Chinese Word Segmentation using Various Dictionaries

Guo-Wei Bian
Department of Information Management
Huafan University, Taiwan, R.O.C.
gwbian@cc.hfu.edu.tw

Abstract

Most of the Chinese word segmentation systems utilizes monolingual dictionary and are used for monolingual processing. For the tasks of machine translation (MT) and cross-language information retrieval (CLIR), another translation dictionary may be used to transfer the words of documents from the source languages to target languages. The inconsistencies resulting from the two types of dictionaries (segmentation dictionary and transfer dictionary) may produce some problems for MT and CLIR. This paper shows the effectiveness of the external resources (bilingual dictionary and word list) for Chinese word segmentations.

1 Introduction

Most of the Chinese word segmentations are used for monolingual processing. In general, the word segmentation program utilizes the word entries, part-of-speech (POS) information (Chen and Liu, 1992) in a monolingual dictionary, segmentation rules (Palmer, 1997), and some statistical information (Sproat, *et al.*, 1994). For the tasks of machine translation (MT) (Bian and Chen, 1998) and cross-language information retrieval (CLIR) (Bian and Chen, 2000), another translation dictionary may be used to transfer the words of documents from the source languages to target languages. Because of the inconsistencies resulting from the two types of dictionaries (segmentation dictionary and transfer dictionary), this approach has the problems that some segmented words cannot be found in the transfer dictionary.

In this paper, we focus on the effectiveness of the Chinese word segmentation using different dictionaries. Four different dictionaries (or word lists) and two different testing collections (testing data) are used to evaluate the results of the Chinese word segmentation.

2 Chinese Word Segmentation System

The segmentation system used only the various dictionaries in this design. In this paper, the other possible resources (POS, segmentation rules, word segmentation guide, and statistical information) are ignored to test the average performance between different testing collections specially followed the different segmented guidelines.

The longest-matching method is adopted in this Chinese segmentation system. The segmentation processing searches for a dictionary entry corresponding to the longest sequence of Chinese characters from left to right. The system provided the approximate matching to search a substring of the input with the entry in the dictionary if no total matching is found. For example, the system will segment the input "看著隨時可能結束生命的妹妹" as "看著 隨時 可能 結束 生命 的 妹妹" which matched the term with the entry "看著辦" in dictionary if no entry "看著" found.

2.1 Various Dictionaries

The word segmentation are evaluated using different dictionaries (or word lists) and different testing collections (testing data). There are four dictionaries are used: the first one is converted from an English-Chinese bilingual dictionary, and the other three are extracted from the training corpora.

The original English-Chinese dictionary (Bian and Chen, 1998), which containing about 67,000 English word entries, is converted to a new Chinese-English dictionary (called CEDIC later). There are 125,719 Chinese word entries in this CEDIC.

The terms in the various training corpora (the Sinica Corpus and the City University Corpus) are extracted to build the different word lists as the segmentation dictionaries (called CKIP and CityU later). The tokens starting with the special

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 166–168,
Sydney, July 2006.

characters or punctuation marks are ignored. The following shows some examples:

(, （０２）, （１）, ㎝, $, %, ☆, , ─,
─Ｗｈｙ, ◢Ｍ４５, 【, ○○○, ..., 「, 」
／ｕｓｒ／ｍａｎ, //, ,, , #, .com,

Table 1 lists the number of tokens (#tokens), the number of ignored tokens (#ignored), the number of words (#words), and the unique words (#unique) for each dictionaries. There are 140,971 unique words are extracted from the training collection of Sinica Corpus, and 75,433 respected to the training set of the City University Corpus. These two dictionaries are combined to another dictionary which containing 174,398 unique words.

	#Tokens	#Ignored	#Words	#Unique
CKIP (CK)	5,468,793	894,686	4,574,107	140,971
CityU (CT)	1,643,421	257,032	1,386,389	75,433
CKIP+CityU (CK + CT)	7,112,214	1,151,718	5,960,496	174,398

Table 1. Statistical Information of the Extracted Dictionaries

3 Experimental Results

To evaluate the results of Chinese word segmentations, we implement 8 experiments (runs) using the 4 different dictionaries (CEDIC, CK, CT, and CK+CT) mentioned in previous section. Two test collections (the Sinica Corpus and the City University Corpus) are used to measure the precision**,** recall**,** and an evenly-weighted F-measure for the Chinese words segmentations.

Table 2 shows the F-measure of the experimental results, and the Figure 1 illustrates the comparisons of the segmentation performances. The symbol (*) indicates that the run is a closed test, which only uses the training material from the training data for the particular corpus. We can find that the larger dictionary (CK+CT) produces better segmentation results even the word lists are combined from the different resources (corpora) and followed the different guidelines of word segmentations.

