Handling Organization Name Unknown Word in Chinese-Vietnamese Machine Translation

Phuoc Tran
Faculty of Information Technology
University of Food Industry
Ho Chi Minh City, Vietnam
phuoctt@cntp.edu.vn

Tan Le
Faculty of Information Technology
Industrial University
Ho Chi Minh City, Vietnam
letan.dhcn@gmail.com

Dien Dinh
Faculty of Information Technology
University of Science
Ho Chi Minh City, Vietnam
ddien@fit.hcmus.edu.vn

Thao Nguyen
Faculty of Information Technology
Phu Lam Technical and Economic College
Ho Chi Minh City, Vietnam
ntthanhthao@ptec.edu.vn

Abstract — **Unknown word (UKW) is an obvious problem of machine translation and named entity (NE) is the most common UKW type. In this paper, we will present a new approach based on the meaning relationship in Chinese and Vietnamese to re-translate organization name UKW. This is the most complicated NE because it consists of other NEs and entities. Applying this approach to Chinese-Vietnamese statistical machine translation (SMT), experimental results show that our approach has significantly improved machine's performance.**

Keywords — Chinese-Vietnamese SMT, unknown word, named entity, organization name

I. INTRODUCTION

New words are frequently generated from activities such as explaining a new concept, a new invention, named newborn children, a new established organization, etc. In SMT problem, we found that even if the training corpus of machine translation system is large, there will not be able to cover all the words of a language. Therefore, instead of finding the way to translate all the words of a language in order not to arise UKWs, here, we see UKWs as obvious part of MT and try to re-translate these UKWs to improve the overall quality of MT.

Most UKWs in Chinese-Vietnamese SMT are NEs. These NEs are divided into the following categories: person name (PN), organizations name (OG), location name (LC) and Number Expression (date, time, percentage, number, phone number) [2] [5].

The PN is formed by the following structure: <Family name> (F) <Given name> (G), both F and G part have a length of one to two characters. For example, the PN "陈子明", "陈" is F part and "子明" is G part. Chinese PNs are translated into Vietnamese to be their Sino-Vietnamese. Normally, a Chinese LC maximum of 10 characters structured as follows: <name part> <keyword>. In particular, <name part> is an item from the list of Chinese LC (approximately 30,000 <name part>), the LC is usually terminated by a <keyword> (approximately 120 keywords), in some cases there is no keyword at the end of LC. For example, 上海市 (or 上海) ("Thành Phố Thượng Hải": Shanghai city) where 上海 is a <name part> and 市 is a <key word>. The Number Expressions are translated into Vietnamese based on grammatical transformation, the translation method is completely based on rule.

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

Based on formation of Chinese OG as well as meaning relation between Chinese and Vietnamese, we build a handling OG-UKW model to re-translate the OG-UKW, improving performance of machine translation (MT). This paper is presented as follows: in section 2, we will present related work about handling UKWs problem in MT as well as some approaches for Chinese NER. Section 3 will present meaning relation between Chinese and Vietnamese. The recognizing as well as re-translating OG-UKW will be presented in section 4. Meanwhile, in section 5, we will present experiments and some discussion will presented in section 6. The conclusion will be presented in section 7.

II. RELATED WORK

In this section, we will present some works related to the named entity recognition (NER) and re-translating UKW.

Chinese NER approaches in recent times are mainly hybrid approaches, combining of statistic and rule. The key point of recognizing Chinese NEs is based on the keywords of each individual NE (rule based approach). For example: To recognize a PN, Jianfeng *et al* [8] analyzed the Chinese PN in the form of <Family name> (F) + <Given name> (G) (rule

978-1-4799-1350-3/13/$31.00 ©2013 IEEE

based approach). In particular, F and G have a length of one or two characters. The authors only considered candidates to be PNs if their F part is stored in the family name list (which contains 373 entries). The next step, the authors recognized the PN's G based on language model (bigram model) (statistic based approach).

Also in this way, the authors Wu [9] *et al* also used the hybrid algorithm which combines a class based statistical model with many different linguistic knowledge to recognize NE. For the authors Chen *et al* [7], they only focused on identifying the OG based on morphological analysis.

In another aspect, Liu *et al* [10] assert that PN recognition based on the keywords in the Chinese family name list is not effective for the wrong-rule PN. Typically, PN does not have family name: 天平 同学 ("bạn cùng lớp tên Thiên Bình": "classmate Thien Binh"), a respected name 陈先生 ("Trần tiên sinh", "ông Tran": "sir Chen"), pen name, stage name, one character PN, etc. To solve this problem, the authors propose PN recognition based on the context around it. For example, the PN often appears in the context having words, such as: 主席 (president), 记者 (journalist), etc.

