ISSN 2007-9737

Building Resources For Vietnamese Clinical Text Processing

Hiep Nguyen Minh[1], Huyen Nguyen Thi Minh[2], Quyen Ngo The[2]

[1] Dalat University, Da Lat,
Vietnam

[2] VNU University of Science, Ha Noi,
Vietnam

hiepnm@dlu.edu.vn, huyenntm@hus.edu.vn, ngoquyenbg@vnu.edu.vn

Abstract. Clinical texts contain textual data recorded by doctors during medical examinations. Sentences in clinical texts are generally short, narrative, not strictly adhering to Vietnamese grammar and contain many medical terms which are not present in general dictionaries. In this paper, we investigate the tasks of lexical analysis and phrase chunking for Vietnamese clinical texts. Although there exist several tools for general Vietnamese text analysis, these tools showed a limited quality in the clinical domain due to the specific grammatical style of clinical texts and the lack of medical vocabulary. Our main contributions are the construction of an annotated corpus (vnEMR) and lexical resources in the medical domain and in consequence the improvement of the quality of the tools for clinical text analysis, including word segmentation, part-of-speech tagging and chunking.

Keywords. Chunking, clinical text, collocation, lexical resources, medical vocabulary, POS tagging, vnEMR, word segmentation.

1 Introduction

For automatic text understanding, we need to process texts through different linguistic levels of language analysis including morpho-syntax, syntax and semantics. For general Vietnamese text, several tools are available for the fundamental analysis tasks like word segmentation, part-of-speech (POS) tagging, chunking, constituency and dependency parsing. Most of these tools are trained on news articles and efficient on this kind of text. However, the application of these tools to domain-specific data such as clinical texts does not have the desired effect. The causes that affect the efficiency of clinical text processing are due to the characteristics of this type of text. For example, electronic medical records (EMR) contain a variety of medical terms, short sentences and possibly not enough syntactic components.

In this paper, we focus on two main tasks of clinical text processing. The first task is the processing at lexical level which includes word segmentation and POS tagging. The second task is shallow parsing, i.e. phrase chunking. For each task, we present the methods and tools available for Vietnamese text, evaluate and choose the best tools for EMR text processing, in building necessary language resources for the task. Enriching language resources is one of the key factors for increasing the quality of clinical text processing.

The rest of the paper is organized as follows. In the next section we present the task of word segmentation for Vietnamese general and medical texts. The construction of a medical vocabulary is also introduced in this section. Section 3 describes the details of POS tagging task for clinical text. The phrase chunking task is presented in Section 4. Finally, we conclude the paper and discuss some future research work.

2 Vietnamese Word Segmentation

In Vietnamese, the spaces in the text are only signs of separating sentences into syllables but not words, as there are many words having more than

one syllable. For example, *"sinh viên"* (student) and *"sân vận động"* (stadium) are words having respectively two and three syllables. Therefore, the word segmentation task is to solve ambiguity in the situation where a sentence has different ways of decomposing into words. For instance, given the sentence *"Học sinh học sinh học"* (the word to word translation is "Pupil learn biology"), there are many possible word segmentation results, such as "Học | sinh | học | sinh | học", "Học | sinh học | sinh học", "Học sinh | học sinh | học", "Học sinh | học | sinh học", but only one correct word segmentation which is "Học sinh | học | sinh học". The quality of this task directly affects other text processing tasks and applications related to language processing. The above sentence has a simple syntax but still it contains ambiguity in the word segmentation. Using the Google Translate tool to translate this sentence from Vietnamese into English, the result is *"Student students learn"*, which is a bad translation: integrated word segmentation information would give a better result.

The approaches to solving the problem of Vietnamese word segmentation often use resources such as dictionaries (including Vietnamese words possibly accompanied by Vietnamese morpho-syntactic information), as well as corpora, raw or word segmented so that a machine learning technique can be applied to solving the ambiguity of the word segmentation.

2.1 Word Segmentation Techniques

There are three main approaches for the word segmentation of Vietnamese texts, namely dictionary based, machine learning, and hybrid techniques combining the first two.

