Normalized Accessor Variety Combined with Conditional Random Fields in Chinese Word Segmentation

Saike HE†
Beijing University of Posts and Telecommunications
Beijing, 100876, China
hsk000@gmail.com

Taozheng ZHANG
Beijing University of Posts and Telecommunications
Beijing, 100876, China
zhangtaozheng@gmail.com

Xue BAI
Beijing University of Posts and Telecommunications
Beijing, 100876, China
bc003@sina.com

Xiaojie WANG
Beijing University of Posts and Telecommunications
Beijing, 100876, China
xjwang@bupt.edu.cn

Yuan DONG
France Telecom R&D Center
Beijing, 100080, China
yuandong@orange-ft.com

Abstract

The word is the basic unit in natural language processing (NLP), as it is at the lexical level upon which further processing rests. The lack of word delimiters such as spaces in Chinese texts makes Chinese word segmentation (CWS) an interesting while challenging issue. This paper describes the in-depth research following our participation in the fourth International Chinese Language Processing Bakeoff [1]. Originally, we incorporate unsupervised segmentation into Conditional Random Fields (CRFs) in the purpose of dealing with unknown words. Normalization is delicately involved in order to cater to problem of small data size. Experiments on CWS corpora from Bakeoff-4 present comparable results with state-of-the-art performance.

Keywords

Unsupervised Segmentation, Conditional Random Fields, Normalized Accessor Variety.

1. Introduction

Words are the basic linguistic units of natural language. However, Chinese texts are character based, not word based. Thus, the identification of lexical words or the delimitation of words in running texts is a prerequisite of NLP.

Chinese word segmentation can be cast as simple and effective formulation of character sequence labeling. A prevailing technique for this kind of labeling task would be Conditional Random Fields1 (CRFs) [1], following the current trend of applying machine learning as a core technology in the field of natural language processing. Based on conditional dependency assumption, CRFs could exert predominant performance on the known words (which refer to those words exist in both the testing and training data), yet further improvement for CWS systems are usually limited by the comparative large fraction of unknown words (which refer to those words exist only in the testing data).

Regarding this nontrivial issue, in this paper, we are intended to provide a semi-supervised methodology: incorporates an unsupervised method into supervised segmentation, following the in-depth research after our participation in Bakeoff-4. Catering to the common case of limited training data, normalization is involved in the unsupervised phrase.

The rest of the paper is organized as follows: Section 2 describes the framework of our CWS system in detail. Section 3 discusses the unsupervised segmentation method based on a modified version of the target function. Section 4 presents and analyzes our experimental results. Finally, we conclude the work in Section 5.

2. Framework of CWS

Our framework of CWS utilizes Conditional Random Fields (CRFs) as the basic statistical model. The Tag set and features used to train CRFs are also introduced briefly in this section.

2.1 Conditional random fields

Conditional random fields (CRFs) for sequence labeling offer advantages over both generative models like HMMs and classifiers applied at each sequence position [2]. CRFs are an undirected graph established on $G = (V, E)$, where V is the set of random variables $Y = \{Y_i | 1 \le i \le n\}$ for each the n tokens in an input sequence and $E = \{(Y_{i-1}, Y_i) | 2 \le i \le n\}$ is the set of (n − 1) edges forming a linear chain. Following [1], 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:

[1] The Fourth International Chinese Language Processing Bakeoff & the First CIPS Chinese Language Processing Evaluation (Bakeoff-4), at: http://www.china-language.gov.cn/bakeoff08/bakeoff-08_basic.html

Student Research Workshop, RANLP 2009 - Borovets, Bulgaria, pages 35–39

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

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 [3]. Recent research indicates that quasi-Newton methods such as L-BFGS [4] are more effective than GIS.

2.2 Tag set

As justified in [5, 6], 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.

Table 1. Illustration of 6-tag format in CWS

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

2.3 Feature templates

Table 2. The features used in CWS systems.

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 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.

3. Unsupervised segmentation

Although CRFs model could tackle the known words accurately based on the information learned from the training data, the segmentation on the unknown words rests on reliable statistical information derived from large amount of running texts. Thus, we resort to unsupervised segmentation method to deal with these unknown words. In general, unsupervised segmentation assumes no label information for training. It rests on statistical information over the whole corpus to identify potential words, each assigned a goodness score to indicate their credibility. In this section we will introduce an existing unsupervised segmentation criterion, whose segmentation results are encoded into additional features to facilitate supervised learning for CWS. To make it more reliable, normalization strategy is involved.

