

The Effectiveness Study of Local Maximum Feature for Chinese Unknown Word Identification

Tao Liu, Bing-Quan Liu, Xiao-Long Wang, Ming-Hui Li
School of Computer Science and Technology
Harbin Institute of Technology, Harbin 150001, China
{tliu, liubq, wangxl, mhli}@insun.hit.edu.cn

Abstract

Word segmentation is a basic step in Chinese text processing. The identification of unknown words is the major bottleneck for the current word segmentation systems. Since a Chinese word is an independent linguistic unit, the component characters of a Chinese word have contextual independency and inner cohesiveness. Most of Chinese words have relatively higher cohesiveness values compared with that of their local context when fair symmetric conditional probability is used to measure the cohesiveness of character sequence. Local maximum feature is used for the Chinese unknown word identification for the first time in this paper and it implicates contextual independency and inner cohesiveness of Chinese word. It has showed its effectiveness in selecting high quality Chinese unknown word candidates. To make up the insufficiency of local maximum feature in depicting word formation history of each Chinese character, heuristic features including word formation power and word formation pattern are used to depict word formation characteristic of different characters and to eliminate candidates which don't accord with Chinese word formation style. Experiments on the second international Chinese word segmentation bakeoff corpora show that our method is effective for recognizing Chinese unknown words. Finally, Chinese unknown word identification is applied to text classification and it can improve the system performance.

Keywords

Chinese word segmentation; unknown word identification; local maximum feature; heuristic feature.

1 Introduction

For many Asian languages, such as Chinese, which are written without explicit delimiters between words to indicate word boundaries, word segmentation (Gao et al., 2005, Sproat et al., 1996) is essential to most of the natural language processing (NLP) tasks. The existence of unknown words or out-of-vocabulary (OOV) words makes word segmentation more difficult. The international Chinese word segmentation bakeoff (Emerson, 2005) shows OOV handling is the Achilles heel of segmentation systems. The development of an automatic unknown words identification system is very important for many NLP applications, such as information retrieval, information extraction and ontology construction.

Much research work has been done on Chinese Unknown Word Identification (UWI) (Sproat and Chilin, 2002, Sun and Tsou, 2001). From the system structure point of view, all the approaches proposed so far for unknown words identification can be classified as embedded approaches (Goh et al., 2005, Peng et al., 2004, Tseng et al., 2005, Xue, 2003) by which, unknown words are detected online along with word segmentation process, and modular approaches (Chen et al., 2005, Gao et al., 2002, Nie et al., 1995) by which, new words acquired from large corpora in an off-line manner are either put into a dictionary before word segmentation starts or used for merging the single characters after word segmentation. Features normally used for UWI include word frequency, in-word probability (IWP), word-formation pattern, mutual information (Church and Hanks, 1990), left/right entropy (Sornlertlamvanich et al., 2000), context dependency (Chien, 1999) and part-of-speech information. Both IWP and word-formation pattern reflect the morphological property of the language which corresponds to the linguistic phenomena occurring in Chinese words. Unknown word candidates which contain functional characters selected by IWP are eliminated in (Nie et al., 1995). (Chen et al., 2005), (Wu and Jiang, 2000) and (Fu, 2001) use IWP to combine adjacent single characters after basic segmentation if the product of their IWP is larger than a preset threshold. In addition, (Wu and Jiang, 2000) and (Fu, 2001) use word formation pattern to depict how likely a character appears in a certain position within a word. Cohesiveness measures such as mutual information estimate the internal associative strength among constituents of character n-gram which tend to be unknown word candidate. Both left/right entropy and context dependency depict the dependency strength of current item (character sequence) on its context. The dependency strength of a current item on its context decreases as the probability of this item to be a Chinese word increases.

Local maximum feature is introduced in this paper, to the best of our knowledge, for the first time to identify Chinese unknown words. Local maximum feature depicts the characteristic of character sequence s with local maximum cohesiveness strength compared with that of longer character sequences which contain s and shorter character sequences contained in s . The cohesiveness strength of s is greater than that of longer character sequences which contain s , which means that s is context independent. The cohesiveness strength of s is not less than that of shorter character sequences contained in s , which means that these character sequences contained in s are context dependent. According to our statistics, most of the Chinese words have local maximum features because Chinese words are fixed character sequences in certain contexts. It is better to use local maximum feature instead of a usually empirically determined global threshold to extract character sequences with comparatively higher cohesiveness strength. In addition, it is more objective to establish Chinese word boundary by local maximum feature than simply merging character sequences with higher IWP product of all component characters after normal word segmentation.

