolution, in this section, we study how to use linguistic findings (anaphora resolution factors) to guide the feature mining for co-reference resolution.

4.1 Linguistic Findings in Anaphora Resolution

Although co-reference resolution and traditional anaphora resolution are different[16], many linguistic

findings in anaphora resolution have been successfully used for rule-based co-reference resolution[17].

Linguistically, the approaches for anaphora resolution usually rely on a set of "anaphora resolution factors"[2,18]. The anaphora resolution factors include various types, such as 1) gender agreement and number agreement between the anaphora and the antecedent; 2) syntactic and semantic parallelism between the anaphora and the antecedent; 3) the selectional restriction between the verb and the arguments, etc. The anaphora resolution factors can be used to filter the impossible noun phrases or to enhance the preference of a noun phrase in the list of candidate antecedents.

The first type of anaphora resolution factors can be easily used in the learning based model for co-reference resolution, which has been described in Section 3 (the baseline system). In the following, we will focus on how to use "syntactic and semantic parallelism factor" and "the selectional restriction factor" in the learning based co-reference framework.

4.1.1 *Syntactic and Semantic Parallelism Preferences*

Main Idea: Two noun phrases with the same syntactic functions or semantic roles are more possible to be co-referent than those with different syntactic functions or semantic roles. Therefore, in anaphora resolution, preference will be given to the candidate antecedents which have the same syntactic function or the same semantic role as the anaphora.

Example 4. 李刚/LI Gang 常/often 找/goes with 刘辉/LIU Hui 打篮球/to play basketball, 徐庆/XU Qing 常/often 找/goes with 他/him 去游泳/to swim.

It is a typical example using the syntactic and semantic parallelism preferences. Although there are 4 mentions, we can also successfully determine that the antecedent of "他/him" should be "刘辉/LIU Hui".

4.1.2 *Selectional Restriction*

Selectional restriction is the constraints a predicate imposes on its arguments. For example, in the sentence fragment "She eats x", the verb "eat" imposes a constraint on the direct object "x": "x" should be something that is usually being eaten. This phenomenon has received considerable attention in the linguistic community and it has been successfully applied as one of the anaphora resolution factors.

Example 5. 约翰/John 买了/bought 一辆马自达轿车/a Mazda car，他/he 每天/everyday 都驾驶/drives 它/it 去上班/to go to work。

In Example 5, by the constrains the verb "drive" imposes on its arguments, we can easily find that the antecedent of "它/it" is "一辆马自达轿车/a Mazda car".

4.2 Mining Multi-Level Contextual Evidence Based on Linguistic Findings

In linguistic theory, applying the anaphora resolution factors in co-reference resolution requires syntactic and semantic analysis of the sentences. However, it is quite difficult to conduct automatic syntactic and semantic analysis for Chinese processing technology currently. Therefore, we try to mine multi-level contextual features in order to approximate the above linguistic factors and use these features in learning based co-reference resolution framework. We hope that, through incorporating multi-level contextual surface features, we could capture some characteristics implied in anaphora resolution factors in linguistic view, which can be helpful for co-reference resolution.

In this paper, we focus on "syntactic and semantic parallelism factor" and "selectional restriction factor". We use the following strategies to simulate and approximate the above two factors.

1) In order to simulate "syntactic and semantic parallelism factor", we extract "bags of word form and POS" feature and "bag of semes" feature from the context of the entity mentions and add them into the baseline feature set.

2) Because it is too coarse to use the feature of bags of word form, POS tag and seme to determine the syntactic and semantic parallelism between two entity mentions, which will introduce a lot of noises, we propose a method for contextual feature reconstruction based on semantic similarity computation. In the contextual feature reconstruction process, at first, we parallel the words in the two contexts through semantic similarity computation based on HowNet[19], then we compute the similarity between the two paralleled contexts. We hope that the reconstructed contextual features could more effectively model the anaphora resolution factor of "Syntactic and Semantic Parallelism Preferences".

3) In order to capture more information related to "selectional restriction factor", we use the strategies of expanding the size of the contextual windows to approximately simulate the selectional restriction factor for anaphora resolution. In addition, we use an entity-mention-based contextual feature representation instead of isolated word-based contextual feature representation. Because the extents of the entity mentions are larger than isolated words, through the representation method, we can more effectively capture long-distance dependency information.

