An Example-Based Chinese Word Segmentation System for CWSB-2

Chunyu Kit **Xiaoyue Liu**
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Tat Chee Ave., Kowloon, Hong Kong
{ctckit, xyliu0}@cityu.edu.hk

Abstract

This paper reports the example-based segmentation system for our participation in the second Chinese Word Segmentation Bakeoff (CWSB-2), presenting its basic ideas, technical details and evaluation. It is a preliminary implementation. CWSB-2 valuation shows that it performs very well in identifying known words. Its unknown word detection module also illustrates great potential. However, proper facilities for identifying time expressions, numbers and other types of unknown words are needed for improvement.

1 Introduction

Word segmentation is to identify lexical items, especially individual word forms, in a text. It involves two fundamental tasks, both aiming at minimizing segmentation errors: one is to infer out-of-vocabulary (OOV) words, also known as unknown (or unseen) word detection, and the other to identify in-vocabulary (IV) words, with an emphasis on disambiguation. OOV words and ambiguities are the two major causes to segmentation errors.

Accordingly，word segmentation approaches can be divided into the categories summarized in Table 1 in terms of the resources in use to tackle these two causes. The closed and open tracks in CWSB correspond, respectively, to the last two categories, both involving inferring OOV words beyond disambiguating IV words. Word discovery and OOV word detection pursue a similar target, i.e., inferring new words. The continuum connecting them is the size of the lexicon in use: the former assumes few words known and the latter an existing lexicon to some scale. Inferring new words is an essential task in word segmentation, for a complete lexicon is rarely a realistic assumption in practice.

Category	Resource in use		Major Task	
	Lexicon	Tr. Corpus	OOV	Disamb.
WD[a]	-	(-)[b]	+	
WS/CL[c]	+	-	-	+
WS/IL[d]	+	-	+	+
WS/TC[e]	(+)[f]	+	+	+
WS/TC+L[g]	+	+	+	+

[a]Word discovery, or unsupervised lexical acquisition
[b]Input data is used for unsupervised training
[c]Word segmentation with a complete lexicon
[d]Word segmentation with an incomplete lexicon
[e]Word segmentation with a pre-segmented training corpus
[f]It can be extracted from the given training corpus.
[g]Word segmentation with a pre-segmented training corpus and an extra lexicon

Table 1: Categories of segmentation approach

This paper presents our segmentation system for participation in CWSB-2. It takes an example-based approach to recognize IV words and follows description length gain (DLG) to infer OOV words in terms of their text compression effect. Sections 2 and 3 below introduce the example-based and DLG-based segmentation respectively. Section 4 presents a strategy to combine their strength and Section 5 reports our system's performance in CWSB-2. Following error analysis in Section 6, Section 7 concludes the paper.

2 Example-based segmentation

How to utilize as much information as possible from the training corpus to adapt a segmentation system towards a segmentation standard has been a critical issue. Kit et al. (2002) and Kit et al. (2003) attempt to integrate case-based learning with statistical models (e.g., n-gram) by extracting transformation rules from the training corpus for disambiguation via error correction; Gao et al. (2004) adopt a similar strategy for adaptive segmentation, with transformation templates (instead of case-based rules) to modify word boundaries (instead of individual words).

The basic idea of example-based segmentation is very simple: existing pre-segmented strings in training corpus provide reliable examples for segmenting similar strings in input texts. In contrast to dictionary checking for locating *possible* words in an input sentence to facilitate later segmentation operations, pre-segmented examples give *exact* segmentation to copy.

The example-based segmentation can be implemented in the following steps.

1. Find all exemplar pre-segmented fragments, with regards to a training corpus, and all possible words, with regards to a lexicon, from each character in an input sentence;
2. Identify the optimal sequence, among all possibilities, of the above items over the sentence following some optimization criterion.

If adopting the minimal number of fragments or words in a sequence as optimization criterion, we have a maximal matching approach to word segmentation. However, it differs remarkably from the previous maximal matching approaches: it matches pre-segmented fragments, instead of dictionary words, against an input sentence. It can be carried out by a best-first strategy: repeatedly select the next longest example or word until the entire sentence is properly covered. Unfortunately, the best-first approach does not guarantee to give the best answer. For CWSB-2, we implemented a program following the Viterbi algorithm to perform a complete search in terms of the number of fragments, and then words, in a sequence.

