How does Dictionary Size Influence Performance of Vietnamese Word Segmentation?

Wuying Liu, Lin Wang

Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies
510420 Guangzhou, Guangdong, CHINA
E-mail: wyliu@gdufs.edu.cn, wanglin@nudt.edu.cn

Abstract

Vietnamese word segmentation (VWS) is a challenging basic issue for natural language processing. This paper addresses the problem of how does dictionary size influence VWS performance, proposes two novel measures: square overlap ratio (SOR) and relaxed square overlap ratio (RSOR), and validates their effectiveness. The SOR measure is the product of dictionary overlap ratio and corpus overlap ratio, and the RSOR measure is the relaxed version of SOR measure under an unsupervised condition. The two measures both indicate the suitable degree between segmentation dictionary and object corpus waiting for segmentation. The experimental results show that the more suitable, neither smaller nor larger, dictionary size is better to achieve the state-of-the-art performance for dictionary-based Vietnamese word segmenters.

Keywords: Vietnamese Word Segmentation, Dictionary-based Segmentation Algorithm, Dictionary Size, Square Overlap Ratio, Relaxed Square Overlap Ratio

1. Introduction

Like Thai, Japanese and Chinese text, Vietnamese text is also a text without any explicit separator between words. Thus, identifying the word boundaries is a challenging basic issue to above oriental languages for natural language processing (Doan, 2008). Vietnamese is a monosyllabic language, whose basic linguistic unit is called 'tiếng', similar to traditional syllables in respect of phonetic form. A Vietnamese word can be made up of a single syllable, or several sequential syllables connected by space symbols. In raw Vietnamese texts, space symbol can be treated as an overload symbol, which is a connector within a word or is a separator between words. Therefore, the Vietnamese word segmentation (VWS) problem can be defined as a binary categorization task for each space symbol. If a space symbol is a connector in a word, we will output a symbol ('_') to replace it. And if a space symbol is a separator between words, we will maintain it as a space symbol (' ') in the segmented result.

Since the early days of Vietnamese information processing researches, VWS problem has been widely investigated, and many effective segmentation algorithms have been proposed (Dinh et al., 2008). The early dictionary-based word segmentation algorithms mainly include maximum matching algorithm and reverse maximum matching algorithm. Subsequently, various kinds of advanced machine learning algorithms (such as maximum entropy (Dinh and Vu, 2006), support vector machines and conditional random fields (Nguyen et al., 2006)) regard word segmentation problem as a sequence labeling task, and related algorithms can obtain preferable performance in VWS. Some investigations show that many affixed resources (such as part of speech tag (Pham et al., 2009)) are helpful in the word segmentation algorithm (Tran et al., 2010). Recently, hybrid and ensemble algorithms have attracted more and more attentions. The hybrid algorithm (Le et al., 2008) combines finite-state automata, regular expression and maximum matching techniques to implement a highly accurate Vietnamese tokenizer (*vnTokenizer*). The ensemble algorithm (Liu and Lin, 2014) combines multiple weak segmenters to form a strong Vietnamese segmenter within the probabilistic ensemble learning framework.

In previous VWS algorithms, the more accurate the model is, generally the more complex and time-consuming it is. However, a real large-scale industry application trends to apply a straightforward and efficient model. Especially, in contemporary big data era, simple models and a lot of data trump more elaborate models based on less data (Halevy et al., 2009). Great minds think alike, we just apply the simple dictionary-based algorithm and a large dictionary in our practical project to deal with large-scale Vietnamese text processing. At beginning, we assume that the larger dictionary size can obtain the more accurate segmentation results as a matter of course. Unfortunately, subsequent project implementation breaks our simple assumption. The dictionary-based VWS algorithm has a straightforward implementation, while whose performance highly depends on a suitable dictionary. "How does dictionary size influence segmentation performance?" and "Which size is suitable?" motivate the current investigation.

2. Re-examination of Dictionary-based VWS Algorithm

2.1 Algorithm and Dictionary

In order to represent the scene like our project implementation, we choose two classical dictionary-based Vietnamese word segmenters: *MMSegmenter* (MM) and *RMMSegmenter*[1] (RMM). The MM and the RMM are implemented from the dictionary-based maximum

[1] http://cbd.nichesite.org/CBD2013S002.htm

matching algorithm and the dictionary-based reverse maximum matching algorithm respectively.