4.2.1 *Expanding Multilevel Contextual Surface Features into the Baseline Feature Set*

In the baseline system, we use the basic features for co-reference resolution. However, these features are far from enough for co-reference resolution, moreover, some of these features are difficult to be obtained in Chinese.

Anaphora resolution factor of "Syntactic and Semantic Parallelism" tells us that two noun phrases with the same syntactic functions or semantic roles are more possible to be co-referent than those with different syntactic functions or semantic roles. We believe that similar contexts can be a kind of approximate representation

of "syntactic and semantic parallelism". Therefore, we try to model syntactic and semantic parallelism to some extent through expanding the basic feature set with multilevel contextual surface features as follows.

• Word Form and POS Features

Word forms and POS (part of speech) of the words in the context of entity mentions 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, trying to capture some syntactic structural information.

• Bag of Semes (BS) Features

Although deep semantic analysis is not available, we can resort to a Chinese-English knowledge base called HowNet[19] to acquire shallow semantic features. HowNet is a bilingual common-sense knowledge base, which uses a set of non-decomposable semes to define a sense of a word. A total of over 1600 semes 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", where there are three semes split by commas, and symbols such as "*" represent specific relations.

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

4.2.2 *Feature Reconstruction Based on Contextual Semantic Similarity*

In Subsection 4.2.1, we just use bag of semes to model the context, losing some useful associated information between semes of the same word. We will make it manifest by analyzing at two different contexts: "校园/schoolyard 歌手/singer" and "学生/student 歌厅/singing hall". The senses for every word and the bag of semes features of the context are shown in Fig.3.

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 semes feature of two contexts:
1. 校园/schoolyard 歌手/singer: (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.3. Bag of semes of the two example contexts.

From Fig.3 we can see that, although the two contexts are semantically discrepant, their BS features are similar to a large extent. The reason lies in that, the semes from different word pairs are mismatched. For example, seme "InstitutePlace|场所" in 1 belongs to word "校园/schoolyard", seme "InstitutePlace|场所" in 2 belongs to word "歌厅/singing hall", however, they are mismatched in bag-of-semes-based contextual similarity computation.

Undoubtedly, BS features can express the semantic information 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. Aiming at this problem, we try to use some strategy to reconstruct those unordered BS features to give a better representation of context.

We regard the two contexts as two sets of words. The reconstruction process parallels the words from two word sets based on word similarity computation. Then, the similarity between the two contexts is computed based on the paralleled contexts.

• Word Similarity Computation

A word similarity computing approach[20] based on HowNet is used for reference in our feature reconstruction method. In our case of word similarity computation, we did not consider the relation seme description and the relation symbol description as Liu Qun *et al.* do[20]. We only consider two aspects: 1) the similarity between the first basic semes of the two words, denoted as $Sim_1(S_1, S_2)$; 2) the similarity of all the other basic semes 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.

$$Sim(w_1, w_2) = \beta_1 Sim_1(S_1, S_2) + \beta_2 Sim_2(S_1, S_2) \quad (4)$$

where $\beta_1 + \beta_2 = 1$. Since the first basic seme indicates the most important semantic feature, we set $\beta_1 = 0.7$. Similarity between two semes is computed according to the distance in the seme hierarchy of HowNet. Details can be found in [20].

• Feature Reconstruction and the Computation of Contextual 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 (we call the process as word paralleling or feature reconstruction), then we compute the arithmetical average of the 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 feature construction and contextual similarity computation is illustrated in Fig.4.

In summary, in the contextual similarity computation process, the words from two contexts are firstly paralleled, then the contextual similarity is computed based on the paralleled word pairs. In other words, the unordered BS features are reconstructed. As a result, the semes between different word pairs are not confused,

