Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005

CHINESE TEXT CHUNKING USING LEXICALIZED HMMS

GUO-HONG FU[1], RUI-FENG XU[2], KANG-KWONG LUKE[1], QIN LU[2]

[1]Department of Linguistics, The University of Hong Kong, Hong Kong SAR, China
[2]Department of Computing, Hong Kong Polytechnic University, Hong Kong SAR, China
E-MAIL: ghfu@hkucc.hku.hk, csrfxu@comp.ployu.edu.hk, kkluke@hkuusa.hku.hk, csluqin@comp.ployu.edu.hk

Abstract:

This paper presents a lexicalized HMM-based approach to Chinese text chunking. To tackle the problem of unknown words, we formalize Chinese text chunking as a tagging task on a sequence of known words. To do this, we employ the uniformly lexicalized HMMs and develop a lattice-based tagger to assign each known word a proper hybrid tag, which involves four types of information: word boundary, POS, chunk boundary and chunk type. In comparison with most previous approaches, our approach is able to integrate different features such as part-of-speech information, chunk-internal cues and contextual information for text chunking under the framework of HMMs. As a result, the performance of the system can be improved without losing its efficiency in training and tagging. Our preliminary experiments on the PolyU Shallow Treebank show that the use of lexicalization technique can substantially improve the performance of a HMM-based chunking system.

Keywords:

Text chunking; base phrase recognition; base phrase structure; lexicalized hidden Markov models (HMMs)

1. Introduction

In general, text chunking consists of identifying non-recursive phrase structures from a sequence of tokens and classifying them into some syntactic categories like base noun phrases (baseNPs) and base verb phrases (baseVPs). As an intermediate step towards full parsing, text chunking has been attracting more and more attention in the NLP community. It is recognized as an important sub-task of many large NLP applications such as machine translation, text mining and question answering. It was also a shared task of the CoNLL-2000 [1].

Current research on text chunking has focused on machine learning approaches, including hidden Markov models (HMMs) [2], transformation-based error-driven learning (TBL) [3][4], maximum entropy (ME) [5], memory-based learning (MBL) [4][6], and support vector machines (SVMs) [7]. In comparison with rule-based methods, machine-learning approaches are more adaptive and robust. However, it is still a challenge for most of them to keep a balance between capacity and computational cost [8]. While a HMM-based tagger has proven to be very speedy in training and processing [9], it usually achieves relatively lower tagging accuracy for it only takes into account contextual category information, and ignores contextual lexical information, which sometimes gives strong evidence for chunking. On the contrary, some learning methods such as ME and SVMs are capable of combining much richer lexical information in a straightforward way. However, they usually need much more time in training and tagging, which will become a serious problem in processing a large amount of data or in some on-line applications such as information retrieval and online question-answering. In order to address these problems, some recent work suggests the utilization of lexicalization techniques to enhance the standard HMMs [9][10][11]. Their experiments have demonstrated that their systems could be improved without increasing much computational cost in training and processing.

Although much progress has been made in the literature, it is still a challenge to develop a practical chunker for Chinese due to the language-specific issues in Chinese. Unlike other languages like English, there are no explicit delimiters to indicate word boundaries in a plain Chinese text. Word segmentation is therefore an essential step to a Chinese chunking task. The second issue concerns unknown words (UWs) in open-ended text. Most current systems need a dictionary to guide their analysis. However, no dictionary could be complete. While a predefined dictionary may cover most words in use, there are many other words in open-ended documents, such as proper nouns and domain-specific terms that cannot be exhaustively listed. On the other hand, unknown word identification (UWI) is still a difficult problem for unknown words are constructed freely and dynamically in Chinese. Furthermore, it is not easy to explore lexical information for chunking from an open set of unknown words. Finally, Chinese language lacks exterior morphological hints for UWI and chunking [6].

0-7803-9091-1/05/$20.00 ©2005 IEEE

In this paper, we propose a lexicalized hidden Markov model (HMM) approach to Chinese text chunking. In order to tackle the problem of unknown words (UWs), Chinese text chunking is modeled as a tagging task on a sequence of known words (KWs). To do this, a tagger is thus developed based on the lexicalized HMMs to assign each known word in input a hybrid tag, which involves four types of information: word boundary, part-of-speech (POS), chunk boundary and chunk type. In this ways, more features, including POS information and contextual information, in particular contextual lexical information can be combined for the recognition of different types of base phrases in Chinese text under the framework of hidden Markov modeling. As a result, the performance of the system can be improved without losing its efficiency in training and chunking.

