Vietnamese Multisyllabic-Word Extraction for Word Segmentation

Wuying Liu[1] and Lin Wang[2(✉)]

[1] Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies
Guangzhou 510420, Guangdong, China

[2] Xianda College of Economics and Humanities, Shanghai International Studies University
Shanghai 200083, China

wyliu@gdufs.edu.cn, lwang@xdsisu.edu.cn

Abstract

The automatic construction of a machine-readable dictionary is a challenging issue for low-resource language processing. In this paper, we address the Vietnamese multisyllabic-word extraction problem and investigate a word extraction algorithm based on unsupervised ensemble learning to detect multisyllabic-words from large-scale Vietnamese text documents. First, we design a syllable-level n-gram gluer to generate potential multisyllabic-words. Then, we calculate two simple statistical features, word frequency and document frequency, and implement three unsupervised word extractors. Subsequently, the ensembler combines several dictionaries extracted by the extractors to obtain the final one. Finally, we evaluate the effectiveness of these individual dictionaries and the final ensemble one through two classical dictionary-based Vietnamese word segmentation algorithms. The experimental results show that our extraction algorithm based on unsupervised ensemble learning is effective, and the two kinds of word segmentation algorithms with automatically extracted dictionaries can achieve comparable results.

Keywords

Multisyllabic-word extraction, Word segmentation, Vietnamese, Unsupervised ensemble learning, Syllable-level n-gram model.

1. Introduction

Vietnamese is a monosyllable-based low-resource language, whose written texts have no

any explicit formal separator between words. Therefore word segmentation, identifying boundaries of words, is crucial to Vietnamese written texts in many applications of Natural Language Processing (Dinh et al., 2008). Machine-readable Vietnamese dictionary for word segmentation needs multisyllabic-words only, which is a basic resource to determine the effectiveness of word segmentation algorithms.

Currently, fast-paced development of computing technology brings the explosive growth of information. A large number of new words have been bloomed in all kinds of languages. While manual updating of dictionaries is a cost-sensitive job, which defeats the timely improvement of word segmentation algorithms. Fortunately, quantitative accumulation leads to qualitative transformation, information explosion produces vast Vietnamese text documents, which provide a new opportunity to extract multisyllabic-words without human intervention (Trung et al., 2013).

Previous investigations have shown that some dictionary-based Vietnamese word segmentation algorithms can achieve high performance (Liu et al., 2014) and a big dictionary can produce big performance by customizing an individual sub-dictionary (Liu et al., 2016). In the following sections, we investigate an unsupervised ensemble learning (UEL) algorithm to detect multisyllabic-words from large-scale Vietnamese text documents.

2. Vietnamese Multisyllabic-Word Extraction

2.1. Framework

Unsupervised learning can give full play to the advantages of large-scale unlabeled corpus, and does not require any manual annotation (Vlachos, 2011). Ensemble learning can combine multiple individual learners to obtain better results (Liu et al., 2012). Therefore, we take the advantages of both two machine learning ideas, and propose an UEL framework. Figure 1 shows our UEL framework, which receives a large-scale text documents and will generate three individual dictionaries and an ensemble one.

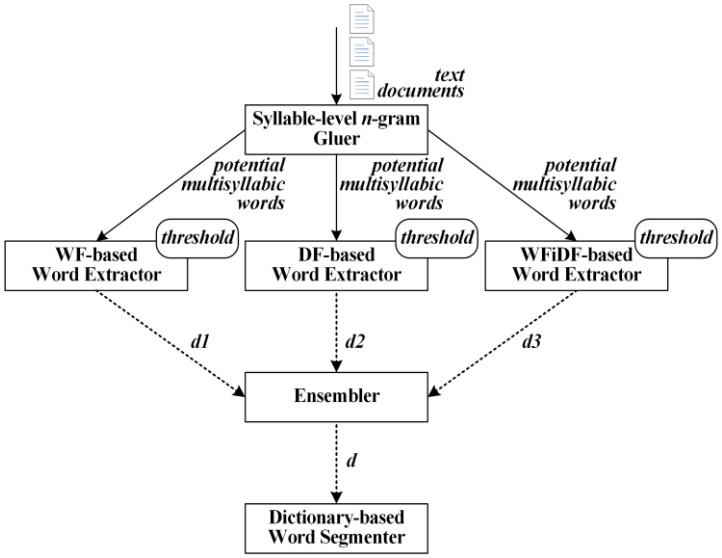


Figure 1: Unsupervised Ensemble Learning Framework.

The core of the UEL framework is the **Syllable-level *n*-gram Gluer**, which can generate many potential multisyllabic-words without any supervised information only according to the *n*-gram lexical model. Subsequently, each word extractor truncates the total potential multisyllabic-words by comparing the rank of statistical feature with a preset top-percentage threshold. Finally, the **Ensembler** takes advantage of complementary virtues from different extractors, and merges several dictionaries extracted by them to form an ensemble one. All above dictionaries from the ensembler and extractors can be used in any dictionary-based word segmenter.

Previous research has shown that the *n*-gram model implicates abundant useful features (Kanaris et al., 2007), and the value of *n* determines the total number of potential multisyllabic-words. Here, we consider four overlapping syllable-level *n*-gram models (2-gram, 3-gram, 4-gram, and 5-gram) to represent potential multisyllabic-words. We will further reveal the word frequency and document frequency of potential multisyllabic-words by statistical analysis in the two corpora of Vietnamese text documents.

