ICIC Express Letters
Part B: Applications
Volume **8**, Number **7**, July **2017**

ICIC International ©2017 ISSN 2185-2766
pp. **1061–1068**

IMPROVING WORD ALIGNMENT BASED ON NAMED ENTITY

Phuoc Tran[1,2], Dien Dinh[2,*] and Hien Thanh Nguyen[1,*]

[1]Faculty of Information Technology
Ton Duc Thang University
19 Nguyen Huu Tho St., Tan Phong Ward, District 7, Ho Chi Minh City 700000, Vietnam
tranthanhphuoc@tdt.edu.vn; *Corresponding author: hien@tdt.edu.vn

[2]Faculty of Information Technology
VNU-HCM University of Science
227 Nguyen Van Cu St., Ward 4, District 5, Ho Chi Minh City 700000, Vietnam
*Corresponding author: ddien@fit.hcmus.edu.vn

Received February 2017; accepted May 2017

Abstract. *Unsupervised word alignments are widely used in phrase-based statistical machine translation. The quality of this model is proportional to the size and quality of a bilingual corpus. However, for low-resource language pairs such as Chinese and Vietnamese, the result of unsupervised word alignment sometimes is of low quality due to the sparse data. In this paper, we integrate the characteristics of named entities into the word alignment model to enhance the quality of Chinese-Vietnamese word alignment. The experimental results show that our method improves the performance of word alignments as well as the quality of machine translation.*
Keywords: Word alignment, Chinese-Vietnamese machine translation, Named entity

1. **Introduction.** The phrase table of phrase-based statistical machine translation (PS MT) is extracted from the results of word alignments (WA) of sentence pairs in a bilingual corpus. Therefore, the quality of the phrase table depends on the performance of a WA model. To obtain a WA model for PSMT, we need to provide the system with a bilingual corpus. The larger and cleaner the bilingual corpus is, the higher the quality of the WA model is. Currently, to achieve a quality WA model for PSMT, GIZA++ toolkits are being widely used.

However, for low-resource language pairs such as Chinese-Vietnamese, the unsupervised WA model will produce some incorrect results due to sparse data. Also, because of the nature of unsupervised learning, this model does not take advantage of the relationships among the languages to increase the quality of WA. Figure 1 shows incorrect alignments of GIZA++. The dash links represent aspects of the imprecise WA.

In this paper, we propose a WA improvement method (WAI) based on named entities (NEs) to increase unsupervised WA models. First, we perform unsupervised WA for a Chinese-Vietnamese bilingual corpus. Then, we adjust incorrect cases. NE is one of the most important linguistic relationships between Chinese and Vietnamese. We used it to correct word alignment errors and enhance PSTM performance. Compared with other methods for improving word alignment, our method is useful for low-resource language pairs having close linguistic relationship, such as Chinese and Vietnamese.

The rest of this paper is structured as follows. Section 2 presents some related work. Section 3 gives a brief background of Chinese-Vietnamese linguistic relationships and named entity. Section 4 provides a detailed description of our proposed method. Section 5 shows and discusses the results of our experiments. Finally, Section 6 summarizes our work and provides our main conclusion.

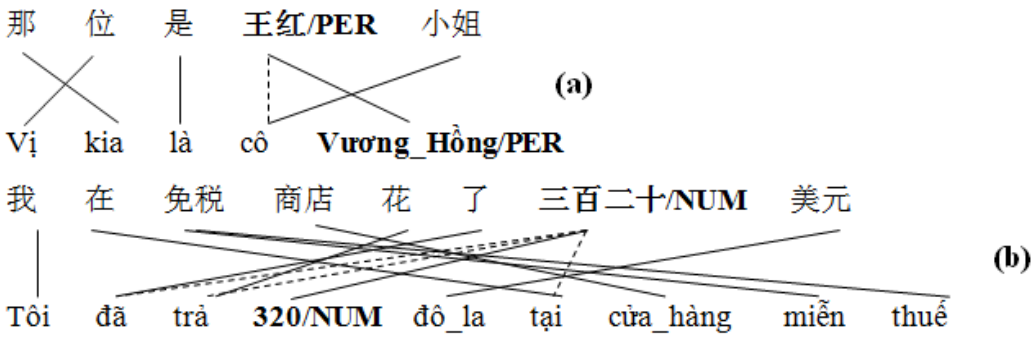

FIGURE 1. An example of the incorrect alignments due to not taking advantages of named entities. The case (a) indicates the incorrect WA of person name (王红→"cô"), and (b) shows the imprecise WA of number (三百二十→"đã", "trả" and "tại").

