A Chunking Strategy Towards Unknown Word Detection in Chinese Word Segmentation

Zhou GuoDong

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
zhougd@i2r.a-star.edu.sg

Abstract. This paper proposes a chunking strategy to detect unknown words in Chinese word segmentation. First, a raw sentence is pre-segmented into a sequence of word atoms[1] using a maximum matching algorithm. Then a chunking model is applied to detect unknown words by chunking one or more word atoms together according to the word formation patterns of the word atoms. In this paper, a discriminative Markov model, named Mutual Information Independence Model (MIIM), is adopted in chunking. Besides, a maximum entropy model is applied to integrate various types of contexts and resolve the data sparseness problem in MIIM. Moreover, an error-driven learning approach is proposed to learn useful contexts in the maximum entropy model. In this way, the number of contexts in the maximum entropy model can be significantly reduced without performance decrease. This makes it possible for further improving the performance by considering more various types of contexts. Evaluation on the PK and CTB corpora in the First SIGHAN Chinese word segmentation bakeoff shows that our chunking approach successfully detects about 80% of unknown words on both of the corpora and outperforms the best-reported systems by 8.1% and 7.1% in unknown word detection on them respectively.

1 Introduction

Prior to any linguistic analysis of Chinese text, Chinese word segmentation is the necessary first step and one of major bottlenecks in Chinese information processing since a Chinese sentence is written in a continuous string of characters without obvious separators (such as blanks) between the words. During the past two decades, this research has been a hot topic in Chinese information processing [1-10].

There exist two major problems in Chinese word segmentation: ambiguity resolution and unknown word detection. While n-gram modeling and/or word co-occurrence has been successfully applied to deal with the ambiguity problems [3, 5, 10, 12, 13], unknown word detection has become the major bottleneck in Chinese

[1] In this paper, word atoms refer to basic building units in words. For example, the word "计算机" (computer) consists of two word atoms: "计算"(computing) and "机"(machine). Generally, word atoms can either occur independently, e.g. "计算"(computing), or only become a part of a word, e.g. "机"(machine) in the word "计算机" (computer).

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 530 – 541, 2005.

word segmentation. Currently, almost all Chinese word segmentation systems rely on a word dictionary. The problem is that when the words stored in the dictionary are insufficient, the system's performance will be greatly deteriorated by the presence of words that are unknown to the system. Moreover, manual maintenance of a dictionary is very tedious and time consuming. It is therefore important for a Chinese word segmentation system to identify unknown words from the text automatically.

In literature, two categories of competing approaches are widely used to detect unknown words[2]: statistical approaches [5, 11, 12, 13, 14, 15] and rule-based approaches [5, 11, 14, 15]. Although rule-based approaches have the advantage of being simple, the complexity and domain dependency of how the unknown words are produced greatly reduce the efficiency of these approaches. On the other hand, statistical approaches have the advantage of being domain-independent [16]. It is interesting to note that many systems apply a hybrid approach [5, 11, 14, 15]. Regardless of the choice of different approaches, finding a way to automatically detect unknown words has become a crucial issue in Chinese word segmentation and Chinese information processing in general.

Input raw sentence:	张杰毕业自交通大学.					
MMA pre-segmentation:	张 杰	毕业	自	交通	大学	.
Unknown word detection:	张杰	毕业	自	交通大学		.
	Zhang Jie	graduate	from	JiaoTong University.		

Fig. 1. MMA and unknown word detection by chunking: an example

This paper proposes a chunking strategy to cope with unknown words in Chinese word segmentation. First, a raw sentence is pre-segmented into a sequence of word atoms (i.e. single-character words and multi-character words) using a maximum matching algorithm (MMA)[3]. Then a chunking model is applied to detect unknown words by chunking one or more word atoms together according to the word formation patterns of the word atoms. Figure 1 gives an example. Here, the problem of unknown word detection is re-cast as chunking one or more word atoms together to form a new word and a discriminative Markov model, named Mutual Information Independence Model (MIIM), is adopted in chunking. Besides, a maximum entropy model is applied to integrate various types of contexts and resolve the data sparseness problem in MIIM. Moreover, an error-driven learning approach is proposed to learn useful

[2] Some systems [13,14] focus on proper names due to their importance in Chinese information processing.