Corpus	Number of Text Documents	Number of Potential Multisyllabic-Words
VWN (Vietnamese Web News)	11,479	942,031
VW (Vietnamese Wikipedia)	1,152,603	10,849,903

Table 1: Two Corpora of Vietnamese Text Documents

Table 1 shows the two corpora of VWN (Vietnamese Web News) and VW (Vietnamese Wikipedia). The VWN corpus is a small set of Vietnamese text documents

crawled from an individual field of web news on the Internet, including 11,479 plain text documents, and from which 942,031 potential multisyllabic-words have been generated according to the above four *n*-gram models. The VW corpus is extracted from Vietnamese Wikipedia corpus (viwiki-20170101-pages-articles.xml.bz2), consisting of 1,152,603 plain text documents, and we have generated 10,849,903 potential multisyllabic-words from the VW corpus according to the same four models.

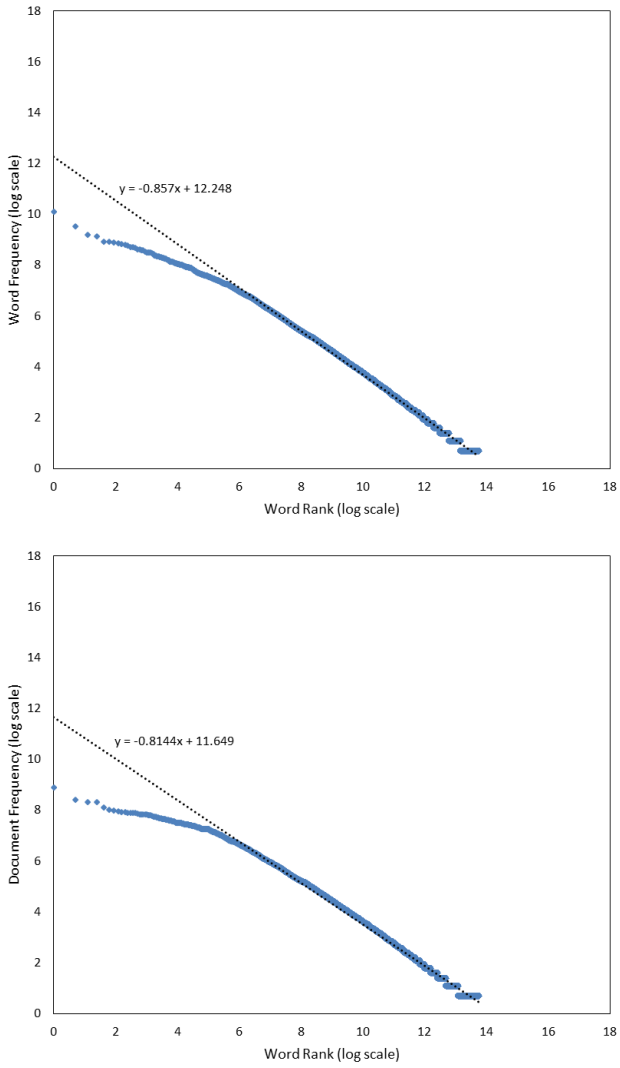


Figure 2: Word Frequency and Document Frequency of Potential Multisyllabic-Words Represented by 2-gram, 3-gram, 4-gram, and 5-gram Models in VWN Corpus.

We calculate the word frequency (WF) and the document frequency (DF) for each

potential multisyllabic-word represented by 2-gram, 3-gram, 4-gram, and 5-gram models in the both two corpora of VWN and VW. Figure 2 shows the above calculating results from the VWN corpus in two sub-figures, while Figure 3 shows the results from the VW corpus.