2. **Related Work.** In this section, we focus on surveying WA methodologies in order to improve the performance of SMT for low-resource language pairs such as Chinese and Vietnamese.

A common approach to improving WA for low-resource language pairs is using a pivot language (PL). In this approach, researchers focus on improving the quality of WA of the language pair X-Y based on the third language Z.

In [2], T. Levinboim and D. Chiang developed a multitasking learning method using French as the PL to improve WA of machine translation of low-resource language pair, i.e., Czech-English. WA results were extracted from the training results of the language pairs including Czech-English, Czech-French, and French-English. The final Czech-English WA result includes initial Czech-English direct WA result, plus Czech-English WA result from combining Czech-French and French-English.

Utilizing the same approach, R. Dabre et al. [3] used many languages to improve performance of Japanese-Hindi SMT. This work focused on exploitation of the phrase tables generated from multiple pivot phrase tables to support the source-target one.

In literature, PL is a third language that is independent from source and target languages. However, in [4], C. Chu et al. used the shared language of Chinese and Japanese as the PL. This is due to the fact that Chinese characters are used in both Chinese and Japanese.

Integrating multiple WA approaches is a common trend to improve the performance of WA alignment for low-resource language pairs. N. Durrani and P. Koehn [5] exploited similarity of the Hindi and Urdu languages simultaneously, using English as the PL to improve Urdu-Hindi machine translation. Hindi and Urdu languages are written in different scripts, but they have a close linguistic relationship. Moreover, they also share a similar grammatical structure and have a high overlapping vocabulary.

Also using the combined method of linguistic similarity and PL to enhance WA performance, P. Nakov and H. T. Ng [1] proposed a translation method from a low-resource language X1 to rich-resource language Y with a limited bilingual corpus. Moreover, they used an available rich-resource bilingual corpus X2-Y, in which X1 has close linguistic relationships with X2. The authors used X2-Y corpus to build a translation model, and then exploited similarity between X1 and X2. They tested two language pairs, including Indonesian (X1)-English (Y) (using Malay as X2) and Spanish (X1)-English (Y) (using Portuguese as X2).

In [6], S. Pal et al. integrated three word alignment models, including the unsupervised model (GIZA++) (A1), the semi-supervised model (Berkeley) (A2), and the rule-based model (A3). We consider this research to be closest to ours. We also use a hybrid model that combines multiple word alignment models. We also use an unsupervised WA model

as well as a rule-based model to advance WA. However, our paper has a few differences, as follows.

- We only improve WA based on the result of the unsupervised WA.
- In the rule-based WA model, we exploit NEs to advance WA. Unlike previous research that only used NE transliteration, our NE translation is a combination of translation, transliteration and transformation-rule-based translation.

3. Background.

3.1. The linguistic relationships between Chinese and Vietnamese.

A Chinese word usually includes many meaningful characters. When translating Chinese into Vietnamese, the meaning of Chinese word is usually divided into three cases. The first case is where the meanings of Chinese characters (in a Chinese word) are their Sino-Vietnamese meanings, usually a 1-1 correspondence. The second case is where the meanings of the Chinese characters are similar or related to the meanings of the Chinese word containing those characters. The final case is where the meanings of Chinese characters are not relevant to the meaning of the Chinese word containing them.

In the first case, Vietnamese words are largely borrowed from Chinese words (often called Sino-Vietnamese, which make up about 65% of the total number of Vietnamese words [9]. Sino-Vietnamese is a reading way of Vietnamese people. For example, the Chinese word 银行 (bank) is pronounced "yín háng" (rendered using Pinyin), with Vietnamese's pronunciation being "ngân hàng". A Chinese character may be pronounced by many Sino-Vietnamese words, but in a specific context, one Chinese character only corresponds to one Sino-Vietnamese. As in the above example, 银行, the corresponding Sino-Vietnamese pronunciation of character 银 is "ngân" and the pronunciation of 行 is "hành" "hạnh" "hàng" "hạng". Nevertheless, when 银 and 行 are combined into a unique word, 银行, we only pronounce it "ngân hàng".