[3] A typical MMA identifies all character sequences which are found in the word dictionary and marks them as words. Those character sequences, which can be segmented in more than one way, are marked as ambiguous and a word unigram model is applied to choose the most likely segmentation sequence. The remaining sequences, i.e. those not found in the dictionary, are called fragments and segmented into single characters. In this way, each Chinese sentence is pre-segmented into a sequence of single-character words and multi-character words. For convenience, we call these single-character words and multi-character words in the output of the MMA algorithm as word atoms.

contexts in the maximum entropy model. In this way, the number of contexts in the maximum entropy model can be significantly reduced without performance decrease. This makes it possible for further improving the performance by considering more various types of contexts in the future. Evaluation on the PK and CTB corpora in the First SIGHAN Chinese word segmentation bakeoff shows that our chunking strategy performs best in unknown word detection on both of the corpora.

The rest of the paper is as follows: In Section 2, we will discuss in details about our chunking strategy in unknown word detection. Experimental results are given in Section 3. Finally, some remarks and conclusions are made in Section 4.

2 Unknown Word Detection by Chunking

In this section, we will first describe the chunking strategy in unknown word detection of Chinese word segmentation using a discriminative Markov model, called Mutual Information Independence Model (MIIM). Then a maximum entropy model is applied to integrate various types of contexts and resolve the data sparseness problem in MIIM. Finally, an error-driven learning approach is proposed to select useful contexts and reduce the context feature vector dimension.

2.1 Mutual Information Independence Model and Unknown Word Detection

Mutual Information Independence Model

In this paper, we use a discriminative Markov model, called Mutual Information Independence Model (MIIM) proposed by Zhou et al [17][4], in unknown word detection by chunking. MIIM is derived from a conditional probability model. Given an observation sequence $O_1^n = o_1 o_2 \cdots o_n$, the goal of a conditional probability model is to find a stochastic optimal state(tag) sequence $S_1^n = s_1 s_2 \cdots s_n$ that maximizes:

$$\log P(S_1^n \mid O_1^n) = \log P(S_1^n) + \log \frac{P(S_1^n, O_1^n)}{P(S_1^n) \cdot P(O_1^n)} \tag{1}$$

The second term in Equation (1) is the pair-wise mutual information (PMI) between S_1^n and O_1^n. In order to simplify the computation of this term, we assume a pair-wise mutual information independence (2):

$$PMI(S_1^n, O_1^n) = \sum_{i=1}^{n} PMI(s_i, O_1^n) \quad \text{or}$$

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n) \cdot P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(s_i, O_1^n)}{P(s_i) \cdot P(O_1^n)} \tag{2}$$

[4] We have renamed the discriminative Markov model in [17] as the Mutual Information Independence Model according to the novel pair-wise mutual information independence assumption in the model. Another reason is to distinguish it from the traditional Hidden Markov Model [18] and avoid misleading.

That is, an individual state is only dependent on the observation sequence O_1^n and independent on other states in the state sequence S_1^n. This assumption is reasonable because the dependence among the states in the state sequence S_1^n has already been captured by the first term in Equation (1). Applying Equation (2) to Equation (1), we have Equation (3)[5]:

$$\log P(S_1^n \mid O_1^n) = \sum_{i=2}^{n} PMI(s_i, S_1^{i-1}) + \sum_{i=1}^{n} \log P(s_i \mid O_1^n) \tag{3}$$

We call the above model as shown in Equation (3) the Mutual Information Independence Model due to its pair-wise mutual information assumption as shown in Equation (2). The above model consists of two sub-models: the state transition model $\sum_{i=2}^{n} PMI(s_i, S_1^{i-1})$ as the first term in Equation (3) and the output model $\sum_{i=1}^{n} \log P(s_i \mid O_1^n)$ as the second term in Equation (3). Here, a variant of the Viterbi algorithm [19] in decoding the standard Hidden Markov Model (HMM) [18] is implemented to find the most likely state sequence by replacing the state transition model and the output model of the standard HMM with the state transition model and the output model of the MIIM, respectively.

