Semi-supervised Chinese Word Segmentation for CLP2012

Sai-ke He
State Key Laboratory of Management
and Control for Complex Systems
Institute of Automation,
Chinese Academy of Sciences
Beijing 100190 China
saike.he@ia.ac.cn

Song-xiang Cen
Baidu, Inc Baidu Campus, No. 10,
Shangdi 10th Street Haidian District,
Beijing 100085 China
censongxiang@baidu.com

Nan He
Nuance Software Technology
(Beijing) Co., Ltd.
hn.ft.pris@gmail.com

Jun Lu
State Key Laboratory of Management
and Control for Complex Systems
Institute of Automation,
Chinese Academy of Sciences
Beijing 100190 China
lujun_tiger@hotmail.com

Abstract

Chinese word segmentation (CWS) lays the essential foundation for Mandarin Chinese analysis. However, its performance is always limited by the identification of unknown words, especially for short text such as Microblog. While local context are helpless in handling unknown words, global context do manifest enough contextual information, and could be used to guide CWS process. Based on this motivation, in this paper, we report our attempt toward building an integrated model in semi-supervised manner. Considering the complexity of model, we design a strategy to manipulate global and local contextual information asynchronously. Though the coverage of unknown words by such integrated model is still small, official results from CLP2012 present promising result.

1 Introduction

Essentially, Chinese is a kind of paratactic language, rather than hypotactic language. This makes it character based, not word based. However, words are the basic linguistic units of natural language. Thus, the identification of lexical words or the delimitation of words in running texts is a prerequisite in Chinese natural language processing (NLP).

Chinese word segmentation can be cast as simple and effective formulation of character sequence labeling. A number of recent papers have examined this problem (Zhang et al., 2003; Xue, 2003; Peng et al., 2004) and could provide relatively good performance. However, these systems are genre or domain specific and use many different segmentation guidelines derived from the training dataset. This characteristic guarantees these systems with good performance on the known words, yet severely deteriorates on unknown words[1] from relatively unfamiliar context. This constitutes the major drawback of supervised segmentation.

In contrast, unsupervised approaches are model-free and more adaptive to unfamiliar context. This provides a potential solution for identify unknown words and have been attracting more attention recent years (Sproat and Shih, 1990; Feng et al., 2004; Goldwater et al., 2006; Mochihashi et al., 2009).

Since super and unsupervised methods excel in different situations, a natural idea would be a combination of these two to overcome drawbacks of both. A myriad of attempts exist and can be roughly categorized into two groups: simultaneous and asynchronous manner.

In simultaneous design, most researchers bind to the theory of transfer learning (or multitask learning, Caruana, 1997), and believe it achieves

[1] Unknown words also refer to out-of-vocabulary (OOV) words in some literature.

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 79–84,
Tianjin, China, 20-21 DEC. 2012

more when all the tasks are solved together. Admitted, this may be true in some situations (Gao et al., 2005; Tou Ng and Low. 2004). However, these achievements are often gained in the cost of complex system design. On the other side, asynchronous system moderate well between performance and simplicity. Thus, it is more favorable for large data processing, especially when real time analysis is primal.

In this paper, we report the integrated system designed for CLP2012 Micro-blog word segmentation subtask[2]. Considering simplicity, we are intended to provide a semi-supervised methodology by execute supervised and unsupervised segmentation asynchronously. In addition, we also design strategies to deal with unknown words: (1) beyond the coverage of training dataset (2) or without obvious segmentation guidelines.

The rest of the paper is organized as follows: Section 2 reviews previous work in the literature. Section 3 describes our integrated framework of CWS in detail. Section 4 presents and analyzes our experimental results. Finally, we conclude the work in Section 5.

2 Related Work

There is a line of research on solving Chinese Word Segmentation in supervised manner. Zhang et al. (2003) use a hierarchical hidden Markov Model (HMMs) to incorporate lexical knowledge. As an advance in this area, Xue (2003) uses a sliding-window maximum entropy classifier to label Chinese characters with one of four position tags, and then covert these labels into final segmentation using rules. Recently, Conditional Random Fields (CRFs) (Lafferty et al., 2001) have been successfully employed in CWS and achieve the state-of-the-art performance (Peng et al., 2004).

At the same time, unknown words gradually develop to be a serious problem that curbs the performance of CWS. As supervised method cannot help much in this situation, researchers begin to resort to new approaches.