and the first seme similarity is given a larger weight, so the BS mismatching phenomena showed in Fig.3 can be overcome to some extent.

```
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
Procedure DeSort (ArrayWordSim) //sort in decreasing
                                 //order of similarity value
ProcessedIndex_1← ∅, ProcessdIndex_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.4. Algorithm for feature construction and contextual similarity computation.

4.2.3 *Contextual Window Enlargement and Entity-Mention-Based Contextual Representation for Obtaining Long Distance Relation*

• Contextual Window Enlargement

Selectional restriction discussed in Subsection 4.1.2 embodies a kind of dependency relation, verb-object relation, which can be held in a long distance in discourse. Without perfect dependency analysis tools available, we try to enlargement the contextual window to obtain some long distance relations.

To avoid much noise, we just enlarge the contextual window for similarity based feature reconstruction described in Subsection 4.2.2. In the feature extraction phase in Subsection 4.2.1, we only consider a 5-width window to extract contextual cues.

• Entity-Mention-Based Contextual Representation

Usually, the granularity of contextual features representation is word-based. This representation limits the capability to capture long distance information. Although we can compensate for it through expanding the contextual window size, however the window expanding will also bring noises in.

Because co-reference resolution is conducted on the basis of entity mention labeling, therefore we try to use entity-mention-based contextual representation instead of word-based representation. Intuitionally, entity-mention-based representation can more effectively grasp the selectional restriction information in the contexts, therefore it is expected to be beneficial for co-reference resolution.

5 Experiments and Analysis

Our experiments are conducted to evaluate whether the following variations are beneficial for co-reference resolution? 1) The multilevel contextual features expanded upon the basic feature set (in Subsection 5.2); 2) Co-Reference Classifier Selection (in Subsection 5.3); 3) feature reconstruction based on Contextual Semantic Similarity (in Subsection 5.4); 4) Contextual Window Enlargement and Entity-Mention-Based Contextual Feature Representation for Obtaining Long Distance Relation (in Subsection 5.5).

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.

5.1 Experimental Data and Evaluation Metric

We evaluate the co-reference system on the standard ACE-05 co-reference data. 80% of the dataset is used for training the co-reference classifier, and other 20% is used for testing. Statistics of train data and test data is shown in Table 2.

The performance of our co-reference system is reported in terms of recall, precision, and F-measure using the model-theoretic MUC scoring program[11].

Table 2. Statistics on Experimental Data

	#Doc	#Entity Mention				#Co-Ref. Chain
		Nam	Nom	Pro	Total	
Train	511	11 649	12 952	2763	27 364	12 258
Test	122	3 048	3 326	583	6 957	3 156

This Vilain algorithm[11] considers the response to be a permutation of the key. For example, the mentions of a single entity class in the key may be found in many entity classes in the response. We count the number of links needed to create the key class. Let $c(S) = (|S| - 1)$ be the number of correct links needed for class S. Let $m(S) = (|p(S)| - 1)$ be the number of missing links where $|p(S)|$ is the number of partitions in which S is broken into in the response. Then Vilain defines recall for the class S as follows[11].

$$\frac{c(S) - m(S)}{c(S)} = \frac{(|S| - 1) - (|p(S)| - 1)}{|S| - 1} = \frac{|S| - |p(S)|}{|S| - 1}. \quad (5)$$

① http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

② http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Recall for all the classes is simply a sum of the individual links:

$$recall = \frac{\sum_{S \in \{key\}} (|S| - |p(S)|)}{\sum_{S \in \{key\}} (|S| - 1)}. \quad (6)$$

Precision is calculated by reversing the role of the key and response.

5.2 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, "WP" denotes the word and POS features, "BS" denotes the semantic feature represented by Bag of Semes.

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, increases non-trivially from 73.84% to 79.74%. The introduction of BS features based on HowNet can further boost the system's performance. Although the recall gets slight degradation, the F-value increases from 79.74% to 79.89%.

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

	Recall	Precision	F-value
Baseline System	61.06 (%)	93.4 (%)	73.84 (%)