However, a serious problem with this example-based approach is the *sparse data* problem. Long exemplar fragments are more reliable but small in number, whereas short ones are large in number but less reliable. In the case of no exemplar fragment available for an input sentence, this approach draws back to the maximal match segmentation with a dictionary. How to incorporate statistical inference into example-based segmentation to infer more reliable optimal segmentation beyond string matching remains a critical issue for us to tackle.

3 DLG-based segmentation

DLG is formulated in Kit and Wilks (1999) and Kit (2000) as an empirical measure for the compression effect of extracting a substring from a given corpus as a lexical item. DLG optimization is applied to detect OOV words for our participation in CWSB-2. It works as follows in two steps.

1. Calculate the DLG for all known words and all new word candidate (i.e., substrings with frequency ≥ 2, preferably, in the test corpus), based on frequency information in the training and the test corpora;
2. Find the optimal sequence of such items over an input sentence with the greatest sum of DLG.

Step 2 above in our system re-implements only the first round of DLG-based lexical learning in Kit (2000). It is implemented by the same algorithm as the one for example-based segmentation, with DLG as optimization criterion. Evaluation results show that this learning-via-compression approach discovers many OOV words successfully, in particular, person names.

4 Integration

The example-based segmentation is good at identifying IV words but incapable of recognizing any new words. In contrast, the DLG-based segmentation performs slightly worse but has potential to detect new words. It is expected that the strength of the two could be combined together for performance enhancement.

However, because of inadequate time we had to take a shortcut in order to catch the CWSB-2 deadline: DLG segmentation is only applied to recognize new words among the sequences of mono-character items in the example-based segmentation output.

Track	P	R	F	OOV	R_{OOV}	R_{IV}
AS_c	.944	.902	.923	.043	.234	.976
PKU_c	.929	.904	.916	.058	.252	.971
MSR_c	.965	.935	.950	.026	.189	.986

Table 2: System performance in CWSB-2

5 Performance

Our group took part in three closed tracks in CWSB-2, namely, AS_c, PKU_c and MSR_c, with a preliminary implementation of the example-based word segmentation presented above. Our system's performance in terms of CWSB-2's official scores is presented in Table 2. Its R_{OOV} scores look undesirable, showing that applying the first round of DLG-based segmentation to sequences of mono-character items is inadequate for the OOV word discovery task. Nevertheless, its R_{IV} scores are, in general, quite close to the top systems in CWSB-2, although it does not have a disambiguation module to polish its maximal matching output.

However, this is not to say that the DLG-based segmentation deserves no credit in unknown word detection. It does recognize many OOV words, as shown in Table 3. The low R_{OOV} rate has to do with our system's incapability in handling time expressions, numbers, and foreign words.

6 Error analysis

Most errors made by our system are due to the following causes: (1) no knowledge, overt or implicit, in use for recognizing time expressions, numbers and foreign words, as restricted by CWSB-2 rules, (2) a premature module for OOV word detection, (3) no further disambiguation besides example application, and (4) significant inconsistency in the training and test data.

The inconsistency exists not only between the training and test corpora for each track but, more surprisingly, also within individual training corpora. Some suspected cases are illustrated in Tables 4, 5 and 6. They are observed to be in a large number in the CWSB-2 corpora. Scoring with a golden standard involving so many of them appears to be problematic, for it penalizes the systems for handling such cases right and rewards the others for producing "correct" answers. What is even more worth noting is that (1) an inconsistent case involves more than one word, and (2) the difference between a correct and an erroneous judgment of a word is 1, in a sense, but the difference between one system that loses it for doing right and another that earns it by doing wrong is surely greater.

AS_c: 小森(106) 左詩雅(45) 馮禮(31) 霍金(29) 瑪雅(21) 丹紐(20) 張庭(18) 英特爾(17) 黑盤鮑(16) 沈時華(15) 證期會(13) 厲莉(12) 黃銘坤(11) 卡門(11) 佛指舍利(11) 陳慶浩(10) 璩女(9) 范曉萱(8) 郭女(8) 邱太三(7) 網站(7) 狄亞尼(7) 張惠妹(6) 米田真澄(6) 鄭儉(5) 丁克華(5) 經發會(5) 蘇盈貴(5) 陳秀玲(5) 八美圖(5) 房貸(5) 蔡信弘(5) 郭玉鈴(4) 地型(4) 御膳(4) 白藜蘆醇(4) 羅麗泰(3) 張玨(3) 米仓涼子(3) 陳兆伸(2) 羅勒(2) 幼發拉底河(2) 溥傑(2) ·····