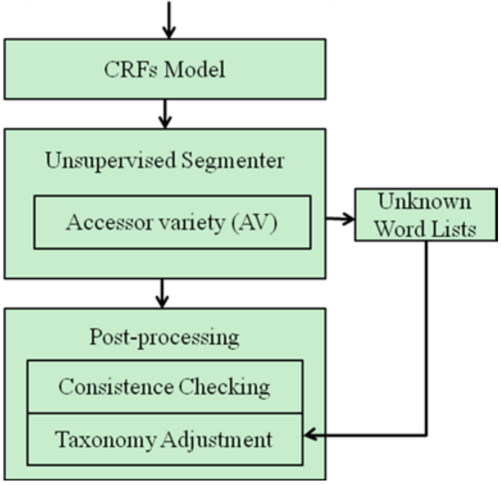
Since Sproat and Shih (1990) introduced mutual information (MI) to word segmentation, there emerges a new line of research on unsupervised approaches. Unsupervised CWS systems tend to use three different types of information: the cohesion of the resulting units (Sproat and Shih, 1990), the degree of separation between the resulting units (Feng et al., 2004, Zhao and Kit, 2008) and the probability of a segmentation given a string (Goldwater et al., 2006; Mochihashi et al., 2009).

As unsupervised approaches can cooperate with supervised ones to achieve mutual enhancement, combination strategies of these two forms the trend. Gao et al. (2005) believe word boundary disambiguation and unknown word identification are not separable in nature, and solve them simultaneously in a pragmatic framework. Mao et al. solve CWS in a by using CRFs and transformation-based error-driven learning (TBL) in a cascaded manner. Evaluation results from Bakeoff-04[3] demonstrate their approach's effectiveness.

3 Framework of CWS

In this section, we define our framework of CWS in three steps (as shown in Figure 1). First, we train a CRFs model based on dataset from Bakeoff-04. This base model is used to segment known words in traditional manner. Then, we use an unsupervised approach to mine out unknown words from the training dataset. Those words will subsequently be used to adjust the segmentation results from CRFs model. Finally, to meet the need from CLP 2012, we also adjust previously segmentation results in the post-processing phase. Those three steps will be illustrated in detail in the following part.

Figure 1: Flow chart of CWS system



3.1 Conditional random fields

Although Chinese Word Segmentation can be solved in many ways, for sequence labeling,

[2] http://www.cipsc.org.cn/clp2012/task1.html

[3] http://www.china-language.gov.cn/bakeoff08/bakeoff-08_basic.html

conditional random fields offer advantages over both generative models like HMMs and classifiers applied at each sequence position (Sha and Pereira, 2003). CRFs are an undirected graph established on G = (V, E), where V is the set of random variables $Y = \{Y_i|1\leqslant i\leqslant n\}$ for each the n tokens in an input sequence and $E = \{(Y_{i-1}, Y_i) |2\leqslant i\leqslant n\}$ is the set of (n-1) edges forming a linear chain. Following (Lafferty et al., 2001), the conditional probability of the state sequence $(s_1,s_2 \ldots s_n)$ given the input sequence $(o_1, o_2 \ldots o_n)$ is computed as follows:

$$P_\Lambda(s|o) = \frac{1}{Z_o}\prod_{c\in C(s,o)} exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}, s_t, o, t)\right) \quad (1)$$

where f_k is an arbitrary feature function; and λ_k is the weight for each feature function; it can be optimized through iterative algorithms like GIS (Darroch and Ratcliff, 1972). Recent research indicates that quasi-Newton methods such as L-BFGS (Byrd and Schnabel, 1994.) are more effective than GIS.

3.2 Tag set

As justified in (Zhao et al., 2007; Zhao et al., 2008), a 6-tag set enables the CRFs learning of character tagging to achieve a better segmentation performance than others. So we adopt this tag set in our CWS framework, namely, B, B2, B3, M, E and S, which respectively indicates the start of a word, the second position within a word, the third position within a word, other positions within a word, and the end of a word. An example is illustrated in Table 1.

Word Length	Tag sequence for a word
1	S
2	BE
3	BB2E
4	BB2B3E
5	BB2B3ME
>=6	BB2B3M … ME

Table 1: Illustration of 6-tag format in CWS

3.3 Feature templates

Table 2 illustrates the features we used in our CWS systems. Where *C* represents character; subscript *n* indicates its relative position taking the current character as its reference; *Pun* derives from the property of the current character: whether it is a punctuation; *T* describes the type of the character: numerical characters belong to class 1, characters whose meanings are date and time represent class 2, English letters represent class 3, punctuation labels represent class 4 while other characters represent class 5. In addition, the tag bi-gram feature is also employed.