The dictionary based approach uses dictionaries to identify possible words, and the words in a sentence are determined by the maximum matching or longest matching methods [1]. Longest matching method result contains words with a bigest number of syllables, and the maximum matching method segments a sentence into the smallest possible number of words. Returning to the example above, there are three ways to segment that sentence, ”Học sinh — học sinh — học”, ”Học sinh — học — sinh học” and ”Học — sinh học — sinh học”. In this example, we also see that even with information of a full list of words, the problem of solving the ambiguity when there are multiple possible segmentations remains. Further, choosing the longest word is not always a good solution.

In fact, the use of dictionaries is not enough to provide a high accuracy for word segmentation, because there are several types of word not in the dictionary, such as proper name, date, time, unit... We can use the regular expressions to solve these problem.

The second approach includes methods using dictionaries that have word category informations combined with machine learning [4], [5]. This approach attempts to resolve the ambiguity between word segmentation possibilities by machine learning, based on information about word categories. Of course these methods require large training data labeled with word categories to obtain good results.

The third approach consists of machine learning methods (HMM in [6], CRF and SVM in [7], and Maximum Entropy in [8]), in which the word segmentation problem is expressed as the problem of labeling the space between two syllables, determining whether the space is the boundary between two words or not. These methods require training data to be word segmented Vietnamese texts.

2.2 Word Segmentation for Clinical Text

We consider in this section the task of word segmentation for clinical texts.

In word segmentation, a text normally is separated into lexical units, where each unit corresponding to a minimal number of syllables that represents a single meaning of a lexeme. However, when dealing with texts in a particular domain, we also need to identify specialized terms containing several lexical units, in order to process them as a whole in the following steps of text semantic analysis. So first, we need to build a vocabulary of medical terms. Then, we conduct an evaluation of two best tools for general text word segmentation (vnTokenizer [1] and DongDu [12])

on clinical records, in order to choose and improve one of them for clinical text processing.

The main difficulty when working with medical literature is the identification of medical terms. In order to solve this problem, we have built additional resources including a medical vocabulary and a medical corpus. More precisely, we obtained a medical corpus with over 10 million Vietnamese words, a candidate vocabulary filter for specialized terms, and a vocabulary with over 1800 medical terms. In addition, we performed the filtering and annotation of a set of acronyms in the existing clinical documents, which is essential for future clinical document analyses.

2.3 Building a Vocabulary of Medical Terms

To build a vocabulary of medicine, we rely on the source of medical literature collected on the Internet. We have collected the list of medicine entries from the "Dictionary of Vietnamese Medicine" published on many websites[1]. From this source, we collected 2142 entries, including 503 entries from the standard dictionary, and 1639 new entries. Besides, we also collected a medical corpus from articles and ebooks related to medicine. From this corpus, we have built a tool to filter the phrases as candidates for medical terms using n-gram statistics and pointwise mutual information (PMI).

A very important data source in this study is the clinical record corpus that we manually annotated. The corpus consists of 375451 words, with annotated information including word segmentation, POS tagging and phrase chunking. In this paper, we use this corpus to conduct experiments and evaluate the text processing tools. Table 1 shows the detail about our corpus (vnEMR).

[1] http://benhvathuoc.com/tu-dien-y-hoc/, https://thietbiysinh.wordpress.com/category/tu-dien-thiet-bi/tu-dien-y-hoc-viet-nam/

Table 1. Clinical record corpus (vnEMR)

File name	DBCS	YLCS	DBDT
Number of characters	901790	971721	695497
Number of syllables	173672	172882	133719
Number of words	142508	120341	112602

2.4 Evaluation of Word Segmentation

We have three file, DienBien_ChamSoc (DBCS), YLenh_Chamsoc (YLCS) and DienBien_DieuTri (DBDT). The evaluation of the word segmentation tools was conducted on two parts of the vnEMR corpus, DienBien_ChamSoc and YLenh_Chamsoc (containing 262849 words). The experiment was divided into two parts; in part 1 we run tools with the available models, and in part 2 we add the lexical resources and medical corpus that we have built. The result of the evaluation is shown on Table 2.