The rest of the paper is organized as follows: Section 2 gives a brief description of our chunking task. Section 3 discusses how to represent Chinese chunks as a sequence of known words with their corresponding hybrid tags. Section 4 presents a lexicalized HMM tagger for Chinese text chunking. Finally, the experimental results and some conclusions on this work are given in section 5 and section 6 respectively.

2. Task definition

Our current task focuses on the recognition of base phrases (BPs) in Chinese text. In our work, a *base phrase* is defined as a minimum non-nesting or non-recursive phrase with a stable internal structure and impendent semantic role. Normally, a base phrase has a lexical word as its headword. Essentially, a base phrase must consist of continuous words and contain no nesting components. It never overlaps with other phrases. Base phrases normally conform to a number of typical patterns, such as *[a+n]->BaseNP*, *[a+a]-> BaseAP*.

As shown in Table 1, we define a total of eleven chunk or base phrase types for Chinese. These types are based on the syntactic phrase categories and the semantic information categories defined in the PolyU Shallow Treebank [12].

Further, our chunking task is slightly different to the CoNLL-2000 shared task [1]. At present, we only concern multi-word chunks. All other single-word chunks and words outside of any chunks are still viewed as common words. For convenience, their POS categories are defined as their types during chunking. In conforming to the PolyU Treebank [12], we adopt the Peking University POS tag-set in this work, which specifies a total of 43 different POS tags [13].

Table 1. Chinese base phrase categories

Category	Description	Example
BNP	Base noun phrase	[市场/n 经济/n]NP *market economy*
BAP	Base adjective phrase	[公正/a 合理/a]BAP *fair and reasonable*
BVP	Base verb phrase	[顺利/a 启动/v]BVP *successfully start*
BDP	Base adverb phrase	[已/d 不再/d]BDP *no longer*
BQP	Base quantifier phrase	[数千/m 名/q]BQP 士兵/n *several thousand soldiers*
BTP	Base time phrase	[早上/t 8 时/t]BTP *8:00 in the morning*
BFP	Base position phrase	[内蒙古/ns 东北部/f]BFP *North-east of Inner Mongolia*
BNT	Name of an organization	[烟台/ns 大学/n]BNT *Yantai University*
BNS	Name of a place	[江苏省/ns 铜山县/ns]BNS *Jiangsu Province, Tongshan Country*
BNZ	Other proper noun phrase	[诺贝尔/nr 奖/n]BNZ *The Nobel Prize*
BSV	S-V-O structure	[领土/n 完整/a]BSV *territorial integrity*

3. Chunk representation

In most chunking tasks, chunk information in a sentence is usually represented by means of tags. In this section, we propose a new representation for Chinese chunks, in which chunks are formulated as a sequence of known words together with their hybrid tags. Each hybrid tag integrates four types of information: word boundary, POS, chunk boundary and chunk type.

3.1. Representation of POS-tagged words

As discussed in [14] and [15], KWs and UWs in a sentence can be represented using word-formation pattern tags. In practice, a lexicon word w has four possible patterns to present itself during word segmentation: (1) w is an independent segmented KW by itself; (2) w is at the beginning of a segmented UW. (3) w is at the middle of a segmented UW. (4) w is at the end of a segmented UW. For convenience, we use four tags *I*, *B*, *M* and *E* to denote these patterns respectively.

With these pattern tags, a POS-tagged word can be represented as a sequence of KWs attached with their

relevant hybrid tags, which usually have the following format: T1-T2. Where T1 denotes a word-formation pattern tag, and T2 denotes a POS tag. For example, the original POS-tagged sentence "*中国/ns 国家/n 主席/n 胡/nr 锦涛/nr 同/p 北朝鲜/ns 领导人/n 金/nr 正日/nr 举行/v 会谈/vn 。/w*" (*Chinese President Hu Jintao held talks with North Korean leader Kim Jong-Il*) can be equally represented as "*中国/I-ns 国家/I-n 主席/I-n 胡/I-nr 锦/B-nr 涛/E-nr 同/I-p 北/B-ns 朝鲜/E-ns 领导人/I-n 金/I-nr 正/B-nr 日/E-nr 举行/I-v 会谈/I-vn 。/I-w*".