NE is a popular UKW type in SMT in general and Chinese-Vietnamese SMT in particular. Currently, there are many studies with different approaches to re-translate UKW and improve MT performance. Based on orthographic cues given by words, Joao et al [2] have proposed two methods to overcome the UKW, namely cognates' detection and logical analogy. This approach has been successfully implemented for inflectional language pair "English – Portugal".

WordNet and International Phonetic Alphabet (IPA) are also used to re-translate UKW. Anwarus Salam *et al.* [3] used these factors to re-translate UKW in Example-Based Machine Translation (EBMT) from English into Bangla. Firstly, the translation system find semantically related English words from WordNet for the UKW. From these related words, the system will choose the semantically closest related word whose Bangla translation exists in English-Bangla dictionary. If no Bangla translation exists, the system uses IPA-based-transliteration. If unknown word is not found in the English IPA dictionary, the system uses Akkhor transliteration mechanism.

Translating UKWs or phrases by analogical learning is performed by Philippe Langlais *et al.* [4]. The concept of analogy is denoted [A : B = C : D], it is a relation between for factors which reads: "A is to B as C is to D", for example, [comfortable : uncomfortable = translatable : untranslatable] in English. The authors found that their approach is able to translate correctly 80% of ordinary UKWs, that is, words that are not proper names, compound words, or numerical expressions. This approach has been made for language pairs, such as French-English, German-English and Spanish-English.

Another handling UKW approach for re-translating UKW is conducted by Eck *et al* [5]. The authors looked for the definitions of UKW in the source language and translate its definitions (instead of translating the UKW). The definitions of UKW will be automatically extracted from online dictionaries and encyclopedias, and then they are translated through the SMT system. The translation result will replace the UKW in previous translation. The approach has been tested on language pair "English – Spanish".

On the other hand, Zhang *et al* [6] translated a Chinese UKW by re-splitting the UKW into sub-words and translating the sub-words (sub-word based translation). Sub-word is a unit in the middle of character and word. In addition, the authors also found that the quality of translation will increase significantly if applying NER to translate UKW before using the sub-word based translation. Our approach is also similar to the approach. The method has been implemented on language pair "Chinese – English".

III. THE MEANING RELATION BETWEEN CHINESE AND VIETNAMESE

A Chinese word usually includes many meaningful characters. When translating it into Vietnamese, its meaning is usually divided into three cases. Firstly, the meanings of Chinese characters are their Sino-Vietnamese meanings, usually 1-1 respectively. Secondly, the meanings of the Chinese characters are similar or related to the meanings of the Chinese word containing those characters. Lastly, 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 and about 65% of the total number of Vietnamese words. The Chinese itself is pronounced differently, even in China, depending on the area where there are many different voices or pronunciations such as Cantonese, Hokkien, Beijing and so on. Neighboring countries such as Korea has its own reading of Korea, known as the Sino-Korean (汉朝); Japanese have their own Chinese reading of the Japanese people, called Sino-Japanese (汉和); and the Vietnamese have their private reading, called Sino-Vietnamese (汉越). Thus, Sino-Vietnamese is a reading way of Vietnamese people. For example, Chinese word 中年 (middle-aged), Chinese (Pinyin) will pronounce "zhong nian" and Vietnamese's pronunciation is "trung niên". A Chinese character may be pronounced by many Sino-Vietnamese words, but in a specific context, one Chinese character is only corresponding to one Sino-Vietnamese. As the above example 中年, corresponding Sino-Vietnamese pronunciation of character 中 is "trung" or "trúng", and pronunciation of 年 is "niên". However, when 中 and 年 combined into a unique word 中年, we only pronounce "trung niên".

In the second case, the meaning of the Chinese word is a combination of Pure-Vietnamese meanings of Chinese characters forming that word. For example, word 白色 ("màu trắng": white color), the corresponding meaning of each character is 白/ "trắng" (white) and 色/"màu" (color).

In the other cases, the words whose 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 proper meaning of “好的” (right).

IV. RE-TRANSLATING OG-UKW

A Chinese OG-UKW is re-translated by our model as figure 1. We use the Stanford Chinese Segmenter[1] to segment Chinese corpus and Vietnamese corpus is segmented by our group’s tool. The tool was implemented by Dinh Dien *et al*, according to Maximum Entropy approach [11]. Chinese corpus continues to be labeled OG through Standford NER[2] tool. A “full OG” usually includes other NEs such as PE, LC, OG and not NE (NNE) components. SMT systems often translate these NNE components successfully. However, in terms of a “full OG”, the translation is often wrong word order. Therefore, assigning OG label helped us identify a “full OG”, after the test corpus is translated by SMT, we not only recognize the OG-UKW but also other components in “full OG”. The identification helps our system re-translate OG-UKW more accurately.