Table 2. Word segmentation result

	A	B	C	D
Recall	91.70%	**95.30%**	89.40%	**95.10%**
Precision	88.20%	**94.90%**	85.50%	**93.70%**

A: vnTokenizer (without additional dictionary)
B: vnTokenzier (with additional dictionary)
C: DongDu (without retrained on new data)
D: DongDu (retrained on new data)

The results show that the toolkit vnTokenizer has a higher precision stability. Therefore, we choose vnToken tools for further development for the purpose of achieving the highest precision word segmentation on the medical text. To improve the quality of vnTokenizer, we added regular expressions to determine unit words that appear quite frequently in the text, such as "lần/phút"(times/minute), "lít/giây"(liter/sec), "$38, 2°C$",... Adding regular expressions significantly increases the efficiency of word segmentation.

We also evaluated the results of the word separation with vnTokenizer on the third data set (DienBien_Dieutri) and the results are good.

When adding 96 new words collected from vnEMR data we obtained very positive results as can be seen in Table 3.

Table 3. EMR word segmentation evaluation result with vnTokenizer

	DBCS	YLCS	DBDT
Recall	98.00%	97.10%	98.90%
Precision	98.90%	98.20%	99.00%

The remaining errors were mainly attributed to ambiguous errors caused by two-syllable corresponding to two words that were seem to be a compound word since the word appears in the dictionary. With these types of errors, we tend to use the machine learning method to learn the error rule by the Brill approach [13].

3 Part-of-Speech Tagging

For a text that has been segmented, the next processing task is POS tagging, which will provide the basis for the parsing step (phrase chunking, sentence parsing) and finally the text semantic analysis. The POS tagging tool is used to determine the word category of each word that appears in the text by the context of this word. Each word corresponds to a certain morpheme and a grammatical role. POS label sets may vary depending on the concept of the lexical unit and language information to be exploited in specific applications. Each word in a language can generally be associated with many word categories, and understanding correctly the meaning of a word depends on whether it is correctly word category identified or not.

In this part, we introduce the POS tagging problem, apply known techniques to our clinical corpus, and evaluate the results.

3.1 Part-of-Speech Tagging Techniques

The process of POS tagging for a word-segmented text can be divided into two steps as follows [14]:

— Step 1. Label prediction, *i.e.* look for every word the set of all categories that it might have. This categories collection can be obtained from a dictionary or manually labeled texts. For a new word that has not yet appeared in the corpus, it can be assigned a default category. For morphological languages, we could also rely on word morphology to predict the corresponding word category, but that approach is not applicable to Vietnamese.

— Step 2. Decide the labeling result, which is the phase of ambiguity resolution, *i.e.*, selecting for each word instance, from its predictive category set, the category that best matches the context where it appears. This ambiguity resolution can be accomplished by a grammatical rule system, or by supervised machine learning methods[15].

There are three main approaches for solving ambiguous word categories: rule-based tagging [19], stochastic tagging [20] and transformation-based tagging [13]. Rule-based tagging uses a set of rules built by hand to determine the label for each word. Stochastic tagging uses a training corpus to determine the probability of a given word being assigned to a label in the given context. Transformation-based tagging uses the characteristics of both approaches. Like the rule-based tagging, it resolves ambiguity in the POS tagging process by using rules. However, these rules are not manually written but are automatically extracted from the pre-labeled training corpus, which makes it similar to the stochastic tagging.

In general, approaches using machine learning methods yield better results for POS tagging problems. Therefore, in order to have a good labeling system, it is necessary to focus on building a good quality manually labeled reference corpus, and ensures consistency in the categories system. In the context of the VLSP project [4], a set of word categories was built. Along with that, a Vietnamese corpus named VietTreeBank has also been developed, including 20000 sentences labeled as a standardized corpus to be used for training as well as evaluating the Vietnamese POS tagging systems [16].

3.2 Part-of-Speech Tagging of Clinical Text

We use the basic word category labels that have been defined in the VLSP project[2]. In addition to this set of labels, we determine the word category for the new medical terms we have collected and add them to the dictionary, along with the category of acronyms and the units of measure.

For the problem of POS tagging in the clinical text, we use existing POS tagging tools, retrained with the VietTreeBank corpus. Our clinical corpus vnEMR was also labeled manually for word category, and a part of it is used for training and testing.

In [13], the authors present some well-known labeling tools for Vietnamese: vnTagger [9], JvnTagger and RDRPOSTagger[3]. We chose to use the RDRPOSTagger tool because its processing speed is faster than the two other tools. Besides, two famous toolkits: ClearNLP[3] and Stanford POSTagger[4] are also used in our experiment. With these three tools, we conducted experiments and evaluations on vnEMR data including 11943 sentences. The accuracy results are shown in Table 4.

