Computational Linguistics and Chinese Language Processing
Vol. 11, No. 4, December 2006, pp. 393-416

A Pragmatic Chinese Word Segmentation Approach Based on Mixing Models[1]

Wei Jiang*, Yi Guan*, and Xiao-Long Wang*

Abstract

A pragmatic Chinese word segmentation approach is presented in this paper based on mixing language models. Chinese word segmentation is composed of several hard sub-tasks, which usually encounter different difficulties. The authors apply the corresponding language model to solve each special sub-task, so as to take advantage of each model. First, a class-based trigram is adopted in basic word segmentation, which applies the Absolute Discount Smoothing algorithm to overcome data sparseness. The Maximum Entropy Model (ME) is also used to identify Named Entities. Second, the authors propose the application of rough sets and average mutual information, etc. to extract special features. Finally, some features are extended through the combination of the word cluster and the thesaurus. The authors' system participated in the Second International Chinese Word Segmentation Bakeoff, and achieved 96.7 and 97.2 in F-measure in the PKU and MSRA open tests, respectively.

Keywords: Word Segmentation, N-Gram, Maximum Entropy Model, Rough Sets, Word Cluster, Machine Learning

1. Introduction

The word is a logical semantic and syntactic unit in natural language. Unlike English, there is no delimiter to mark word boundaries in Chinese language, so, in most Chinese NLP tasks, word segmentation is the foundational task which transforms the Chinese character string into a word sequence. It is a prerequisite to POS tagging, parser or further applications, such as Information Extraction, and the Question Answer system.

Word segmentation has attracted long-term attention in the research community for more

[1] This investigation is supported by the Key Program Projects of National Natural Science Foundation of China (60435020), and the National Natural Foundation of China (60504021).

* School of Computer Science and Technology, Harbin Institute of Technology, Heilongjiang Province, 150001, P. R. China
E-mail: jiangwei@insun.hit.edu.cn

[Received September 5, 2006; Revised December 3, 2006; Accepted December 6, 2006]

than two decades. Various methods have been proposed, which fall into two main categories. The first category is made up of rule-based approaches that make use of linguistic knowledge. Cheng [1999] and Liang [1993] described Maximum Forward Match and Maximum Backward Match segmentation. Hockenmaier [1998] and Palmer [1997] used transformation-based error-driven learning. Wu [1998] combined segmentation with a parser and word segmentation became a by-product of the sentence parser. The second category is made up of statistical methods that make use of machine learning algorithms and training on corpus. The typical language model is n-gram [Gao 2002]. Zhang [2003] used the Hierarchical Hidden Markov Model (HMM). In addition, there are some other machine learning methods, such as EM [Peng and Schuurmans 2001], and the channel noise model [Gao 2003]. Sproat [1996] used the WFST method. At present, many state-of-the-art systems use hybrid approaches. Gao [2004] proposed a unified method via the class-based model, and Zhang [2003] presented a unified approach using the Hierarchical Hidden Markov Model. Xue [2003] used Maximum Entropy. Peng [2004] used the Conditional Random Fields model.

Though many methods have been proposed and many improvements have been achieved, as a challenge task, word segmentation is not well-performed. The disambiguation and the out-of-vocabulary (OOV) identification are the main bottlenecks. Due to Zipf's Law, the sparse data problem is rarely avoided, while this problem brings great difficulties in improving the performance of the disambiguation and OOV identification. A meaningful direction for exploration to overcome the sparse data problem is to collect more linguistic knowledge or features and incorporate them into the processing systems.

In this paper, the authors propose to solve the Chinese word segmentation task based on mixing models. The "No Free Lunch Theorem" and "Ugly Duckling Theorem" in Machine Learning theory have indicated that domain knowledge is essential for improving the processing performance. For this reason, different language models will be applied to solve each special sub-task, which is classified according to its linguistic phenomenon and the Natural Language Processing (NLP) technology used in the word segmentation. Another consideration is the pragmatic attribution, *e.g.* some successive processing may require different kinds of balance between precision and efficiency. So, this approach is a pragmatic one, which may incorporate several delicate processing modules, some of which can improve precision by introducing complicated models and utilizing more linguistic knowledge. However, this does result in a decrease in efficiency. Based on the assumption that more delicate linguistic knowledge or some fine linguistic statistical phenomenon can bring information gain to the segmentation task, the authors propose to apply Rough Set Theory and Average Mutual Information, etc. to extract complicated and long distance features. and the authors will also explore combining the word cluster and the thesaurus to extend the features so as to overcome the sparse data problem. This system participated in the Second

International Chinese Word Segmentation Bakeoff (SIGHAN 2005), and a simplified version participated in the SIGHAN 2006.

Section 2 describes the structure of the system. Section 3 describes in detail Named Entity Recognition, which is one of the difficult tasks in word segmentation. Section 4 presents experimental results obtained with the authors' system. Finally, some conclusions will be drawn and direction for future work will be given in Section 5.

2. System Description

All words in this system are categorized into five types: Lexicon words (LW), Factoid words (FT), Morphological derived words (MDW), Named entities (NE), and New words (NW). Table 1 shows the tag, description, and some examples for each word type.

Table 1. The tag, description and examples for each word category

TAG	Description	Examples
LW	The word in the Lexicon	最近(recent),博士(doctor), 学位(degree)
FT	Number, Date, Time etc.	2910, 46.12%, 2004年05月12日, 01:06
MDW	Morphological derived words	朋友们(friends), 高高兴兴(happily), 进出口(imports and exports)
NE	Named Entities	孙桂平(Sun Gui-Ping), 哈尔滨(harbin)
NW	The other OOV except FT, MDW, NE	海风牌(sea breeze brand), 古典式 (classical), 景观灯(sighting lamp)

To the sentence "同学们下午两点三十分到孙桂平家做客" (Some students visit Sun Gui-Ping in his home at 2:30 p.m.), the segmentation result is "{同学们/[MR_Suffix]} {下午两点三十分/[TIME]} {到} {孙桂平/[PER]} {家} {做客}". where the word "同学们/[MR_Suffix]" is a morphologically derived word, and "下午两点三十分/[TIME]" is a factoid word, all of which can be detected by Segmentation module while "孙桂平/[PER]" is a named entity, and detected in NE Recognition module. Figure 1 shows the structure of this system.