Unknown Word Detection

For unknown word detection by chunking, a word (known word or unknown word) is regarded as a chunk of one or more word atoms and we have:

- $o_i =< p_i, w_i >$; w_i is the $i-th$ word atom in the sequence of word atoms $W_1^n = w_1 w_2 \cdots w_n$; p_i is the word formation pattern of the word atom w_i. Here p_i measures the word formation power of the word atom w_i and consists of:
  - o The percentage of w_i occurring as a whole word (round to 10%)
  - o The percentage of w_i occurring at the beginning of other words (round to 10%)
  - o The percentage of w_i occurring at the end of other words (round to 10%)
  - o The length of w_i
  - o The occurring frequency feature of w_i , which is mapped to max(log(Frequency), 9).
- s_i: the states are used to bracket and differentiate various types of words. In this way, Chinese unknown word detection can be regarded as a bracketing process while differentiation of different word types can help the bracketing process. s_i is structural and consists of three parts:

[5] Details about the derivation are omitted due to space limitation. Please see [17] for more.

- **Boundary Category (B):** it includes four values: {O, B, M, E}, where O means that current word atom is a whOle word and B/M/E means that current word atom is at the Beginning/in the Middle/at the End of a word.
- **Word Category (W):** It is used to denote the class of the word. In our system, words are classified into two types: pure Chinese word type and mixed word type (i.e. including English characters and Chinese digits/numbers/symbols).
- **Word Atom Formation Pattern (P):** Because of the limited number of boundary and word categories, the word atom formation pattern described above is added into the structural state to represent a more accurate state transition model in MIIM while keeping its output model.

Problem with Unknown Word Detection Using MIIM

From Equation (3), we can see that the state transition model of MIIM can be computed by using ngram modeling [20, 21, 22], where each tag is assumed to be dependent on the N-1 previous tags (e.g. 2). The problem with the above MIIM lies in the data sparseness problem raised by its output model: $\sum_{i=1}^{n} \log P(s_i \mid O_1^n)$. Ideally, we would have sufficient training data for every event whose conditional probability we wish to calculate. Unfortunately, there is rarely enough training data to compute accurate probabilities when decoding on new data. Generally, two smoothing approaches [21, 22, 23] are applied to resolve this problem: linear interpolation and back-off. However, these two approaches only work well when the number of different information sources is very limited. When a few features and/or a long context are considered, the number of different information sources is exponential. This makes smoothing approaches inappropriate in our system. In this paper, the maximum entropy model [24] is proposed to integrate various context information sources and resolve the data sparseness problem in our system. The reason that we choose the maximum entropy model for this purpose is that it represents the state-of–the-art in the machine learning research community and there are good implementations of the algorithm available. Here, we use the open NLP maximum entropy package[6] in our system.

2.2 Maximum Entropy

The maximum entropy model is a probability distribution estimation technique widely used in recent years for natural language processing tasks. The principle of the maximum entropy model in estimating probabilities is to include as much information as is known from the data while making no additional assumptions. The maximum entropy model returns the probability distribution that satisfies the above property with the highest entropy. Formally, the decision function of the maximum entropy model can be represented as:

$$P(o,h) = \frac{1}{Z(h)} \prod_{j=1}^{k} \alpha_j^{f_j(h,o)} \tag{4}$$

[6] http://maxent.sourceforge.net

where *o* is the outcome, *h* is the history (context feature vector in this paper), *Z(h)* is a normalization function, {$f_1, f_2, ..., f_k$} are feature functions and {$\alpha_1, \alpha_2, \ldots, \alpha_k$} are the model parameters. Each model parameter corresponds to exactly one feature and can be viewed as a "weight" for that feature. All features used in the maximum entropy model are binary, e.g.

$$f_j(h,o) = \begin{cases} 1, & if \quad o = Independen\ tWord\ , \qquad CurrentWor\ dAtom\ = \text{我们}(we); \\ 0, & \quad otherwise\ . \end{cases} \tag{5}$$

In order to reliably estimate $P(s_i \mid O_1^n)$ in the output model of MIIM using the maximum entropy model, various context information sources are included in the context feature vector:

- p_i: current word atom formation pattern
- $p_{i-1}p_i$: previous word atom formation pattern and current word atom formation pattern
- $p_i p_{i+1}$: current word atom formation pattern and next word atom formation pattern
- $p_i w_i$: current word atom formation pattern and current word atom
- $p_{i-1}w_{i-1}p_i$: previous word atom formation pattern, previous word atom and current word atom formation pattern
- $p_i p_{i+1} w_{i+1}$: current word atom formation pattern, next word atom formation pattern and next word atom
- $p_{i-1}p_i w_i$: previous word atom formation pattern, current word atom formation pattern and current word atom
- $p_i w_i p_{i+1}$: current word atom formation pattern, current word atom and next word atom formation pattern
- $p_{i-1}w_{i-1}p_i w_i$: previous word atom formation pattern, previous word atom, current word atom formation pattern and current word atom
- $p_i w_i p_{i+1} w_{i+1}$: current word atom formation pattern, current word atom, next word atom formation pattern and next word atom

However, there exists a problem when we include above various context information in the maximum entropy model: the context feature vector dimension easily becomes too large for the model to handle. One easy solution to this problem is to only keep those frequently occurring contexts in the model. Although this frequency filtering approach is simple, many useful contexts may not occur frequently and be filtered out while those kept may not be useful. To resolve this problem, we propose an alternative error-driven learning approach to only keep useful contexts in the model.

2.3 Context Feature Selection Using Error-Driven Learning

Here, we propose an error-driven learning approach to examine the effectiveness of various contexts and select useful contexts to reduce the size of the context feature vector used in the maximum entropy model for estimating $P(s_i \mid O_1^n)$ in the output model of MIIM. This makes it possible to further improve the performance by incorporating more various types of contexts in the future.

Assume Φ is the container for useful contexts. Given a set of existing useful contexts Φ and a set of new contexts $\Delta\Phi$, the effectiveness of a new context $C_i \in \Delta\Phi$, $E(\Phi, C_i)$, is measured by the C_i-related reduction in errors which results from adding the new context set $\Delta\Phi$ to the useful context set Φ :

$$E(\Phi, C_i) = \#Error(\Phi, C_i) - \#Error(\Phi + \Delta\Phi, C_i) \tag{6}$$

Here, $\#Error(\Phi, C_i)$ is the number of C_i-related chunking errors before $\Delta\Phi$ is added to Φ and $\#Error(\Phi + \Delta\Phi, C_i)$ is the number of C_i-related chunking errors after $\Delta\Phi$ is added to Φ. That is, $E(\Phi, C_i)$ is the number of the chunking error corrections made on the context $C_i \in \Delta\Phi$ when $\Delta\Phi$ is added to Φ. If $E(\Phi, C_i) > 0$, we declare that the new context C_i is a useful context and should be added to Φ. Otherwise, the new context C_i is considered useless and discarded.

Given the above error-driven learning approach, we initialize $\Phi = \{p_i\}$ (i.e. we assume all the current word atom formation patterns are useful contexts) and choose one of the other context types as the new context set $\Delta\Phi$, e.g. $\Phi = \{p_i w_i\}$. Then, we can train two MIIMs with different output models using Φ and $\Phi + \Delta\Phi$ respectively. Moreover, useful contexts are learnt on the training data in a two-fold way. For each fold, two MIIMs are trained on 50% of the training data and for each new context C_i in $\Delta\Phi$, evaluate its effectiveness $E(\Phi, C_i)$ on the remaining 50% of the training data according to the context effectiveness measure as shown in Equation (6). If $E(\Phi, C_i) > 0$, C_i is marked as a useful context and added to Φ. In this way, all the useful contexts in $\Delta\Phi$ are incorporated into the useful context set Φ. Similarly, we can include useful contexts of other context types into the useful context set Φ one by one. In this paper, various types of contexts are learnt one by one in the exact same order as shown in Section 2.2. Finally, since different types of contexts may have cross-effects, the above process is iterated with the renewed useful context set Φ until very few useful contexts can be found at each loop. Our experiments show that iteration converges within four loops.