In the second case, meaning of a Chinese word is a combination of Pure-Vietnamese meanings of Chinese characters forming it. Vietnamese vocabulary, apart from words borrowed from other languages, is called Pure-Vietnamese. The word "Pure" in "Pure-Vietnamese" means vernacular (the native language). A Chinese character is often translated into a one-syllable Vietnamese word, and few remaining can be translated into a Vietnamese word with more syllables. Some examples are 天/trời (heaven), 市/"thành phố" (city). Another feature of translation from Chinese to Pure-Vietnamese is that meaning of Chinese characters can be reordered in Pure-Vietnamese translation. For example, the Chinese word 零钱 with 零/lẻ (loose) and 钱/tiền (cash, money) is translated into Vietnamese as "tiền lẻ" (loose cash) (instead of "lẻ tiền").

In other cases, the words which their meanings are not related to the characters forming them. 好的 ("đúng": right) is a typical instance. The corresponding Sino-Vietnamese meanings of characters are "hảo", "đích" and their Pure-Vietnam meanings are "tốt (good), "của" (of). Clearly, 'hảo đích" and "tốt của" (good of) are not relevant to the correct meaning of "好的" (right).

3.2. A brief introduction of Chinese NE.

Chinese NE consists of four categories, i.e., person's name (PER), location name (LOC), organization name (ORG), and number expression (NumExp). PER is formed by the following structure: <Family name> (F) <Given name> (G), where both F and G have a length from one to two characters. For example, the PER "赵经生", "赵" is F part and "经生" is G part. Chinese PER is translated into Sino-Vietnamese.

Normally, a Chinese LOC has a maximum of 10 characters, structured as follows: <name part> <keyword> [10]. LOC is usually terminated by a <keyword>. For example:

北京市 (or 北京) ("Thành Phố Bắc Kinh": Beijing city) where 北京 is a <name part> and 市 is a <keyword>. In some cases there is no <keyword> in LOC.

ORG (or "full OG") is more complicated than PER and LOC because it usually includes PER, LOC, and many other entities combined together; its maximum length is 15 characters. Its common structure is: { [PER] [ORG] [LOC] [kernel name] }* [organization type] <keyword> [11]. Symbol {}* means selecting at least one of items, for example, 北京语言学校 ("Học viện Ngôn ngữ Bắc Kinh": Beijing Language Institute), in which 北京 is a LOC and 学院 is a <keyword>.

NumExp includes numerical characters associated with the representative keywords for each NumExp, for example, NumExp 十二月 ("tháng 12": December), in which, 十二 are numerical characters and 月 is a <keyword>.

4. **Word Alignment Improvement Based on NE.** Figure 2 shows our NE-WAI framework.

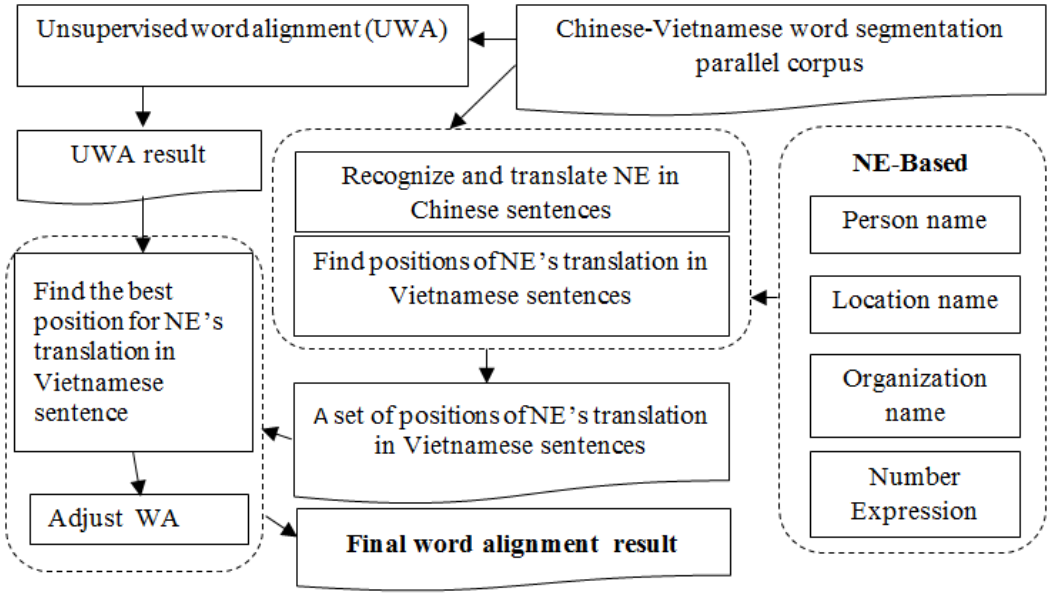

FIGURE 2. Our NE-WAI framework

4.1. **Word segmentation and word alignment for the corpora.** The words in Chinese and Vietnamese are not distinguished by spaces [7]. In the work, we used Stanford Segmenter[1] to segment words, and the Vietnamese corpus is segmented by the CLC_VN_WS[2] toolkit. Then, these corpora are words aligned by the GIZA++ toolkit.