The input character sequence is converted into one or several sentences, which is the basic dealing unit. The internal encoding is UNICODE, and the "Code Convert" module is used to convert the permitted encoding, such as GB2312 and BIG5, into UNICODE. "Basic Segmentation" is used to deal with the LW, FT, MDW words, and "Named Entity Recognition" is used to detect NW words. The authors adopt the New Word Detection algorithm to detect suffix-based new words. The "Disambiguation" module is performed to classify complicated ambiguous words, and all the above results are connected to the final result, namely "word sequence", which is denoted by XML format. The sequence of each

applied component is decided by the performance of the system. In the following part of this section, the authors will detail the basic theory and the implementation of the system.

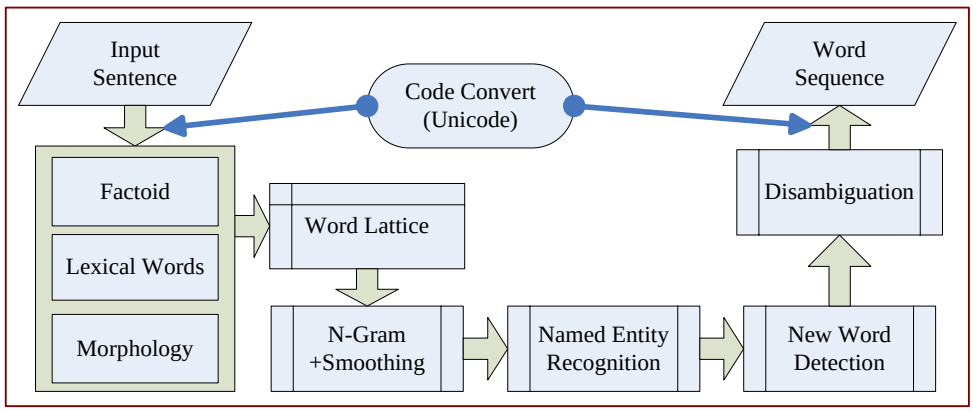

Figure 1. The structure of the proposed system

2.1 Trigram and Smoothing Algorithm

The authors apply the Trigram model to the word segmentation task [Jiang 2005; Jiang 2007], and make use of the Absolute Discount Smoothing algorithm to overcome the sparse data problem.

The Trigram model is used to convert the sentence into a word sequence. Let $\mathbf{w} = w_1 w_2 \ldots w_n$ be a word sequence, then the most likely word sequence w* in Trigram is:

$$\mathbf{w}^* = \arg\max_{w_1 w_2 \cdots w_n} \prod_{i=1}^{n} P(w_i \mid w_{i-2} w_{i-1}) , \quad (1)$$

where let $P(w_0|w_{-2}\ w_{-1})$ be $P(w_0)$ and let $P(w_1|w_{-1}\ w_0)$ be $P(w_1|w_0)$, and w_i represents LW or a type of FT or MDW. In order to search for the best segmentation way, all the word candidates are filled into the word lattice [Jiang 2006B], as shown in Figure 2, and the Viterbi algorithm is used to search for the best word segmentation path over the built word lattice.

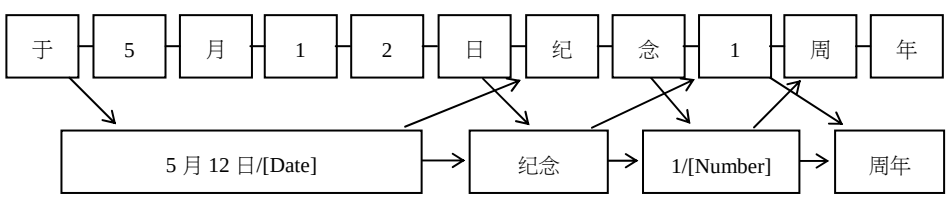

Figure 2. The class-based word lattice in the segmentation task

FT and MDW need to be detected when constructing a word lattice (detailed in section 2.2). The data structure of the lexicon can affect the efficiency of word segmentation, so

lexicon words are represented as a set of TRIEs, which are tree-like structures. Words starting with the same character are represented as a TRIE, where the root represents the first Chinese character, and the children of the root represent the second character, and so on (detailed in section 2.3).

When searching a word lattice, there is a zero-probability phenomenon due to the sparse data problem. For instance, if there is no co-occurrence pair "我们/吃/香蕉"(we eat bananas) in the training corpus, then P(香蕉|我们，吃) = 0. According to formula (1), the probability of the whole candidate path, which contains "我们/吃/香蕉", is zero as a result of the local zero probability. In order to overcome the sparse data problem, this system has applied the Absolute Discounting Smoothing algorithm [Chen 1999].

$$N_{1+}(w_{i-n+1}^{i-1}\bullet) = |\{w_i : c(w_{i-n+1}^{i-1}w_i) > 0\}| \tag{2}$$

The notation N_{1+} is meant to evoke the number of words that have one or more counts, and the $\bullet$ is meant to evoke a free variable that is summed over. The function $c()$ represents the count of one word or the co-occurrence count of multi-words. In this case, the smoothing probability can be calculated by the Equation 3.

$$p(w_i \mid w_{i-n+1}^{i-1}) = \frac{\max\{c(w_{i-n+1}^{i}) - D, 0\}}{\sum_{w_i} c(w_{i-n+1}^{i})} + (1-\lambda)\, p(w_i \mid w_{i-n+2}^{i-1}) \tag{3}$$

where

$$1-\lambda = \left(\frac{D}{\sum_{w_i} c(w_{i-n+1}^{i})} N_{1+}(w_{i-n+1}^{i-1}\bullet) \right) \tag{4}$$

In this trigram model, the maximum n may be 3. A fixed discount D ($0 \le D \le 1$) can be set through the deleted estimation on the training data. They arrive at the estimate

$$D = \frac{n_1}{n_1 + 2n_2} \tag{5}$$

where n_1 and n_2 are the total number of n-grams with exactly one and two counts, respectively [Jiang 2006B; Jiang 2007].

After basic segmentation, some complicated ambiguous segmentation can be further disambiguated. In the Trigram model, only the previous two words are considered as context features, while in disambiguation processing (detailed in section 2.4), one can use the Maximum Entropy Model-fused features [Jiang 2006A] or a rule-based method.

2.2 Factoid and Morphological Words

As all of the Factoid words can be represented as regular expressions, the detection of factoid words can be achieved by Finite State Automaton (FSA).The categories of factoid words, which can be detected [Jiang 2006B; Jiang 2006D] by this system, are shown in Table 2.