3 Experimental Results

All of our experiments are evaluated on the PK and CTB benchmark corpora used in the First SIGHAN Chinese word segmentation bakeoff[7] with the closed configuration. That is, only the training data from the particular corpus is used during training. For unknown word detection, the chunking training data is derived by using the same Maximum Matching Algorithm (MMA) to segment each word in the original training data as a chunk of word atoms. This is done in a two-fold way. For each fold, the

[7] http://www.sighan.org/bakeoff2003/

MMA is trained on 50% of the original training data and then used to segment the remaining 50% of the original training data. Then the MIIM is used to train a chunking model for unknown word detection on the chunking training data. Table 1 shows the details of the two corpora. Here, OOV is defined as the percentage of words in the test corpus not occurring in the training corpus and indicates the out-of-vocabulary rate in the test corpus.

Table 1. Statistics of the corpora used in our evaluation

Corpus	Abbreviation	OOV	Training Data	Test Data
Beijing University	PK	6.9%	1100K words	17K words
UPENN Chinese Treebank	CTB	18.1%	250K words	40K words

Table 2 shows the detailed performance of our system in unknown word detection and Chinese word segmentation as a whole using the standard scoring script[8] on the test data. In this and subsequent tables, various evaluation measures are provided: precision (P), recall (R), F-measure, recall on out-of-vocabulary words (R_{OOV}) and recall on in-vocabulary words (R_{IV}). It shows that our system achieves precision/recall/F-measure of 93.5%/96.1%/94.8 and 90.5%/90.1%/90.3 on the PK and CTB corpora respectively. Especially, our chunking approach can successfully detect 80.5% and 77.6% of unknown words on the PK and CTB corpora respectively.

Table 2. Detailed performance of our system on the 1st SIGHAN Chinese word segmentation benchmark data

Corpus	P	R	F	R_{OOV}	R_{IV}
PK	93.5	96.1	94.8	80.5	97.3
CTB	90.5	90.1	90.3	77.6	92.9

Table 3 and Table 4 compare our system with other best-reported systems on the PK and CTB corpora respectively. Table 3 shows that our chunking approach in unknown word detection outperforms others by more than 8% on the PK corpus. It also shows that our system performs comparably with the best reported systems on the PK corpus when the out-of-vocabulary rate is moderate(6.9%). Our performance in Chinese word segmentation as a whole is somewhat pulled down by the lower performance in recalling in-vocabulary words. This may be due to the preference of our chunking strategy in detecting unknown words by wrongly combining some of in-vocabulary words into unknown words. Such preference may cause negative effect in Chinese word segmentation as a whole when the gain in unknown word detection fails to compensate the loss in wrongly combining some of in-vocabulary words into unknown words. This happens when the out-of-vocabulary rate is not high, e.g. on the

[8] http://www.sighan.org/bakeoff2003/score

PK corpus. Table 4 shows that our chunking approach in unknown word detection outperforms others by more than 7% on the CTB corpus. It also shows that our system outperforms the other best-reported systems by more than 2% in Chinese word segmentation as a whole on the CTB corpus. This is largely due to the huge gain in unknown word detection when the out-of-vocabulary rate is high (e.g. 18.1% in the CTB corpus), even though our system performs worse on recalling in-vocabulary words than others. Evaluation on both the PK and CTB corpora shows that our chunking approach can successfully detect about 80% of unknown words on corpora with a large range of the out-of-vocabulary rates. This suggests the powerfulness of using various word formation patterns of word atoms in detecting unknown words. This also demonstrates the effectiveness and robustness of our chunking approach in unknown word detection of Chinese word segmentation and its portability to different genres.

Table 3. Comparison of our system with other best-reported systems on the PK corpus

Corpus	P	R	F	R_{OOV}	R_{IV}
Ours	93.5	96.1	94.8	80.5	97.3
Zhang et al [25]	94.0	96.2	95.1	72.4	97.9
Wu [26]	93.8	95.5	94.7	68.0	97.6
Chen [27]	93.8	95.5	94.6	64.7	97.7

Table 4. Comparison of our system with other best-reported systems on the CTB corpus

Corpus	P	R	F	R_{OOV}	R_{IV}
Ours	90.5	90.1	90.3	77.6	92.9
Zhang et al [25]	87.5	88.6	88.1	70.5	92.7
Duan et al [28]	85.6	89.2	87.4	64.4	94.7