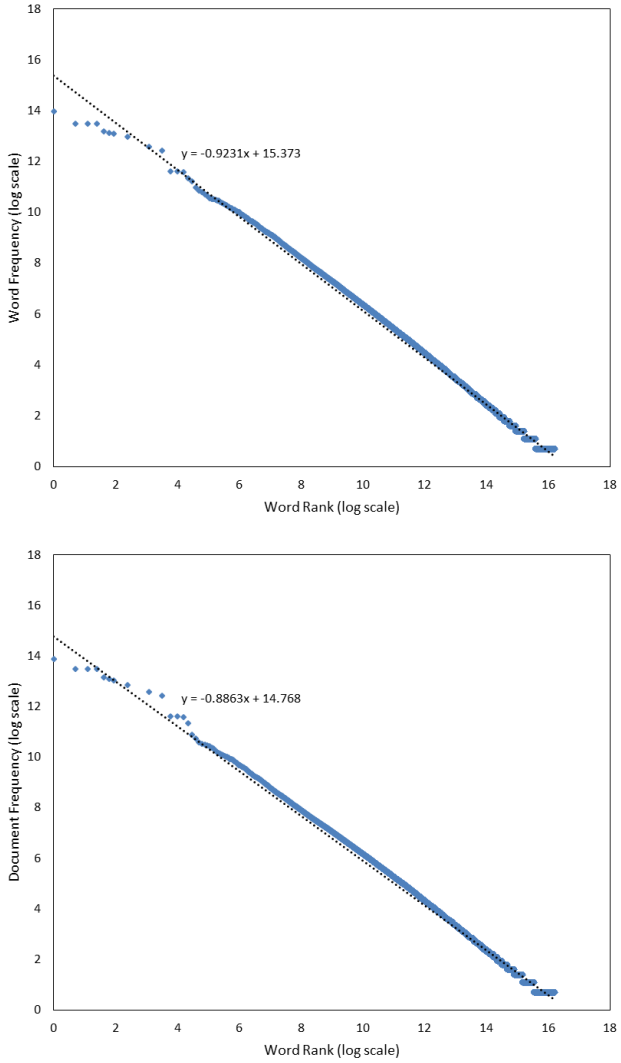


Figure 3: Word Frequency and Document Frequency of Potential Multisyllabic-Words Represented by 2-gram, 3-gram, 4-gram, and 5-gram Models in VW Corpus.

In each sub-figure of Figure 2 and Figure 3, the horizontal-axis (***x***-axis) indicates the word rank (log scale), and the vertical-axis (***y***-axis) indicates the word frequency or the document frequency (log scale). We also fit out a trendline (***y***=a***x***+b) for each sub-figure respectively, which presents that the WF and the DF distributions both follow the power

law approximately. By comparing Figure 2 and Figure 3, we can find that the power law distribution will be more obvious with the increasing of the number of documents and words. The ubiquitous power law brings a possibility to cut low frequency words down (Liu et al., 2013).

2.2. Algorithm

Normally, a Vietnamese word is 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 extraction task of multisyllabic-words can be defined as a co-occurrence detection of multiple syllables. If several syllables trend to occur together frequently, we will predict them as a multisyllabic-word. Supported by the power law distribution, we can straightforwardly calculate two unsupervised features (WF, DF) for a potential word and use a top-percentage threshold to extract some multisyllabic-words. The word with a high frequency or a high document frequency trends to be a common word. We can also calculate another unsupervised WFiDF feature by WF/DF, widely used in Natural Language Processing, and use a top-percentage threshold to extract specialized words. Consequently, within the UEL framework, we propose a UEL-based word extraction (WE) algorithm. Figure 4 shows the detailed UEL-based WE algorithm.

```
1.//UEL-based Word Extraction Algorithm
2.Input (string[] tds, float tt) //text documents, threshold
3.Output (string[] d) //dictionary of multisyllabic-words
4.Main Function uelwe()
5.    For int n←2 To 5 Do
6.        string[] pmw←ngram(tds, n);
7.    End For
8.
9.    string[] pmw1←wfextractor.rank(pmw);
10.   string[] d1←wfextractor.truncate(pmw1, tt);
11.
12.   string[] pmw2←dfextractor.rank(pmw);
13.   string[] d2←dfextractor.truncate(pmw2, tt);
14.
15.   string[] pmw3←wfidfextractor.rank(pmw);
16.   string[] d3←wfidfextractor.truncate(pmw3, tt);
17.
18.   d←ensembler.merge(d1, d2, d3);
19.   Return d.
```

Figure 4: Pseudo-code for the UEL-based WE Algorithm.

In Figure 4, the UEL-based WE algorithm firstly calls the ***ngram***() function in loops to glue potential multisyllabic-words (*pmw*) from Vietnamese fragments separated by punctuations, Arabic numerals, and loanwords in raw large-scale text documents (*tds*).

Subsequently, three word extractors (**wfextractor**, **dfextractor**, **wfidfextractor**) are called separately. There are two functions in each extractor: the ***rank***() function sorts the potential multisyllabic-words according to the value of statistical feature, and the ***truncate***() function cuts low frequency words down according to the preset top-percentage threshold (*tt*). Finally, by dereplicating the ***merge***() function of the ensembler combines the three dictionaries (*d1*, *d2*, *d3*) to form an ensemble one (*d*). The UEL-based WE algorithm is space-time-efficient for the potential parallelism among word extractors within the UEL framework.

3. Experiment

3.1. Corpus and Evaluation

In order to prove the effectiveness of the UEL-based WE algorithm both in an individual field and in the encyclopedia field, we run the algorithm twice using the VWN corpus and the VW corpus respectively as the training corpus of text documents to the **Syllable-level *n*-gram Gluer**. After each run, we will extract three individual dictionaries and an ensemble one. In the following experiments, we will further evaluate the effectiveness of our UEL-based WE algorithm through a direct evaluation and an indirect evaluation.

In the direct evaluation, we have a big manual dictionary with 159,214 multisyllabic-words, which will be used as a golden standard to calculate the precisions of generated dictionaries. The Precision at a Threshold (P@T) is the evaluation measure.

In the indirect evaluation, we run two dictionary-based Vietnamese word segmenters: the MM and the RMM with different dictionaries. The MM and the RMM segmenters are implemented from the dictionary-based maximum matching algorithm and the dictionary-based reverse maximum matching algorithm respectively. We use a publicly available benchmark dataset (Corpus for Vietnamese Word Segmentation, CVWS) as a golden standard. The CVWS dataset contains total 7,807 sentences with word boundary labels from 305 Vietnamese news articles in various domains. The international Bakeoff (Sproat et al., 2003) evaluation measure and associated evaluation methodology are applied. We report the classical Precision (P), Recall (R), F1-measure (F1) and Error Rate (ER). 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.

$$\mathrm{P@T} = V / W \quad (1)$$
$$\mathrm{P} = C / (C + M) \quad (2)$$
$$\mathrm{R} = C / N \quad (3)$$
$$\mathrm{F1} = 2\mathrm{PR}/(\mathrm{P} + \mathrm{R}) \quad (4)$$
$$\mathrm{ER} = M / N \quad (5)$$

The above five measures are computed as Eq. (1) to Eq. (5) separately. Where the *W* denotes the number of multisyllabic-words extracted by the automatic extractor, the *V* denotes the number of multisyllabic-words both in manual dictionary and automatic dictionary, the *N* denotes the total number of words in the manual segmented corpus, the *C* denotes the number of correctly segmented words by the word segmenter, and the *M* denotes the number of mistakenly segmented words by the word segmenter.