Table 2. Factoid word categories

FT type	Factoid word description	Examples
Number	Integer, percent, real etc.	2203, 25.78%, 零点五, 20.542
Date	Date	2004 年 5 月 12 日, 2004-06-06
Time	Time	8:00, 十点二十分, 晚上 6 点
English	English word,	Hello, How, are, you
www	Website, IP address	http://www.hit.edu.cn; 192.168.140.133
email	Email	jiangwei@insun.hit.edu.cn
phone	Phone, fax	+86-451-86413322; (0451)86413322

Deterministic FSA (DFA) is efficient because a unique "next state" is determined when given an input symbol and the current state. However, it is common for a linguist to write rules, which can be represented directly as a non-deterministic FSA (NFA), *i.e.* which allow several "next states" to follow a given input and state.

Since every NFA has an equivalent DFA, an FT rule compiler was build to convert all the FT generative rules into a DFA [Jiang 2007]. The rule description is in Table 3.

Table 3. The demonstration of partial ELUSLex rules

```
<digit> -> [0..9]| [０..９];          //define Arabic numerals
<integer> ::= {<digit>+};          // define Arabic Integer
<real> ::= <integer>(.|・|·|点)<integer>; // decimal fraction
<day> -> <integer>日;               // define day
<month> -> <integer>月;            // define month
<year> -> <digit><integer>年;          // define year
<date> ::= <year><month><day>;    // define date
```

In order to provide a kind of convenient and powerful description ability, some meta descriptions are assigned to the meta language.

- ✓ Permitted meta rules: <Non-terminator>, terminator, {Loop block}, {Loop block+}, {Loop block*}, [Range block] (e.g. [a..z], ["a".."z"]), |, (Optional block), (Optional block +), (Optional block *).
- ✓ Transferred meaning : if the token in the meta rule is the terminator, one needs to transfer its meaning, so one can use double quotation marks to bracket the terminator when it present ambiguity. *e.g.* "(", "|", ")".

- ✓ Rule type: "->" is a temporary generative rule, and "::=" is a real generative rule or a detected rule. This method makes the rule easily written.

The authors built an FT rule compiler to convert all the FT generative rules into a DFA. Obviously, this method makes the system easy to be transferred into a different word segmentation definition, such as from PKU to MSRA. In fact, the authors have used it in SIGHAN 2005 and SIGHAN 2006. Correspondingly, the DFA is represented by the matrix [Jiang 2007], and a run API is provided to make this method easily used. FT detection is important in building the word lattice in word segmentation and also important in the POS tagging task.

The proposed system tries to deal with five main categories of morphologically derived words in real application, the same as Wu [2003] and Gao [2004]: 1) Affixation：老师们(teachers), 朋友们(friends); 2) Reduplication: 高高兴兴(happily); 3) Splitting:玩会球(play ball for a while)，洗了澡(already wash), 吃了饭(already ate); 4) Merging: "进出口" comes from "进口"（importation）and "出口"(exportation); 5) Head Particle: "走出去"comes from "走"(walk) and "出去"(out).

The authors collate the possible MDW into a morphological dictionary from a large corpus, according to the morphological categories mentioned above. Then, some manual selections are needed to select fitting MDW words. As the segmentation specifications of all kinds of corpora are usually different, one needs to collect the corresponding MDW words.

2.3 The Data Structure of Lexicon

The data structure of a lexicon affects the efficiency of word segmentation, as the word candidate in the word lattice is generated through searching the lexicon. When given a sentence string, the candidate comes from matching the substring (starting from the current Chinese character), and judging whether this substring exists in the lexicon. The authors represent lexicon words as a set of TRIEs, which is a tree-like structure. Words starting with the same character are represented as a TRIE, where the root represents the first Chinese character, and the children of the root represent the second characters, and so on, as shown in Figure 3.

The lexical word starts from the "Start state", and ends in the "End state". When matching the input sentence and generating the word candidate in the word lattice, each time "End State" is passed, a word candidate is formed and the properties of the current word represented in the "End State" are filled into word lattice.

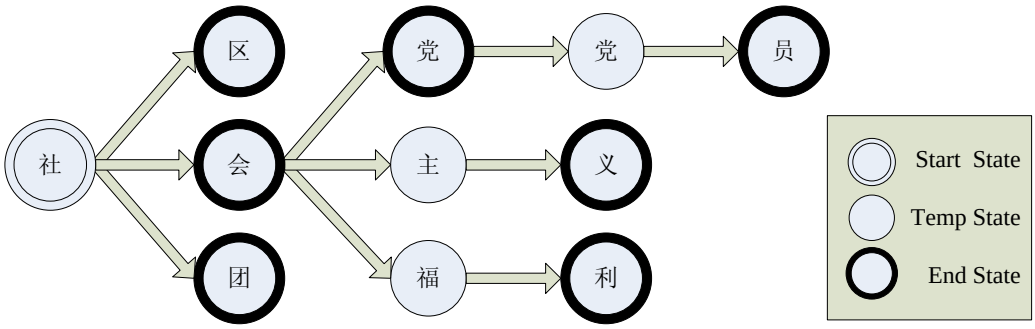

Figure 3. The example of data structure in the lexicon (TRIEs)

Since each Chinese character in the input sentence needs to match the word candidate, the authors build many TRIEs, as shown in Figure 3, to form a lexicon. The example in Figure 3, "社会主义" (socialism), is a word, and this tree is used to match the candidate from the start to the end in the sentence. If one constructs a word lattice in the opposite direction, the tree needs to be built correspondingly, *e.g.* "义主会社". This data structure can improve speed in generating the word lattice.

2.4 The Disambiguation

It is necessary to effectively exploit the context in the disambiguation process. The authors have proposed using rough sets to extract complicated features and long distance features for disambiguation, which has been reported in previous work [Jiang 2006A]. In that paper, the authors proposed introducing a variable precision Rough Set in feature extraction, in order to acquire a balance of features in disambiguation processing, along with attempting to process complicated and consecutive ambiguity segmentation in the paper. In this paper, the ambiguity segmentations come from the error-total results after evaluating the system.

In Rough Set theory, knowledge is represented via relational tables. An Information System can be defined as follows: $I = (U, A, V_a, f_a)_{a \in A}$, where U is a non-empty set of objects; A is a non-empty set of attribute a's; for each attribute $a \in A$, there is an attribute value V_a set and an information function $f_a : U \to V_a$. An equivalence θ on set U is called an indiscernible relation, and lower approximation for an object set $X \subseteq U$ is defined as $\underline{X}\theta = \{\theta x : \theta x \subseteq X\}$. However, this formula is too strict to fit the requirements of Natural Language Processing. For this reason, the concept of α *-approximation* is provided: $\underline{X}\theta(\alpha) = \bigcup \left\{ \theta x : \frac{|\theta x \cap X|}{|\theta x|} \geq \alpha \right\}$, where α is an external parameter [Jiang 2006A].