2 Unknown word identification

2.1 Problem statement

Chinese words can be classified into five types of words: lexicon words, morphologically derived words (MDW), factoids, named entities and new words (Gao et al., 2002, Gao et al., 2005), of which, new words are time-sensitive concepts or domain-specific terms which have not been collected into a dictionary, and MDW, factoids, named entities and new words are all beyond the cover of a dictionary. Since methods for detecting factoids is a trivial task,

generally unknown words refer to the other three types. In this paper we focus on these three types including morphologically derived words, named entities and new words.

2.2 Local maximum feature for Chinese words

In a broad sense, Chinese words include single character words and multi-character words. Chinese unknown word identification refers to the identification of unknown Chinese multi-character word. Unless otherwise stated, Chinese words means Chinese multi-character words in this paper. Chinese words are linguistic units which can be independently used in text. This definition contains two properties of Chinese words: contextual independency and inner cohesion. Contextual independency means a Chinese word is a context independent unit which has a weak cohesiveness with its surrounding characters. Inner cohesion of a Chinese word means the component characters of a word are aggregate and fixed sequence which makes the sub sequences of the word context-dependent. Chinese words tend to be character sequences with relatively higher contextual independency and inner cohesion. Local maximum feature (Dias et al., 2000, Silva et al., 1999) is introduced in this paper to depict this characteristic of Chinese words.

If we use some statistical measures to depict the cohesiveness or stability of a character sequence, contextual dependency of a character sequence can be determined by comparing the statistical values of current character sequences with those of other character sequences which contain the current character sequence in text. Relatively higher contextual independency and inner cohesion are expressed by local maximum value of cohesiveness measure.

Given an n-gram (a sequence with n elements) of characters, W, we denote the set all the (n-1)-gram contained in the n-gram W by Θ_{n-1} and the set of all (n+1)-grams containing W by Θ_{n+1} . W processes local maximum feature if and only if:

$$\text{For } \forall x \in \Theta_{n-1}, \forall y \in \Theta_{n+1} \text{ W satisfies } \begin{cases} v(W) > v(y) & n = 2 \\ v(x) \leq v(W) > v(y) & n > 2. \end{cases}$$

where, $v(\cdot)$ denotes the cohesiveness value of a sequence.

For a continuous sequence of characters, $\dots c_{i-1}c_i c_{i+1} \dots c_{i+n-1}c_{i+n} \dots$, suppose current n-gram as $c_i c_{i+1} \dots c_{i+n-1}$, Θ_{n-1} is simplified as $\{(c_i c_{i+1} \dots c_{i+n-2}), (c_{i+1} \dots c_{i+n-2} c_{i+n-1})\}$; Θ_{n+1} is simplified as $\{(c_{i-1} c_i \dots c_{i+n-1}), (c_i c_{i+1} \dots c_{i+n-1} c_{i+n})\}$. i.e. the cohesiveness of a character sequence is local maximum if its cohesiveness value is larger than that of longer character sequences which contain it and not less than that of shorter character sequences contained in it.

Local maximum contextual independency is thus expressed by local maximum value of cohesiveness measure. In fact a statistical measure can only approximately reflect the contextual independency since it is obtained from corpora with a very limited size. In real condition, most of Chinese words have local maximum value of cohesiveness measure.