3.2. Representation of chunks

Up to now, two major types of representation are proposed to formulate chunking: OIB-representation and initial/final representation (also referred to as Start/End representation [7]). OIB-representation is originally presented by Ramshaw and Marcus [3], which consists of three tags: words inside a base phrase were marked an *I* tag, words outside any base phrases received an *O* tag, and a special tag *B* was used for the first word inside a base phrase immediately following another base phrase. In contrast, the initial/final representation generally consists of four tags: words inside a chunk receive an *I* tag, words outside any chunks receive an *O* tag, all chunk-initial words receive an *B* tag, all chunk-final words receive an *E* tag. Kudo and Matsumoto also use an *S* tag to mark single-word chunks [7].

In our system, we use initial/final representation to formulate Chinese chunking for it is similar to the above representation for segmented words. Furthermore, O-tag and S-tag are merged into a new tag *I* in that our current task only concerns multi-word chunks. For simplicity, we use the four tags used in word representation, i.e. I/B/M/E to mark independent words outside any chunks, chunk-initial words, words at the middle of chunks, chunk-final words, respectively. Since our task consists of chunk identification and classification, the four tags are further attached a category tag shown in Table 1.

Thus, a base-phrase bracketed sentence can be fully reformulated as a sequence of KWs together with their hybrid tags. As shown in Table 2, the original chunked sentence "*[中国/ns 旅游年/n]BNP 是/v [一/m 次/q]BQP 国家级/b 的/u [宣传/vn 促销/vn 活动/vn]BNP 。/w*" (*China Tourism Year is a national-level promotion and marketing activity.*) can be represented as follows:

中国/I-ns-B-BNP 旅游/B-n-M-BNP 年/E-n-E-BNP 是/I-v-I-v 一/I-m-B-BQP 次/I-q-E-BQP 国家/B-b-I-b 级/E-b-I-b 的/I-u-I-u 宣传/I-vn-B-BNP 促销/I-vn-M-vn 活动/I-vn-E-BNP 。/I-w-I-w (e.g.1)

Instead of common segmented words, we consider KWs (viz. words listed in the system dictionary) to be the basic tokens in chunking, because: (1) a segmented word may be out of the system dictionary. It is not convenient to explore the important lexical information for chunking from such UWs. On the contrary, chunking based on KWs does not have this problem. (2) The input for chunking may be a plain text. By comparison, it is easier to segment a sequence of Chinese characters to a sequence of KWs [14].

Table 2. Example for chunk representation

Word	POS	Word level		Chunk level		Hybrid tag
		KW	Tag	Boundary	Type	
中国	ns	中国	I-ns	B	BNP	I-ns-B-BNP
旅游年	n	旅游	B-n	M	BNP	B-n-M-BNP
		年	E-n	E	BNP	E-n-E-BNP
是	v	是	I-v	I	v	I-v-I-v
一	m	一	I-m	B	BQP	I-m-B-BQP
次	q	次	I-q	E	BQP	I-q-E-BQP
国家级	b	国家	B-b	I	b	B-b-I-b
		级	E-b	I	b	E-b-I-b
的	u	的	I-u	I	u	I-u-I-u
宣传	vn	宣传	I-vn	B	BNP	I-vn-B-BNP
促销	vn	促销	I-vn	M	BNP	I-vn-M-BNP
活动	vn	活动	I-vn	E	BNP	I-vn-E-BNP
。	w	。	I-w	I	w	I-w-I-w

4. Lexicalized HMM tagger

On the basis of the above formulation, Chinese text chunking can be formalized as a task of tagging a sequence of KWs with a proper sequence of tags. With a view to the convenience in implementation, we employ the uniformly lexicalized models to perform this task.

4.1. Lexicalized HMMs

Given a sequence of KWs $W = w_1 w_2 \cdots w_n$, the goal of a tagger for Chinese text chunking is to find an appropriate sequence of hybrid tags $\hat{T} = t_1 t_2 \cdots t_n$ that maximizes the conditional probability $P(T \mid W)$, namely

$$\hat{T} = \arg\max_T P(T \mid W) = \arg\max_T \frac{P(W \mid T)P(T)}{P(W)} \quad (1)$$

For a specific sequence of KWs W, the probability $P(W)$ is fixed. Therefore, it can be dropped from the above equation. Thus, we have a general statistical model for KW tagging as follows:

$$\hat{T} = \arg\max_{T} P(W \mid T)P(T) = \arg\max_{T} \prod_{i=1}^{n} P(w_i \mid w_{1,i-1}, t_{1,i})P(t_i \mid w_{1,i-1}, t_{1,i-1}) \quad (2)$$

In theory, the general model in Equation (2) can provide the tagging system with a powerful capacity of disambiguation. However, it is not computable in practice for it involves too many parameters. Thus, two types of approximations are usually employed to simplify it.