Table 4. POS tagging evaluation result

Training data	A	B	C
Only Viet-Treebank	78.80%	80.43%	78.82%
Additional 1/10 of the vnEMR data	93.08%	91.73%	92.35%
Additional 9/10 of the vnEMR data	98.90%	98.67%	98.70%

A: RDRPOSTagger
B: ClearNLP
C:Stanford POS Tagger

So we realize that the vnEMR corpus plays an important role. Specifically, when applying word category labeling for medical text, if only use VietTreebank training data, all three tools are quite low accuracy(less than 81% for all 3 tools). However, after adding 1/10 of vnEMR data to the training data, the results improved significantly. The results add 9/10 EMR data to the training data for very high accuracy, over 98.5% for all tagging tools.

In the above test, we found that the RDR-POSTagger kit yielded the lowest result when training data is only VietTreeBank (newspaper text), but yielded the best results when training data is supplemented with vnEMR data. The opposite happens with the ClearNLP. Thus, the RDRPOSTagger is a suitable POS tagging kit for clinical text. We also tried using less EMR data for training, with a 4/10 ratio, the accuracy achieved with RDRPOSTagger was 95.5%.

We have presented two basic text processing steps, word segmentation and POS tagging. In the next section, we will describe the phrase chunking problem for clinical text.

4 Phrase Chunking

After word segmentation and POS tagging, the next step is phrase chunking. In this step, a sentence will be chunked into phrases. Phrase chunking can be seem as a shallow syntax parsing step, its results are used for deeper text analysis steps such as syntax parsing, dependency parsing, and semantic analysis.

The following example describes the results of the word clustering process on an English text:

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September].

In particular, the phrase labels used in the popular word clustering problem for languages are:

— NP (noun phrase): is the phrase in which nouns play a central role.

— VP (verb Phrase): is the phrase in which verbs play a central role.

— ADVP and ADJP: adjective- and adverb-phrases.

[2] http://vlsp.vietlp.org:8080/demo/vcl/PoSTag.htm
[3] https://github.com/clir/clearnlp
[4] https://nlp.stanford.edu/software/tagger.shtml

— PP and SBAR: prepositional phrases and sub clauses in the sentence.

— CONJC: conjunctions.

4.1 Clinical Text Chunking

For the chunking problem, the basic approach is to use supervised learning machine methods or construct regular expressions that describe the rules.

In a rule-based method, the chunker tool consists of a set of regular expressions. A rule-based systems can be developed relatively easily, without training corpus. However, these systems are difficult to apply and adapt to a new text type.

For the machine learning methods, the most important task of a phrase chunking system is to build a sample corpus, which is used to train the model. To implement phrase chunking for clinical documents, we continue to use the VietTreeBank corpus with defined phrase labels in combination with the manually labeled vnEMR corpus as training data, and to evaluate results.

Chunking tools are quite abundant. In [22], the authors tested and compared the results of some chunking tools, using the GENIA TreeBank medical corpus [23] as training data and test data. Results showed that, when conducting analyses on noun and verb phrase, OpenNLP tools, based on the maximum entropy model, always achieved the highest efficiency (F-measure of 89.7% and 95.7% for noun and verb phrases respectively). Two other tools that achieved similar performance to OpenNLP, but a little lower, are Genia Tagger and Yamcha.

Ginea Tagger is an integration of several tools: POS tagging, chunking, named entity identifier. This tool is based on the maximum entropy model. The downside of this tool is that only certain corpus can be used to train the model.

Yamcha is an easy-to-customize, open source chunking tool built on SVM algorithms. Based on the evaluation of the tool presented, we chose to use OpenNLP and Yamcha to test the phrase chunking on vnEMR corpus in cases using only VietTreeBank and add a part of vnEMR corpus to the training data.

4.2 Evaluation Results

The accuracy of each case using OpenNLP and Yamcha tools is shown in the following Table 5.