3.2. Result and Discussion of Running in the VWN Corpus

We run the UEL-based WE algorithm in the VWN corpus. Table 2 presents the detailed results of word number and corresponding precision after the direct evaluation at different top-percentage thresholds from 10% to 90%. For instance, according to the word frequency, we rank the 942,031 potential multisyllabic-words generated from the VWN corpus, and use a top-percentage threshold of 10% to extract 276 high rank words, in which there are more than 66.67% words hitting in the big manual dictionary with 159,214 words. Table 2 also shows a similar trend of the results from the three word extractors and the ensembler. The precision is not very high, but there will be many new words detected in the remaining words. Another reason lies in that the training corpus is independent with the manual dictionary.

	10%	**30%**	**50%**	**70%**	**90%**
d1	276	614	1,106	2,411	40,412
	0.6667	0.5945	0.5479	0.4355	0.1435
d2	219	489	850	1,767	7,523
	0.6210	0.5665	0.5224	0.4414	0.2499
d3	17,792	60,199	99,212	129,461	162,650
	0.1197	0.0837	0.0786	0.0782	0.0764
d	17,975	60,521	99,594	130,007	174,920
	0.1241	0.0856	0.0795	0.0787	0.0745

Table 2: Word Number and Precision from VWN Corpus

In the indirect evaluation, we run the MM segmenter in three individual dictionaries and an ensemble one respectively. Table 3 presents the experimental result, which shows that the value of the four measures in *d3* dictionary generated from the VWN corpus is the best among those in other individual dictionaries. For instance, the best P value (0.6761) is at 50% in *d3* dictionary.

		P	R	F1	ER
	10%	0.5833	0.7242	0.6461	0.5174
	30%	0.6143	0.7337	0.6687	0.4606
d1	**50%**	0.6329	0.7305	0.6782	0.4236
	70%	0.6425	0.7094	0.6743	0.3947
	90%	0.6283	0.5833	0.6050	0.3450
	10%	0.5671	0.7123	0.6315	0.5437
	30%	0.5913	0.7188	0.6488	0.4968
d2	**50%**	0.6090	0.7185	0.6592	0.4614
	70%	0.6268	0.7070	0.6645	0.4210
	90%	0.6356	0.6600	0.6476	0.3784
	10%	0.5882	0.7308	0.6518	0.5117
	30%	0.6479	**0.7356**	0.6889	0.3998
d3	**50%**	**0.6761**	0.7086	**0.6920**	0.3395
	70%	0.6701	0.6536	0.6618	**0.3218**
	90%	0.6313	0.5708	0.5996	0.3333

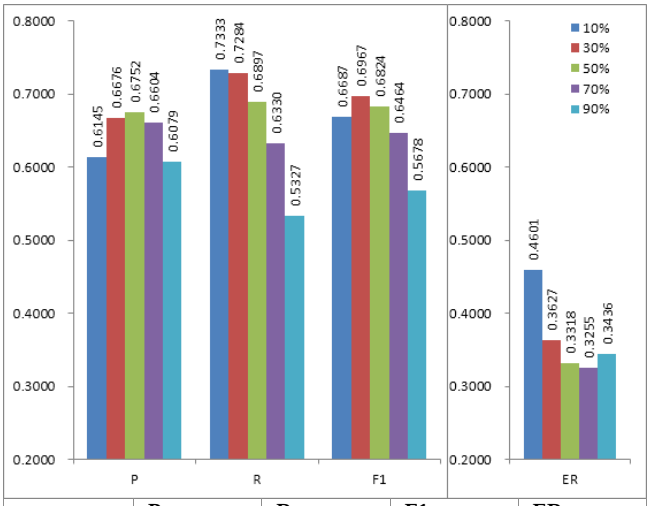
Table 3: Result of MM Segmenter in Individual Dictionaries from VWN Corpus

Table 4 presents the experimental result of RMM segmenter in three individual dictionaries, which shows a similar trend with that of the MM segmenter. The optimal value of the four measures also belongs to *d3* dictionary generated from the VWN corpus. For instance, the best R value (0.7390) is at 30% in *d3* dictionary. The experimental result proves that the WFiDF is a more optimized unsupervised feature than the WF feature and the DF feature individually for the small corpus of an individual field.