2.3 Fair symmetric conditional probability (FSCP)

Fair symmetric conditional probability (Silva and Lopes, 1999) is used to measure the cohesiveness of generic n-gram ($n > 2$). For an n-gram $c_1 \dots c_n$ ($n \geq 2$), the fair symmetric conditional probability of it has the following form:

$$FSCP([c_1 \dots c_n]) = \frac{p(c_1 \dots c_n)^2}{Avp} \quad (1)$$

$$Avp = \frac{1}{n-1} \sum_{i=1}^{i=n-1} p(c_1 \dots c_i) \times p(c_{i+1} \dots c_n) \quad (2)$$

The fair symmetric conditional probability of generic n-grams is obtained from the Symmetric Conditional Probability (SCP) of bigrams using the fair dispersion point normalization method. There are two ways in making a cohesiveness measure of bigrams applicable to generic n-grams: 1) regard the extracted bigrams as a single character and compute the cohesiveness value of new bigrams. Continue this process until no more bigram is discovered; 2) use a normalization method to make the formula of bigram applicable to generic n-gram ($n > 2$). One limitation of the first method is that they depend on the identification of suitable bigram for the iterative procedure. The second method is adopted in this paper to transform SCP with the following form to FSCP.

$$SCP([x, y]) = p(x | y) \times p(y | x) = \frac{p(x, y)^2}{p(x) \times p(y)} \quad (3)$$

where $p(x, y)$, $p(x)$ and $p(y)$ are the probabilities of the bigram $[x, y]$ and the unigrams $[x]$ and $[y]$ to occur in the corpus respectively. As a pseudo bigram, n-gram $c_1 \dots c_n$ ($n \geq 2$) in equation (1) can be split into pseudo bigrams at every split point ranging from the first character to the $(n-1)$ th character. Then denominator of equation (3) is replaced by the average value denoted by Avp of original values for all these pseudo bigrams. FSCP which is applicable to generic n-gram is obtained from SCP in this way.

To illustrate the local maximum feature of Chinese words computed by FSCP measure, we present a sentence from the U. Penn Chinese Treebank corpus provided in the first international Chinese word segmentation bakeoff: 薰衣草可以舒缓镇定 (one kind of grass can make people comfortable and unflappable). The proper segmentation of this sentence is 薰衣草/可以/舒缓/镇定. Figure 1 shows that all the words of this sentence have their local maximum FSCP values.

n-gram	FSCP	n-gram	FSCP	n-gram	FSCP	n-gram	FSCP
薰衣草可	0.0199	草可以	0.0067	以舒缓	0.0040	缓镇定	0.0303
薰衣草	0.3636	可以	0.1042	舒缓镇	0.0952	镇定	0.0333
薰衣草	0.0615	可以舒	0.0090	舒缓	0.1777		
衣草	0.0513						

Figure 1. FSCP value for n-grams of sentence “薰衣草可以舒缓镇定”

2.4 Heuristic filtering features

The local maximum feature obtained using an unsupervised procedure makes the best use of statistical information of testing corpora and is effective in discovering new linguistic phenomena of corpora. In fact, pure statistics of unsegmented corpora omits the word formation history of each character indicated by a preexisted dictionary, which contributes to

unknown word identification. It is seldom for such characters as “的”, “了”, “都” to form multi-character words and the probability of these characters to form a multi-character word is low in the future. Such a character as “昨” tends to take the first position in such words as “昨天”, “昨夜”, while such characters as “们”, “者” tend to take the last position of such words as “朋友们”, “孩子们”, “学者”. i.e. some characters always take the same position in different words. Heuristic features (Nie et al., 1995), which include word formation power and word formation pattern, are used to depict word formation characteristic of different characters and to eliminate non-word candidates. Some frequently co-occurred single-character word sequences, such as “我在”, “不到” which are wrongly extracted by local maximum feature, are eliminated in this process because of the low word formation probability of their component characters. Some sequences such as “但以她” which don't accord with the Chinese word formation pattern are also eliminated.

As shown in Figure 2, local maximum feature is obtained from test corpus and used to get possible words candidates. To increase the processing speed, a suffix array(Yamamoto and Church, 2001) based data structure with a sorting algorithm(Bentley and Sedgewick, 1997) that blends quick sort with radix sort is used to compute character n-grams. We also use heuristic features to remove candidates not in match with the Chinese word formation style and to obtain unknown word candidates. Finally, we use these unknown word candidates to post-process the documents which have been segmented by in-vocabulary segmentation modular.

