Omni-word Feature and Soft Constraint for Chinese Relation Extraction

Yanping Chen[†] **Qinghua Zheng**[†] **Wei Zhang**[‡]
[†]MOEKLINNS Lab, Department of Computer Science and Technology
Xi'an Jiaotong University, China
ypench@gmail.com, qhzheng@mail.xjtu.edu.cn
[‡]Amazon.com, Inc.
wzhan@amazon.com

Abstract

Chinese is an ancient hieroglyphic. It is inattentive to structure. Therefore, segmenting and parsing Chinese are more difficult and less accurate. In this paper, we propose an Omni-word feature and a soft constraint method for Chinese relation extraction. The Omni-word feature uses every potential word in a sentence as lexicon feature, reducing errors caused by word segmentation. In order to utilize the structure information of a relation instance, we discuss how soft constraint can be used to capture the local dependency. Both Omni-word feature and soft constraint make a better use of sentence information and minimize the influences caused by Chinese word segmentation and parsing. We test these methods on the ACE 2005 RDC Chinese corpus. The results show a significant improvement in Chinese relation extraction, outperforming other methods in F-score by 10% in 6 relation types and 15% in 18 relation subtypes.

1 Introduction

Information Extraction (IE) aims at extracting syntactic or semantic units with concrete concepts or linguistic functions (Grishman, 2012; McCallum, 2005). Instead of dealing with the whole documents, focusing on designated information, most of the IE systems extract named entities, relations, quantifiers or events from sentences.

The relation recognition task is to find the relationships between two entities. Successful recognition of relation implies correctly detecting both the relation arguments and relation type. Although this task has received extensive research. The performance of relation extraction is still unsatisfactory with a F-score of 67.5% for English (23 subtypes) (Zhou et al., 2010). Chinese relation extraction also faces a weak performance having F-score about 66.6% in 18 subtypes (Dandan et al., 2012).

The difficulty of Chinese IE is that Chinese words are written next to each other without delimiter in between. Lacking of orthographic word makes Chinese word segmentation difficult. In Chinese, a single sentence often has several segmentation paths leading to the segmentation ambiguity problem (Liang, 1984). The lack of delimiter also causes the Out-of-Vocabulary problem (OOV, also known as *new word detection*) (Huang and Zhao, 2007). These problems are worsened by the fact that Chinese has a large number of characters and words. Currently, the state-of-the-art Chinese OOV recognition system has performance about 75% in recall (Zhong et al., 2012). The errors caused by segmentation and OOV will accumulate and propagate to subsequent processing (e.g. part-of-speech (POS) tagging or parsing).

Therefore, the Chinese relation extraction is more difficult. According to our survey, compared to the same work in English, the Chinese relation extraction researches make less significant progress.

Based on the characteristics of Chinese, in this paper, an Omni-word feature and a soft constraint method are proposed for Chinese relation extraction. We apply these approaches in a maximum entropy based system to extract relations from the ACE 2005 corpus. Experimental results show that our method has made a significant improvement.

The contributions of this paper include

1. Propose a novel Omni-word feature for Chinese relation extraction. Unlike the traditional segmentation based method, which is a partition of the sentence, the Omni-word feature uses every potential word in a sentence as lexicon feature.

2. Aiming at the Chinese inattentive structure, we utilize the soft constraint to capture the local dependency in a relation instance. Four constraint conditions are proposed to gener-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 572–581,
Baltimore, Maryland, USA, June 23-25 2014.

ate combined features to capture the local dependency and maximize the classification determination.

The rest of this paper is organized as follows. Section 2 introduces the related work. The Omni-word feature and soft constrain are proposed in Section 3. We give the experimental results in Section 3.2 and analyze the performance in Section 4. Conclusions are given in Section 5.

2 Related Work

There are two paradigms extracting the relationship between two entities: the Open Relation Extraction (ORE) and the Traditional Relation Extraction (TRE) (Banko et al., 2008).

Based on massive and heterogeneous corpora, the ORE systems deal with millions or billions of documents. Even strict filtrations or constrains are employed to filter the redundancy information, they often generate tens of thousands of relations dynamically (Hoffmann et al., 2010). The practicability of ORE systems depends on the adequateness of information in a big corpus (Brin, 1999). Most of the ORE systems utilize weak supervision knowledge to guide the extracting process, such as: Databases (Craven and Kumlien, 1999), Wikipedia (Wu and Weld, 2007; Hoffmann et al., 2010), Regular expression (Brin, 1999; Agichtein and Gravano, 2000), Ontology (Carlson et al., 2010; Mohamed et al., 2011) or Knowledge Base extracted automatically from Internet (Mintz et al., 2009; Takamatsu et al., 2012). However, when iteratively coping with large heterogeneous data, the ORE systems suffer from the "semantic drift" problem, caused by error accumulation (Curran et al., 2007). Agichtein, Carlson and Fader et al. (2010; 2011; 2000) propose syntactic and semantic constraints to prevent this deficiency. The soft constraints, proposed in this paper, are combined features like these syntactic or semantic constraints, which will be discussed in Section 3.2.