Baseline+WP	74.82 (%)	85.37 (%)	79.74 (%)
Baseline+WP+BS	74.56 (%)	86.06 (%)	79.89 (%)

The experimental results are largely consistent with our hypothesis. System performance improves about 6% by applying diverse contextual features. This can be explained that combining multiple contextual features can capture some syntactic constrains information which is definitely helpful for co-reference resolution according to traditional linguistic findings. On the other hand, although BS features will introduce some noises, they can model some semantic properties of the context to some extent and lead to better performance.

5.3 Performance Comparison Between Different Classifiers

We consider two classifiers, ME and SVM, in our machine learning co-reference resolution framework. In this section, performance comparison between them will be conducted and the combination strategy is employed. For combination, we compare the confidences of two classifiers' predication and selected the predication result with the larger confidence value. 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 regression while SVM is classification model based on large-margin principle. Maybe for this reason, SVM can outperform ME when the training data is not significantly sufficient. Also we can see that the noise of BS overwhelms its usefulness when using ME and fail to improve the performance.

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

5.4 Performance Improvement by Feature Reconstruction

In this section we try to validate the effectiveness of our motivation to better model context similarity by feature reconstruction. Fig.5 shows the improvement results, among which "RC" denotes the feature reconstruction.

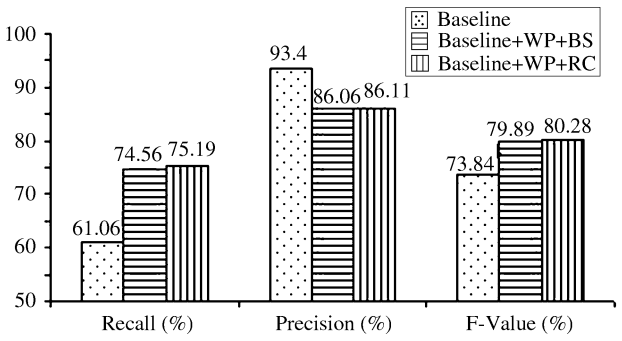

Fig.5. Performance improvement from feature reconstruction and collocation features.

From Fig.5, we can see that RC feature outperforms BS feature and get increase both in recall and precision. The resulting F-value increases from 79.89% to 80.28%. This verifies our analysis in Section 4. Through contextual similarity, the semantic semes information can be better organized and can model the context in a more reasonable way.

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

5.5 Impacts of Different Contextual Windows and Entity-Mention-Based Representation

The performance (F-value) curve with enlarged contextual window is illustrated in Fig.6. It shows that extending the contextual window is effective for acquiring long distance relation such as selectional restriction, and the system's performance obtains a slight improvement. When the window is 9, it gets the best F-value of 80.35%.

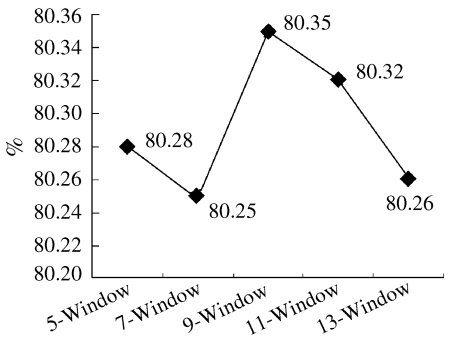

Fig.6. Performance curve with difference contextual window sizes.

In addition, we also conduct the experiment to evaluate the performance of entity-mention-based contextual feature representation. We respectively fix the window sizes as 6, 7, 8, 9, and compare the F-value between isolated-word-based representation and entity-mention-based representation. The result shows that, when window size is 7, the entity-mention-based representation gets the best F-value as 80.49%. The result implies that, window size expansion and entity-mention-based representation are beneficial for co-reference resolution.

6 Conclusions and Future Work

Under a learning based framework for Chinese co-reference resolution, we investigate multiple contextual features to improve the system performance based on related linguistic theory.

First, we incorporate diverse contextual features into the baseline statistical co-reference resolution model, in order to capture some syntactic structural information and semantic information in the contexts, which is inspired by "syntactic and semantic parallelism factor" and "selectional restriction factor". Second, a feature 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. Furthermore, performance comparison between two classifiers in the learning based framework is conducted and a simple combining method is tried. Finally, impacts of different sizes of context window on system performance are investigated empirically based on feature reconstruction, trying to obtain longer distance relation in discourse. Experimental results prove that our approach performs well on the standard ACE datasets without deep syntactic and semantic analysis.

Nevertheless, there is substantial room for improvement. Our future work will focus on the following:

- to investigate an effective sense disambiguation strategy to reduce the noise of introducing semantic features;
- to find a more appropriate similarity computation methods to model the semantic context better;