Originally, the MM and the RMM have integrated a Vietnamese dictionary with 87,399 multi-syllable words. Furthermore, we also examine another two dictionaries. One dictionary is extracted from the *JVnSegmenter*[2] tool, which contains 64,546 multi-syllable words. The other is an actual dictionary used in our practical project, which contains 122,727 multi-syllable words.

2.2 Corpus and Evaluation

In this section, we use a publicly available benchmark dataset (Corpus for Vietnamese Word Segmentation[3], CVWS), which contains total 7,807 sentences with word boundary labels from 305 Vietnamese newspaper articles in various domains.

The international Bakeoff (Richard and Thomas, 2003) evaluation measure and associated evaluation methodology are applied. Here, we report the classical Precision (P), Recall (R), F1-measure (F1) and Error Rate (ER) to re-exam the performance of dictionary-based segmenters. The value of P, R, F1 belongs to [0, 1], where 1 is optimal, while the value of ER belongs to [0, 1], where 0 is optimal.

$$\mathrm{P} = C/(C+M) \tag{1}$$

$$\mathrm{R} = C/N \tag{2}$$

$$\mathrm{F1} = 2\mathrm{PR}/(\mathrm{P}+\mathrm{R}) \tag{3}$$

$$\mathrm{ER} = M/N \tag{4}$$

The above four measures are computed as Eq. (1) to Eq. (4) separately. Where the N denotes the total number of words in the manually segmented text, the C denotes the number of correctly segmented words by an automatic segmenter, and the M denotes the number of mistakenly segmented words by an automatic segmenter.

2.3 Result and Discussion

We run the two segmenters with above three different dictionaries respectively. Table 1 presents the experimental result, which shows that four measures from the dictionary of 87,399 words are the best ones in three MM's runs and in three RMM's runs respectively. For instance, the F1 value (0.9477) of MM with 87,399 words dictionary is best among 0.9321, 0.9477 and 0.9423; and the ER value (0.0396) of RMM with 87,399 words dictionary is best among 0.0506, 0.0396 and 0.0432.

	DictSize	P	R	F1	ER
MM	122,727	0.9515	0.9135	0.9321	0.0466
	87,399	**0.9625**	**0.9332**	**0.9477**	**0.0363**
	64,546	0.9587	0.9264	0.9423	0.0399
RMM	122,727	0.9473	0.9094	0.9280	0.0506
	87,399	**0.9591**	**0.9299**	**0.9443**	**0.0396**
	64,546	0.9553	0.9230	0.9389	0.0432

Table 1: Experimental result in different DictSize.

Moreover, the performance of MM excels that of RMM with the optimal dictionary. For instance, the P value of MM and RMM is 0.9625 and 0.9591, and the R value of MM and RMM is 0.9332 and 0.9299 with the optimal dictionary of 87,399 words.

The experimental result verifies that dictionary size influences the performance of dictionary-based VWS algorithm, and neither smaller nor larger dictionary size is the best one. How to select a suitable dictionary with an optimal size? Our corresponding investigation motivates the following dictionary performance prediction methods.

3. Dictionary Performance Prediction Method

3.1 Supervised Prediction Method

The dictionary-based VWS algorithm is straightforward, whose performance depends on two factors: segmentation dictionary and object corpus waiting for segmentation. For a given Vietnamese corpus, the optimal dictionary is just made up of the total multi-syllable words occurring in the corpus. During the VWS procedure of the corpus, each multi-syllable word can be retrieved in the optimal dictionary and each word may be segmented correctly at greatly reduced combinatorial ambiguities and overlapping ambiguities.

Under the supervised condition, the labeled training corpus, with the same distribution to the unlabeled testing corpus, can help to predict dictionary performance. Therefore, we propose a square overlap ratio (SOR) measure to predict the performance of dictionary. The SOR value is the product of dictionary overlap ratio (DOR) and corpus overlap ratio (COR). The value of DOR, COR and SOR belongs to [0, 1], where 1 is optimal.

$$\mathrm{DOR} = W_o / W_d \tag{5}$$

$$\mathrm{COR} = W_o / W_c \tag{6}$$

$$\mathrm{SOR} = \mathrm{DOR} \cdot \mathrm{COR} \tag{7}$$

The above SOR measure is computed as Eq. (5) to Eq. (7). Where the W_o denotes the number of multi-syllable words co-occurred in dictionary and corpus, the W_d denotes the total number of words in dictionary, and the W_c denotes the total number of multi-syllable words in corpus.