The TRE paradigm takes hand-tagged examples as input, extracting predefined relation types (Banko et al., 2008). The TRE systems use techniques such as: Rules (Regulars, Patterns and Propositions) (Miller et al., 1998), Kernel method (Zhang et al., 2006b; Zelenko et al., 2003), Belief network (Roth and Yih, 2002), Linear programming (Roth and Yih, 2007), Maximum entropy (Kambhatla, 2004) or SVM (GuoDong et al., 2005). Compared to the ORE systems, the TRE systems have a robust performance. Disadvantages of the TRE systems are that the manually annotated corpus is required, which is time-consuming and costly in human labor. And migrating between different applications is difficult. However, the TRE systems are evaluable and comparable. Different systems running on the same corpus can be evaluated appropriately.

In the field of Chinese relation extraction, Liu et al. (2012) proposed a convolution tree kernel. Combining with external semantic resources, a better performance was achieved. Che et al. (2005) introduced a feature based method, which utilized lexicon information around entities and was evaluated on Winnow and SVM classifiers. Li and Zhang et al. (2008; 2008) explored the position feature between two entities. For each type of these relations, a SVM was trained and tested independently. Based on *Deep Belief Network*, Chen et al. (2010) proposed a model handling the high dimensional feature space. In addition, there are mixed models. For example, Lin et al. (2010) employed a model, combining both the feature based and the tree kernel based methods.

Despite the popularity of kernel based method, Huang et al. (2008) experimented with different kernel methods and inferred that simply migrating from English kernel methods can result in a bad performance in Chinese relation extraction. Chen and Li et al. (2008; 2010) also pointed out that, due to the inaccuracy of Chinese word segmentation and parsing, the tree kernel based approach is inappropriate for Chinese relation extraction. The reason of the tree kernel based approach not achieve the same level of accuracy as that from English may be that segmenting and parsing Chinese are more difficult and less accurate than processing English.

In our research, we proposed an Omni-word feature and a soft constraint method. Both approaches are based on the Chinese characteristics. Therefore, better performance is expected. In the following, we introduce the feature construction, which discusses the proposed two approaches.

3 Feature Construction

In this section, the employed candidate features are discussed. And four constraint conditions are proposed to transform the candidate features into combined features. The soft constraint is the

method to generate the combine features[1].

3.1 Candidate Feature Set

In the ACE corpus, an *entity* is an object or set of objects in the world. An *entity mention* is a reference to an entity. The entity mention is annotated with its full *extent* and its *head*, referred to as the *extend mention* and the *head mention* respectively. The extent mention includes both the head and its modifiers. Each *relation* has two entities as arguments: Arg-1 and Arg-2, referred to as E1 and E2. A *relation mention* (or instance) is the embodiment of a relation. It is referred by the sentence (or clause) in which the relation is located in. In our work, we focus on the detection and recognition of relation mention.

Relation identification is handled as a classification problem. Entity-related information (e.g. head noun, entity type, subtype, CLASS, LDC-TYPE, etc.) are supposed to be known and provided by the corpus. In our experiment, the entity type, subtype and the head noun are used.

All the employed features are simply classified into five categories: *Entity Type and Subtype*, *Head Noun*, *Position Feature*, *POS Tag* and *Omni-word Feature*. The first four are widely used. The last one is proposed in this paper and is discussed in detail.

Entity Type and Subtype: In ACE 2005 RDC Chinese corpus, there are 7 entity types (Person, Organization, GPE, Location, Facility, Weapon and Vehicle) and 44 subtypes (e.g. Group, Government, Continent, etc.).

Head Noun: The head noun (or head mention) of entity mention is manually annotated. This feature is useful and widely used.

Position Feature: The position structure between two entity mentions (extend mentions). Because the entity mentions can be nested, two entity mentions may have four coarse structures: "E1 is before E2", "E1 is after E2", "E1 nests in E2" and "E2 nests in E1", encoded as: '`E1_B_E2`', '`E1_A_E2`', '`E1_N_E2`' and '`E2_N_E1`'.

POS Tag: In our model, we use only the adjacent entity POS tags, which lie in two sides of the entity mention. These POS tags are labelled by the ICTCLAS package[2]. The POS tags are not used independently. It is encoded by combining the POS tag with the adjacent entity mention information. For example '`E1_Right_n`' means that the right side of the first entity is a noun ("n").