4.2. **NE's recognition and translation.** We used the method in [9] to recognize and translate PER, LOC, and ORG. Particularly for NumExp, the system identified and translated this NE based on [13]. In it, PER, LOC and ORG are translated based on combination of rule method (Sino-Vietnamese) and statistical one (2-gram language model). As for NumExp, the system uses some transformation rules to transcript them into Vietnamese numbers.

4.3. **NE-based word alignment improvement.** The system uses Stanford NER and CLC_VN_NER toolkits to tag NE labels for Chinese and Vietnamese corpora, respectively.

The system selects the candidate WA and adjusts them based on NE in both Chinese and Vietnamese. The method is as follows.

[1]Download at: http://nlp.stanford.edu/software/segmenter.shtml
[2]Download at: http://www.clc.hcmus.edu.vn/?page_id=471&lang=en

Assumption:

- LR is a set of Chinese NE.
- (C, V) is a Chinese-Vietnamese sentence pair with $C = c_1c_2 \dots c_n$ and $V = v_1v_2 \dots v_m$; c_i and v_i are the words in Chinese and Vietnamese sentences, respectively.
- A is a set of WA of the (c, v) sentence pair.
- $AV(i)$ is a set of Vietnamese positions that are aligned with the Chinese word c_i:

$$AV(i) = \{j \in [1 \dots m], (i-j) \in A, c_i \in C\} \tag{1}$$

- $AC(j)$ is a set of Chinese positions that are aligned with the Vietnamese word v_j:

$$AC(j) = \{i \in [1 \dots n], (i-j) \in A, v_j \in V\} \tag{2}$$

- $VC(i)$ is a set of Vietnamese words that are aligned with the Chinese word c_i:

$$VC(i) = \{v_j \in V, j \in AV(i)\} \tag{3}$$

Finding the best position k_{best} of the Vietnamese word that is a translation of the Chinese word $c_i \in LR$:

- Given that $tr(c_i)$ is a Vietnamese translation of c_i, the set of positions POS of $tr(c_i)$ in the set $VC(i)$ is determined as follows:

$$POS = Position(tr(c_i), VC(i)) = \begin{cases} \{k \in AV(i), & \text{if } tr(c_i) \in VC(i)\} \\ \emptyset, & \text{if else} \end{cases} \tag{4}$$

- *Finding the best position* k_{best}.

$$k_{best} = \begin{cases} -1, & \text{if } POS = \emptyset \\ k \in POS, & \text{if } |POS| = 1 \\ \arg\max_{k \in POS} \left(1 - abs(rel(i, C) - rel(k, V))\right), & \text{if } |POS| > 1 \end{cases} \tag{5}$$

where $rel(i, C)$ and $rel(k, V)$ are relative positions of c_i and v_k in the Chinese sentence C and the Vietnamese sentence V, respectively; abs is the modulus.

$$rel(i, C) = \frac{1}{n} \text{ and } rel(k, V) = \frac{k}{m} \tag{6}$$

where n and m are total number of words of the Chinese sentence C and the Vietnamese sentence V, respectively.

Adjusting WA.

- If $k_{best} = -1$, do not customize the WA.
- If $k_{best} \neq -1$,

$$\begin{cases} +\text{Store alignment } (i - k_{best}) \\ +\text{Delete all alignments } (i-j),\ j \in AV(i) \text{ and } j \neq k_{best} \\ +\text{Delete all alignments } (i' - k_{best}),\ i' \in AC(k_{best}) \text{ and } i' \neq i \text{ and } |AV(i')| \geq 2 \end{cases} \tag{7}$$

5. **Experiments.**

5.1. **Toolkits in experiment.** We used Stanford Segmenter and Stanford NER[3] to segment words and recognize NE in the Chinese corpus, respectively. As for Vietnamese, we used the CLC_VN_WS toolkit to segment words and utilize the CCL_VN_NER[4] toolkit to recognize NE.