When extracting features, α -Approximation will probably cause unbalanced support, since each segmentation of the ambiguities possibly has disproportionate distribution. In order to let all the features that were added in provide more evidence in guiding toward the correct segmentation, λ -Approximation is introduced in this model. Let filter parameter $\alpha_d \in [0,1]$,

and the n-order rough rule set of keyword t be noted as R_t^n, then $R_t^n \in G_{t,n}$, and defined as: $R_t^n = \left\{ r \in G_{t,n} \mid r \in \underline{X}_{d,\theta}^{(i)}(\alpha_d) \right\}$, where $n = \left| A_f \right| - 1$, $i \in [1, K]$ and $G_{t,n}$ represents generalized LIT. In $G_{t,n}$, indiscernible objects are merged, the objects of each equivalence classes are counted and potential rule precision is calculated. If one lets each α_d have the same value, namely, let $\alpha_d = \alpha$ to the decision attribute d, then λ-Approximation will revert back to the conventional definition of α-Approximation.

In order to make effective use of contextual knowledge, the authors adopt the Maximum Entropy model (ME), which is a conditional probabilistic model, and relax the feature independent assumption. Disambiguation is regarded as a classifying problem in ambiguous words by the Maximum Entropy model, which is defined over $H \times T$ in segmentation disambiguation, where H is the set of possible contexts around the target word that will be tagged, and T is the set of allowable tags. Then, the model's conditional probability is defined as:

$$p(t \mid h) = \frac{p(h,t)}{\sum_{t' \in T} p(h,t')} \tag{6}$$

where,

$$p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)} \tag{7}$$

It has been pointed out that two kinds of ambiguities were dealt with. One is the simple two categories problem, such as "从/小学"(from elementary school) and "从小/学"(study since youth), where the tags are 0 and 1; here 0 represents the first segmentation and 1 represents the second. H includes the near context and long distance context. The former is comprised of two words around the target word, and the latter features can be obtained by Average Mutual Information, Information Gain, etc.

In fact, a rough statistical result showed that the "one segmentation error" occupied more than 90% of all errors when not considering the errors caused by Named Entity Recognition. Here, "one segmentation error" means that the segmentations surrounding this segmentation error are correct. So, the authors focus on "one segmentation error", which may be seen in two types of Chinese segmentation ambiguities: overlapping ambiguity and combining ambiguity.

Rough rule features are added in the ME model as a new kind of feature:

$$f_j(a,b) = \begin{cases} 1 & if\,((w = KeyWord)and(A_f(r) = b)and(a = d)) \\ 0 & \text{others} \end{cases} \tag{8}$$

where the formula $A_f(r) = b$ represents that the conditional attribute of r can be reconstructed in the current context, and $a = d$ represents the decision attribute of d is equal

to the tag of ambiguous word. (More details were reported in the paper "[Jiang 2006A]".)

2.5 The Suffix Based New Word Detection

New word (NW) in this system refers to the out-of-vocabulary word that isn't an FT word, MDW word or NE word. The authors do not try to detect all the NW words, since the precision is not satisfactory based on the existing methods in some applications.

On the other hand, in some applications, it is meaningful to recognize some special new words. For instance, "景观+灯" (sightseeing light), "海风+牌" (Sea Breeze brand). Since some prefixes or some suffixes are paid attention to by this system, such as "现代 + 化"(modernization), "x + 式"(x + way), "x + 灯"(x + light), the authors propose to apply a variance algorithm to acquire the prefix or suffix candidate, leaving some minor manual selections possibly required. Hereafter, this paper takes the suffix as an instance, and collects the new words, *e.g.* "日光+灯" (sunlight), "霓虹+灯" (neon light), "景观+灯" (sightseeing light), etc. Table 4 illustrates the method.

Table 4. The variance method to obtain the suffix

	S_1	S_2	…	S_m
W_1	c_{11}	c_{21}	…	S_{m1}
W_2	c_{12}	c_{22}	…	S_{m2}
….	…	…	…	…
W_n	c_{1n}	c_{2n}	…	S_{mn}

Use $S_1..S_m$ to represent m candidate suffixes, $W_1..W_n$ represent n remained word with the suffix being razed. *e.g.* S_1 is "灯" (light), then W_1 represents "景观" (sightseeing), W_1S_1 is the W_1+S_1="景观灯" (sightseeing light). Now, suppose C_{xy}=Count(S_x,W_y)[2], and N_{xy} is the existence of a co-occurring pair (S_x,W_y)[3], then, one gets the following formula:

$$CV(S_x) = \sum_{i=1}^{m} N_{xi}\ ,\quad Sum(S_x)=\sum_{i=1}^{m} C_{xi}\ ,\ avg(S_x)= Sum(S_x)/CV(S_x),$$

$$p_{xi}=C_{xi}/Sum(S_x), V_{xi}=p_{xi}*(C_{xi}-avg(S_x)) *(C_{xi}-avg(S_x))$$

So, the variance $V(S_x)= \sum_{i=1}^{m} V_{xi}$.

Besides the variance, one also needs to consider two other factors: (1) the occurrence count in the corpus; (2) the type count that this suffix has constructed words in the lexicon. By considering the above two factors in Sighan2005 evaluation [Jiang 2005], the researchers selected 25 new word suffixes, *e.g.* 制 (method), 牌(brand), 型 (type)、式 (way). These

[2] Here, Count(x,y) represents taking count of the co-occurrence of pair (x,y).

[3] Namely, if C_{xy}>0 then N_{xy} is 1, else N_{xy} is 0.

suffixes also seem to be useful in the Information Retrieval task.

The detection process adopts the Local Maximum Entropy Model, and this process is similar to the NER module [Jiang 2007].

3. Named Entity Recognition

Named Entity Recognition (NER) is one of the common message understanding tasks. The objective is to identify and categorize all members of certain categories of "proper names". In MUC-7, there are seven categories: person, organization, location, date, time, percentage, and monetary amount. Named Entities (NE) are broadly distributed in original texts from many domains. In this work, the authors only focus on those more difficult, yet commonly used categories: PER, LOC and ORG. Other NE, such as times and quantities can be recognized simply via Finite State Automata (Section 2.2), and do not need to be aided by a disambiguation algorithm (Section 2.4).

