# A Search-based Chinese Word Segmentation Method

Xin-Jing Wang
IBM China Research Center
Beijing, China
wangxinj@cn.ibm.com

Wen Liu
Huazhong Univ. of Sci. & Tech.
Wuhan, China
tmac@smail.hust.edu.cn

Yong Qin
IBM China Research Center
Beijing, China
qinyong@cn.ibm.com

## ABSTRACT

In this paper, we propose a novel Chinese word segmentation method which leverages the huge deposit of Web documents and search technology. It simultaneously solves ambiguous phrase boundary resolution and unknown word identification problems. Evaluations prove its effectiveness.

## Categories and Subject Descriptors

I.2.7. [**Artificial Intelligence**]: Natural Language Processing– *Language parsing and understanding*. H.3.1. [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Linguistic processing*.

**General Terms:** Performance, Algorithms.

**Keywords:** Chinese word segmentation, search.

## 1. INTRODUCTION

Automatic Chinese word segmentation is an important technique for many areas including speech synthesis, text categorization, etc [3]. It is challenging because 1) there is no standard definition of words in Chinese, 2) word boundaries are not marked by spaces. Two research issues are mainly involved: ambiguous phrase boundary resolution and unknown word identification.

Previous approaches fall roughly into four categories: 1) Dictionary-based methods, which segment sentences by matching entries in a dictionary [3]. Its accuracy is determined by the coverage of the dictionary, and drops sharply as new words appear. 2) Statistical machine learning methods [1], which are typically based on co-occurrences of character sequences. Generally large annotated Chinese corpora are required for model training, and they lack the flexibility to adapt to different segmentation standards. 3) Transformation-based methods [4]. They are initially used in POS tagging and parsing, which learn a set of n-gram rules from a training corpus and then apply them to the new text. 4) Combining methods [3] which combine two or more of the above methods.

As the Web prospers, it brings new opportunities to solve many previously "unsolvable" problems. In this paper, we propose to leverage the Web and search technology to segment Chinese words. Its typical advantages include:

1) Free from the Out-of-Vocabulary (OOV) problem, and this is a typical feature of leveraging the Web documents.

2) Adaptive to different segmentation standards since ideally we can obtain all valid character sequences by searching the Web.

3) Can be entirely unsupervised that need no training corpora.

*WWW 2007,* May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

## 2. THE PROPOSED APPROACH

The approach contains three steps: 1) segments collecting, 2) segments scoring, and 3) segmentation scheme ranking.

### 2.1 Segments Collecting

The segments are collected in two steps:

1) Firstly, the query sentence is semantically segmented by punctuation which gives several sub-sentences.

2) Then each sub-sentence is submitted to a search engine for segments collecting. Technically, if the search engine's inverted indices are inaccessible as commercial search engines do, e.g. Google and Yahoo!, we collect the highlights (the red words in Figure 1) from the returned snippets as the segments. Otherwise, we check the characters' positions indicated by the inverted indices and find those that neighbor each other in the query.

Although search engines generally have local segmentors, we argue that their performance normally will not affect our results, e.g. Figure 1 shows the search results of "他高兴地说" (he said happily), our method assumes that the highlight "他高兴地" (he happily) is a segment. However, by checking the HTML source, we found that Yahoo!'s local segmentor gives "<b>他</b><b>高兴</b><b>地</b>", which cut it into three segments. Consider an extreme case that the local segmentor segments each sentences into unigrams, intuitively, segments collected will still be n-grams since the unigrams neighbor each other in the retrieved documents as they are written in natural language. This shows that our results are generally independent to search engines' local segmentors.

### 2.2 Segments Scoring

Each segment is scored so that we can select a subset of segments as the final segmentation which, when reconstructing the query, scores the highest. Obviously various methods can be used. Here we try two of them, namely frequency-based and Support Vector Machine (SVM)-based method.

#### *2.2.1 Frequency-based*

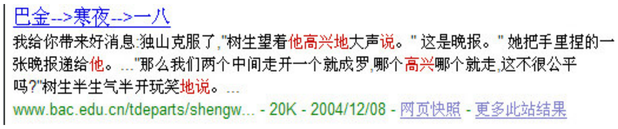
This method uses term frequency as the scoring function, which is defined as the ratio of the number of occurrences of the segment to the total number of occurrences of all the segments.

#### *2.2.2 SVM-based*

This method uses SVM classifier with RBF kernel and maps the outputs into probabilities as the scores [2].

### 2.3 Segments Selecting

We call a subset "valid" if its member segments can reconstruct exactly the query, and the score of a valid subset is the average score of its member segments. We select the valid subset which scores the highest as the final segmentation. For efficiency consideration, we use greedy search rather than dynamic programming to find valid subsets.



巴金-->寒夜-->一八
我给你带来好消息:独山克服了,"树生望着他高兴地大声说。" 这是晚报。" 她把手里捏的一张晚报递给他。..."那么我们两个中间走开一个就成罗,哪个高兴哪个就走,这不很公平吗?"树生半生气半开玩笑地说。...
www.bac.edu.cn/tdeparts/shengw... - 20K - 2004/12/08 - 网页快照 - 更多此站结果

**Figure 1. Yahoo! search result of "他高兴地说" (He said happily). Red words are the segments.**

## 3. EVALUATIONS

We evaluate our method on the benchmark MSR dataset provided by SIGHAN'05 workshop (www.sighan.org/bakeoff2005/) and also compare to IBM full-parser, a state-of-the-art dictionary-based method adopting maximum matching strategy.