		P	R	F1	ER
	10%	0.5833	0.7242	0.6462	0.5173
	30%	0.6153	0.7348	0.6698	0.4594
d1	**50%**	0.6353	0.7333	0.6808	0.4209
	70%	0.6487	0.7162	0.6808	0.3879
	90%	0.6345	0.5890	0.6109	0.3393
	10%	0.5670	0.7121	0.6313	0.5439
	30%	0.5925	0.7202	0.6502	0.4953
d2	**50%**	0.6110	0.7209	0.6614	0.4590
	70%	0.6313	0.7121	0.6693	0.4159
	90%	0.6425	0.6671	0.6546	0.3712
	10%	0.5883	0.7309	0.6519	0.5116
	30%	0.6509	**0.7390**	0.6921	0.3964
d3	**50%**	**0.6799**	0.7129	**0.6960**	0.3356
	70%	0.6760	0.6592	0.6675	**0.3160**
	90%	0.6369	0.5758	0.6048	0.3283

Table 4: Result of RMM Segmenter in Individual Dictionaries from VWN Corpus

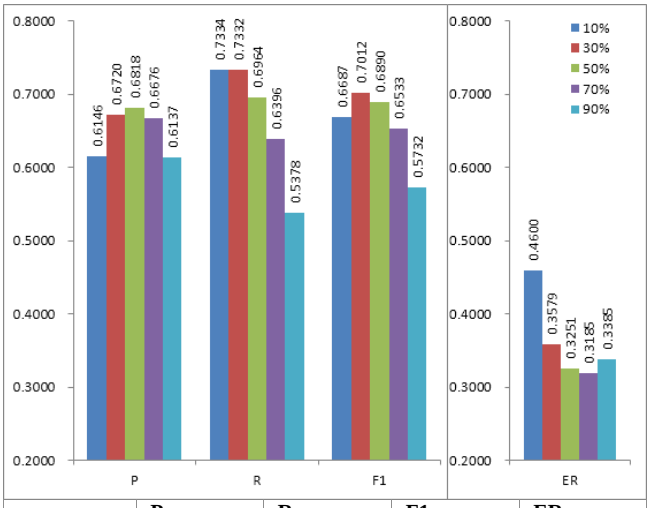
Moreover, we will detailedly analyze the experimental results of MM and RMM segmenters in the ensemble dictionary generated from the VWN corpus. Figure 5 presents the experimental result of the MM segmenter in the ensemble dictionary, which shows that a more suitable top-percentage threshold (30%), neither smaller nor larger, is better to achieve the comparable performance for MM segmenter. Though there is no significant improvement, the best F1 value (0.6967) in the ensemble dictionary is beyond those in three individual dictionaries yet.



	P	**R**	**F1**	**ER**
10%	0.6145	**0.7333**	0.6687	0.4601
30%	0.6676	0.7284	**0.6967**	0.3627
50%	**0.6752**	0.6897	0.6824	0.3318
70%	0.6604	0.6330	0.6464	**0.3255**
90%	0.6079	0.5327	0.5678	0.3436

Figure 5: Result of MM Segmenter in Ensemble Dictionary from VWN Corpus.

Figure 6 presents the experimental result of the RMM segmenter in the ensemble dictionary generated from the VWN corpus, which shows a similar trend with that of the MM segmenter. The best four measures of the RMM segmenter excel that of the MM segmenter. For instance, the best F1 value of the MM segmenter is 0.6967, while the best F1 value of the RMM segmenter is 0.7012. The experimental result verifies that the UEL-based extraction algorithm is effective for the small corpus of an individual field.



	P	R	F1	ER
10%	0.6146	**0.7334**	0.6687	0.4600
30%	0.6720	0.7332	**0.7012**	0.3579
50%	**0.6818**	0.6964	0.6890	0.3251
70%	0.6676	0.6396	0.6533	**0.3185**
90%	0.6137	0.5378	0.5732	0.3385

Figure 6: Result of RMM Segmenter in Ensemble Dictionary from VWN Corpus.

Through detailed analysis of the above experimental results, we find that the straightforward calculating of unsupervised features can extract most common Vietnamese multisyllabic-words efficiently. That is the reason of why no more than 1,200 words can reach the 0.6808 F1 value in the RMM run. Though the WFiDF feature can detect some specialized words, unfortunately, the CVWS dataset is independent with the training corpus. So, there are still a great many of long tail words cut by top-percentage thresholds. If you want to obtain a more precision dictionary, you can firstly use our algorithm as a preprocessing and then add a manual filtering.

3.3. Result and Discussion of Running in the VW Corpus

After running our algorithm in the VW corpus, we can obtain the results of word number and corresponding precision in Table 5. Comparing the results in Table 5 with that in Table 2, we can find that there is a similar distribution of result values from the Wikipedia corpus and the web news corpus. In the direct evaluation, we rank the 10,849,903 potential multisyllabic-words by word frequency, and also use the threshold of 10% to truncate the top 618 words as the *d1* dictionary. Compared with the results in Table 2, the hit rate nearly loses a half (0.3576 / 0.6667 ≈ 53.64%), but we can detect extra 37 (618 * 0.3576 - 276 * 0.6667 ≈ 37) multisyllabic-words from the VW corpus more than from the VWN corpus. Although this achievement is not so proud, a possible reason is that our big manual

dictionary only has constant 159,214 words.