The extensive evaluation of NER systems in recent years (such as CoNLL-2002 and CoNLL-2003) indicates the best statistical systems are typically achieved by using a linear (or log-linear) classification algorithm, such as the Maximum Entropy model, together with a vast amount of carefully designed linguistic features. This still seems true at present in terms of statistics based methods.

In this section, the authors adopt the ME model, which is a linear (or log-linear) classification, to identify the Named Entities, and the focus will be on the utilization of the features [Jiang 2006C]. In addition, the authors propose to build double-layer fixing models to detect the Named Entities, which has also been reported in another paper [Jiang 2007].

The authors use w_i (i=0,1,…n) to denote the input sequence, then every token w_i should be assigned a tag t_i. B-I-O encoding, *e.g.*, B-CPN, I-CPN as the beginning of Chinese person's name and the continued part of person's name, respectively, is adopted. Furthermore, in order to improve the ability of describing the rich tagging knowledge, part of the role tags [Zhang 2003] is appended, including the Named Entity prefix, suffix and infix. For example:

- ✓ 我/O 荣幸/O 地/O 拜访/B-PER_PREFIX 孙/B-PER 桂/I-PER 平/I-PER 女士/B-PER_SUFFIX (Note: It's my honor to visit Ms. Sun Gui-Ping.)

As there are distinct differences between a Chinese person's name and the translation of the person's name in terms of the person construction, the person name is divided into Chinese Person Name (CPN) and the Translation Person Name (FPN). In addition, the authors do not distinguish the type of infix, so the tag number for NER in this system is: 4 * 4 + 1 (O) + 1 (INFIX) = 18.

3.1 The Context Features

The "Ugly Duckling Theorem" has denoted that there is no generic feature extraction method suitable for all kinds of tasks. The basic feature template is shown in Table 5.

Table 5. Feature templates for Named Entity Recognition

Type	Feature Template
one order feature	w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}
two order feature	$w_{i-1:i}$, $w_{i:i+1}$
NER tag feature	t_{i-1}

In addition, in order to solve the unstable feature collection problem caused by having no delimiters to separate Chinese words, inspired by the term extraction in text classification, the authors construct a novel feature template of "word->tag" to extract the trigger features, which have a flexible distance between the two units [Jiang 2006C].

Mutual Information (MI) measures the interdependence between a trigger word and a NE type, being defined as:

$$\mathrm{MI}(W,C) = \log \frac{P(W \Lambda C)}{P(W) \times P(C)} \tag{9}$$

where P(W) represents the probability of the trigger word, and P(C) is the probability of the corresponding NE category. However, this method does not consider the influence of lacking one point. In contrast, average mutual information (AMI) is defined as:

$$\begin{aligned}\mathrm{AMI}(W,C) = {} & P(W,C)\log\frac{P(C|W)}{P(C)} + P(W,\overline{C})\log\frac{P(\overline{C}|W)}{P(\overline{C})} \\ & + P(\overline{W},C)\log\frac{P(C|\overline{W})}{P(C)} + P(\overline{W},\overline{C})\log\frac{P(\overline{C}|\overline{W})}{P(\overline{C})}\end{aligned} \tag{10}$$

MI in fact is point wise information, while AMI can look like a Kullback-Leibler (KL) divergence:

$$\mathrm{AMI}(X,Y) = D(P(X,Y)||P(X) \times P(Y)) \tag{11}$$

Equation 11 measures the two different probability distributions between $P(X,Y)$ and $P(X) \times P(Y)$. However, MI is only a point in the whole set of distributions.

Let m be the number of the possible categories count, the average mutual information is

$$\mathrm{AMI}_{\mathrm{avg}}(W) = \sum_{i=1}^{m} P(C_i) \times \mathrm{AMI}(W, C_i) \tag{12}$$

or another optional formula adopted in this paper:

$$AMI_{max}(W) = MAX_{i=1}^{m} AMI(W, C_i) \quad (13)$$

The authors select the top triggers with higher AMI value, and acquire the trigger words.

3.2 The Entity Features

Besides context features, entity features are also very important in the NER task, such as the suffix of Location or Organization. The authors performed statistical analysis of foreign resources, including the corpora and the collected entity name on the Internet. The authors built 8 kinds of dictionaries:

Table 6. The resource dictionary for the Named Entity Recognition

List Type	Lexicon	Examples
Word list	Place lexicon	北京, 纽约, 马家沟
	Chinese surname	张, 王, 赵, 欧阳
String list	Prefix of PER	老, 阿, 小
	Suffix of PLA	山, 湖, 寺, 台, 海
	Suffix of ORG	会, 联盟, 组织, 局
Character list	Character for CPER	军,刚, 莲, 茵, 倩
	Character for FPER	科, 曼, 斯, 娃, 贝
	Rare character	滠, 胨, 薷

Table 6 gives several kinds of resource dictionaries used in this system. Take the “Suffix of ORG” as an example, the suffix “局”, “组织” is a good hint to detect the Organization Name, so the authors collected them into a “Suffix of ORG” dictionary. When used in the Maximum Entropy Model, this dictionary is used to judge the existing cases of the specified context feature.

3.3 The Feature Extension

Feature extension is used to overcome the sparse data problem and to increase robustness. In addition, semantic and pragmatic knowledge is useful in language processing, *e.g.*, if one knows “教授” (professor) is a good hint to label a person’s name, the similar words {老师 teacher), 助教(assistant), 讲师(lecturer)}, should have the same effect. So, one can build a semantic class by combining word clusters and using a thesaurus.

A vector for word *w* is derived from the close neighbors of *w* in the corpus. Close neighbors are all words that co-occur with *w* in a sentence or a larger context. The entry for word *v* in the vector for *w* records the number of times that word *v* occurs close to *w* in the corpus. The authors refer this vector space to as Word Space.

Figure 4 gives a schematic example of two words being represented in a two-dimensional space. This vector representation captures the typical topic or subject matter of a word. By looking at the amount of overlap between two vectors, one can roughly determine how closely they are related semantically. This is because related meanings are often expressed by similar sets of words. Semantically related words will, therefore, co-occur with similar neighbors and their vectors will have considerable overlap.

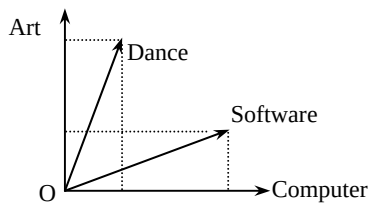

Figure 4. A demonstration of word vectors

The authors combine the basic semantic word in a thesaurus -- HOWNET2005 -- with the TF-IDF algorithm [Zhao 2005B], and use a frequency cutoff to select the 2000 words to serve as the dimensions of Word Space. Compared with the traditional TF-IDF method, this method increases the taxonomical information, so this method can give a better measure of the word similarity.