3.1 Accessor variety

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 [7] 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 (AV value) of a string *s* is defined as:

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

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.

3.2 Unsupervised segmentation

Given the formula for calculating the AV value of a certain string within a sentence, the segmentation problem is then cast as an optimization problem to maximize the target function of the AV value over all word candidates in a sentence. For the sake of convenient, we use a *segmentation* to denote a segmented sentence, a *segment* to denote a continuous substring in the segmentation, and *f* to denote the target function. We use *s* to represent a string (e.g. a sentence), *S* to represent a segmentation of *s*, n to represent the number of characters in *s*, and m to denote the number of segmentation in *S*. The sentence *s* can be displayed as the concatenation of n characters, and S as the concatenation of m strings:

$$s = c_1c_2c_3\ldots c_i\ldots c_n$$

$$S = w_1w_2w_3\ldots w_i\ldots w_m$$

where c_i stands for a character and w_i stands for a *segment*. The target functions *f* is given below [8]:

$$f(\mathrm{S}) = \sum_{i=1}^{m} f(\mathrm{w}_i) \quad (3)$$

Given a target function f and a particular sentence s, we need to choose the *segmentation* that maximizes the values of f(S) over all the possible segmentations. In formulation function f(w) ,we consider two factors: one is the segment length, denoted as |w|, and the other is the AV value of a segment, denoted as AV(w). Then, f(w) can be formulated as a function of |w| and AV(w), thus the target function can be regarded as a choice of normalization for the AV(w) to balance the segmentation length and the AV value for each segmentation. Theoretically, the choice of f(w) is arbitrary, among the most representative types of functions (namely, polynomial, exponential, and logarithmic functions), we choose polynomial function for f(w) (hereafter, referred as AV), since it proves to be the best in our CWS system, and it is defined as:

$$f(\mathrm{w}) = |\mathrm{w}|^c \times \mathrm{AV}^d(\mathrm{w}) \quad (4)$$

where c and d are integer parameters that are used to define the target function f(w), whose performance has been justified in [8].

As the training is usually too limited, then there would be a great chance that fluctuation exists in the AV value of a string consist of extreme number of characters, that is to say: there should be a disparity between dealing with strings with very few characters and that with much more characters when calculating AV values. Such fluctuation may deteriorate the reliability of AV value in that: single-character candidate, such as stop word or interrogative marker, may receive comparatively low AV value, though considering them as an isolate word is actually much better; multi-character potential word, which carries no practical meaning is highly possible to obtain a relatively high AV value just because there is a high concurrence frequency among those characters. Unfortunately, both of these flaws inherent in formula (4) are overlooked in either [6] or [8], at least not mentioned in detail. To deal with this special case, as well as alleviate the fluctuation in AV values, we introduce a normalized version of formulation function f_N(w) (hereafter, referred as NAV) in accessor variety, as formulated below:

$$f_N(\mathrm{w}) = \frac{|\mathrm{w}|^c \times \mathrm{AV}^d(\mathrm{w})}{1 + \left(\frac{|\mathrm{w}|}{\mathrm{Norm}}\right)} \quad (5)$$

A real-value normalizer, named as *Norm* is involved in (4) to obtain (5). The modified formulation function f_N is based on the following consideration: on the one hand, when |w| is large enough, unless its accessor variety is relative high, it would not be considered as a potential word, thereby a low value would be assigned to the current segment strategy; on the other hand, when |w| is too small, unless its accessor variety is also relative low, it would still enjoy high favor, the current segment strategy receives comparably high value accordingly. This measure coincides with that proposed in [8], with a superiority of the absence of special consideration for single character or multi-character candidates.

With all the information above prepared, here comes the computation of f(S) for a given sentence s. Since the value of each segment can be computed independently from the other segments in S, f(S) can be computed using a dynamic programming technique, in which the time complexity is linear to sentence length. Let us use f_i to denote the optimal target function value for the sub-sentence $c_1c_2\ldots c_i$ and $w_{j\ldots i}$ to denote the segment $c_{j+1}c_{j+2}\ldots c_i$ (for $j \leqslant i$). Then we have the following dynamic equations:

$$f_0 = 0;$$

$$f_1 = f(w_{1\ldots1} = c1);$$

$$f_i = \max_{0<i-j<7} f_j + f(w_{j\ldots i}), \text{ for } i > 1;$$

$$f(S) = f_n.$$

It is worth noticing that in each iteration, there are at most N(in our experiment $N = 6$) possible choice, where N is the maximum length of a word.