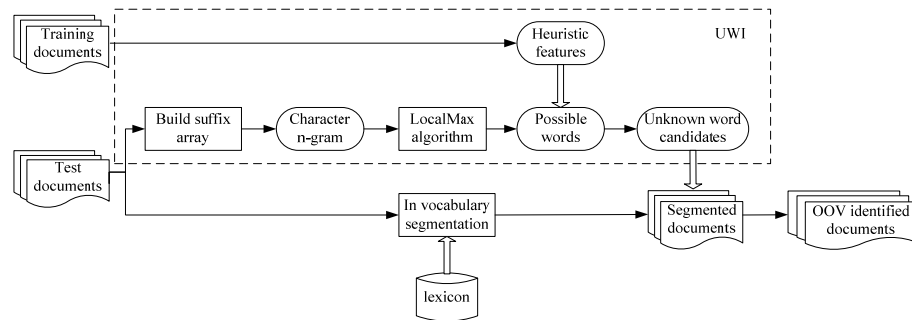


Figure 2. System overview

3 Experiment and result

In order to show the effectiveness of the UWI module, the system is first evaluated on the second international Chinese word segmentation bakeoff corpora and then applied in a text classification task.

3.1 Word segmentation evaluation

3.1.1 Test local maximum feature for its unknown word identification ability

As shown in Table 1 below, One GB encoded corpora of the second international Chinese word segmentation bakeoff provided by Microsoft Research (MSR) is used for the test and it includes a training corpus and a test corpus.

Corpus		Encoding	Word Types	Words	Character Types	Characters
MSR	Training	CP936	88,119	2,368,391	5,167	4,050,469
	Test	CP936	-	-	2882	169247

Table 1. Details of training and test corpora

There are two kinds of test: open test and closed test. For closed test, it is not allowed to use any other materials or knowledge except training material from a particular corpus. A closed test is adopted in this paper. The IV word segmentation model includes the bigram model which uses absolute smoothing and a rule-based numeric expression recognizer. Through several test, we select an empirical optimal value (-1.8) for weight threshold of UWI modular. Table 2 show results of baseline system (IV word segmentation) and system with UWI modular on MSR corpus. OOV recall is improved by 22 percent with the UWI modular. IV recall decreased a little since some words are wrongly identified by the UWI modular. The value of F measure is improved by 0.6 percent with the UWI modular.

MSR corpus	OOV Recall	Precision	Recall	IV Recall	F Measure
Baseline	0.361	0.936	0.971	0.988	0.953
Baseline+UWI	0.581	0.951	0.967	0.977	0.959

Table 2. Results of word segmentation system on MSR corpus

Table 3 shows some wrongly extracted sequences, original sentence of these sequences and corresponding segmentation in MSR corpus. Both “唱响” and “憋足” are composed of a verb and an adverb which modifies the verb. “铸宝刀” is a verb phrase. “中堡岛”, the name of an island is wrongly split by local maximum feature into “中/堡岛” since “中” always occurs independently. Local maximum feature has a low ability to recognize long named entity such as “毛乌素大沙漠” since long named entity always includes one or more sub words such as “沙漠”. So the more fixed sequence “毛乌素” in this named entity is extracted by local maximum feature.

Table 4 shows some plausible words of MSR corpus. People always have inconsistent views on the way of segmenting Chinese words, which leads to different standards for word segmentation. Even under the same word segmentation standard, word segmentation is inevitably inconsistent. For example, in the sentence “守信者得到酬赏, 失信者受到惩罚” (A man of one's word will be rewarded, a man breaks one's word will be punished) of MSR corpus, “酬赏” (reward) isn't a word, while “惩罚” (punish) which is an antonym of “酬赏” (reward) is a word. We think this is inconsistent word segmentation. For different applications, there are different requirements for the granularity of word segmentation. Non-word sequences such as “崇仁麻鸡”, “茶鲜菇”, “巨型机”, “贫油史”, “早蕾” are all meaningful units which provide abundant information for a particular application.