	10%	**30%**	**50%**	**70%**	**90%**
d1	618	1,946	3,630	6,809	22,388
	0.3576	0.3962	0.3601	0.3166	0.2286
d2	543	1,680	3,179	6,136	21,513
	0.2210	0.2976	0.2960	0.2704	0.2012
d3	75,602	280,059	787,219	1,120,100	1,511,570
	0.0658	0.0608	0.0364	0.0335	0.0306
d	76,209	281,422	788,979	1,122,432	1,516,266
	0.0675	0.0613	0.0366	0.0335	0.0306

Table 5: Word Number and Precision from VW Corpus

We also run the MM segmenter in three individual dictionaries and an ensemble one generated from the VW corpus in the indirect evaluation. Table 6 shows that the value of the four measures in *d3* dictionary generated from the VW corpus is the best. For instance, the best F1 value (0.7011) is at 30% in *d3* dictionary.

		P	**R**	**F1**	**ER**
	10%	0.5666	0.7071	0.6291	0.5409
	30%	0.6246	0.7259	0.6714	0.4364
d1	**50%**	0.6436	0.7174	0.6785	0.3973
	70%	0.6546	0.7017	0.6773	0.3703
	90%	0.6532	0.6485	0.6509	0.3443
	10%	0.5412	0.6906	0.6068	0.5856
	30%	0.5892	0.7037	0.6414	0.4905
d2	**50%**	0.6188	0.7044	0.6588	0.4340
	70%	0.6398	0.6949	0.6662	0.3912
	90%	0.6460	0.6447	0.6453	0.3533
	10%	0.5628	0.7201	0.6318	0.5593
	30%	0.6772	**0.7266**	**0.7011**	0.3463
d3	**50%**	**0.6862**	0.6842	0.6852	**0.3129**
	70%	0.6498	0.5957	0.6216	0.3210
	90%	0.5904	0.4932	0.5374	0.3421

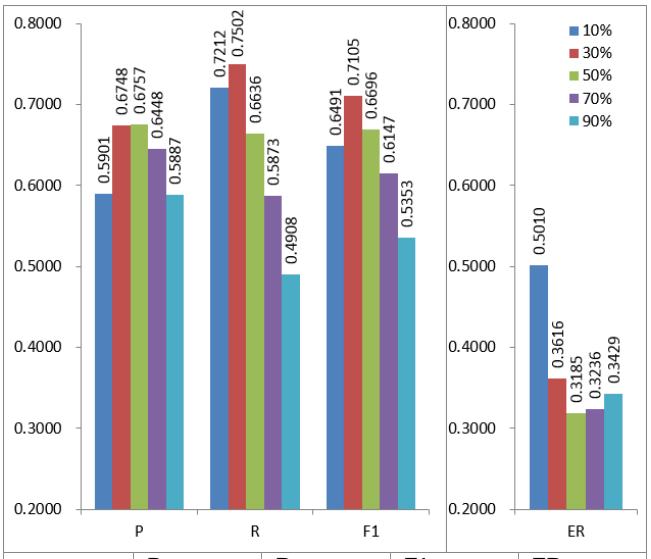
Table 6: Result of MM Segmenter in Individual Dictionaries from VW Corpus

Table 7 presents the experimental result of RMM segmenter in three individual dictionaries, which shows a similar trend with that of the MM segmenter. The optimal value of the four measures also belongs to *d3* dictionary generated from the VW corpus. For instance, the best ER value (0.3046) is at 50% in *d3* dictionary. The experimental result also proves the superiority of the WFiDF feature for the big corpus in the encyclopedia field.

		P	R	F1	ER
	10%	0.5676	0.7084	0.6303	0.5396
	30%	0.6272	0.7289	0.6742	0.4333
d1	**50%**	0.6482	0.7225	0.6833	0.3922
	70%	0.6616	0.7093	0.6847	0.3627
	90%	0.6616	0.6569	0.6593	0.3359
	10%	0.5408	0.6902	0.6065	0.5860
	30%	0.5919	0.7069	0.6443	0.4873
d2	**50%**	0.6224	0.7085	0.6627	0.4298
	70%	0.6465	0.7021	0.6731	0.3839
	90%	0.6544	0.6531	0.6538	0.3449
	10%	0.5629	0.7202	0.6319	0.5592
	30%	0.6826	**0.7323**	**0.7065**	0.3405
d3	**50%**	**0.6946**	0.6928	0.6937	**0.3046**
	70%	0.6574	0.6025	0.6287	0.3141
	90%	0.5951	0.4970	0.5417	0.3381

Table 7: Result of RMM Segmenter in Individual Dictionaries from VW Corpus

We will also analyze the experimental results of MM and RMM segmenters in the ensemble dictionary generated from the VW corpus. Figure 7 shows that a suitable top-percentage threshold (30%) is better to achieve the comparable performance for MM segmenter in the ensemble dictionary. The best F1 value (0.7105) is beyond those in three individual ones.