Omni-word Feature: The notion of "word" in Chinese is vague and has never played a role in the Chinese philological tradition (Sproat et al., 1996). Some Chinese segmentation performance has been reported precision scores above 95% (Peng et al., 2004; Xue, 2003; Zhang et al., 2003). However, for the same sentence, even native peoples in China often disagree on word boundaries (Hoosain, 1992; Yan et al., 2010). Sproat et al. (1996) has showed that there is a consistence of 75% on the segmentation among different native Chinese speakers. The word-formation of Chinese also implies that the meanings of a compound word are made up, usually, by the meanings of words that contained in it (Hu and Du, 2012). So, fragments of phrase are also informative.

Because high precision can be received by using simple lexical features (Kambhatla, 2004; Li et al., 2008). Making better use of such information is beneficial. In consideration of the Chinese characteristics, we use *every potential word in a relation mention* as the lexical features. For example, relation mention '台北大安森林公园' (Taipei Daan Forest Park) has a "PART-WHOLE" relation type. The traditional segmentation method may generate four lexical features {'台北', '大安', '森林', '公园'}, which is a partition of the relation mention. On the other hand, the Omni-word feature denoting all the possible words in the relation mention may generate features as:

> {'台', '北', '大', '安', '森', '林', '公', '园', '台北', '大安', '森林', '公园', '森林公园', '大安森林公园'}[3]

Most of these features are nested or overlapped mutually. So, the traditional character-based or word-based feature is only a subset of the Omni-word feature. To extract the Omni-word feature, only a lexicon is required, then scan the sentence to collect every word.

Because the number of lexicon entry determines the dimension of the feature space, performance of Omni-word feature is influenced by the lexicon being employed. In this paper, we generate the lexicon by merging two lexicons. The first lexicon

[1] If without ambiguity, we also use the terminology of "soft constraint" denoting features generated by the employed constraint conditions.

[2] `http://ictclas.org/`

[3] The generated Omni-word features dependent on the employed lexicon.

is obtained by segmenting every relation instance using the ICTCLAS package, collecting very word produced by ICTCLAS. Because the ICTCLAS package was trained on annotated corpus containing many meaningful lexicon entries. We expect this lexicon to improve the performance. The second lexicon is *the Lexicon Common Words in Contemporary Chinese*[4].

Despite the Omni-word can be seen as a subset of n-Gram feature. It is not the same as the n-Gram feature. N-Gram features are more fragmented. In most of the instances, the n-Gram features have no semantic meanings attached to them, thus have varied distributions. Furthermore, for a single Chinese word, occurrences of 4 characters are frequent. Even 7 or more characters are not rare. Because Chinese has plenty of characters[5], when the corpus becoming larger, the n-Gram (n¿4) method is difficult to be adopted. On the other hand, the Omni-word can avoid these problems and take advantages of Chinese characteristics (the word-formation and the ambiguity of word segmentation).

3.2 Soft Constraint

The structure information (or dependent information) of relation instance is critical for recognition. However, even in English, "deeper" analysis (e.g. logical syntactic relations or predicate-argument structure) may suffer from a worse performance caused by inaccurate chunking or parsing. Hence, the local dependency contexts around the relation arguments are more helpful (Zhao and Grishman, 2005). Zhang et al. (2006a) also showed that Path-enclosed Tree (PT) achieves the best performance in the kernel based relation extraction. In this field, the tree kernel based method commonly uses the parse tree to capture the structure information (Zelenko et al., 2003; Culotta and Sorensen, 2004). On the other hand, the feature based method usually uses the combined feature to capture such structure information (GuoDong et al., 2005; Kambhatla, 2004).

In the open relation extraction domain, syntactic and semantic constraints are widely employed to prevent the "semantic drift" problem. Such constraints can also be seen as structural constraint. Most of these constraints are hard constraints. Any relation instance violating these constraints (or below a predefined threshold) will be abandoned. For example, Agichtein and Gravano (2000) generates patterns according to a *confidence threshold* (τ_t). Fader et al. (2011) utilizes a *confidence function*. And Carlson et al. (2010) filters candidate instances and patterns using the number of times they co-occurs.

Deleting of relation instances is acceptable for open relation extraction because it always deals with a big data set. But it's not suitable for traditional relation extraction, and will result in a low recall. Utilizing the notion of combined feature (GuoDong et al., 2005; Kambhatla, 2004), we replace the hard constraint by the soft constraint. Each soft constraint (combined feature) has a parameter trained by the classifier indicating the discrimination ability it has. No subjective or priori judgement is adopted to delete any potential determinative constraint (except for the reason of dimensionality reduction).

Most of the researches make use of the combined feature, but rarely analyze the influence of the approaches we combine them. In this paper, we use the soft constraint to model the local dependency. It is a subset of the combined feature, generated by four constraint conditions: *singleton*, *position sensitive*, *bin sensitive* and *semantic pair* . For every employed candidate feature, an appropriate constraint condition is selected to combine them with additional information to maximize the classification determination.