Table 5. Chunking evaluation result

		Precision	Recall	F1
A	**OpenNLP**	36.92	49.82	42.41
	Yamcha	55.51	82.24	66.28
B	**OpenNLP**	63.22	73.04	67.7
	Yamcha	72.62	88.77	79.89
C	**OpenNLP**	90.06	92.14	91.08
	Yamcha	93.37	94.79	94.08
A: Only VietTreeBank B: Additional 1/10 of the EMR data C: Additional 9/10 of the EMR data				

In general, with all three cases using these training data sets, the Yamcha tool always produces better word chungking results than OpenNLP. The above results also show that with the addition of 9/10 vnEMR data for training improves significantly the quality of phrase chunking compared to only adding 1/10 of vnEMR data or using only VietTreeBank as training data. We also tested using Yamacha tool with the addition of 4/10 of EMR data for training with VietTreeBank for the F1 measure is 82.11%, relatively low. The detailed clustering results for each phrase types using Yamcha are presented in the tables below.

Table 6. Phrase chunking result of NP and VP

	NP			VP		
	P	**R**	**F1**	**P**	**R**	**F1**
A	56.5	94.59	70.74	61.01	72.27	66.16
B	71.15	93.05	80.64	78.07	85.72	81.72
C	91.36	94.36	92.83	95.53	95.56	95.54
A: Only VietTreeBank						
B: Additional 1/10 of the EMR data						
C: Additional 9/10 of the EMR data						

Table 7. Phrase chunking result of AP and PP

	AP			PP		
	P	**R**	**F1**	**P**	**R**	**F1**
A	30.04	74.62	42.83	97.07	99.61	98.32
B	50.98	82.4	62.99	96.7	98.97	97.82
C	90.14	91.85	90.99	96.02	95.96	95.99
A: Only VietTreeBank						
B: Additional 1/10 of the EMR data						
C: Additional 9/10 of the EMR data						

5 Conclusion

In this paper, we present the basic text processing steps for clinical text. We do not go into technical details, and focus on exploring existing text processing tools, thereby identifying problems encountered by these tools with clinical text. Our main contribution is the building of language resources such as medical terminology, medical data corpus, and a annotated clinical corpus. These data make the text processing tools work well in clinical text. We aim to bring the processing to a higher level in the next study, such as syntax parsing and semantic parsing. Along with that is the research and development of medical applications related to natural language processing.

Acknowledgements

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2016-42-01.

References

1. **Hông-Phuong, L., Thi Minh Huyên, N., Roussanaly A., & Vinh H.T. (2008)**. A Hybrid Approach to Word Segmentation of Vietnamese Texts. In:**Martín-Vide, C., Otto, F., Fernau, H. (eds)**. *Language and Automata Theory and Applications, LATA Lecture Notes in Computer Science*, Vol. 5196, Springer, Berlin, Heidelberg.
2. **Cam-Tu, N. & Xuan-Hieu, P. (2007)**. *JVnSegmenter: A Java-based Vietnamese Word Segmentation Tool*.
3. **Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham, & Son Bao Pham (2014)**. RDRPOSTagger: A Ripple Down Rules-based Part-Of-Speech Tagger. *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 17-20.
4. **Dinh, D., Kiem, H., & Toan, N.V. (2001)**. Vietnamese Word Segmentation. *The 6th Natural Language Processing Pacific Rim Symposium*, pp. 749–756.
5. **Pham, DD., Tran, GB., & Pham, SB. (2009)**. A hybrid approach to Vietnamese word segmentation using part of speech tags. *International Conference on Knowledge*.
6. **Nguyen, P.T., Nguyen, V.V., & Le, A.C. (2003)**. Vietnamese word segmentation using hidden markov model. *International Workshop for Computer, Information, and Communication Technologies on State of the Art and Future Trends of Information techonologies in Korea and Vietnam*.
7. **Nguyen, C.T., Nguyen, T.K., Phan, X.H., Nguyen, L.M, & Ha, Q.T. (2006)**. Vietnamese word segmentation with CRFs and SVMs. *An investigation, Proceedings of the 20t PACLIC*, pp. 215–222.
8. **Dinh, D. & Vu, T. (2006)**. A maximum entropy approach for Vietnamese word segmentation. *Proceedings of 4th RIVF VietNam*, pp. 12–16.
9. **Dinh, Q.T., Nguyen, T.M.H., Vu, X.L., Rossignol, M., Le-Hong, P., & Nguyen, C.T. (2008)**. Word segmentation of Vietnamese texts: a comparison of approaches. *Proceedings of The Sixth International Conference on Language Resources and Evaluation*, Marrakech.
10. **Le, H.P, Nguyen, T.M.H, Azim, R., & Ho, T.V. (2008)**. A hybrid approach to Word Segmentation of Vietnamese texts. *Language and automata theory and applications 2ndinternational conference, LATA*.
11. **VLSP project (2012)**. *Vietnamese Language Processing*. http://vlsp.vietlp.org.
12. **JNLP. (2010)**. http://viet.jnlp.org/dongdu.
13. **Nguyen, T., Minh, H., Vu-Xuan, L., & Le-Hong, P. (2003)**. Sử dụng bộ gán nhãn từ loại xác suất QTAG cho văn bản tiếng Việt. *Kỷ yếu hội thảo ICT.rda'03, Hà Nội*.
14. **Le-Hong, P. & Nguyen, T. M. H. (2013)**. Part-of-Speech Induction for Vietnamese. *The Fifth*