Wrongly extracted sequence	Original sentence or sentence snippet	Segment in MSR corpus
唱(sing)响(ringingly)	唱响国企志气歌	唱/响
憋(hold back)足(adequately)	职工憋足了劲	憋/足
铸(cast)宝刀(precious sword)	锻得新钢铸宝刀	铸/宝刀
堡(fort)岛(island)	一向喧闹的中堡岛今夜格外宁静	中堡岛(name of an island)
毛(hair)乌(black)素(plain)	穿过茫茫的毛乌素大沙漠东南边缘	毛乌素大沙漠(name of a desert)

Table 3. Examples of some wrongly extracted sequences

Meaningful units	Original sentence or sentence snippet	Segment in MSR corpus
酬赏(reward)	即守信者得到酬赏,失信者受到惩罚	酬/赏
崇仁麻鸡(a kind of chook)	崇仁麻鸡和东乡黑鸡为主的饲养量	崇仁/麻/鸡
茶鲜菇(a kind of fungus)	以茶鲜菇为主的食用菌加工企业	茶/鲜/菇
巨型机(supercomputer)	新一代超高性能巨型机	巨型/机
贫油史(the history of lean oil)	以改写中国贫油史而成为举世闻名的“功勋城”	贫油/史
早蕾(early bud)	摘除7月10日前的早蕾	早/蕾

Table 4. Some plausible words but meaningful units

3.1.2 Comparison of our system with other systems for unknown word identification

The comparison between our system and top ten systems in the second international Chinese word segmentation bakeoff (Emerson, 2005) is summarized as Table 5. The definition of Run ID and effectiveness measures can be found in (Emerson, 2005). The performance of our system (Baseline+UWI) is in the middle of these top ten systems. For named entities, the proposed method is not as effective as special named entity tagging tools which use machine learning methods. If we use some certain technique to process named entities, the result can be better.

3.1.3 Test the word extraction ability of local maximum feature

In fact, local maximum feature can be used as a word extraction method in the absence of a dictionary or be used for expanding a small dictionary. To show this, we use some small scale training sets with different OOV rate to train the heuristic features of the UWI modular. Training sets are subsets of MSR corpora of the second international Chinese word segmentation bakeoff with their details shown in Table 6. We can see from Fig.3 that UWI

modular is also very useful in the case of small training set. When the OOV rate increases from 0.22 to 0.5, the improvement of OOV recall (dashed curve) ranges from 23 to 26 percent and the improvement of F measure ranges from 7 to 13 percent.

Run ID	OOV Recall	IV Recall	Recall	Precision	F measure
14	0.717	0.968	0.962	0.966	0.964
7	0.592	0.972	0.962	0.962	0.962
27(a)	0.379	0.985	0.969	0.952	0.960
27(b)	0.381	0.984	0.968	0.953	0.960
4	0.323	0.991	0.973	0.945	0.959
15(b)	0.718	0.958	0.952	0.964	0.958
5	0.210	0.995	0.974	0.940	0.957
13	0.496	0.972	0.959	0.956	0.957
12	0.673	0.960	0.952	0.960	0.956
24	0.503	0.970	0.958	0.952	0.955
Our system	0.581	0.977	0.967	0.951	0.959

Table 5. Closed test result on MSR corpus

Corpus ID	OOV rate	Words type	Words
1	0.224	4622	27625
2	0.269	3374	21360
3	0.300	2757	15871
4	0.334	2131	10678
5	0.426	1189	4995
6	0.494	750	3060

Table 6. Small scale training corpus information

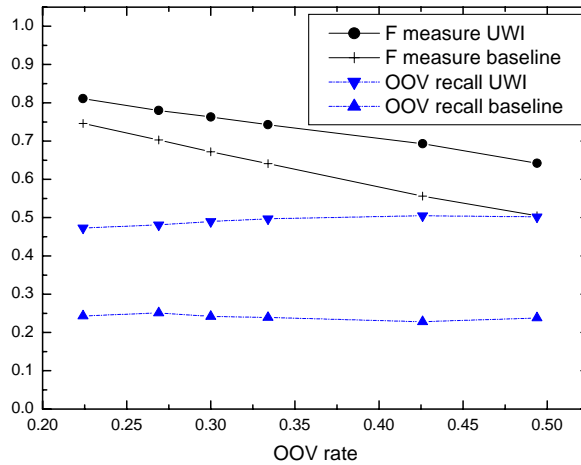


Figure 3. UWI performance at different OOV rate

3.2 Text classification evaluation

3.2.1 Impact of unknown word identification on text classification

The UWI method is applied in a text classification system which is a modular of our web news browser for mobile phone. We use web corpora of tourism domain to test the effect of UWI on text classification. The corpora are divided into eight categories: city general situation, hotel, custom, shopping, traffic, service and entertainment. We build a K nearest neighbor (KNN) classifier (K=20) on 2000 training documents from web. Feature selection method based on domain information (Liu et al., 2005) is used in the classifier. We use the MSR dictionary with 88,119 words of the second international Chinese word segmentation bakeoff. Then we tested the classifier on 1000 testing web documents. The baseline system is a text classification system with normal word segmentation and there is no unknown word detection in the process.