Singleton: A feature is employed as a singleton feature when it is used without combining with any information. In our experiments, only the *position feature* is used as singleton feature.

Position Sensitive: A position sensitive feature has a label indicating which entity mention it depends on. In our experiment, the *Head noun* and *POS Tag* are utilized as position sensitive features, which has been introduced in Section 3.1. For example, '台北_E1' means that the head noun '台北' depend on the first entity mention.

Semantic Pair: Semantic pair is generated by combining two semantic units. Two kinds of semantic pair are employed. Those are generated by combining two entity types or two entity subtypes into a semantic pair. For example, 'Person_Location' denotes that the type of the first relation argument is a "Person" (entity

[4]Published by *Ministry of Education of the People's Republic of China* in 2008, containing 56,008 entries.

[5]Currently, at least 13000 characters are used by native Chinese people. *Modern Chinese Dictionary*: http://www.cp.com.cn/

type) and the second is a "Location" (entity type). Semantic pair can capture both the semantic and structure information in a relation mention.

Bin Sensitive: In our study, *Omni-word feature* is not added as "bag of words". To use the Omni-word feature, we segment each relation mention by two entity mentions. Together with the two entity mentions, we get five parts: "FIRST", "MIDDLE", "END", "E1" and "E2" (or less, if the two entity mentions are nested). Each part is taken as an independent bin. A flag is used to distinguish them. For example, '台北_Bin_F', '台北_Bin_E1' and '台北_Bin_E' mean that the lexicon entry '台北' appears in three bins: the FIRST bin, the first entity mention (E1) bin and the END bin. They will be used as three independent features.

To sum up, among the five candidate feature sets, the position feature is used as a singleton feature. Both head noun and POS tag are position sensitive. Entity types and subtypes are employed as semantic pair. Only Omni-word feature is bin sensitive. In the following experiments, focusing on Chinese relation extraction, we will analyze the performance of candidate feature sets and study the influence of the constraint conditions.

sectionExperiments

In this section, methodologies of the Omni-word feature and the soft constraint are tested. Then they are compared with the state-of-the-art methods.

3.3 Settings and Results

We use the ACE 2005 RDC Chinese corpus, which was collected from newswires, broadcasts and weblogs, containing 633 documents with 6 major relation types and 18 subtypes. There are 8,023 relations and 9,317 relation mentions. After deleting 5 documents containing wrong annotations[6], we keep 9,244 relation mentions as positive instances.

To get the negative instances, each document is segmented into sentences[7]. Those sentences that do not contain any entity mention pair are deleted. For each of the remained sentences, we iteratively extract every entity mention pair as the arguments of relation instances for predicting. For example, suppose a sentence has three entity mentions: A,B and C. Because the relation arguments are order sensitive, six entity mention pairs can be generated: [A,B], [A,C], [B,C], [B,A], [C,A] and [C,B]. After discarding the entity mention pairs that were used as positive instances, we generated 93,283 negative relation instances labelled as "OTHER". Then, we have 7 relation types and 19 subtypes.

A maximum entropy multi-class classifier is trained and tested on the generated relation instances. We adopt the five-fold cross validation for training and testing. Because we are interested in the 6 annotated major relation types and the 18 subtypes, we average the results of five runs on the 6 positive relation types (and 18 subtypes) as the final performance. F-score is computed by

$$\frac{2 \times (Precision \times Recall)}{Precision + Recall}$$

To implement the maximum entropy model, the toolkit provided by Le (2004) is employed. The iteration is set to 30.

Five candidate feature sets are employed to generate the combined features. The *entity type and subtype*, *head noun*, *position feature* are referred to as $\mathcal{F}_{thp}$[8]. The POS tags are referred to as $\mathcal{F}_{pos}$. The Omni-word feature set is denoted by $\mathcal{F}_{ow}$.

Table 1 gives the performance of our system on the 6 types and 18 subtypes. Note that, in this paper, bare numbers and numbers in the parentheses represent the results of the 6 types and the 18 subtypes respectively.

Table 1: Performance on Type (Subtype)

Features	P	R	F
$\mathcal{F}_{thp}$	61.51 (52.92)	48.85 (36.92)	54.46 (43.49)
$\mathcal{F}_{ow}$	80.16 (66.98)	75.45 (54.85)	77.74 (60.31)
$\mathcal{F}_{thp} \cup \mathcal{F}_{pos}$	83.93 (69.83)	77.81 (61.63)	80.76 (65.47)
$\mathcal{F}_{thp} \cup \mathcal{F}_{ow}$	92.40 (81.94)	88.37 (70.69)	90.34 (75.90)
$\mathcal{F}_{thp} \cup \mathcal{F}_{pos} \cup \mathcal{F}_{ow}$	92.26 (80.52)	88.51 (70.96)	90.35 (75.44)

In Row 1, because $\mathcal{F}_{thp}$ are features directly obtained from annotated corpus, we take this per-

[6] DAVYZW_{20041230.1024, 20050110.1403, 20050111.1514, 20050127.1720, 20050201.1538}.