Finally, Table 5 and Table 6 compare our error-driven learning approach with the frequency filtering approach in learning useful contexts for the output model of MIIM on the PK and CTB corpora respectively. Due to memory limitation, at most 400K useful contexts are considered in the frequency filtering approach. First, they show that the error-driven learning approach is much more effective than the simple frequency filtering approach. With the same number of useful contexts, the error-driven learning approach outperforms the frequency filtering approach by 7.8%/0.6% and 5.5%/0.8% in R_{OOV} (unknown word detection)/F-measure(Chinese word segmentation as a whole) on the PK and CTB corpora respectively. Moreover, the error-driven learning approach slightly outperforms the frequency filtering approach with the best configuration of 2.5 and 3.5 times of useful contexts. Second, they show that increasing the number of frequently occurring contexts using the frequency filtering approach may not increase the performance. This may be due to that some of frequently occurring contexts are noisy or useless and including them may have negative effect. Third, they show that the error-driven learning approach is effective in learning useful contexts by reducing 96-98% of possible contexts. Finally, the figures inside parentheses show the number of useful patterns shared between the error-driven learning approach and the frequency filtering approach. They show that about 40-50% of useful contexts selected using the error-driven learning approach do not occur frequently in the useful contexts selected using the frequency filtering approach.

Table 5. Comparison of the error-driven learning approach with the frequency filtering approach in learning useful contexts for the output model of MIIM on the PK corpus (Total number of possible contexts: 4836K)

Approach	#useful contexts	F	R_{OOV}	R_{IV}
Error-Driven Learning	98K	94.8	80.5	97.3
Frequency Filtering	98K (63K)	94.2	72.7	97.4
Frequency Filtering (best performance)	250K (90K)	94.7	80.2	97.3
Frequency Filtering	400K (94K)	94.6	79.1	97.1

Table 6. Comparison of the error-driven learning approach with the frequency filtering approach in learning useful contexts for the output model of MIIM on the CTB corpus (Total number of possible contexts: 1038K)

Approach	#useful contexts	F	R_{OOV}	R_{IV}
Error-Driven Learning	43K	90.3	77.6	92.9
Frequency Filtering	43K (21K)	89.5	72.1	92.8
Frequency Filtering (best performance)	150K	90.1	76.1	93.0
Frequency Filtering	400K (40K)	89.9	75.8	92.9

4 Conclusion

In this paper, a chunking strategy is presented to detect unknown words in Chinese word segmentation by chunking one or more word atoms together according to the various word formation patterns of the word atoms. Besides, a maximum entropy model is applied to integrate various types of contexts and resolve the data sparseness problem in our strategy. Finally, an error-driven learning approach is proposed to learn useful contexts in the maximum entropy model. In this way, the number of contexts in the maximum entropy model can be significantly reduced without performance decrease. This makes it possible for further improving the performance by considering more various types of contexts. Evaluation on the PK and CTB corpora in the First SIGHAN Chinese word segmentation bakeoff shows that our chunking strategy can detect about 80% of unknown words on both of the corpora and outperforms the best-reported systems by 8.1% and 7.1% in unknown word detection on them respectively. While our Chinese word segmentation system with chunking-based unknown word detection performs comparably with the best systems on the PK corpus when the out-of-vocabulary rate is moderate(6.9%), our system significantly outperforms others by more than 2% when the out-of-vocabulary rate is high(18.1%). This demonstrates the effectiveness and robustness of our chunking strategy in unknown word detection of Chinese word segmentation and its portability to different genres.