[3]Download at: http://nlp.stanford.edu/software/CRF-NER.shtml

[4]Download at: http://www.clc.hcmus.edu.vn/?page_id=471&lang=en

In addition, we also used GIZA++[5] toolkit to align words, Chinese characters and Vietnamese spelling words. The SRILM[6] toolkit is used to train the language model and a state-of-the-art Moses[7] toolkit is used for phrase-based SMT.

5.2. **Experimental corpora.** Our experiment bilingual corpus consists of 35,623 Chinese-Vietnamese sentence pairs, which were extracted from Chinese conversational textbooks, online Chinese-Vietnamese forums and Chinese-Vietnamese bilingual websites. Documents in the corpus are mostly communicative text, so the length of the sentences is relatively short. We used 90% of the sentences for training, 5% of the sentences for testing, and the remaining 5% of the sentences for developing. We used these corpora to perform two experiments including WS translation and NE based WAI (NE-WAI) translation.

Table 1 shows the number of words (NW), number of sentences (NS) and number of words per sentence (NW/NS) in the experimental corpora of the two translation systems. These corpora are divided as follows: in each group of 20 sentences, the first 18 sentences are for training, the 19th sentence is for developing and the last one is for testing.

TABLE 1. Illustration of number of words as well as number of words per sentence

	Chinese			Vietnamese		
	Training	Developing	Testing	Training	Developing	Testing
NoW	236,598	13,071	13,186	296,256	16,328	16,509
NoS	32,061	1,781	1,781	32,061	1,781	1,781
NoW/NoS	7.4	7.3	7.4	9.2	9.2	9.3

Here NoW: number of words; NoS: number of sentences.

5.3. **Experimental result.** We evaluated the impact of NE-based WAI on translation results based on the BLEU score [14]. Table 2 shows that BLEU scores of the NE-WAI translation system are higher than the ones of the WS translation system.

TABLE 2. BLEU score of five translation systems

	Case 1	Case 2	Case 3	Case 4	Case 5	Average
WS	35.49	35.03	34.50	35.47	35.07	35.11
NE-WAI	35.70	35.11	34.73	35.52	35.21	35.25

5.4. **Analysis.** The improved WA increases final SMT performance. Below are two cases of incorrect alignments by the WS translation system using GIZA++ in the training corpus and their results after the improved WA (Figure 3).

The dash arrows in Figure 3 are incorrect alignments which should be deleted. Considering Case 1, the Chinese PER 王红 must be translated into SINO "Vương Hồng" and the characters in the Chinese PER must be mapped "1-1" to the spelling words in Vietnamese PER (王→"Vương", and 红→"Hồng"). Therefore, WA 王红→"cô" is erroneous and needs to be deleted. In Case 2, 三百二十 is a Chinese number and is transliterated as "320". So, three incorrect alignments (三百二十→"đã", 三百二十→"trả", and 三百二十→"tại") should be deleted.

According to [8], performance of WA directly affects quality of phrase extraction and word order in PSMT. So, with a significant improvement in alignment result, the NE-WAI translation system is also improved.

[5]Download at: Download at: http://www.opentag.com/okapi/wiki/index.php?title=GIZA%2B%2B_Installation_and_Running_Tutorial