[7] The five punctuations are used as sentence boundaries: Period (。), Question mark (？), Exclamatory mark (！), Semicolon (；) and Comma (，).

[8] "thp" is an acronym of "**t**ype, **h**ead, **p**osition". Features in $\mathcal{F}_{thp}$ are the candidate features combined with the corresponding constraint conditions. The following $\mathcal{F}_{pos}$ and $\mathcal{F}_{ow}$ are the same.

formance as our referential performance. In Row 2, with only the $\mathcal{F}_{ow}$ feature, the F-score already reaches 77.74% in 6 types and 60.31% in 18 subtypes. The last row shows that adding the $\mathcal{F}_{pos}$ almost has no effect on the performance when both the $\mathcal{F}_{thp}$ and $\mathcal{F}_{ow}$ are in use. The results show that $\mathcal{F}_{ow}$ is effective for Chinese relation extraction.

The superiorities of Owni-word feature depend on three reasons. First, the specificity of Chinese word-formation indicates that the subphrases of Chinese word (or phrase) are also informative. Second, most of relation instances have limited context. The Owni-word feature, utilizing every possible word in them, is a better way to capture more information. Third, the entity mentions are manually annotated. They can precisely segment the relation instance into corresponding bins. Segmentation of bins bears the sentence structure information. Therefore, the Owni-word feature with bin information can make a better use of both the syntactic information and the local dependency.

3.4 Comparison

Various systems were proposed for Chinese relation extraction. We mainly focus on systems trained and tested on the ACE corpus. Table 2 lists three systems.

Table 2: Survey of Other Systems

System	P	R	F
Che et al. (2005)	76.13	70.18	73.27
Zhang et al. (2011)	80.71 (77.75)	62.48 (60.20)	70.43 (67.86)
Liu et al. (2012)	81.1 (79.1)	61.0 (57.5)	69.0 (66.6)

Che et al. (2005) was implemented on the ACE 2004 corpus, with 2/3 data for training and 1/3 for testing. The performance was reported on 7 relation types: 6 major relation types and the none relation (or negative instance). Zhang et al. (2011) was based on the ACE 2005 corpus with 75% data for training and 25% for testing. Performances about the 7 types and 19 subtypes were given. Both of them are feature based methods. Liu et al. (2012) is a kernel based method evaluated on the ACE 2005 corpus. The five-fold cross validation was used and declared the performances on 6 relation types and 18 subtypes.

The data preprocessing makes differences from our experiments to others. In order to give a better comparison with the state-of-the-art methods, based on our experiment settings and data, we implement the two feature based methods proposed by Che et al. (2005) and Zhang et al. (2011) in Table 2. The results are shown in Table 3.

In Table 3, Ei ($i \in 1, 2$) represents entity mention. "Order" in Che et al. (2005) denotes the position structure of entity mention pair. Four types of order are employed (the same as ours). $Word_{Ei\pm k}$ and $POS_{Ei\pm k}$ are the words and POS of Ei, "$\pm k$" means that it is the kth word (of POS) after (+) or before (-) the corresponding entity mention. In this paper, $k = 1$ and $k = 2$ were set.

In Row 2, the "Uni-Gram" represents the Unigram features of internal and external character sequences. Internal character sequences are the four entity extend and head mentions. Five kinds of external character sequences are used: one In-Between character sequence between E1 and E2 and four character sequences around E1 and E2 in a given window size w_s. The w_s is set to 4. The "Bi-Gram" is the 2-gram feature of internal and external character sequences. Instead of the 4 position structures, the 9 position structures are used. Please refer to Zhang et al. (2011) for the details of these 9 position structures.

In Table 3, it is shown that our system outperforms other systems, in F-score, by 10% on 6 relation types and by 15% on 18 subtypes.

For researchers who are interested in our work, the source code of our system and our implementations of Che et al. (2005) and Zhang et al. (2011) are available at `https://github.com/YPench/CRDC`.

4 Discussion

In this section, we analyze the influences of employed feature sets and constraint conditions on the performances.

Most papers in relation extraction try to augment the number of employed features. In our experiment, we found that this does not always guarantee the best performance, despite the classifier being adopted is claimed to control these features independently. Because features may interact mutually in an indirect way, even with the same feature set, different constraint conditions can have significant influences on the final performance.