Type	Feature
Unigram	C_n(n=-2,-1,0,1,2)
Bigram	C_nC_{n+1}(n=-2,-1,0,1)
Jump	$C_{-1}C_1$
Punctuation	Pun(C0)
Date,Digit,letter	$T_{-1}\ T_0T_1$

Table 2: The features used in CWS systems

3.4 Unsupervised segmentation

Due to the inherent Markovian assumption, sequence models, including CRFs, could only capture local structure, and thereby encode local context, i.e. labels directly depend only on the labels and observations within small window around them. This constraint hinders us from exploiting the global contextual information presents in natural language, such as information concerning label assigned at a long distance from a given character string, or even crucial textual information from the whole text.

Such global contextual information play key roles in two-fold: (1) serves to warrant that same or similar character sequences receive the same segmentation label; (2) enhance weak context by leveraging contextual information globally – essential to unknown word detection. Thus, to capture and utilize global contextual information, we employ an unsupervised segmentation approach in our system, as described below.

In Chinese text, each substring of a whole sentence can potentially form a word, but only some substrings carry clear meanings and thus form a correct word. Accessor variety (AV), sparked by (Feng, 2004) is used to evaluate how independent a string is from the rest of the text. The more independent it is, the higher the possibility that it is a potential word carrying a certain kind of meaning. The accessor variety value (AV value) of a string s is defined as:

$$AV(s) = \min\{Lav(s), Rav(s)\} \quad (2)$$

where Lav(s) is the left accessor variety of s, which is defined as the number of its distinct predecessors, plus the number of distinct sentences in which s appears at the beginning, while Rav(s) is the right accessor variety of s, which is defined as the number of its distinct successors, plus the number of distinct sentences in which s appears at the end.

Given the definition in formula (2), the segmentation problem is then cast as an opti-

LW	Lexical Word	教授,朋友,高兴,吃饭
MDW	Morphologically Derived Word	
MP_, MS_	Affixation (Prefix, Suffix)	朋友们
MR_	Reduplication	高高兴兴
ML_	Splitting	吃了饭
MM_	Merging	上下班
MHP_	Head + Particle	走出去
FT	Factoid word	
Dat	Date	1983年, 10月11日
Dur	Duration	2个月
Tim	Time	12点30分
Per	Percent and fraction	百分之十 , 1/4
Mon*	Money	1000(美元)
NUMBER*	Frequency, integer, decimal, ordinal, rate, etc.	(每秒)5(次), 33.8, 第一(届), 三比三
MEASURE*	Age, weight, length, area, capacity, speed, temperature, angle, etc.	二十二(岁), 19(摄氏度), 360(米), 600(公顷)
Ema	E-mail	annoymous@sina.com
Pho	Phone, fax, telex	(0086)12345678
WWW	WWW	http://weibo.com
NE*	Named Entity	
P	Person name	白(岩松) 杨(幂)
L	Location name	天河(体育场)
O	Organization name	新闻(纵横), 百度
NW	New Word	三通, 非典

Table 3: Taxonomy of Chinese words used in CLP2012
* indicates adjustment specified for CLP2012 subtask. Note, pair-wised brackets represent delimitation among character strings here, yet such delimitation rule may not hold under other segmentation guidelines.

Category	Original Words	Gazetter Words		Volume	
		First Name	Last Name	First Name	Last Name
Person name	刘翔, 吴奇隆, 司马义 …	吴, 刘,司马, 吴刘, 刘吴 …	翔, 奇隆, 义 …	4138	7326
Location name	涿州市, 广西壮族自治区 …	涿州, 市, 广西, 壮族, 自治区 …		66461	
Organization name	剑桥大学, 社区管理委员会 …	剑桥, 大学, 社区, 管理, 委员会 …		21351	

Table 4: Gazetteer collected for
For person names, we mainly statistic elites from China, Japan, Europe, and Northern America.

mization problem to maximize the target function of the AV value over all word candidates in a sentence. The target function takes two factors: the segment length and the corresponding AV value. Theoretically, the choice of target function is arbitrary. Here, we choose polynomial function for its simplicity yet good generalize ability.

Since the value of each segment can be computed independently from the other segments in the same sentence, the optimal segmentation strategy for a sentence can be computed using a dynamic programming technique, in which the time complexity is linear to sentence length. After this procedure, we can obtain a plausible segmentation of the text as well as candidate unknown word lists.

3.5 Post-processing

In the pos-processing phase, we mainly utilize two techniques: consistence checking and taxonomy adjustment.

Consistence Checking: Label inconsistency is ubiquitous in context with great variance, especially in short text scenarios. To solve this problem, we use consistency checking inspired by (Ng and Low, 2004). The mechanism is to guarantee same word stings occur at different places labeled consistently. To this aim, we design the following rule：

Class-majority: Assign the majority label to the token sequence which is matched with the potential word list exactly. This rule enables us to capture the long distance dependencies between identical words, so that the same candidate words of different occurrences can be recalled favorably.