For text categorization evaluation, the effectiveness measures of precision, recall and F1 are defined respectively as shown below.

$$precision_i = \frac{right_i}{predict_i} \times 100\%$$

$$recall_i = \frac{right_i}{test_i} \times 100\%$$

$$F1_i = \frac{recall_i \times precision_i \times 2}{recall_i + precision_i}$$

where $right_i$ is the number of correctly classified texts in category $_i$ and $predict_i$ is the number of texts that are classified into category $_i$. $test_i$ is the number of examples of category $_i$ in test set. Thus for all categories, Macro-recall, Macro-precision and Macro averaged F1 score are respectively defined. Table 7 shows the result of the baseline system and system with UWI. The improvement is satisfactory since it is the impact of word level information to document level.

method	Macro-Precision	Macro-Recall	Macro-F1
baseline	0.866	0.818	0.841
UWI	0.874	0.828	0.850

Table 7. Test classification result with MSR dictionary

3.2.2 Impact of word extraction on text classification

In the case of small scale original dictionary obtained from data sets of Table 6, the performance of a text classification system shown by Figure 4 is improved greatly by the UWI modular. The corpus ID in the figure denotes the training corpus we select for word segmentation of text classification. We can see in the case of the smallest original dictionary provided by the first corpus (ID=1), the text classification system performance is improved by 19 percent.

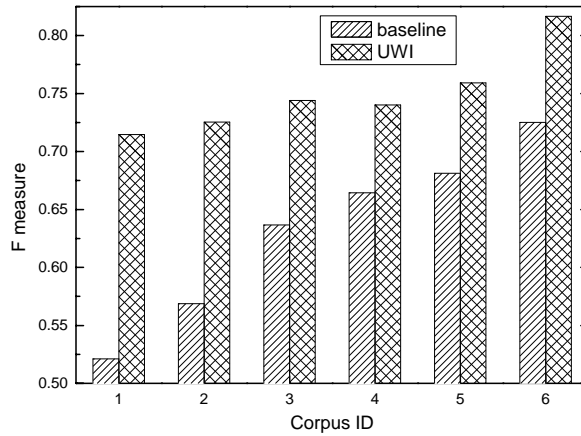


Figure 4. Text classification result with small original dictionary

4 Conclusion and future work

This paper aims at showing the effectiveness of local maximum feature for Chinese unknown word identification. Local maximum feature, which depicts contextual independency and inner cohesiveness of Chinese words, is introduced to select unknown word candidates with statistical information. It is more objective to select unknown word candidate in this way since most of Chinese words have local maximum cohesiveness value. Local maximum feature is also effective for word extraction in the absence of dictionary or in the case of small scale dictionary. Experiments on a Chinese word segmentation task and a text classification task show effectiveness of the proposed method for Chinese unknown word identification and word extraction.

For named entities, the proposed method is not as effective as special named entity tagging tools which use machine learning methods. Our future work will focus on integrating the local maximum feature and machine learning methods for Chinese unknown word identification and getting a unified approach for the identification of new words and named entity. The whole test corpus is used in the current system to compute local maximum feature. We will use corpora in a more reasonable size to compute local maximum feature. For example, the corpus of the same semantic domain can be taken as a computing unit for domain-specific term extraction.

5 Acknowledgements

We would like to thank National Natural Science Foundation of China (60435020, 60673037) and the high Technology Research and Development Programme of China (2006AA01Z197) for their support, and the reviewers for their valuable comments. Special thank goes to Chengjie Sun for his help during the research.