In Section 3, we introduced five candidate feature sets. Instead of using them as independent features, we combined them with additional in-

Table 3: Comparing With the State-of-the-Art Methods

System	Feature Set	P	R	F
(Che et al., 2005)	Ei.Type, Ei.Subtype, Order, $Word_{Ei\pm1}$, $Word_{Ei\pm2}$, $POS_{Ei\pm1}$, $POS_{Ei\pm2}$	84.81 (64.89)	75.69 (52.99)	79.99 (58.34)
(Zhang et al., 2011)	Ei.Type, Ei.Subtype, 9 Position Feature, Uni-Gram, Bi-Gram	79.56 (66.78)	72.99 (54.56)	76.13 (60.06)
Ours	$\mathcal{F}_{thp} \cup \mathcal{F}_{pos} \cup \mathcal{F}_{ow}$	92.26 (80.52)	88.51 (70.96)	90.35 (75.44)

formation. We proposed four constraint conditions to generate the soft constraint features. In Table 4, the performances of candidate features are compared when different constraint conditions was employed.

In Column 3 of Table 4 (**Constraint Condition**), (1), (2), (3), (4) and (5) stand for the referential feature sets[9] in Table 1. Symbol "/" means that the corresponding candidate features in the referential feature set are substituted by the new constraint condition. **Par** in Column 4 is the number of parameters in the trained maximum entropy model, which indicate the model complexity. **I** in Column 5 is the influence on performance. "-" and "+" mean that the performance is decreased or increased.

The first observation is that the combined features are more powerful than used as *singletons*. Model parameters are increased by the combined features. Increasing of parameters projects the relation extraction problem into a higher dimensional space, making the decision boundaries become more flexible.

The named entities in the ACE corpus are also annotated with the CLASS and LDCTYPE labels. Zhou et al. (2010) has shown that these labels can result in a weaker performance. Row 1, 2 and 3 show that, no matter how they are used, the performances decrease obviously. The reason of the performance degradation may be caused by the problem of over-fitting or data sparseness.

At most of the time, increase of model parameters can result in a better performance. Except in Row 8 and Row 11, when two *head nouns* of entity pair were combined as *semantic pair* and when *POS tag* were combined with the entity type, the performances are decreased. There are 7356 head nouns in the training set. Combining two head nouns may increase the feature space by $7356 \times (7356 - 1)$. Such a large feature space makes the occurrence of features close to a random distribution, leading to a worse data sparseness.

In Row 4, 10 and 13, these features are used as *singleton*, the performance degrades considerably. This means that, the missing of sentence structure information on the employed features can lead to a bad performance.

Row 9 and 12 show an interesting result. Comparing the reference set (5) with the reference set (3), the *Head noun* and *adjacent entity POS tag* get a better performance when used as *singletons*. These results reflect the interactions between different features. Discussion of this issue is beyond this paper's scope. In this paper, for a better demonstration of the constraint condition, we still use the *Position Sensitive* as the default setting to use the *Head noun* and the *adjacent entity POS tag*.

Row 13 and 14 compare the *Omni-word feature* (*By-Omni-word*) with the traditional segmentation based feature (*By-Segmentation*). *By-Segmentation* denotes the traditional segmentation based feature set generated by a segmentation tool, collecting every output of relation mention. In this place, the ICTCLAS package is adopted too.

Conventionally, if a sentence is perfectly segmented, *By-Segmentation* is straightforward and effective. But, our experiment shows different observations. Row 13 and 14 show that the *Omni-word* method outperforms the traditional method. Especially, when the *bin* information is used (Row 15), the performance of *Omni-word feature* increases considerably.

Row 14 shows that, compared with the traditional method, the Omni-word feature improves the performance by about 8.79% in 6 relation types and 11.83% in 18 subtypes in F-core. Such improvement may reside in the three reasons discussed in Section 3.3.

In short, from Table 4 we have seen that the *en-*

[9](1), (2), (3), (4) and (5) denote $\mathcal{F}_{thp}$, $\mathcal{F}_{ow}$, $\mathcal{F}_{thp} \cup \mathcal{F}_{pos}$, $\mathcal{F}_{thp} \cup \mathcal{F}_{ow}$ and $\mathcal{F}_{thp} \cup \mathcal{F}_{pos} \cup \mathcal{F}_{ow}$ respectively.