References

1. Jie CY, Liu Y and Liang NY. (1989). On methods of Chinese automatic segmentation, *Journal of Chinese Information Processing,* 3(1):1-9.
2. Li KC, Liu KY and Zhang YK. (1988). Segmenting Chinese word and processing different meanings structure, *Journal of Chinese Information Processing,* 2(3):27-33.
3. Liang NY, (1990). The knowledge of Chinese word segmentation*, Journal of Chinese Information Processing,* 4(2):29-33.
4. Lua KT, (1990). From character to word - An application of information theory, *Computer Processing of Chinese & Oriental Languages,* 4(4):304-313.
5. Lua KT and Gan GW. (1994). An application of information theory in Chinese word segmentation. *Computer Processing of Chinese & Oriental Languages*, 8(1):115-124.
6. Wang YC, SU HJ and Mo Y. (1990). Automatic processing of Chinese words. *Journal of Chinese Information Processing*. 4(4):1-11.
7. Wu JM and Tseng G. (1993). Chinese text segmentation for text retrieval: achievements and problems. *Journal of the American Society for Information Science*. 44(9):532-542.
8. Xu H, He KK and Sun B. (1991) The implementation of a written Chinese automatic segmentation expert system, *Journal of Chinese Information Processing,* 5(3):38-47.
9. Yao TS, Zhang GP and Wu YM. (1990). A rule-based Chinese automatic segmentation system, *Journal of Chinese Information Processing,* 4(1):37-43.
10. Yeh CL and Lee HJ. (1995). Rule-based word identification for Mandarin Chinese sentences - A unification approach, *Computer Processing of Chinese & Oriental Languages,* 9(2):97-118.
11. Nie JY, Jin WY and Marie-Louise Hannan. (1997). A hybrid approach to unknown word detection and segmentation of Chinese, *Chinese Processing of Chinese and Oriental Languages*, 11(4): pp326-335.
12. Tung CH and Lee HJ. (1994). Identification of unknown word from a corpus, *computer Processing of Chinese & Oriental Languages,* 8(Supplement):131-146.
13. Chang JS et al. (1994). A multi-corpus approach to recognition of proper names in Chinese Text, *Computer Processing of Chinese & Oriental Languages,* 8(1):75-86
14. Sun MS, Huang CN, Gao HY and Fang J. (1994). Identifying Chinese Names In Unrestricted Texts, *Communications of Chinese and Oriental Languages Information Processing Society*, 4(2):113-122.
15. Zhou GD and Lua KT, (1997). Detection of Unknown Chinese Words Using a Hybrid Approach, *Computer Processing of Chinese & Oriental Language*, 11(1):63-75.
16. Eugene Charniak, *Statistical language learning,* The MIT Press, ISBN 0-262-03216-3
17. Zhou GDong and Su J. (2002). Named Entity Recognition Using a HMM-based Chunk Tagger, *Proceedings of the Conference on Annual Meeting for Computational Linguistics (ACL'2002)*. 473-480, Philadelphia.

18. Rabiner L. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *IEEE 77(*2), pages257-285.
19. Viterbi A.J. 1967. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory,* IT 13(2), 260-269.
20. Gale W.A. and Sampson G. 1995. Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*. 2:217-237.
21. Jelinek F. (1989). Self-Organized Language Modeling for Speech Recognition. In Alex Waibel and Kai-Fu Lee(Editors). *Readings in Speech Recognitiopn*. Morgan Kaufmann. 450-506.
22. Katz S.M. (1987). Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics. Speech and Signal Processing.* 35: 400-401.
23. Chen and Goodman. (1996). An Empirical Study of Smoothing Technniques for Language Modeling. In *Proceedings of the 34th Annual Meeting of the Association of Computational Linguistics (ACL'1996).* pp310-318. Santa Cruz, California, USA.
24. Ratnaparkhi A. (1996). A Maximum Entropy Model for Part-of-Speech Tagging. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*, 133-142.
25. Zhang HP, Yu HK, Xiong DY and Liu Q. (2003). HHMM-based Chinese Lexical Analyzer ICTCLAS. *Proceedings of 2nd SIGHAN Workshop on Chinese Language Processing.* 184-187. Sapporo, Japan.
26. Wu AD. (2003). Chinese Word Segmentation in MSR-NLP. *Proceedings of 2nd SIGHAN Workshop on Chinese Language Processing.* 172-175. Sapporo, Japan.
27. Chen AT. (2003). Chinese Word Segmentation Using Minimal Linguistic Knowledge. *Proceedings of 2nd SIGHAN Workshop on Chinese Language Processing.* 148-151. Sapporo, Japan.
28. Duan HM, Bai XJ, Chang BB and Yu SW. (2003). Chinese Word Segmentation at Peking University. *Proceedings of 2nd SIGHAN Workshop on Chinese Language Processing.* 152-155. Sapporo, Japan.