	CEDIC	CK + CT	CK	CT
CKIP	0.710	0.695	0.692*	0.611
CityU	0.481	0.589	0.547	0.513*

Table 2. The F-measure results of segmentation performances using various dictionaries (*: closed test)

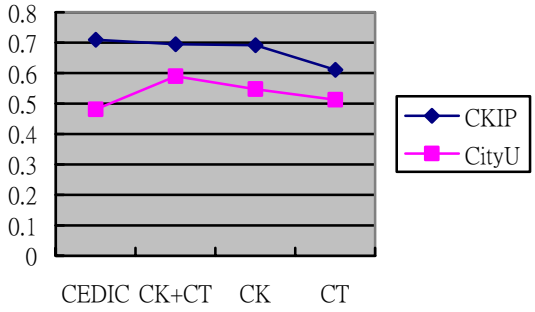

Figure 1. The comparison of segmentation performances using various dictionaries (*: close test)

3.1 Error Analysis

3.1.1 Format Error of Result File

The results file for word segmentation is required to appear with one line for each sentence/line in the test file with words and punctuation separated by whitespace. Our system makes some mistakes to produce no whitespace before English terms and Arabic numbers, and produce no whitespace after Chinese punctuation marks. This formatting problem has made many adjacent segmented words to be evaluated as errors. A sentence with such errors is listed below

(Our Answer)
與 大珠三角 相鄰 、相互 間 經貿 關係 密切 的 福建 、江西 、湖南 、廣東 、廣西 、海南 、四川 、貴 州 、雲南 9 個 省區 ，以及 香港 、澳門 特別 行政區 （簡 稱 “9+2 ”） 。

(Standard)
與 大珠三角 相鄰 、 相互 間 經貿 關係 密切 的 福建 、 江西 、 湖南 、 廣東 、 廣西 、 海南 、 四川 、 貴州 、 雲南 9 個 省區 ， 以及 香港 、 澳門 特別 行政區 （ 簡稱 “ 9+2 ” ） 。

The standard answer of the testing collection (CityU) of the City University Corpus has 7,512 sentences and 220,147 words. The total number of English terms, Arabic numbers, and Chinese punctuation marks is 37,644. Such formatting problem makes the error rate of about 30% for the City University Corpus.

3.1.2 Different Viewpoints of Segmentations

In our experiments, there are different word lists extracted from the different training corpora. Some errors are produced because of the differ-

ent results of word segmentations in the training corpora according to the different guidelines. Table 3 shows some different results. The first column (CKIP) is the standard answer of the testing collection of Sinica Corpus, and the second column (HFUIM) is our answer. The third and fourth columns are the words with their frequencies appeared in the training collections of Sinica Corpus and City University Corpus. For example, our system produces the word "心中", but the standard answer of Sinica Corpus is "心" and "中". However, the word "心中" appear 61 times in the training collection of City University Corpus.

CKIP	HFUIM	CKIP-Training	CityU-Training
林婦	林 婦	林婦 (0)	
整 夜	整夜	整 (1839) 夜 (366)	整夜 (2)
看 著	看著	看著辦 (4) 眼看著 (20)	
心 中	心中	心 (2551) 中 (16694)	心中 (61)
這 個	這個	這 (32409) 個 (39558)	這個 (714)
死 後	死後	死 (984) 後 (7967)	死後 (18)
所 需	所需	所 (9012) 需 (963)	所需 (35)

Table 3. The Different Segmentation Results

3.1.3 Inconsistency of Word Segmentation

Some errors of word segmentations are reported because of the inconsistency of word segmentations. The following shows such a problem. For example, the word "還有" appears 317 times in the training data, but it has been treated as two terms ("還" and "有") 19 times in the golden standard of the testing data.

(Training data)

- 歐盟 委員會 設置 等 問題 上 還有 一些 不同 聲音

(Golden Standard)

- 目前 兩 地 機場 還有 一些 商業 問題 要 談
- 鍾麗緹 透露 身上 還 有 一 個 紋身 圖案
- 還 有 其他 歐國盃 的 有趣 專題 ， 萬勿錯過 已 出版 的 《 明報 歐洲 國家盃 特刊 》 。

4 Conclusion

In this paper, we discuss the effectiveness of the Chinese word segmentation using various dictionaries. In the experimental results, we can find that the larger dictionary will produce better segmentation results even the word lists are combined from the different resources (corpora) and followed the different guidelines of word segmentations. Some results show that the external resource (e.g., the bilingual dictionary) can perform the task of Chinese word segmentation better than the monolingual dictionary which extracted from the training corpus.

Reference

Bian, G.W. and Chen, H.H. (2000). "Cross Language Information Access to Multilingual Collections on the Internet." *Journal of American Society for Information Science & Technology (JASIST), Special Issue on Digital Libraries*, 51(3), 2000, 281-296.

Bian, G.W. and Chen, H.H. (1998). "Integrating Query Translation and Document Translation in a Cross-Language Information Retrieval System." *Machine Translation and the Information Soap (AMTA '98),* D. Farwell, L Gerber, and E. Hovy (Eds.), Lecture Notes in Computer Science, Vol. 1529, Springer-Verlag, pp. 250-265, 1998

Chen, K.J and Liu, S.H (1992), "word identification for Mandarin Chinese sentences" Proceedings of the 14th conference on Computational linguistics, pp. 101-107, France, 1992

Palmer, D. (1997), "A trainable rule-based algorithm for word segmentation", Proceeding of ACL'97, 321-328, 1997.

Sproat, R., *et al*. (1994) "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese", *Proceeding of 32nd Annual Meeting of ACL*, New Mexico, pp. 66-73.