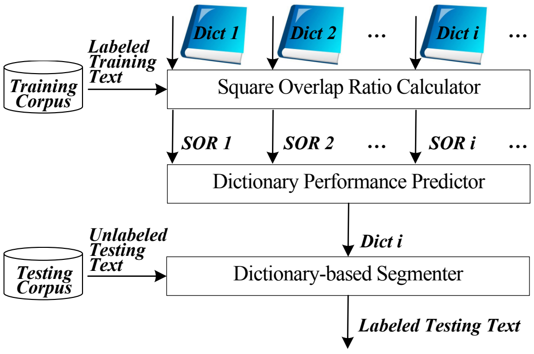

Figure 1: Supervised prediction framework.

[2] http://jvnsegmenter.sourceforge.net

[3] http://www.jaist.ac.jp/~hieuxuan/vnwordseg/data

Supported by the SOR measure, we propose a supervised prediction framework to predict dictionary performance. Figure 1 shows the framework, which mainly includes a square overlap ratio calculator (SORC), a dictionary performance predictor (DPP) and a dictionary-based segmenter (DS). The SORC receives the labeled training text from training corpus, and calculates a SOR value for each dictionary. The DPP receives several SOR values, and selects the corresponding dictionary by the maximal SOR value. The DS segments the unlabeled testing text from testing corpus according to the suitable dictionary, and outputs the labeled testing text.

3.2 Unsupervised Prediction Method

Supposing to obtain a label without any cost, the supervised prediction method is an ideal implement. However, in practice, it is costly to obtain a label for a real-world segmenter. Especially, there is not any label under the unsupervised condition, which defeats the supervised prediction method.

In order to cope with the unsupervised condition, we relax the calculation restriction of SOR measure, and propose a relaxed square overlap ratio (RSOR) measure to predict the performance of dictionary. The RSOR value is the product of relaxed dictionary overlap ratio (RDOR) and relaxed corpus overlap ratio (RCOR). The value of RDOR, RCOR and RSOR belongs to [0, 1], where 1 is optimal.

$$\mathrm{RDOR} = S_o / S_d \tag{8}$$

$$\mathrm{RCOR} = S_o / S_c \tag{9}$$

$$\mathrm{RSOR} = \mathrm{RDOR} \cdot \mathrm{RCOR} \tag{10}$$

The above RSOR measure is computed as Eq. (8) to Eq. (10). Where the S_o denotes the number of syllables co-occurred in dictionary and corpus, the S_d denotes the total number of syllables in dictionary, and the S_c denotes the total number of syllables in corpus.

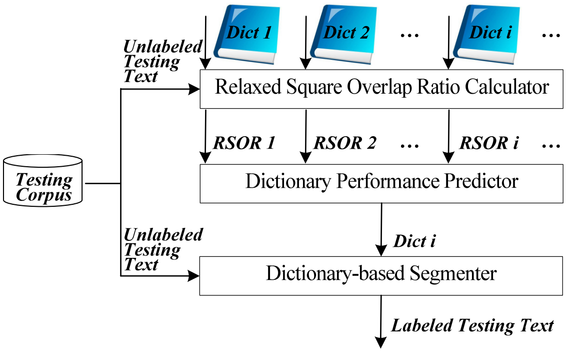

Figure 2: Unsupervised prediction framework.

Supported by the RSOR measure, we propose an unsupervised prediction framework without any label. Figure 2 shows the framework, which mainly includes a relaxed square overlap ratio calculator, a dictionary performance predictor and a dictionary-based segmenter. The crucial difference is the counting object within above two frameworks, one is multi-syllable word, and the other is syllable.

4. Experiment

4.1 Supervised Prediction Result

In the supervised experiment, we try to predict a suitable dictionary from four dictionaries, among which there are three ones (122,727 words, 87,399 words and 64,546 words) have been mentioned in Section 2.1, and the remaining one is man-made particularly as a dictionary of reference, which is just made up of the total 9,113 multi-syllable words occurring in the CVWS dataset.

We use three-fold cross validation by evenly splitting the CVWS dataset into three parts and use two parts for training and the remaining third for testing. We perform the training-testing procedure three times and use the average of the three performances as the final result.