Table 4: Influence of Feature Set

No.	Feature	Constraint Condition	Par	P	R	F	I
1	entity CLASS and LDCTYPE	(1)/as singleton	21,112 21,910	60.29 (41.70)	42.82 (25.18)	50.07 (31.40)	-4.39 -12.09
2		(1)/combined with positional Info	21,159 22,013	63.02 (41.61)	44.47 (26.31)	52.15 (32.24)	-2.31 -11.25
3		(1)/as semantic pair	21,207 22,068	63.35 (42.98)	47.67 (31.34)	54.40 (36.25)	-0.06 -7.24
4	Type, Subtype semantic pair	(1)/as singleton	19,390 147,435	51.37 (32.8)	29.16 (18.97)	37.20 (24.06)	-17.26 -19.43
5		(1)/combined with positional info	19,524 20,297	61.77 (41.13)	43.67 (26.83)	51.17 (32.47)	-3.29 -11.02
6		(5)/as singleton	105,865 121,218	91.39 (79.32)	87.92 (68.73)	89.62 (73.65)	-0.73 -1.79
7	head noun	(3)/as singleton	21,450 22,409	85.66 (64.38)	75.74 (57.14)	80.40 (60.55)	-0.36 -0.34
8		(3)/as semantic pair	77,333 77,947	83.05 (59.70)	73.14 (51.70)	77.78 (55.41)	-2.54 -5.48
9		(5)/as singleton	100,963 115,499	92.50 (82.63)	88.90 (71.67)	90.66 (76.76)	+0.31 +1.32
10	adjacent entity POS tag	(3)/as singleton	21,450 22,409	72.66 (62.42)	61.16 (45.69)	66.41 (52.76)	-13.91 -8.13
11		(3)/combined with entity type	22,151 23,357	80.66 (63.41)	71.67 (53.16)	75.90 (57.83)	-4.42 -3.06
12		(5)/as singleton	106,931 121,194	92.50 (82.04)	88.66 (71.36)	90.54 (76.33)	+0.19 +0.89
13	Omni-word feature	(2)/*By-Segmentation* as singleton	36,916 41,652	67.19 (55.85)	60.12 (44.50)	63.46 (49.54)	-14.28 -10.77
14		(2)/*By-Segmentation* with bins	79,430 84,715	71.12 (54.76)	66.90 (43.50)	68.95 (48.48)	-8.79 -11.83
15		(2)/*By-Omni-word* as singleton	47,428 57,702	69.67 (54.85)	63.77 (48.84)	66.59 (51.67)	-11.15 -8.64
16		(5)/as singleton	57,321 67,722	91.43 (76.43)	86.37 (69.57)	88.83 (72.84)	-1.52 -2.60

tity type and subtype maximize the performance when used as *semantic pair*. *Head noun* and *adjacent entity POS tag* are employed to combine with positional information. *Omni-word feature* with bins information can increase the performance considerably. Our model (in Section 3.3) uses these settings. This insures that the performances of the candidate features are optimized.

5 Conclusion

In this paper, We proposed a novel Omni-word feature taking advantages of Chinese sub-phrases. We also introduced the soft constraint method for Chinese relation recognition. The soft constraint utilizes four constraint conditions to catch the structure information in a relation instance. Both the Omni-word feature and soft constrain make better use of information a sentence has, and minimize the deficiency caused by Chinese segmentation and parsing.

The size of the employed lexicon determines the dimension of the feature space. The first impression is that more lexicon entries result in more power. However, more lexicon entries also increase the computational complexity and bring in noises. In our future work, we will study this issue. The notion of soft constraints can also be extended to include more patterns, rules, regexes or syntac-

tic constraints that have been used for information extraction. The usability of these strategies is also left for future work.

Acknowledgments

The research was supported in part by NSF of China (91118005, 91218301, 61221063); 863 Program of China (2012AA011003); Cheung Kong Scholar's Program; Pillar Program of NST (2012BAH16F02); Ministry of Education of China Humanities and Social Sciences Project (12YJC880117); The Ministry of Education Innovation Research Team (IRT13035).

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of DL '00*, pages 85–94. ACM.

Michele Banko, Oren Etzioni, and Turing Center. 2008. The tradeoffs between open and traditional relation extraction. *Proceedings of ACL-HLT '08*, pages 28–36.

Sergey Brin. 1999. Extracting patterns and relations from the world wide web. *The World Wide Web and Databases*, pages 172–183.

Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr, and Tom M. Mitchell. 2010. Coupled semi-supervised learning for information extraction. In *Proceedings of WSDM '10*, pages 101–110.

Wanxiang Che, Ting Liu, and Sheng Li. 2005. Automatic entity relation extraction. *Journal of Chinese Information Processing*, 19(2):1–6.

Yu Chen, Wenjie Li, Yan Liu, Dequan Zheng, and Tiejun Zhao. 2010. Exploring deep belief network for chinese relation extraction. In *Proceedings of CLP '10*, pages 28–29.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of ISMB '99*, pages 77–86.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of ACL '04*, page 423.

James R. Curran, Tara Murphy, and Bernhard Scholz. 2007. Minimising semantic drift with mutual exclusion bootstrapping. In *Proceedings of PACLING '07*, pages 172–180.