After constructing word vectors, the similarity can be measured by the cosine between two vectors. The cosine is equivalent to the normalized correlation coefficient:

$$corr(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2 \sum_{i=1}^{N} w_i^2}} \tag{14}$$

The word cluster algorithm in the word vectors is used to measure the similarity by totaling the pragmatic knowledge from the corpora.

4. Evaluation and Discussion

The authors evaluated the system with two kinds of corpora: 1) The corpora in the International Chinese Word Segmentation Bakeoff; 2) The prior six-month corpora of Peoples' Daily (China) in 1998, which came from Peking University, and have been annotated with lexical tags, including word segmentation, POS tagging, and Named Entity Recognition tags.

4.1 The International Chinese Word Segmentation Bakeoff

This system participated in the Second International Chinese Word Segmentation Bakeoff (SIGHAN-2005) held in 2005, and also participated in SIGHAN-2006.

The performance of ELUS in the SIGHAN-2005 bakeoff is presented in Table 7 and Table 8 respectively, in terms of Recall (R), Precision (P) and F score in percentages. The score software is standard and open by SIGHAN.

Table 7. Closed test, in percentages (%)

Closed	R	P	F	OOV	R_{oov}	R_{iv}
PKU	95.4	92.7	94.1	5.8	51.8	98.1
MSR	97.3	94.5	95.9	2.6	32.3	99.1
CITYU	93.4	86.5	89.8	7.4	24.8	98.9
AS	94.3	89.5	91.8	4.3	13.7	97.9

This system has good performance in terms of F-measure in the simplified Chinese open test, including the PKU and MSR open tests. In addition, its In-vocabulary word (IV, namely, Lexical words) identification performance is remarkable, ranging from 97.7% to 99.1%, standing at the top or near the top in all the tests in which it has participated.

Table 8. Open test, in percentages (%)

Open	R	P	F	OOV	R_{oov}	R_{iv}
PKU	96.8	96.6	96.7	5.8	82.6	97.7
MSR	98.0	96.5	97.2	2.6	59.0	99.0
CITYU	94.6	89.8	92.2	7.4	41.7	98.9
AS	95.2	92.0	93.6	4.3	35.4	97.9

This good performance in the R_{iv} is due to the class-based Trigram, Absolute Discount Smoothing and Word Disambiguation module with the rough rule features. In this bakeoff, the Name Entity Recognition is a two layer mixing approach, which is reported in detail in a previous paper [Jiang 2007]. The Maximum Entropy Model in the mixing method is similar to that found in Section 3.

The performance of this system in the SIGHAN-2006 bakeoff is presented in Table 9.

Table 9. MSRA test in SIGHAN2006 (%)

MSRA	R	P	F	OOV	R_{oov}	R_{iv}
Close	96.3	91.8	94.0	3.4	17.5	99.1
Open	97.7	96.0	96.8	3.4	62.4	98.9

The system has good performance in terms of R_{iv} measure. The R_{iv} measure in a closed test and in an open test was 99.1% and 98.9%, respectively. This good performance is due to a

class-based Trigram with the Absolute Smoothing and Word Disambiguation algorithm.

In this system, the following reasons illustrate why the open test had better performance than the closed test:

(1) Named Entity Recognition module is added into the open test system. And Named Entities, including PER, LOC, ORG, occupy the most of the out-of-vocabulary words.

(2) The system of closed test can only use the dictionary that is collected from the given training corpus, while the system of open test can use a better dictionary, which includes the words that exist in MSRA training corpus in SIGHAN-2005. As is known, the dictionary is one of the important factors that affects the performance, because the LW candidates in the word lattice are generated from the dictionary.

As for the dictionary, the authors compare the two collections in SIGHAN-2005 and SIGHAN2006, and in evaluating the SIGHAN-2005 MSRA closed test. There are less training sentences in SIGHAN-2006. As a result, there is at least a 1.2% performance decrease. So, this result indicates that the dictionary can have an important impact in a system.

4.2 The Detailed Evaluation of the System

In this section, some detailed evaluation results are presented. The authors mainly focus on two difficult sub-tasks in the word segmentation task, namely disambiguation and Named Entity Recognition. The measurements in the following experiments include: the precision P = the right count / the model count, the recall rate R = the right count / the corpus count, and F-measure = (2 * P * R) / (P + R).

Table 10. The comparison experiment for some ambiguities

Ambiguity	Type	Train Count	Test Count	ME Precision	RS model Precision
才能	才能	704	190	90%	93%
	才/能	7612	300		
不要	不要	1421	150	91%	95%
	不/要	497	80		
从小学	从小/学	170	40	88%	91%
	从/小学	260	70		
将来	将来	1200	200	92%	97%
	将/来	35	10		
个人	个人	1016	150	89%	94%
	个/人	819	120		

The authors firstly evaluate the disambiguation performance. Training was done with the preceding five month's Corpus of the People's Daily Newspaper, 1998, including 664,805 sentences, and the test corpus was the sixth month corpus, including 136,647 sentences. The authors applied the Rough Set (RS) theory to extract the rough rule features, and fused this theory into the Maximum Entropy Model. The basic feature templates are the w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, furthermore, the rough rule features were fused into the ME disambiguation model [Jiang 2006A], the results are shown in Table 10.

Table 10 demonstrates that RS model may achieve improvement over the baseline ME model. There are at least two main advantages in the proposed method: 1) As a conditional probabilistic model, ME can be fused to more effective features, which relaxes the features independent assumption that is suffered from by the N-Gram model; 2) The authors apply the rough set theory to extract complicated and long distance features. Due to how more effective features are utilized, the new method overcomes the sparse data problem to a certain extent.

Now, the authors evaluate the second group of difficult sub-tasks, namely, the NER module. The experimental corpora also came from the Chinese People's Daily Newspaper in the first half-year of 1998. The overview of the entity distribution is shown in Table 11.

Table 11. The entity distribution in People's Daily

Named Entity	CPN	FPN	LOC	ORG
By entities	27.54%	8.86%	41.53%	22.07%
By corpus	1.29%	0.41%	1.94%	1.03%
Occur Count	92941	29912	140162	74483

Figure 5 shows that the distribution of the entities complies with the Zipf's law. As a result, the entities exhibit the sparse property; thereby bringing trouble to the model.