DictSize	DOR	COR	SOR
122,727	0.0445	0.7397	0.0329
87,399	0.0613	0.7257	**0.0445**
64,546	0.0698	0.6108	0.0427
9,113	0.8098	1.0000	0.8098

Table 2: Overlap ratio in different DictSize.

Table 2 shows the final result of three overlap ratios in the four **DictSize**s, which shows that (I) the SOR value (0.8098) of dict9113 excels that of others obviously; and (II) the SOR value (0.0445) of dict87399 is optimal among the remaining three dictionaries. The result predicts that the performance rank of dictionaries will be dict9113, dict87399, dict64546 and dict122727.

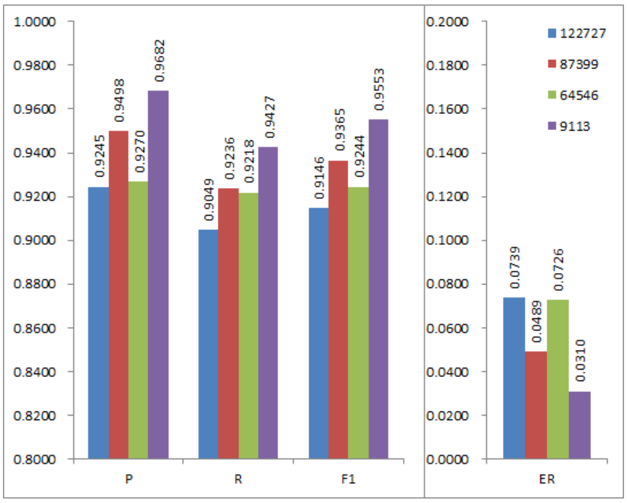

Figure 3: Experimental result of the MM segmenter in different dictionary size.

Figure 3 presents the experimental result of the MM segmenter in the four dictionaries, which shows that (I) the four measures of dict9113 excel that of others, for instance, the F1 value of dict9113 is 0.9553, while that of dict87399, dict64546 and dict122727 is 0.9365, 0.9244 and 0.9146 respectively; and (II) the four measures of dict87399 is optimal except the man-made reference, for instance, the ER value of dict87399 is 0.0489, while that of dict64546 and dict122727 is 0.0726 and 0.0739. The result proves that above prediction of the performance rank is effective.

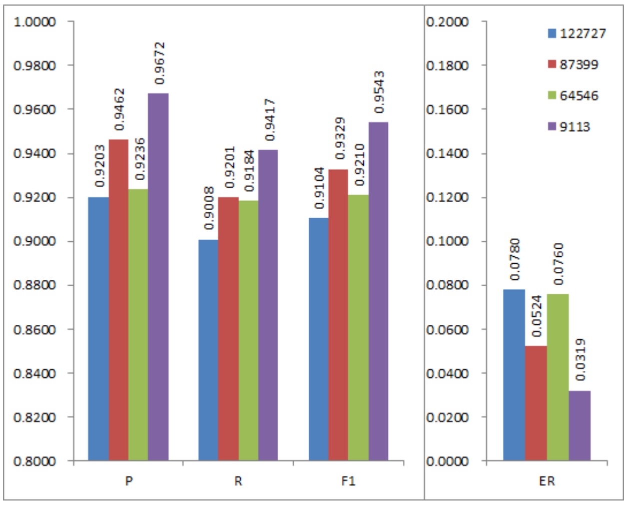

Figure 4: Experimental result of the RMM segmenter in different dictionary size.

Figure 4 presents the experimental result of the RMM segmenter in the four dictionaries, which shows a similar situation with above MM segmenter, and also proves that the performance rank prediction is correct.

4.2 Unsupervised Prediction Result

In the unsupervised experiment, we also predict a suitable dictionary from the four dictionaries without training corpus. So, we calculate the RSOR value according to the four dictionaries and the full CVWS dataset directly without three-fold cross validation.

DictSize	RDOR	RCOR	RSOR
122,727	0.2979	0.5307	0.1581
87,399	0.3326	0.5205	**0.1731**
64,546	0.4444	0.3776	0.1678
9,113	1.0000	0.5606	0.5606

Table 3: Relaxed overlap ratio in different DictSize.