[6]Download at: http://www.speech.sri.com/projects/srilm/download.html

[7]Download at: http://www.statmt.org/moses/?n=Moses.Releases

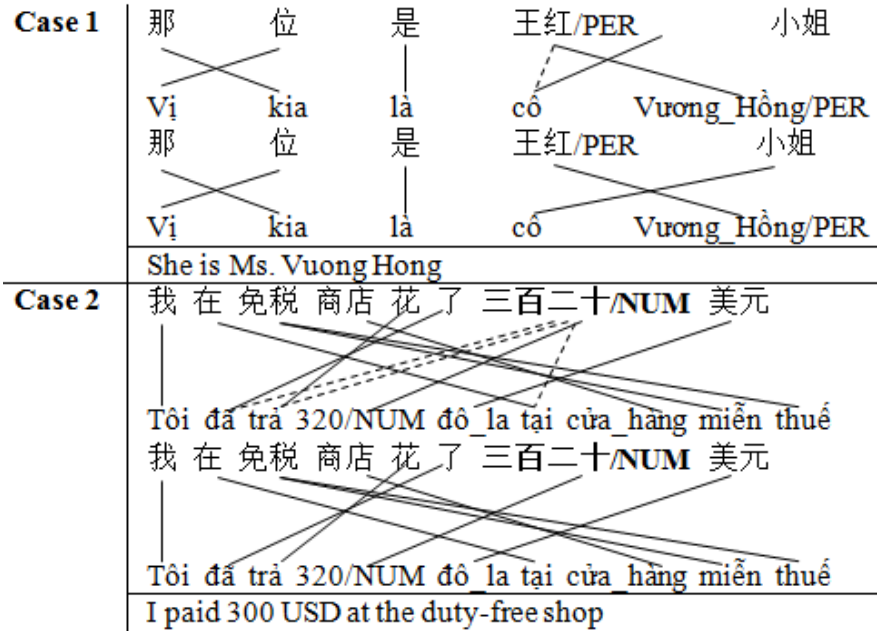

FIGURE 3. The two incorrect alignment cases and their improved results

6. **Conclusions and Perspectives.** In this paper, we improved a WA model for Chinese-Vietnamese SMT based on named entity. This approach is suitable for low-resource language pairs that have close linguistic relationships. For such language pairs, the sparse data problem is inevitable, and leads to a low-quality WA.

The purpose of word alignment improvement is to increase the accuracy of word alignments that are related to named entities. The experimental results indicated that our model enhanced the quality of the word alignment and improved the SMT's performance as compared to the performance of the WS translation system using GIZA++.

Given these results, we plan to integrate more linguistic information (such as POS, and chunking) into the system so as to continue increasing the quality of Chinese-Vietnamese MT.

Acknowledgment. This research is funded by Foundation for Science and Technology Development of Ton Duc Thang University (FOSTECT), website: http://fostect.tdt.edu.vn, under Grant FOSTECT.2014.BR.10.

REFERENCES

[1] P. Nakov and H. T. Ng, Improved statistical machine translation for resource-poor languages using related resource-rich languages, *Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp.1358-1367, 2009.

[2] T. Levinboim and D. Chiang, Multi-task word alignment triangulation for low-resource languages, *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pp.1221-1226, 2015.

[3] R. Dabre, F. Cromieres, S. Kurohashi and P. Bhattacharyya, Leveraging small multilingual corpora for SMT using many pivot languages, *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pp.1192-1202, 2015.

[4] C. Chu, T. Nakazawa and S. Kurohashi, Japanese-Chinese phrase alignment using common Chinese characters information, *Proc. of MT Summit XIII*, pp.475-482, 2011.

[5] N. Durrani and P. Koehn, Improving machine translation via triangulation and transliteration, *Proc. of the 17th Annual Conference of the European Association for Machine Translation*, pp.71-78, 2014.

[6] S. Pal, S. K. Naskar and S. Bandyopadhyay, A hybrid word alignment model for phrase-based statistical machine translation, *Proc. of the 2nd Workshop on Hybrid Approaches to Translation*, pp.94-101, 2013.

[7] Y. M. Oh, F. Pellegrino, E. Marsico and C. Coupé, A quantitative and typological approach to correlating linguistic complexity, *The 5th Conference on Quantitative Investigations in Theoretical Linguistics*, 2013.

[8] X. Wang, M. Utiyama, A. Finch and E. Sumita, Refining word segmentation using a manually aligned corpus for statistical machine translation, *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.1654-1664, 2014.

[9] D. K. Le, *Vietnamese Vocabulary Having Chinese Origin*, National University of HCMC Press, 2002 (in Vietnamese).

[10] J. Gao, M. Li and C.-N. Huang, Improved source-channel models for Chinese word segmentation, *Proc. of the 41st Annual Meeting on Association for Computational Linguistics*, pp.272-279, 2003.

[11] Y. Wu, J. Zhao and B. Xu, Chinese named entity recognition combining a statistical model with human knowledge, *Proc. of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition*, vol.15, pp.65-72, 2003.

[12] P. Tran, D. Dinh and L. Tran, Resolving named entity unknown word in Chinese-Vietnamese machine translation, *The 5th International Conference on Knowledge and Systems Engineering*, pp.273-285, 2013.

[13] P. Tran and D. Dinh, Retranslating number expression unknown word in Chinese-Vietnamese statistical machine translation, *Journal of Computer Science and Cybernetics*, vol.30, no.2, pp.127-138, 2014 (in Vietnamese).

[14] K. A. Papineni, S. Roukos, T. Ward and W. J. Zhu, BLEU: A method for automatic evaluation of machine translation, *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.311-318, 2002.