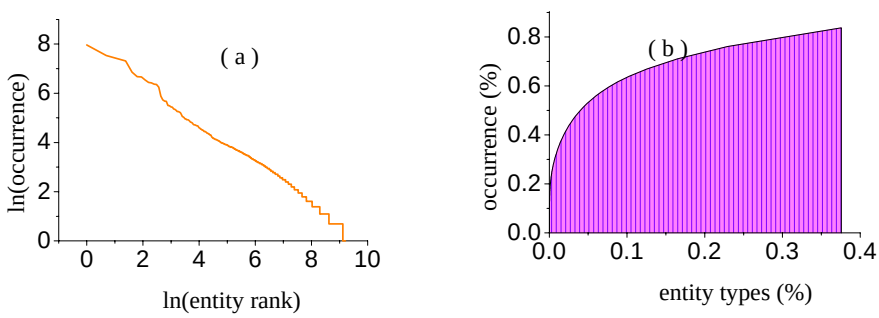

Figure 5. The entities that exhibit Zipf's law

The authors compared several Named Entity Recognition Models, and Table 12 gives the evaluation result. The baseline result is obtained by selecting the NER tag that is most frequently associated with the current word. The authors add several tags in the tag set (Called adding "role"), including the entity prefix, infix and suffix. These tags are used to enhance the ability of the context repetition. In this experiment, HMM is one order model, and ME, CRF use the feature template:W_{-2},W_{-1},W_0,W_1,W_2,$W_{-1:0}$,$W_{0:1}$,T_{-1}.

Table 12 indicates that the ME + Role has achieved the best performance. Compared with Hidden Markov Model (HMM), ME can fuse more context features.

Table 12. The comparison of several NER models

Model	Precision	Recall	F-measure
BaseLine	68.99%	73.54%	71.19%
HMM	79.20%	79.96%	79.58%
ME	84.77%	83.23%	83.99%
HMM + Role	83.68%	85.20%	84.43%
ME + Role	87.95%	84.62%	86.25%

Table 13. Trigger pairs draw from corpus

Pair	AMI		MI	
	Value	Rank	Value	Rank
同志 CPN	3.9e-4	6	2.71	144
说 CPN	2.3e-4	11	1.85	885
主任 ORG	1.2e-4	23	2.63	181
会见 CPN	1.1e-4	27	2.43	269
举行 LOC	9.5e-5	34	1.61	1279
北部 LOC	3.9e-5	80	2.45	271
会议 ORG	3.8e-5	83	1.39	1650
教授 CPN	3.1e-5	96	2.21	463

In another experiment, the authors selected the pairs using two methods, one is to filter by the threshold, such as AMI>0.001, the other method is to select the top pair after ranking the pair in descending order, *e.g.* selecting the top 500 pairs, having the maximum value. The partial pairs are shown in Table 13, including the MI, AMI value and their rank.

Then, the trigger features were collected, respectively, from above corpora. Taking AMI as an example, after being put in descending order, the top 500 features were selected. Table 14 shows the compared performance with trigger selected by AMI.

Table 14. The performance with AMI trigger

Entity type	ME (%)			ME+AMI(%)		
	P	R	F	P	R	F
CPN	84.54	77.71	80.98	86.36	82.41	84.34
FPN	73.27	53.21	61.65	78.50	56.90	65.97
LOC	86.95	76.53	81.41	87.57	77.62	82.30
ORG	74.87	55.29	63.61	74.08	60.95	66.88
Overall	82.81	69.74	75.71	83.60	72.97	77.92

Table 14 gives the detailed comparison between ME and ME with AMI trigger features. The overall improvement is 2.21% in terms of F-measure. Another experiment is done to compare ME with ME + MI model trained by five month corpora. The result is exhibited in Figure 6.

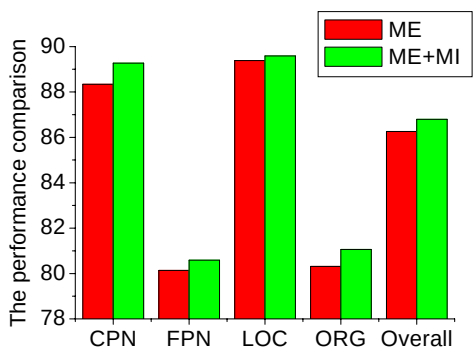

Figure 6. The comparison about MI in F-measure

The effectiveness of the proposed method has been confirmed. A similar result is also achieved for the IG approach. Experimental results show that the new method is more efficient.

In the last part of this section, the authors evaluate the word cluster performance. The word vectors method is performed in the large-scale corpora, in the 1998 and 2000 People's Daily Newspaper, the window of size k =8 being used in this experiment.

The hierarchical cluster analysis or other cluster analysis methods can be used to obtain the word cluster result. Table 15 demonstrates the proximity matrix, and Figure 7 gives its corresponding hierarchical cluster result. The authors used a synonym dictionary "Word Forest of The Synonym" to reduce the cluster space and increase prior knowledge. For instance, there are about 63 synonyms to the word "教授" (professor).

Though it is helpful to build the word classes for the NER task by combining the word cluster and the thesaurus, some manual correction is also needed, because the linguistic phenomenon is too complicated, therefore making it impossible to acquire all the perfect word

classes by only making statistical analysis of some corpora.

Table 15. The proximity matrix [4]

Case	Cosine of Vectors of Values					
	学生	教授	副教授	导师	大学生	中学生
学生	1.000	.352	.280	.288	.433	.331
教授	.352	1.000	.722	.815	.310	.174
副教授	.280	.722	1.000	.641	.216	.136
导师	.288	.815	.641	1.000	.226	.139
大学生	.433	.310	.216	.226	1.000	.674
中学生	.331	.174	.136	.139	.674	1.000

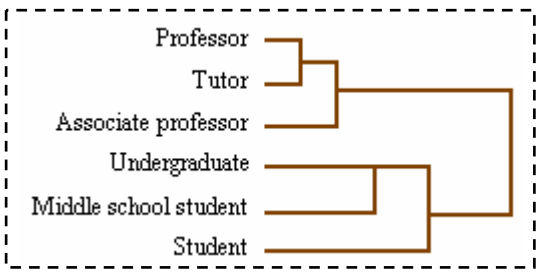

Figure 7. The demonstration about hierarchical cluster

Based on the analysis of the errors, one finds that the sparse data problem is the main problem [Jiang 2006A; Jiang 2007]. In this paper, the authors apply the Smoothing Algorithm, Word Cluster Method, etc. to overcome the sparse data problem.