International Conference on Knowledge and Systems Engineering (KSE'13), Vol. 2, Springer-Verlag, pp. 261–272.

15. **Nguyen, P.T., Xuan, L.V., Nguyen, T.M.H., Nguyen, V.H. & Le-Hong, P. (2009)**. Building a Large Syntactically-Annotated Corpus of Vietnamese. *Proceedings of the 3rd Linguistic Annotation Workshop*, Singapore, pp. 182–185.
16. **Le-Hong, P., Nguyen, T.M.H., & Rossingnol, M.A. (2010)**. An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts. *Actes du Traitement Automatique des Langues Naturelles (TALN'10)*.
17. **Nguyen Minh Hiep, Nguyen Thi Minh Huyen, & Ngo The Quyen. (2016)**. Nghiên cứu về tập từ loại tiếng Việt sử dụng kĩ thuật phân cụm. *Kỷ yếu của Hội thảo Quốc gia về CNTT&TT lần thứ 18*.
18. **Nguyen, M.L. & Cao, T.H. (2008)**. Constructing a Vietnamese Chunking System. *The 4rd National Symposium on Research, Development and Application of Information and Communication Technology (ICTrda'08), Science and Technics Publishing House*, pp. 249–257.
19. **Nguyen Huong Thao, Nguyen Phuong Thai, Nguyen Le Minh, & Ha Quang Thuy. (2009)**. Vietnamese Noun Phrase Chunking based on Conditional Random Fields. *Proceedings of The first International Conference on Knowledge and Systems Engineering (KSE'09)*.
20. **Brill, E. (1995)**. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Journal of Computational Linguistics*, Vol. 21, No. 4, pp. 543–565.
21. **Nguyen Thi Minh Huyen, Vu Xuan Luong, & Le Hong Phuong. (2003)**. Sử dụng bộ gán nhãn từ loại xác suất QTAG cho văn bản tiếng Việt. *Kỷ yếu hội thảo ICT.rda'03*.
22. **Le-Hong, P. & NGUYEN, T.M.H. (2013)**. Part-of-Speech Induction for Vietnamese. *The Fifth International Conference on Knowledge and Systems Engineering (KSE 2013)*, Vol. 2, Springer-Verlag, pp. 261–272.
23. **Nguyen, P.T., Xuan, L.V., Nguyen, T.M.H., Nguyen, V.H. & Le-Hong, P. (2009)**. Building a Large Syntactically-Annotated Corpus of Vietnamese. *Proceedings of the 3rd Linguistic Annotation Workshop*, Singapore, pp. 182–185.
24. **Brown, P.F., Della-Pietra, V.J., Desouza, P.V., Lai, J.C. & Mercer, R.L. (1992)**. Classbased n-gram models of natural language. *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479.
25. **Clark, A. (2003)**. Combining distributional and morphological information for part of speech induction. *Proceedings of (EACL'03)*, pp. 59–66.
26. **Jianfeng, G., Joshua, T., & Goodman-Jiangbo, M. (2005)**. *The Use of Clustering Techniques for Language Modeling – Application to Asian Languages*. Microsoft Research, China.
27. **Biemann, C. (2011)**. *Unsupervised Part-of-Speech Tagging Employing Efficient Graph Clustering*.

Article received on 14/12/2017; accepted on 15/02/2018. Corresponding author is Hiep Nguyen Min.