The first approximation is based on the independent hypothesis in the standard HMMs: The appearance of current word w_i depends only on current tag t_i during tagging, and the assignment of current tag t_i depends only on its previous K $(1 \le K \le i-1)$ tags $t_{i-K,i-1}$. Based on these assumptions, the general model in Equation (4) can be rewritten as:

$$\hat{T} = \arg\max_{T} \prod_{i=1}^{n} P(w_i \mid t_i)P(t_i \mid t_{i-K,i-1}) \quad (3)$$

Equation (3) presents a *K*-order HMMs for known word tagging. Where, $P(w_i \mid t_i)$ denotes the so-called lexical probability; and $P(t_i \mid t_{i-K,i-1})$ denotes the contextual tag probability. In view of data sparseness, we use the first-order HMMs, i.e. $P(t_i \mid t_{i-K,i-1}) \approx P(t_i \mid t_{i-1})$.

The second approximation follows the notion of the uniform lexicalization technique, where two main hypotheses are made: The appearance of current word w_i is assumed to depend not only on the tag t_i of itself but also its K $(1 \le K \le i-1)$ previous words $w_{i-K,i-1}$ and $K-1$ tags $t_{i-K+1,i-1}$; The assignment of current tag t_i is supposed to depend both on its K previous words $w_{i-K,i-1}$ and tags $t_{i-K,i-1}$. Thus, the K^{th} order uniformly lexicalized HMMs can be formulated as follows:

$$\hat{T} = \arg\max_{T} \prod_{i=1}^{n} P(w_i \mid w_{i-K,i-1}, t_{i-K+1,i})P(t_i \mid w_{i-K,i-1}, t_{i-K,i-1}) \quad (4)$$

Equation (6) gives a general form of the *K*-order uniformly lexicalized HMMs. Similarly, we only consider the first-order lexicalized HMMs, i.e. $K = 1$.

In contrast to the standard HMMs, lexicalized HMMs can handle richer contextual information, both contextual words and contextual tags for the assignment of tags to known words, which will result in improvement of accuracy in chunking.

4.2. Estimation and data smoothing

If a large chunked corpus is available, the parameters in Equation (3) and (4) can be easily estimated with their relative frequencies counted directly from the training corpus under the framework of maximum likelihood estimation (MLE). However, MLE will yield zero probabilities for any cases that are not observed in the training data. To tackle this problem, we employ the linear interpolation smoothing technique in our implementation. As shown in equation (5), higher-order parameters in HMMs are smoothed with the relevant lower-order probabilities.

$$\begin{cases} P'(w_i \mid t_i) = \lambda P(w_i \mid t_i) + \dfrac{1-\lambda}{Count(t_i)} \\ P'(t_i \mid t_{i-1}) = \mu P(t_i \mid t_{i-1}) + (1-\mu)P(t_i) \end{cases} \quad (5)$$

Similarly, the lexicalized models are smoothed with the relevant non-lexicalized models, namely

$$\begin{cases} P'(w_i \mid w_{i-1}, t_i) = \lambda P(w_i \mid w_{i-1}, t_i) + (1-\lambda)P(w_i \mid t_i) \\ P'(t_i \mid w_{i-1}, t_{i-1}) = \mu P(t_i \mid w_{i-1}, t_{i-1}) + (1-\mu)P(t_i \mid t_{i-1}) \end{cases} \quad (6)$$

where λ and μ denote the interpolation coefficients.

4.3. The tagging algorithm

Based on the models in equation (3) or (4), the tagging algorithm aims to score all candidate sequences of tags and find the best one that has the maximum score. In our system, the classical Viterbi algorithm is employed to perform this task, which consists of three major steps as follows:

(1) Preprocessing: The task of preprocessing is to convert different types of input for chunking into a sequence of known word-based tokens using different techniques [14][15]: If the input is a sequence of characters, the system will segment it to a sequence of KWs using a known word-based n-gram segmenter; If the input is a sequence of segmented words, the system will convert the input into a sequence of KWs with their pattern tags using the maximum matching technique. If the input is a sequence of POS-tagged words, the system will apply the maximum matching technique to convert it to a sequence of KWs together with their hybrid tags shown in Section 3.1.