### 3.1 Evaluation on SIGHAN'05 Benchmark Data

The training data used is 3,000 randomly selected sentences (Note that in the case of using frequency-based scoring function, our method needs no training and is unsupervised segmentation). And the entire testing dataset (about 4,500 sentences) is used for testing. The feature space is three-dimensional: {TF, DF, LEN}. TF is defined as in Section 2.2.1. DF is the number of documents indexed by a segment, and LEN indicates the number of characters in a segment.

Figure 2 shows the performance of our approach which is output by SIGHAN'05 benchmark evaluation. The dotted and blocked columns correspond to frequency- and SVM-based approaches separately. Although they are worse than those reported by SIGHAN'05, the approach is effective because we used only 3,000 training sentences (in the case of SVM-based method) while SIGHAN'05 groups used about 86,000. Moreover, out method avoids OOV problem.

Interestingly, frequency-based method performs better than SVM-based method in precision and F-measure. A possible reason is that the feature space is too simple to fully describe the data, so that the power of SVM models was not fully taken advantages of.

We argue that a better performance can be achieved with more search results provided. Since currently only Google search is used and it returns only about 800 snippets whose highlighted character sequences (i.e. segments) are generally long and contain multiple semantic concepts due to the great search power of Google, these limit the effectiveness of the segments extracted. In fact, based on a rough evaluation, much better performance can be achieved if we combine search results of Yahoo! and Google. However, since Yahoo! prohibits frequent query (to prevent DDOS attack), we were not able to collect enough training data from Yahoo!, but it inspires us that with a local search engine and a large document set, we can expect a much better performance.

### 3.2 Comparison to IBM Full-parser (FP)

Figure 3 gives examples of the comparison results between our method and IBM Full-Parser, which show four cases that our method is superior to the dictionary-based methods. The correct segmentation is boldfaced, and "<>" and "[]" quoted character sequences show separately the wrong and correct output by IBM Full-parser and our method.

The first two examples contain one location name "止锚湾" (Zhimao Bay) and a Chinese newly proposed social sense "八荣八耻" (Eight-Honors-and-Eight-Disgraces) which are not included in FP's corpus, thus it separates the two proper nouns as independent characters. Example (3) has an idiom which contains a phrase, "苦难" (tribulation), that happens to be an entry in FP's corpus, hence FP separates this idiom into three words. Example (4) shows an ambiguous query "和尚未有". It can either be parsed as "[和尚(monk)] [未有(has not)]" or "[和(and)] [尚未(not yet)] [有(have)]" if no context information (here is "职称" (technical title)) is given. Since FP adopts the maximum matching strategy and "和尚" (monk) is also an entry in its corpus, it takes the former segmentation. Contrarily, leveraging document information and search technology, the context information "职称" is taken into consideration which directs us to select the latter and correct segmentation, as monks never have technical titles.

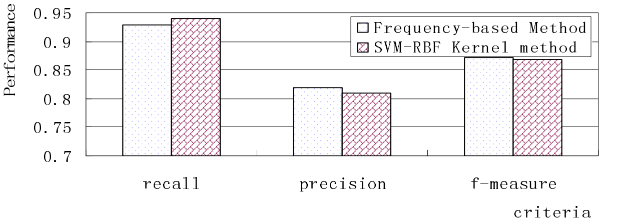


**Figure 2. Evaluation on SIGHAN'05 data with the two different segment scoring methods**

| | |
|---|---|
| (1) 我明天要去[**止锚湾**]玩<br>IBM: 我 明天 要 去 <止 锚 湾> 玩<br>Our: 我 明天 要 去 [止锚湾] 玩 | (2) 胡锦涛说[**八荣八耻**]很重要<br>IBM: 胡锦涛 说 <八 荣 八 耻> 很重要<br>Our: 胡锦涛 说 [八荣八耻] 很重要 |
| (3) 老百姓[**有苦难言**]<br>IBM: 老百姓 <有 苦难 言><br>Our: 老百姓 [有苦难言] | (4) 有职称的和[**尚未**]有职称的<br>IBM: 有 职称 的 <和尚 未有> 职称 的<br>Our: 有 职称 的 和 [尚未] 有 职称 的 |

**Figure 3. Examples of the superiority to IBM Full-Parser**

## 4. CONCLUSIONS

Chinese word segmentation is a widely requested Chinese information processing step. In this paper, we propose a novel solution which leverages the Web data and search technology. It contains three steps: 1) collecting segments from search results, 2) scoring segments, and 3) ranking segmentations. It is good at discovering new words (no OOV problem) and adapting to different segmentation standards, and can be entirely unsupervised which saves labors to labeling training data.

There are many possible future works, such as finding more effective scoring methods, combining current approach to other types of segmentation methods to give a better performance, etc.

## 5. REFERENCES

[1] Gao, J.F., Li, M., Wu A., Huang, C.N. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. Computational Linguistics. MIT Press. 2005.

[2] Platt, J., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Large Margin Classifiers, MIT Press, 1999.

[3] Sproat, R., and Shih C. Corpus-based Methods in Chinese Morphology and Phonology. COOLING, 2002.

[4] Xue, N.W. Chinese Word Segmentation as Character Tagging. Computational Linguistics and Chinese Language Processing. Vol. 8, No. 1, Feb. 2003, pp.29-48.