3.3 AV feature

Having nailed down the definition of accessor variety and target function, we could conduct the unsupervised segmentation. However, we now confront two choices to utilize the AV feature: (1) using the unsupervised segmentation result (in the form of 6-tag set as mentioned in section 2 as auxiliary feature for each character within a sentence s in training CRFs. (hereafter, referred as 'Auxiliary Seg') (2) directly assigning the AV value calculated by formula (5) to a string under the best segmentation S for sentence s (hereafter, referred as 'NAV value'). In the latter case, we need to define a feature function to narrow down the value span of AV feature to avoid the problem of data sparsity. Here, we adopt the same feature function in [6], which is defined as

$$f_n(s) = t, \text{ if } 2^t \le \mathrm{AV}(s) < 2^{t+1} \quad (6)$$

where t is an integer to logarithmize the score.

Without any single piece of proof that either of two methods of utilizing AV feature is superior to the other, controlled experiment is conducted in section 4 to seek for an explicit conclusion to this issue.

4. Evaluation results

This section reports the experiment result based on CWS corpora from Bakeoff-4. The corpora consists of 5 data sets, namely, CITYU, CKIP, CTB, NCC and SXU on both closed and open tracks. The corpus from MSRA is simplified Chinese text while the other corpora are in traditional Chinese. The original label for the training data set is IOB-2. Here, we convert all the corpora to 6-tag set as introduced in section 2.2.

4.1 Subsections experiment setting

In the unsupervised method (both AV and NAV), maximal segment length of potential word is set to 6. The two parameters c, d in formula (4) and (5) are set to 1, and 2 respectively, followed by the best setting achieved in our CWS system. Notice, the calculation of AV values in the phrase of unsupervised segmentation are derived from both training and testing corpus (in unsupervised segmentation, the training data is utilized as unlabeled data as well).

4.2 Two ways of utilizing AV value

To find out the better strategy to utilize accessor variety, we conduct a controlled experiment on the close tracks, that is： CWS with AV, CWS with NAV, and the result is shown in Table 3.

Table 3. Comparison between two ways of utilizing AV value

Run ID	F-Score	
	Auxiliary Seg	NAV value
CITYU	94.50	94.93
CKIP	93.21	94.04
CTB	94.89	95. 39
NCC	92.41	93.93
SXU	95.63	96.19

(Note: the parameter *Norm* in formula (5) for NAV is set to 2.5)

The final result indicates that the strategy with 'NAV value' presents better performance. This may be explained as the error brought in by the 'Auxiliary Seg' which promulgates through the whole sentence thus misguides the CRFs learner.

4.3 'Norm' parameter setting in NAV

Table 4. The result of NAV on CWS closed tracks with different settings of parameter Norm

Run ID	F-Score		
	Norm=2	Norm =2.5	Norm =3
CITYU	94.92	94.93	94.87
CKIP	93.94	94.04	94.05
CTB	95.50	95.39	95.35
NCC	93.91	93.93	93.94
SXU	96.15	96.19	96.08

As we can see from Table 4, NAV achieves comparatively higher performance when Norm is set to 2.5. Our experiment implies that when parameter Norm is set within the span between 2 to 3, relatively performance promotion can be obtained. For the sake of convenience, the parameter *Norm* in formula (5) for NAV is set to 2.5 in the following experiments.

4.4 Performance of four systems

For the purpose of comparison, Table 5 lists the performance of four systems on the close tracks.

Table 5. The results of four systems on CWS closed tracks[2]

Run ID	F-Score			
	baseline	+AV	+NAV	best
CITYU	94.43	94.78	94.93	95.10
CKIP	93.17	93.90	94.04	94.70
CTB	94.86	95. 45	**95.39**	95.89
NCC	92.99	93.00	93.93	94.05
SXU	95.46	96.15	96.19	96.23

Where 'baseline' presents our CWS system participating in Bakeoff-4, which only utilizes the feature defined in Table 2. '+AV' indicates AV features are applied; '+NAV' indicates normalized NAV features are involved; while 'best' indicate the topline achieved in Bakeoff-4. Close scrutiny to Table 5 indicates '+AV' can lift the performance of the original CWS ('baseline') to a comparatively higher position, while '+NAV' performs best and are really comparative to the topline result. For the performance improvement of NAV, the normalization mechanism in formula (5) plays a key role. However, it is necessary to point out that the performance of CTB is slightly drawn down by NAV feature compared to that of AV , yet still higher than the 'baseline' system. The value, 2.5 for Norm may not be a proper setting, which can serve as a reasonable explanation for this abnormal phenomenon.