Table 3 shows the result of three relaxed overlap ratios in the four **DictSize**s, which shows that the RSOR value rank is 0.5606, 0.1731, 0.1678 and 0.1581. Being identical with the supervised prediction, the result predicts that the performance of dict9113 is optimal, and the performance rank of remaining three dictionaries will be dict87399, dict64546 and dict122727. The P, R, F1 and ER measures of the MM and the RMM in the remaining three dictionaries are presented in Table 1, for instance, the MM's P value of dict87399, dict64546 and dict122727 is 0.9625, 0.9587 and 0.9515 respectively, and RMM's R value of dict87399, dict64546 and dict122727 is 0.9299, 0.9230 and 0.9094 respectively, which prove that the unsupervised performance prediction is effective too.

5. Conclusion

This paper investigates the influence of dictionary size to VWS, and suggests the supervised and the unsupervised prediction methods, which can select a suitable dictionary and make the simple VWS algorithm to solve the complex VWS issue efficiently. If there is a big dictionary, our prediction methods can automatically customize an individual sub-dictionary for each object corpus waiting for segmentation. Just like Albert Einstein's wisdom "everything should be made as simple as possible, but no simpler", our idea of using a simple algorithm affixed suitable data will produce big performance for real industry application in big data age.

Further research will concern the influence of combinatorial ambiguity and overlapping ambiguity to dictionary selection. We will transfer above research productions to other suitable oriental languages like Thai, Japanese, Chinese, and so on.

6. Acknowledgements

The research is supported by the Featured Innovation Project of Guangdong Province (No.2015KTSCX035) and the National Natural Science Foundation of China (No.61402119).

7. References

Doan Nguyen. (2008). Query preprocessing: improving web search through a Vietnamese word tokenization approach. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore, July 20-24, 2008). SIGIR '08. ACM New York, NY, USA, 765-766.

Quang Thang Dinh, Hong Phuong Le, Thi Minh Huyen Nguyen, Cam Tu Nguyen, Mathias Rossignol, Xuan Luong Vu. (2008). Word segmentation of Vietnamese texts: a comparison of approaches. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (Marrakech, Morocco, May 28-30, 2008). LREC '08. European Language Resources Association, 1933-1936.

Dinh Dien and Vu Thuy. (2006). A maximum entropy approach for Vietnamese word segmentation. In *Proceedings of the 4th International Conference on Computer Sciences: Research, Innovation and Vision for the Future* (Ho Chi Minh City, Vietnam, February 12-16, 2006). RIVF '06. IEEE, 248-253.

Cam Tu Nguyen, Trung Kien Nguyen, Xuan Hieu Phan, Le Minh Nguyen, Quang Thuy Ha. (2006). Vietnamese word segmentation with CRFs and SVMs: an investigation. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation* (Wuhan, China, November 2-4, 2006). PACLIC 2006. Tsinghua University Press, 215-222.

Dang Duc Pham, Giang Binh Tran, Son Bao Pham. (2009). A hybrid approach to Vietnamese word segmentation using part of speech tags. In *Proceedings of the 1st International Conference on Knowledge and Systems Engineering* (Hanoi, Vietnam, October 13-17, 2009). KSE '09. IEEE Computer Society Washington, DC, USA, 154-161.

Thi Oanh Tran, Anh Cuong Le, Quang Thuy Ha. (2010). Improving Vietnamese word segmentation and POS tagging using MEM with various kinds of resources.

Information and Media Technologies. 5, 2 (2010), 890-909.

Hong Phuong Le, Thi Minh Huyen Nguyen, Azim Roussanaly, Tuong Vinh Ho. (2008). A hybrid approach to word segmentation of Vietnamese texts. *Language and Automata Theory and Applications*. Lecture Notes in Computer Science, Volume 5196, 240-249.

Wuying Liu and Li Lin. (2014). Probabilistic ensemble learning for Vietnamese word segmentation. In *Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Gold Coast, QLD, Australia, July 06-11, 2014). SIGIR '14. ACM New York, NY, USA, 931-934.

Alon Halevy, Peter Norvig, Fernando Pereira. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*. 24, 2 (2009), 8-12.

Richard Sproat and Thomas Emerson. (2003). The first international Chinese word segmentation Bakeoff. In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing* (Sapporo, Japan, July 11-12, 2003). SIGHAN '03. Association for Computational Linguistics, Stroudsburg, PA, USA, 133-143.