(2) Generation of candidate tags: This step aims to generate a lattice of candidate hybrid tags for each KW token produced in the first step. As shown in Table 2, there are two levels of tags: word-level tags and chunk-level tags. For the generation of word-level candidate tags, we use the same strategy as shown in [14]. With respect to chunk-level candidates, a KW may take one of the four boundary tags and the eleven category tags as its potential candidates. Given a KW, its candidate set of hybrid tags is a combination of all its word-level candidates and chunk-level candidates. All these candidate tags are stored in a lattice structure.

Lattice pruning is crucial for efficient chunking, which aims at preventing some improper candidates from entering

the lattice for decoding. In our implementation, whether a hybrid tag is an eligible candidate for a given KW depends on the relevant lexical probability shown in equation (3). If the lexical probability is larger than a given threshold, then the hybrid tag is considered as an eligible candidate. In general, the threshold is empirically determined. The larger threshold usually results in higher precision and lower recall in chunking. In current system, we set the threshold to 0, which implies that if a hybrid tag is observed to be attached to a given KW, then it is an eligible candidate hybrid tag of the KW.

(3) Decoding of the best tag sequence: In this step, the Viterbi algorithm is used to score all candidate tags with the proposed language models, and then searches the best path through the candidate lattice that has the maximal score. This path contains the most probable chunks of the input.

5. Experiments

In order to examine the effectiveness of our approach, we test our system using the PolyU Shallow Treebank. This section reports the relevant results.

5.1. Experimental data and evaluation measures

Table 3. Distributions of different types of base phrases in the experimental data

Type	Training data		Test data	
	Count	%	Count	%
BNP	46,676	48.15	4,456	47.19
BVP	25,214	26.01	2,506	26.54
BQP	8,599	8.87	568	6.02
BNT	5,532	5.71	716	7.58
BTP	3,340	3.45	567	6.01
BAP	3,294	3.40	313	3.31
BFP	1,916	1.98	98	1.04
BSV	1,121	1.16	79	0.84
BNS	1,115	1.15	83	0.88
BDP	136	0.14	21	0.22
BNZ	0	0.00	35	0.37
Total	96,943	100	9,442	100

The experimental corpus is derived from the PolyU Shallow Treebank, which contains 2,639 articles totaling 1,035,058 words. The PolyU Treebank has annotated three levels of phrases: maximal phrase, mid-phrase and base phrase. In our work, we only use the base phrase annotations. As shown in Table 3, we further divide this corpus into two parts: 90% for training and 10% for testing.

In addition to the PolyU Shallow Treebank, we also use a POS-lexicon in our system, which is mainly derived from the *Grammatical Knowledge-base of Contemporary Chinese* developed by the Peking University. In order to make this lexicon complete, we also add all GBK Hanzi and non-Hanzi characters to it. Consequently, the final dictionary contains about 65,270 different word-forms in all.

We evaluate our system in terms of three measures: *recall* (R), *precision* (P) and *F-score* (F). Here, recall (R) is defined as the number of correctly recognized base phrases divided by the total number of base phrases in the manually annotated corpus, and precision (P) is defined as the number of correctly recognized base phrases divided by the total number of output base phrases by the system. F-score is a weighted harmonic mean of precision and recall. In our experiments, we use the balanced F-score to evaluate the overall chunking performance, i.e. *2RP/(R+P)*.

5.2. Experimental results and discussions

In principle, a lexicalized HMM-based tagger should have a more powerful capacity to achieve correct tagging for text chunking than a standard HMM-based tagger because lexicalized HMMs can handle richer contextual information for tagging, in particular the contextual lexical information. Consequently, our current experiment is conducted to examine how the use of the lexicalization technique improves the chunking performance. In current experiment, the input is a sequence of POS-tagged words.

Table 4 shows the experimental results. It should be noted that the first line of each row stands for the non-labeled recall, precision and F-score, and the second line stands for the labeled recall, precision and F-score.

Table 4. Experimental results on the PolyU Treebank

Methods	R (%)	P (%)	F (%)
Standard	59.39	68.20	63.49
HMMs	60.50	69.47	64.67
Lexicalized	85.58	90.82	88.12
HMMs	86.03	91.30	88.59

As can be seen in Table 4, the lexicalized HMM based tagger performs better than the standard HMM-based system. It can be observed that the lexicalized HMMs improve the labeled F-measure in chunking by 14.63 percents and the non-labeled F-score by 13.92 percents. Furthermore, there is less difference between labeled measures and non-labeled measures for the lexicalized HMM tagger than for the standard HMM tagger, which indicates that the use of lexicalization technique is helpful to improve the performance in chunk classification as well as chunk identification.