PKU_c: 罢免(38) 世清(23) 任免(21) 拉姆斯菲尔德(20) 哈苏(19) 海合会(17) 军级(16) 加纳(15) 水心村(12) 网友(11) 小阜村(10) 宋双(10) 吐逊江(9) 申奥(9) 库福尔(8) 金税(8) 亚布力(8) 金门(7) 三阳镇(7) 教主(6) 罢免书(6) 法轮功(6) 香客(6) 普京(6) 福清市(5) 柯尔克孜(5) 华能(5) 罗林斯(5) 松诺斯(5) 洋务(5) 马祖(5) 甲午(5) 哈斯达(4) 奥申委(4) 高唐县(4) 文津(4) 门市部(4) 妈祖(4) 刁民(4) 闭会(4) 商机(4) 临武县(4) 湄洲岛(4) 圈套(4) 法塔赫(3) 流星雨(3) 堆龙德庆县(3) 阿族(3) 哥德堡(3) 信众(3) 钱柜(3) 衡水(3) 内坦亚(3) 双丰村(3) ·····

MSR_c: 博古(26) 玲英(19) 游景玉(19) 任尧森(17) 秦机(15) 亚仿(14) 进占(14) 刘积仁(13) 东宝(13) internet(12) 猴王(12) 双保(11) 张肇群(10) 抚州(10) 南丁格尔(10) 彭珮云(10) 海塘(10) 王常力(9) 穆守家(9) 大关村(8) 赞皇(8) 张钧(8) 嶂石岩(7) 三老四严(7) 局域网(7) 透支(6) 秦家山(6) 东软(6) 后金(6) 八旗(6) 弄堂(5) 黎秀芳(5) 提灌(5) 大关(5) 王丙乾(5) 棉铃虫(5) 百亿次(4) 西文(4) 米夫(4) 陆冰(4) 郑守仁(4) 秦邦宪(4) 关小瑛(3) 二进制(3) 顺延(3) 李朋朋(3) 瞿秋白(3) 黄华平(3) 汪赛进(3) 汪延(3) 何元亮(3) 陆佑楣(3) 阜平县(3) 中信所(3) 导流(3) 徐殿龙(3) 重男轻女(2) ·····

Table 3: Illustration of new words successfully detected, with frequency in parentheses

7 Conclusions

In the above sections we have reported the example-based word segmentation system for our participation in CWSB-2, including its basic ideas, technical details and evaluation results. It has illustrated an excellent performance in IV word identification and nice potential in OOV word discovery. However, its weakness in handling time expressions, numbers and other types of unknown words has hindered it from performing better. We are expecting to implement a full-fledged version of the system for improvement.

Acknowledgements

The work described in this paper was supported by the RGC of HKSAR, China, through the CERG grant 9040861. We wish to thank Alex Fang and Robert Neather for their help.

Training & Answer	f_T/f_A	Golden Standard	f_T/f_A
繁殖 場	4/8	繁殖場	0/0
老 歌	28/7	老歌	0/0
準備 率	5/7	準備率	0/0
小 火車	11/6	小火車	0/0
本 文	186/5	本文	0/0
第三 者	41/4	第三者	0/0
新 一 代	29/4	新一代	0/0
稱 之 為	129/4	稱之為	0/0
夠 不 夠	23/3	夠不夠	0/0
短 時間	47/3	短時間	0/0
中正 機場	33/2	中正機場	0/0
台北 市長	32/2	台北市長	0/0
民意 代表	85/2	民意代表	0/0
清醒過來	10/2	清醒 過來	0/0
統治 者	62/2	統治者	0/0
新 生命	23/2	新生命	0/0
也就是說	192/1	也 就是 說	0/0
不景氣	149/1	不 景氣	0/0
中央 政府	66/1	中央政府	0/0
挑戰性	31/1	挑戰 性	0/0
看 電影	80/1	看電影	0/0
研究 者	68/1	研究者	0/0
面 無 表情	13/1	面無表情	0/0
曾文 水庫	13/1	曾文水庫	0/0
無 意識	20/1	無意識	0/0
無 話 不 談	6/1	無話不談	0/0
另 一 面	29/1	另一面	0/0
紅透 半 邊 天	4/1	紅透半邊天	0/0
民間團體	24/7	民間 團體	25/0
女性主義	17/3	女性 主義	53/0
是 不 是	1201/2	是不是	2/0