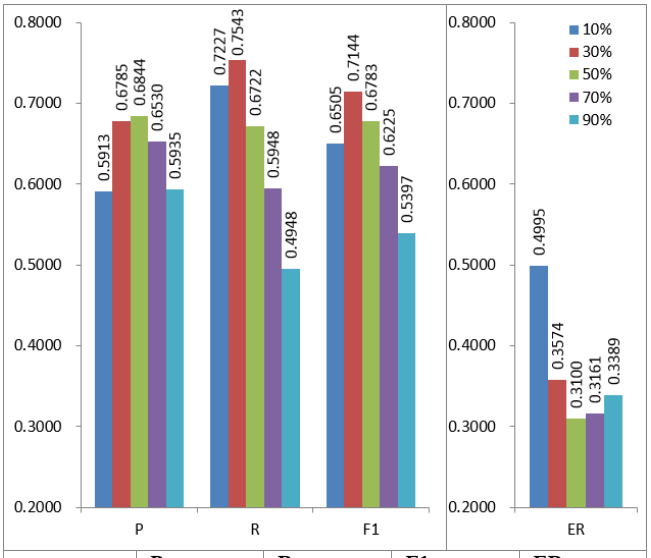


	P	R	F1	ER
10%	0.5901	0.7212	0.6491	0.5010
30%	0.6748	**0.7502**	**0.7105**	0.3616
50%	**0.6757**	0.6636	0.6696	**0.3185**
70%	0.6448	0.5873	0.6147	0.3236
90%	0.5887	0.4908	0.5353	0.3429

Figure 7: Result of MM Segmenter in Ensemble Dictionary from VW Corpus.

Figure 8 shows that there is an approximate trend of results between the RMM

segmenter and the MM segmenter. And the RMM segmenter's optimal values of four measures are all better than the MM segmenter's. For instance, the best F1 value of the MM segmenter is 0.7105, while the best F1 value of the RMM segmenter is 0.7144. The experimental result also verifies the effectiveness of unsupervised ensemble learning for the big corpus in the encyclopedia field.



	P	**R**	**F1**	**ER**
10%	0.5913	0.7227	0.6505	0.4995
30%	0.6785	**0.7543**	**0.7144**	0.3574
50%	**0.6844**	0.6722	0.6783	**0.3100**
70%	0.6530	0.5948	0.6225	0.3161
90%	0.5935	0.4948	0.5397	0.3389

Figure 8: Result of RMM Segmenter in Ensemble Dictionary from VW Corpus.

Overall, comparing the value of the four measures from *d1*, *d2*, and *d3* dictionaries, we will find that the value from the *d3* dictionary is always optimal under the same condition. And comparing the F1 value of ensemble dictionary from the VW corpus with that from the VWN corpus, we will list a rank from the best to the worst of 0.7144/RMM/VW, 0.7105/MM/VW, 0.7012/RMM/VWN, and 0.6967/MM/VWN. From all above experimental results, we can draw the following conclusions:

(1) The WFiDF feature is an optimized comprehensive feature than the WF feature and the DF feature for the unsupervised word extraction not only in the small corpus of an individual field, but also in the big corpus in the encyclopedia field. Manual selection is essential for the extraction of multisyllabic-words from long tail potential multisyllabic-words.

(2) Applying automatically generated dictionary of multisyllabic-words from the same

corpus, there will be comparable performance between the dictionary-based reverse maximum matching algorithm and the dictionary-based maximum matching algorithm for Vietnamese word segmentation. On the same premise, the word segmentation result of RMM segmenter is slightly better than that of MM segmenter.

(3) The UEL-based WE algorithm is not only effective in an individual field for Vietnamese multisyllabic-word extraction, but also more generalizable in the encyclopedia field. The more generalizable the corpus is, the more multisyllabic-words can be extracted. The effectiveness of unsupervised ensemble learning lies in the statistical, computational and representational advantages (Thomas, 2000).

anh chàng	biên giới	biểu diễn xiếc	bà lão	bàng quang
anh chồng	biên giới phía bắc	biểu quyết	bà mẹ	bành trướng
anh dũng	**biên hòa**	biểu tình	bà mối	bá thước
anh hào	biên kịch	biểu tượng	bà rịa	bác hồ
anh hùng	biên nhận	biểu đồ	bà triệu	bác sĩ
anh hùng liệt sĩ	biên phòng	biệt thự	bà ấy	bách khoa
anh linh	biên soạn	biệt tài	bài diễn văn	bách khoa toàn thư
anh quân	biên tập	bong bóng	bài giảng	bám biển
anh rể	biến số	buôn bán	bài hát	bán dâm
anh sơn	biến thế	buôn lậu	bài học	bán dạo
anh thư	biến thể	buôn ma thuột	bài thơ	bán hàng
anh trai	biến áp	buôn người	bài tập	bán kết
anh tuấn	biếu quà	buồn nôn	bài viết	bán lẻ
anh tú	biển hiệu	buổi làm	bài văn	bán nhà
anh việt	biển quảng cáo	buổi tọa đàm	bàn chân	bán nước
anh vũ	biển số	bà chủ	bàn phím	bán thịt
anh vợ	biển động	bà cô	bàn thắng	bán độ
anh ấy	biển đồ	bà cụ	bàn thờ	bánh mì
biên bản	biển động	bà già	bàn tròn	bánh pháo
biên dịch	biểu diễn	bà hoàng	bàn ủi	bánh tráng

Table 8: Data Sample of *d*30% from the VWN Corpus

In order to show the details of our dictionaries more intuitively, we give some data samples of them. Table 8 shows the first 100 multisyllabic-words in alphabetical order, which are partial data samples of our generated ensemble dictionary, totally including 60,521 words, under the threshold of 30% from the VWN corpus. While table 9 shows the first 100 multisyllabic-words of the ensemble dictionary under the threshold of 30% from the VW corpus. There are total 281,422 multisyllabic-words in the *d*30% dictionary from the VW corpus.