- to try to combine with existing linguistic analysis tools in the statistical framework to boost the system performance;
- to leverage global contextual features as well as local contextual features useful for Chinese co-reference resolution.

References

[1] Mitkov R. Anaphora Resolution. London: Longman Press, 2002.
[2] NIST. The Official Evaluation Plan for the ACE 2005 Evaluation. 2005, http://www.nist.gov/speech/tests/ace/ace05/.
[3] Soon W M, Ng H T, Lim D. A machine learning approach to co-reference resolution of noun phrases. *Computational Linguistics*, 2001, 27(4): 521~544.
[4] Ng V, Cardie C. Improving machine learning approaches to co-reference resolution. In *Proc. the 40th Anniversary Meeting of the Association for Computational Linguistics* (*ACL-02*), Philadelphia, PA, USA, 2002, pp.104~111.
[5] Vincent Ng. Machine learning for coreference resolution: From local classification to global ranking. In *Proc. the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL-05*), Ann Arbor, MI, 2005, pp.157~164.
[6] Yang X, Zhou G, Su J, Tan C L. Improving noun phrase co-reference resolution by matching strings. In *Proc. IJCNLP-04*, Hainan, China, *Lecture Notes in Computer Science*, Volume 3248, 2004, pp.22~31.
[7] Strube M, Rapp S, Muller C. The influence of minimum edit distance on reference resolution. In *Proc. the Conference on Empirical Methods in Natural Language Processing* (*EMNLP-2002*), Philadelphia, USA, 2002, pp.312~319.
[8] Houfeng Wang, Tingting He. Research on Chinese pronominal anaphora resolution. *Chinese Journal of Computers*, 2001, 24(2): 136~143.
[9] Houfeng Wang, Zheng Mei. Robust pronominal resolution within Chinese text. *Journal of Software*, 2005, 16(5): 700~707.
[10] Chinchor N, Marsh E, MUC-7 Information Extraction Task Definition, In *Proc. the Seventh Message Understanding Conference* (*MUC-7*), San Diego, CA, USA, Chinchor NA (ed.), Science Applications International Corporation, 1998.
[11] Vilain M, Burger J, Aberdeen J *et al.* A model-theoretic coreference scoring scheme. In *Proc. the Sixth Message Understanding Conference* (*MUC-6*), Columbia, Maryland, USA, Morgan Kaufmann, 1995, pp.45~52.
[12] Doddington G, Mitchell A, Przybocki M *et al.* Automatic Content Extraction (ACE) program — Task definitions and performance measures. In *Proc. the Fourth International Conference on Language Resources and Evaluation* (*LREC 2004*), Lisbon, Portugal, 2004, pp.837~840.
[13] Florian R, Hassan H, Ittycheriah A *et al.* A statistical model for multilingual entity detection and tracking. In *Proc. the Human Language Technology Conference — North American Chapter of the Association for Computational Linguistics An-*

nual Meeting (*HLT/NAACL-2006*), Boston, Massachusetts, USA, 2004, pp.1∼8.

[14] Iida R, Inui K, Takamura H *et al.* Incorporating contextual cues in trainable models for coreference resolution. In *Proc. the EACL'03 Workshop on the Computational Treatment of Anaphora*, Budapest, Hungary, 2003, pp.23∼30.

[15] John Bryant. Combining feature based and semantic information for co-reference resolution. Research Report at U.C. Berkeley and ICSI.

[16] Van Deemter K, Kibble R. On Coreferring: Coreference in MUC and Related Annotation Schemes 2000. *Computational Linguistics*, 2004, 26(4): 629∼637.

[17] Aone C, Halverson L, Hampton T, Ramos-Santacruz M. SRA: Description of the IE^2 System Used for MUC-7. In *Proc. the Seventh Message Understanding Conference* (*MUC-7*), Chinchor N A (ed). San Diego, CA, Science Applications International Corporation, 1998.

[18] Jurafsky Dan, James Martin. Speech and Language Processing. Prentice-Hall, Englewood Cliffs NJ, 2000.

[19] Zhendong Dong, Qiang Dong. HowNet and the Computation of Meaning. Singapore: World Scientific 2006.

[20] Qun Liu, Sujian Li. Word similarity computing based on Hownet. *Journal of Computational Linguistics and Chinese Language Processing*, 2002, 7(2): 59∼76.

Jun Zhao is an associated professor and Ph.D. supervisor at Institute of Automation, Chinese Academy of Sciences. His research interest includes natural language processing, information extraction and web mining. He received his Ph.D. degree in computer application from Tsinghua Univ. He is a senior member of China Computer Federation.

Fei-Fan Liu received his Ph.D. degree in pattern recognition and intelligent system from Institute of Automation, Chinese Academy of Sciences in 2006. Now he is a postdoctoral research fellow in Human Language Technology Research Institute, Computer Science Department, University of Texas at Dallas, USA. His research interest includes natural language processing, information extraction, text mining and speech summarization.