Taxonomy Adjustment: In taxonomy adjustment, we develop a taxonomy redefined from (Gao et al., 2005) where Chinese words are categorized into five types: lexicon words (LW), morphologically derived words (MDW), factoids (FT), named entities (NE), and new words (NW)[4]. The detail is shown in Table 3.

In taxonomy adjustment, we carry out a fine-tuned design.

For words following into category LW, MDW, and NW, we mainly use the semi-supervised method introduced previously.

Results Run ID	Precision Rate	Recall Rate	F Score	#Total Correct Sentences	Ratio of Correct Sentences
Our Result	0.9195	0.9085	0.914	1414	28.28%
Best	0.946	0.9496	0.9478	2244	44.88%

Table 5: Evaluation Results
‘Best’ indicate the high score achieved in CLP2012 Micro-blog word segmentation subtask.

For those belongs to FT, we rely on rule-based method, which could be considered as a simplified version of deterministic finite automaton (DFA) approach (Sipser, 1997). For each subgroup from FT, we design segmentation rules accordingly. To avoid conflicts among these rules, they are launched in a cascaded manner with dedicatedly specified execution order.

For those belong to NE, we use a dictionary matching method and collect word lists for each subgroups from NE category accordingly. However, words collected in this manner could not be used directly in exact matching way, for this is not the segmentation granularity needed for CLP2012. To solve this conflict, we further segment the collected words into more subtle linguistic units, as exemplified in Table 4.

4 Evaluation Results

This section reports the experiment result based on CWS corpora from CLP2012 Micro-blog word segmentation subtask. The corpora consists of 5000 messages crawled from Sina Weibo[5], a Twitter-like Micro-blog system in China. All the corpora are simplified Chinese text encoded in UTF-8 format. Table 5 lists the official results.

5 Conclusions

In this paper, we report our work on CLP2012 Micro-blog word segmentation subtask. Specific to the characteristics of short text, we design our system in three steps. First, we train a statistical model to mainly segment known words. Then, we utilize an unsupervised segmentation method to indentify unknown words. Third, for the words beyond knowledge of the training data, we employed a dictionary based approach. Generally, our system design is easy to implement and presents good segmentation results.

References

Haodi Feng, Kang Chen, Xiaotie Deng, and Weiming Zheng. 2004. Accessor variety criteria for Chinese word extraction. *Computational Linguistics*, 30(1):75–93.

Jianfeng Gao, Mu Li, Changning Huang, Andi Wu. 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computational Linguistics*, 31(4): 531-574.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the*

[4] New words are identical to unknown words, but more suitable in the taxonomy. These words are identified in the unsupervised segmentation phase.

[5] http://weibo.sina.com/

21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, page 673–680.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conf. on Machine Learning*, pages 282–289.

Xinnian Mao, Yuan Dong, Saike He, Sencheng Bao, and Haila Wang. 2008. Chinese word segmentation and named entity recognition based on conditional random fields. In *Proceedings of IJCNLP 2008*.

Daichi Mochihashi, Takeshi. Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*: Volume 1-Volume 1, page 100–108.

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese Part-of-Speech Tagging: One-at-a-Time or All at Once? Word-based or Character based? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Spain.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *COLING 2004*, pages 562–568, Geneva, Switzerland, August 23-27.

Caruana, R. 1997. Multitask Learning. Ph.D. thesis, School of Computer Science, CMU.

Sproat, Richard and Chilin Shih. 1990. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4):336–351.

F. Sha and F. Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of HLT/NAACL-2003*, 134-141. Edmonton, Canada.

Michael Sipser. 1997. Introduction to the Theory of Computation. PWS, Boston. Section 1.1: Finite Automata, pp.31–47.

N. Xue. 2003. Chinese Word Segmentation as Character Tagging. *International Journal of Computational Linguistics and Chinese Language Processing*, 8(1).

H. Zhang, Q. Liu, X. Cheng, H. Zhang, and H. Yu. 2003. Chinese Lexical Analysis Using Hierarchical Hidden Markov Model. *In Proceedings of the Second SIGHAN Workshop*, pages 63–70, Japan.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2006. Effective tag set selection in Chinese word segmentation via conditional random field modeling. In *PACLIC-20*, pages 87–94, Wuhan, China, November 1-3.

Hai Zhao and Chunyu Kit. 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition, *The Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6)*, pp.106-111, Hyderabad, India, January 11-12.