anh hùng	biên dịch	biến số	biển đen	biện pháp
anh linh	biên dịch địa chỉ mạng	biến thể	biển đông	biện tài
anh quân	biên giới	biến tính	biểu diễn	biệt cách
anh sơn	**biên hoà**	biến tần	biểu diễn số âm	biệt cư
anh thư	**biên hòa**	biến tốc	biểu hiện	biệt dược
anh tuấn	biên khu	biến áp	biểu kiến	biệt hoá
anh tú	biên phòng	biến điệu	biểu mô	biệt hải
anh việt	biên soạn	biến đổi	biểu mô lát	biệt khu
anh vũ	biên tái	biến đổi khí hậu	biểu quyết	biệt kích
anh đào	biên tập	biến đổi tuyến tính	biểu thức	biệt thức
anh ấy	biên tế	biến đổi tích phân	biểu trưng	biệt thự
axit amin	biên độ	biển báo	biểu tình	biệt điện
axit béo	biên đội	biển chết	biểu tượng	biệt đội
axít béo	biến chất	biển hồ	biểu tượng thất truyền	biệt đội thần tốc
axít clohiđric	biến chứng	biển khơi	biểu đạt	biệt động
axít flohiđric	biến dạng	biển số	biểu đồ	biệt động quân
axít phốtphoric	biến dị	biển số ô tô	biện chứng	bong bóng
axít prôpionic	biến hình	biển thước	biện chứng duy vật	bong bóng cá
axít selenơ	biến ngẫu nhiên	biển thắm	biện hàn	buôn bán
biên chế	biến nhiệt	biển tiến	biện hi	buôn làng

Table 9: Data Sample of *d*30% from the VW Corpus

There are so many multisyllabic-words in our ensemble dictionary, which are really out of expectation. After careful identification, it is not difficult to find that the difference between the old and new spelling forms of Vietnamese compound vowels also adds to the scale of words. For instance, there is only one spelling form of **biên hòa** in Table 8, while Table 9 contains two spelling forms of **biên hoà** and **biên hòa**. The two synonyms represent the same place name of Vietnam, which has motivated us that it will further improve the effectiveness of Vietnamese segmentation dictionary through a mapping rule between the old and new spelling forms of Vietnamese compound vowels.

4. Conclusion

This paper investigates how to unsupervisedly build a Vietnamese dictionary containing multisyllabic-words from a lot of text documents. The proposed UEL-based extraction algorithm takes full advantage of high frequency co-occurrence of multiple syllables. Using automatically extracted dictionaries, the dictionary-based MM and RMM algorithms can achieve comparable results.

Further research will concern the influence of other affixed resources, such as stop words, morphologic rules, semantic contexts, and so on. It is undeniable that manual work is indispensable for language experts to obtain a more precision dictionary. Therefore, the semi-unsupervised learning will be more expected for optimal multisyllabic-word extraction. We will also transfer above research productions to other suitable Asian languages like Thai, Japanese, Chinese, and so on.

5. Acknowledgment

The research is supported by the Key Project of State Language Commission of China (No. ZDI135-26), the Featured Innovation Project of Guangdong Province (No. 2015KTSCX035), and the Bidding Project of Guangdong Provincial Key Laboratory of Philosophy and Social Sciences (No. LEC2017WTKT002).

6. References

Dinh Q. T., Le H. P., Nguyen T. M., Nguyen C. T., Rossignol M., and Vu X. L., 2008, Word Segmentation of Vietnamese Texts: a Comparison of Approaches, In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC '08), pp. 1933–1936, Marrakech, Morocco.

Trung H. L., Anh V. L., Dang V.-H., and Hoang H. V., 2013, Recognizing and Tagging Vietnamese Words Based on Statistics and Word Order Patterns, *Advanced Methods for Computational Collective Intelligence*, Springer Berlin Heidelberg, pp. 3–12.

Liu W. Y. and Lin L., 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* (SIGIR '14), ACM, pp. 931–934.

Liu W. Y. and Wang L., 2016, How does Dictionary Size Influence Performance of Vietnamese Word Segmentation? In *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC '16), ELRA, pp. 1079–1083.

Vlachos A., 2011, Evaluating Unsupervised Learning for Natural Language Processing Tasks, In *Proceedings of the 1st Workshop on Unsupervised Learning in NLP* (UNSUP '11), ACL, pp. 35–42.

Liu W. Y. and Wang T., 2012, Online Active Multi-Field Learning for Efficient Email Spam Filtering, *Knowledge and Information Systems*, 33(1):117–136.

Kanaris I., Kanaris K., Houvardas I., and Stamatatos E., 2007, Words versus Character N-grams for Anti-Spam Filtering, *International Journal on Artificial Intelligence Tools*, 16:1047–1067.

Liu W. Y., Wang L., and Yi M. Z., 2013, Power Law for Text Categorization, In *Proceedings of the 12th China National Conference on Computational Linguistics* (CCL '13), Springer, LNAI 8202, pp. 131–143.

Sproat R. and Thomas E., 2003, The First International Chinese Word Segmentation Bakeoff, In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing* (SIGHAN '03), ACL, pp. 133–143.

Thomas G. D., 2000, Ensemble Methods in Machine Learning, In *Proceedings of the 1st International Workshop on Multiple Classifier Systems* (MCS '00), LNCS 1857, pp. 1–15.