4.5 Performance of CWS open tracks

In this experiment group, we will report the performance of NAV on the open tracks.

In the open tracks, corpus from previous bakeoffs are involved to train CRFs. Additionally, transformation-based error-driven learning (TBL) is also involved and used in

[2] The evaluation tool can be downloaded from http://www.china-language.gov.cn/bakeoff08/

the post-processing phrase. Table 6 lists the corpora used to train the CRFs and TBL learner in the open tracks.

Table 6. Corpora used to train the CRFs classifier and the TBL learner

Run ID	CRFs	TBL
CityU	2005,2006,2007	2003
CKIP	2007	2006
CTB	2006,2007	2007

This experiment group aims at clarifying whether NAV could bring further performance promotion for CRFs in open tracks. As a great amount of external resource is involved, the space for improvement left for NAV is really limited, thus proves to be a challenging task for NAV. Table 7 lists the result of NAV and four comparison systems on the CWS open tracks.

Table 7. The results of four systems on CWS open tracks

Run ID	**F-Score**			
	baseline	**+AV**	**+NAV**	**best**
CITYU	96.97	97.00	**96.99**	**96.97**
CKIP	93.64	94.48	94.53	95.63
CTB	97.93	97.94	97.96	99.20
NCC	-	-	-	
SXU	-	-	-	

(In this experiment setting, we did not conduct experiments on NCC or SXU since no extra data are available for us on these two data sets.)

With a stronger CRFs model and an additional TBL learner, the performance of 'baseline' system are boosted to a much higher level, as we can see from the comparison of Table 6 and Table 7. Still, performance promotion does occur under such circumstance, and the result brought by NAV (96.99) even surpass the topline (96.97) on CITYU data set. Thus, it demonstrates that accessor variety is also useful in the case of open tracks where large amount of external resource are involved, and Normalized accessor variety turns out to be more effective than original AV value.

5. Conclusions

In this paper, we have proposed an effective method of incorporating unsupervised segmentation method into CRFs model. To make the unsupervised strategy more reliable, normalization strategy is involved. Our experiments justify that accessor variety used as 'NAV value' presents better performance over 'Auxiliary Seg' strategy. Although a core parameter Norm, which if differ in diverse settings, will bring about different results in the final evaluation, creditable performance promotion can be obtained within a certain span. In the closed tracks of Bakoff-4, CRFs model with NAV method achieves comparable performance with the topline; while in the open tracks, NAV is still useful when large amount of external resource are involved. Thus, NAV provides us with a effective way to further boost the performance of Chinese Word Segmentation.

6. References

[1] J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th ICML, 282–289, San Francisco, CA.

[2] F. Sha and F. Pereira. 2003. Shallow parsing with conditional random fields. In Proc. of HLT/NAACL-2003, 134-141. Edmonton, Canada.

[3] J.N. Darroch and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. The Annals of Mathematical Statistics, 43 (5):1470-1480.

[4] R.H. Byrd, J. Nocedal and R.B. Schnabel. 1994. Representations of quasi-Newton matrices and their use in limited memory methods. Mathematical Programming, (63):129-156.

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

[6] 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.

[7] Haodi Feng, Kang Chen, Chunyu Kit, and Xiaotie Deng. 2005. Unsupervised segmentation of Chinese corpus using accessor variety. In K.-Y. Su, J. Tsujii, J. H. Lee, and O. Y. Kwong, editors, Natural Language Processing- IJCNLP 2004, volume 3248 of Lecture Notes in Computer Science, pages 694–703, Sanya, Hainan Island, China. Springer Berlin / Heidelberg.

[8] Haodi Feng, Kang Chen, Chunyu Kit, and Xiaotie Deng. 2005. Unsupervised segmentation of Chinese corpus using accessor variety. In K.-Y. Su, J. Tsujii, J. H. Lee, and O. Y. Kwong, editors, Natural Language Processing - IJCNLP 2004, volume 3248 of Lecture Notes in Computer Science, pages 694–703, Sanya, Hainan Island,China. Springer Berlin / Heidelberg.