6 References

- Bentley, J. and Sedgewick, R., 1997, Fast Algorithms for Sorting and Searching Strings, *Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 360-369, New Orleans.
- Chen, A., Zhou, Y., Zhang, A. and Sun, G., 2005, Unigram Language Model for Chinese Word Segmentation, *Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 138-141, Korea.
- Chien, L.-F., 1999, PAT-Tree-Based Adaptive Keyphrase Extraction for Intelligent Chinese Information Retrieval, *Information Processing and Management*, vol. 35, no. 4, pp. 501-521.
- Church, K. and Hanks, P., 1990, Word association norms mutual information and lexicography, *Computational Linguistics*, vol. 16, no. 1, pp. 23-29.
- Dias, G., Guilloré, S. and Lopes, J.G.P., 2000, Normalization of Association Measures for Multiword Lexical Unit Extraction, *International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications*, pp. 207-216, Monastir, Tunisia.
- Emerson, T., 2005, The Second International Chinese Word Segmentation Bakeoff, *Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 123-133, Korea.
- Fu, G. 2001, Research on Statistical Methods of Chinese Syntactic Disambiguation, Doctor Thesis, Harbin Institute of Technology, Harbin, China.
- Gao, J., Goodman, J., Li, M. and Lee, K.-F., 2002, Toward a Unified Approach to Statistical Language Modeling for Chinese, *ACM Transactions on Asian Language Information Processing*, vol. 1, no. 1, pp. 3-33.
- Gao, J., Li, M., Huang, C.-N. and Wu, A., 2005, Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach, *Computational Linguistics*, vol. 31, no. 4, pp. 531-574.
- Goh, C.-L., Asahara, M. and Matsumoto, Y., 2005, Chinese Word Segmentation by Classification of Characters, *Computational Linguistics and Chinese Language Processing*, vol. 10, no. 3, pp. 381-396.
- Liu, T., Wang, X., Xu, Z. and Wang, Q., 2005, Domain-Specific Term Extraction and Its Application in Text Classification, *8th Joint Conference on Information Sciences*, pp. 1481-1484, USA.
- Nie, J.-Y., Hannan, M.-L. and Jin, W., 1995, Unknown Word Detection and Segmentation of Chinese using Statistical and Heuristic Knowledge, *Communications of the Chinese and Oriental Languages Information Processing Society*, vol. 5, no. 1&2, pp. 47-57.
- Peng, F., Feng, F. and McCallum, A., 2004, Chinese Segmentation and new word detection using conditional random fields, *The 20th International Conference on Computational Linguistics*, pp. 562-568, Geneva, Switzerland.
- Silva, J.F.d., Dias, G., Guilloré, S. and Lopes, J.G.P., 1999, Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units, *The 9th Portuguese Conference in Artificial Intelligence, Lecture Notes in Artificial Intelligence*, pp. 113-132, Universidade de Evora, Evora, Portugal.

- Silva, J.F.d. and Lopes, J.G.P., 1999, A local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units, *The 6th Meeting on the Mathematics of Language*, pp. 369-381, Orlando.
- Sornlertlamvanich, V., Potipiti, T. and Charoenporn, T., 2000, Automatic corpus-based Thai word extraction with the c4.5 learning algorithm, *Proceedings of the 18th conference on Computational linguistics*, pp. 802 - 807.
- Sproat, R. and Chilin, S., 2002, Corpus-based methods in Chinese morphology and phonology, *19th International Conference on Computational Linguistics*, Taiwan.
- Sproat, R., Shih, C., Gale, W. and Chang, N., 1996, A Stochastic Finite-State Word-Segmentation Algorithm for Chinese, *Computational Linguistics*, vol. 22, no. 3, pp. 377-404.
- Sun, M. and Tsou, B.K., 2001, A review and evaluation on automatic segmentation of Chinese, *Contemporary Linguistics*, vol. 3, no. 1, pp. 22-32.
- Tseng, H., Chang, P., Andrew, G., Jurafsky, D. and Manning, C., 2005, A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005, *Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 168-171, Jeju Island, Korea.
- Wu, A. and Jiang, Z., 2000, Statistically-Enhanced New Word Identification in a Rule-Based Chinese System, *The Second Chinese Language Processing Workshop*, pp. 46-51, China.
- Xue, N., 2003, Chinese Word Segmentation as Character Tagging, *Computational Linguistics and Chinese Language Processing*, vol. 8, no. 1, pp. 29-48.
- Yamamoto, M. and Church, K.W., 2001, Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus, *Computational Linguistics*, vol. 27, no. 1, pp. 1-30.