Liu Dandan, Hu Yanan, and Qian Longhua. 2012. Exploiting lexical semantic resource for tree kernel-based chinese relation extraction. *Natural Language Processing and Chinese Computing*, pages 213–224.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of EMNLP '11*, pages 1535–1545.

Ralph Grishman. 2012. Information extraction: Capabilities and challenges. *Notes prepared for the*.

Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. 2005. Exploring various knowledge in relation extraction. In *Proceedings of ACL '05*, pages 427–434.

Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. 2010. Learning 5000 relational extractors. In *Proceedings of ACL '10*, volume 10, pages 286–295.

Rumjahn Hoosain. 1992. Psychological reality of the word in chinese. *Advances in psychology*, 90:111–130.

He Hu and Xiaoyong Du. 2012. Radical features for chinese text classification. In *Proceedings of FSKD '12*, pages 720–724.

Changning Huang and Hai Zhao. 2007. Chinese word segmentation : A decade review. *Journal of Chinese Information Processing*, 21(3):8–19.

Ruihong Huang, Le Sun, and Yuanyong Feng. 2008. Study of kernel-based methods for chinese relation extraction. *Information Retrieval Technology*, pages 598–604.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of ACL-demo '04*, page 22.

Zhang Le. 2004. Maximum entropy modeling toolkit for python and c++. *Natural Language Processing Lab, Northeastern University, China*.

Wenjie Li, Peng Zhang, Furu Wei, Yuexian Hou, and Qin Lu. 2008. A novel feature-based approach to chinese entity relation extraction. In *Proceedings of HLT-Short '08*, pages 89–92.

Nanyuan Liang. 1984. Written chinese word segmentation system-cdws. *Journal of Beijing Institute of Aeronautics and Astronautics*, 4.

Ruqi Lin, Jinxiu Chen, Xiaofang Yang, and Honglei Xu. 2010. Research on mixed model-based chinese relation extraction. In *Proceedings of ICCSIT '10*, volume 1, pages 687–691.

Andrew McCallum. 2005. Information extraction: Distilling structured data from unstructured text. *Queue*, 3(9):48–57.

Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 1998. Algorithms that learn to extract information: Bbn: Tipster phase iii. In *Proceedings of TIPSTER '98*, pages 75–89.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL '09*, pages 1003–1011.

Thahir P Mohamed, Estevam R Hruschka Jr., and Tom M Mitchell. 2011. Discovering relations between noun categories. In *Proceedings of EMNLP '11*, pages 1447–1455.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of COLING '04*.

Dan Roth and Wen-tau Yih. 2002. Probabilistic reasoning for entity & relation recognition. In *Proceedings of COLING '02*, pages 1–7.

Dan Roth and Wen-tau Yih. 2007. Global inference for entity and relation identification via a linear programming formulation. *Introduction to Statistical Relational Learning*, pages 553–580.

Richard Sproat, William Gale, Chilin Shih, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for chinese. *Computational linguistics*, 22(3):377–404.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of ACL '12*, pages 721–729.

Fei Wu and Daniel S. Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of CIKM '07*, pages 41–50.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48.

Ming Yan, Reinhold Kliegl, Eike Richter, Antje Nuthmann, and Hua Shu. 2010. Flexible saccade-target selection in chinese reading. *The Quarterly Journal of Experimental Psychology*, 63(4):705–725.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of SIGHAN '03*, pages 184–187.

Min Zhang, Jie Zhang, and Jian Su. 2006a. Exploring syntactic features for relation extraction using a convolution tree kernel. In *Proceedings of HLT-NAACL '06*, pages 288–295.

Min Zhang, Jie Zhang, Jian Su, and Guodong Zhou. 2006b. A composite kernel to extract relations between entities with both flat and structured features. In *Proceedings of ACL '06*, pages 825–832.

Peng Zhang, Wenjie Li, Furu Wei, Qin Lu, and Yuexian Hou. 2008. Exploiting the role of position feature in chinese relation extraction. In *Proceedings of LREC '08*.

Peng Zhang, Wenjie Li, Yuexian Hou, and Dawei Song. 2011. Developing position structure-based framework for chinese entity relation extraction. *ACM Transactions on Asian Language Information Processing (TALIP)*, 10(3):14.

Shubin Zhao and Ralph Grishman. 2005. Extracting relations with integrated information using kernel methods. In *Proceedings of ACL '05*, pages 419–426.

Ming Zhong, Sheng Wang, and Ming Wu. 2012. Revising word lattice using support vector machine for chinese word segmentation. In *Proceedings of IIWAS '12*, pages 352–355.

Guodong Zhou, Longhua Qian, and Jianxi Fan. 2010. Tree kernel-based semantic relation extraction with rich syntactic and semantic information. *Information Sciences*, 180(8):1313–1325.