6. Conclusions

In this paper, we have presented a lexicalized HMM approach to Chinese text chunking. In this work, Chinese text chunking is formalized as a tagging task on a sequence of known words. To do this work, a lexicalized HMM tagger is further developed to assign each known word in input an appropriate hybrid tag that involves four types of information: word boundary, POS, chunk boundary and chunk type. In comparison with standard HMMs, the lexicalized HMMs can handle richer contextual information, both contextual words and tags for correct tagging of known words. The preliminary experiment on the PolyU Shallow Treebank shows that the lexicalized HMMs can substantially improve chunking performance than the standard HMMs. In practice, the proposed approach also provides an input-adaptive framework, which is workable for different types of input, including plain text, segmented text and POS-tagged text. For future work, we hope to conduct more experiments to examine how different types of input affect chunking performance.

Acknowledgements

We would like to thank the Institute of Computational Linguistics of the Peking University for their part-of-speech tagset, lexicon and corpus.

References

[1] Erik F. Tjong Kim Sang, and S. Buchholz, "Introduction to the CoNLL-2000 shared task: Chunking", Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, pp. 127-132, 2000.

[2] Tie-jun Zhao, Mu-yun Yang, Fang Liu, Jian-min Yao, Hao Yu, ""Statistics based hybrid approach to Chinese base phrase identification", Proceedings of the 2nd Workshop on Chinese Language Processing, Hong Kong, pp.73-76, 2000.

[3] Lance A. Ramshaw, and Mitch P. Marcus, "Text chunking using transformation-based learning", Proceedings of the Third ACL Workshop on Very Large Corpora, Somerset, New Jersey, USA, pp.82-94, 1995.

[4] Qin Lu, Jing Zhou, and Ruifeng Xu, "Machine learning approaches for Chinese shallow parsers", Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi'an, China, pp.2309-2314, Nov. 2003.

[5] Wojciech Skut, and Thorsten Brants, "A maximum entropy partial parser for unrestricted text", Proceedings of the Sixth ACL Workshop on Very Large Corpora, Montreal, Canada, pp.143-151, August 1998.

[6] Yuqi Zhang, and Qiang Zhou, "Chinese base-phrases chunking", Proceedings of the First SIGHAN Workshop on Chinese Language Processing, Taipei, 2002.

[7] Taku Kudo, and Yuji Matsumoto, "Chunking with support vector machines", Proceedings of NAACL 2001, pp.192-199, 2001.

[8] Tetsuji Nakagawa, Taku Kudo, and Yuji Matsumoto, "Revision learning and its application to part-of-speech tagging", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), Philadelphia, USA, pp.497-504, July 2002.

[9] Perran Pla, and Antonio Molina, "Improving part-of-speech tagging using lexicalized HMMs", Natural Language Engineering, Vol.10, No.2 pp.167-189, 2004.

[10] Antonio Molina, and Perran Pla, "Shallow parsing using specialized HMMs", Journal of Machine Learning Research, Vol. 2, pp.595-613, March 2002.

[11] Sang-Zoo Lee, Jun-ichi Tsujii, and Hae-Chang Rim, "Lexicalized hidden Markov models for part-of-speech tagging", Proceeding of The 18th Conference on Computational Linguistics (COLING 2000), Saarbruken, Germany, pp. 481-487, July 2000.

[12] Ruifeng Xu, Qin Lu, Yin Li, and Wanyin Li, "The design and construction of the PolyU Shallow TreeBank", Computational Linguistics and Chinese Language Processing, to appear.

[13] Shiwen Yu, Huiming Duan, Suefeng Zhu, Bin Swen, and Baobao Chang, "Specification for corpus processing at Peking University: Word segmentation, POS tagging and phonetic notation", Journal of Chinese Language and Computing, Vol.13, No.2, pp. 121-158, 2003.

[14] Guohong Fu, and Kang-Kwong Luke, "Chinese unknown word identification as known word tagging", Proceedings of the Third IEEE International Conference on Machine Learning and Cybernetics (ICMLC 2004), Shanghai, China, pp. 2612-2617, August 2004.

[15] Guohong Fu, and Kang-Kwong Luke, "Chinese unknown word identification using class-based LM", Lecture Notes in Artificial Intelligence (IJCNLP 2004), No. 3248, pp.704-713, Jan. 2005.

[16] Erik F. Tjong Kim Sang, and Jorn Veenstrra, "Representing text chunks", Proceedings of EACL'99, pp. 173-179, 1999.