5. Conclusion

A pragmatic Chinese word segmentation approach having balance between the precision, efficiency and model complication is described in this paper. The disambiguation and out-of-vocabulary detection are the two main difficulties found in the Word Segmentation task. Accordingly, a lot of work is done in order to improve the performance of the above two problems. The contributions of this research are:

1) Apply multiple models to build a word segmentation model, and a special sub-task can be effectively solved via an optimized language model.

2) The authors propose to apply Average Mutual Information, etc. to extract stable entity features, and also present a novel method to provide an auxiliary function in extending the

[4] 学生 student, 教授 professor, 副教授 associate professor, 导师 tutor, 大学生 undergraduate, 中学生 middle school student.

features by combining the word cluster and the thesaurus.

3) Rough Set theory is present to extract the complicated features and the long distance features for the segmentation disambiguation and for the Named Entity Recognition.

The work in the future will concentrate on two sides: improving the NER performance and exploring New Word Detection Algorithm.

Acknowledgements

The authors thank Dr. Yan Zhao and Dr. Jian Zhao for their valuable suggestions in the proposed system. The authors also thank the members of the Natural Language Computing Group at School of Computer Science and Technology of the Harbin Institute of Technology. The authors especially thank the anonymous reviewers for their insightful comments and suggestions, based on which the paper has been improved.

References

Chen, S. F., and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, 13(4) 1999, pp. 359-394.

Cheng, K.-S., G. Young, K.-F Wong, "A study on word–based and integral-bit Chinese text compression algorithms," *Journal of the American Society for Information Science*, 50(3) 1999, pp. 218-228.

Gao, J.-F., A.-D. Wu, M. Li, and C.-N. Huang, "Chinese word segmentation and named entity recognition: a pragmatic approach in Computational Linguistics," *Computational Linguistics*, 31(4) 2005, pp.531-574.

Gao, J.-F., M. Li, A.-D. Wu, and C.-N. Huang, "Chinese Word Segmentation: A Pragmatic Approach," *Microsoft Research,Technical Report: MSR-TR-2004-123*, November 2004.

Gao, J.-F., J. Goodman, M. Li, and K.-F. Lee, "Toward a unified approach to statistical language modeling for Chinese," *ACMTrans, Asian Language Information Process*, 1(1) 2002, pp. 3-33.

Gao, J.-F., M. Li, and C.-N. Huang, "Improved source-channel model for Chinese word segmentation," *In the 41nd Annual Meeting of the Association for Computational Linguistics*, 2003, Sapporo, Japan, pp. 272-279.

Hockenmaier, J., and C. Brew, "Error-driven Learning of Chinese word segmentation," *In the 12th Pacific Conference on Language and Information*, 1998, Singapore, pp. 218-229.

Jiang, W., X.-L. Wang, Y. Guan, and J. Zhao, "Research on Chinese Lexical Analysis System by Fusing Multiple Knowledge Sources," *Chinese Journal of Computer*, January, 2007.

Jiang, W., X.-L. Wang, Y. Guan, and G.-H. Liang, "Applying Rough Sets in Word Segmentation Disambiguation Based on Maximum Entropy Model," *Journal of Harbin Institute of Technology (New Series)*, 13(1) 2006A, pp. 94-98.

Jiang, W., J. Zhao, Y. Guan, and Z.-M. Xu, "Chinese Word Segmentation based on Mixing Model," *In The 4th SIGHAN Workshop*, 2005, Jeju Island, Korea, pp. 180-182.

Jiang, W., Y. Guan, and X.-L. Wang, "A Pragmatic Chinese Word Segmentation System," *In proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 2006B, Sydney, pp. 189–192.

Jiang, W., Y. Guan, and X.-L. Wang, "Improving Feature extraction in Named Entity Recognition based on Maximum Entropy Model," *In the 2006 International Conference on Machine Learning and Cybernetics (ICMLC2006)*, 2006C, China, pp. 2630-2635.

Jiang, W., Y. Guan, and X.-L. Wang, "An Improved Unknown Word Recognition Model based on Multi-Knowledge Source Method," *In the 6th International Conference on Intelligent Systems Design and Applications (ISDA'06)*, vol 2, 2006D, China, pp. 825-830

Liang, N.-Y., "automatic word segmentation in written Chinese and an auto match word segmentation system-CDWS," (in Chinese) *Journal of Chinese information processing*, 1(2), 1987, pp. 44-52.

Palmer, D., "A trainable rule-based algorithm to word segmentation," *In proceedings of the 35th Annual Meeting of the Association of Computational Linguistics*, 1997, Madrid, Spain, pp. 321-328.

Peng, F.-C., F.-F Feng, and A. McCallum, "Chinese segmentation and new word detection using conditional random fields," *In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, 2004, Geneva, Switzerland, pp. 562-568.

Peng, F.C., and D. Schuurmans, "A hierarchical EM approach to word segmentation," *In 6th Natural Language Processing Pacific Rim Symposium (NLPRS-2001)*, 2001, pp. 475-480.

Schutze, H.. "Automatic word sense discrimination," *Computational Linguistics*, 24(1) 1998, pp. 97-123.

Sproat, R. C. Shih, G. William, and N. Chang, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Computational linguistics*, 22(3) 1996, pp. 377-404.

Wu, A.-D., and Z.-X. Jiang, "Word Segmentation in Sentence Analysis," *In 1998 International Conference on Chinese Information Processing*, 1998, Beijing, China, pp. 169-180.

Xue, N.-W., and L.-B. Shen, "Chinese Word Segmentation as LMR Tagging," *In the Second SIGHAN Workshop on Chinese Language Processing*, 2003, Japan, pp. 176-179.

Zhang, H.-P., Q. Liu, X.-Q. Cheng, H. Zhang, and H.-K. Yu, "Chinese Lexical Analysis Using Hierarchical Hidden Markov Model," In the *Second SIGHAN workshop affiliated with 4th ACL*, 2003, Sapporo Japan, pp. 63-70.

Zhao, J., "Research on Conditional Probabilistic Model and Its Application in Chinese Named Entity Recognition," PhD thesis, *Harbin Institute of Technology, China*, 2006.

Zhao, Y., "Research on Chinese Morpheme Analysis Based on Statistic Language Model," PhD thesis, *Harbin Institute of Technology, China*, 2005A.

Zhao, Y., X.-L. Wang, B.-Q. Liu, and Y. Guan, "Solution Strategies for Word Sense Problems Based On Vector Space Model and Maximum Entropy Model," (In Chinese), *Chinese High Technology Letters*, 15(1) 2005B, pp. 1-6.