Table 4: Some inconsistent cases in AS corpus

Training & Answer	f_T/f_A	Golden Standard	f_T/f_A
乡政府	14/26	乡 政府	0/0
区政府	6/1	区 政府	0/0
镇政府	5/21	镇 政府	0/0
世纪之交	24/19	世纪 之 交	0/0
改 为	23/18	改为	0/0
生产总值	66/15	生产 总值	0/0
个人所得税	10/9	个人 所得税	0/0
无党派人士	10/5	无 党派 人士	0/0
新年伊始	45/5	新年 伊始	0/0
组成部分	42/5	组成 部分	0/0
固定资产	27/4	固定 资产	0/0
世界各地	21/4	世界 各地	0/0
乡镇企业	126/4	乡镇 企业	0/0
至关重要	20/4	至关 重要	0/0
综合国力	15/4	综合 国力	0/0
总参谋长	25/4	总 参谋长	0/0
经济社会	25/3	经济 社会	0/0
劳动生产率	13/3	劳动 生产率	0/0
去年底	32/3	去年 底	0/0
世界纪录	30/3	世界 纪录	0/0
无党派	11/3	无 党派	0/0
议事日程	15/3	议事 日程	0/0
中小企业	22/3	中小 企业	0/0
不同 于	11/2	不 同于	0/0
此事	25/2	此 事	0/0
火树银花不夜天	3/1	火树银花 不 夜 天	0/0
雨夹雪	13/1	雨 夹 雪	0/0
县政府	24/5	县 政府	1/0
意味着	49/4	意味 着	1/0
很多	112/3	很 多	14/0
发达国家	48/1	发达 国家	1/0

Table 5: Some inconsistent cases in PKU corpus

Training & Answer	f_T/f_A	Golden Standard	f_T/f_A
营业 部	12/7	营业部	0/0
高性能	16/6	高 性能	0/0
“ 八五	29/5	“八五	0/0
大型 企业集团	6/3	大型企业 集团	0/0
径流 量	3/3	径 流量	0/0
海信电器	1/2	海信 电器	0/0
投资银行?	4/2	投资 银行	0/0
一 版	10/2	一版	0/0
浙东	3/2	浙 东	0/0
２０００ 多 人	7/1	２０００多人	0/0
采访团	2/1	采访 团	0/0
产油国 家	1/1	产 油 国家	0/0
第一、二、三?	4/1	第一 、 二 、 三	0/0
葛洲坝大江	1/1	葛洲坝 大江	0/0
国际互联网 络	1/1	国际 互联网络	0/0
计算机 应用软件	1/1	计算机应用 软件	0/0
家 门口?	10/1	家门 口	0/0
两点?	16/1	两 点	0/0
吕梁?	4/1	吕 梁	0/0
一口气?	16/1	一 口 气	0/0
中国人民	122/1	中国 人民	0/0
中外合资 经营	3/1	中外 合资经营	0/0

Table 6: Some inconsistent cases in MSR corpus

References

E. Brill. 1993. *A Corpus-Based Approach to Language Learning*. PhD thesis, University of Pennsylvania, Philadelphia.

J. Gao, A. Wu, M. Li, C. Huang, H. Li, X. Xia and H. Qin. 2004. Adaptive Chinese word segmentation. In *ACL-04*. Barcelona, July 21-26.

C. Kit and Y. Wilks. 1999. Unsupervised learning of word boundary with description length gain. In M. Osborne and E. T. K. Sang (eds.), *CoNLL-99*, pp.1-6. Bergen, Norway, June 12.

C. Kit 2000. *Unsupervised Lexical Learning as Inductive Inference*. PhD thesis, University of Sheffield.

C. Kit, H. Pan and H. Chen. 2002. Learning case-based knowledge for disambiguating Chinese word segmentation: A preliminary study. *SIGHAN-1*, pp.33－39. Taipei, Sept. 1, 2002.

C. Kit, Z. Xu and J. J. Webster. 2003. Integrating n-gram model and case-based learning for Chinese word segmentation. In Q. Ma and F. Xia (eds.), *SIGHAN-2*, pp.160-163. Sapporo, 11 July, 2003.

D. Palmer. A trainable rule-based algorithm for word segmentation. In *ACL-97*, pp.321-328. Madrid.