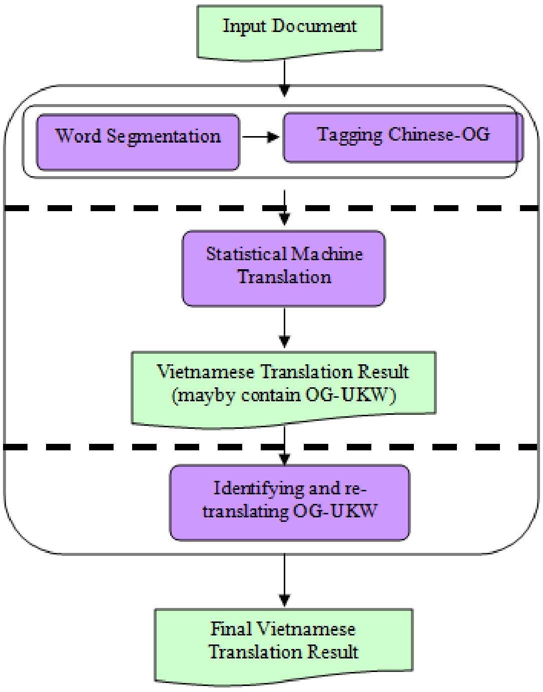

Input Document
Word Segmentation
Tagging Chinese-OG
Statistical Machine Translation
Vietnamese Translation Result (mayby contain OG-UKW)
Identifying and re-translating OG-UKW
Final Vietnamese Translation Result

Fig. 1. Identifying and re-translating OG-UKW model.

The following example illustrates the OG marked by Stanford NER in our system:

Chinese	Word Segmentation	Labeled OG
北京语言学院	[北京] [语言] [学院]	<ORG>[北京] [语言] [学院]</ORG>

In which, [北京] is a LC, [语言] is a NNE component and [学院] is an OG.

Next, we use MOSES[3] tool to train these corpora (creating translation model and language model) and translate Chinese sentences. The translation result (Vietnamese translation) continues to be identified OG-UKW which are tagged in previous stage. As mentioned above, a “full OG” usually includes many other entities combining together, both NEs and NNEs. Word order of “full OG” is reversed in the opposite direction when it is translated into Vietnamese. For example: [北京] [语言] [学院] is translated into Vietnamese in order like [学院] [语言] [北京] ([Học viện] [ngôn ngữ] [Bắc Kinh]).

A noticeable point is that the reordering in “full OG” is not character based reordering but sub-word based reordering. The sub-word can be a character, a word consisting of many characters or maybe the “full OG” (the sub-word segmentation is done by Stanford Segmenter in “word segmentation” stage). As above example 北京语言学院, it is not reordered to be 院学言语京北 (character based) that the correct order is [学院] [语言] [北京], in which, [学院] [语言] and [北京] are sub-words in “full OG”. The OG-UKW re-translating process is conducted as follows:

- Identifying and tagging sub-words in OG including NEs, NNEs and keyword (KEY). We continue to use Standford NER tool to tag NE sub-words in OG.
- Reordering and re-translating the sub-words. The sub-words translation is presented in following sections (section *A*, *B* and *C*). The general process and illustrative example are illustrated in Figure 2 and 3.

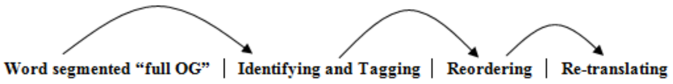


Fig. 2. Method of re-translating OG-UKW.

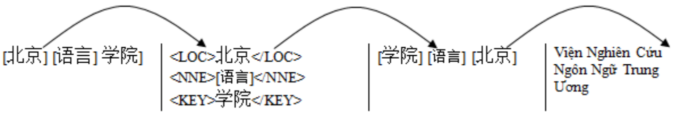


Fig. 3. Example of re-translating OG-UKW

A. Re-translating PN-UKW sub-word

A PN-UKW consists of two parts: F (Family name) and G (Given name), in the form F + G. F and G have a length from one to two characters. We only consider candidates to be PN if their F part is stored in the Chinese Family list (our system includes 484 Family Names). The purpose of the filter is to avoid some error cases due to word segmentation and NER.

- Generating translations of PN-UKW: Chinese PN is translated into Sino-Vietnamese. Therefore, we are only interested in the Sino-Vietnamese meaning of Chinese characters. Normally, each character will have one Sino-Vietnamese respectively. In some cases, a Chinese character is translated into many

[1] http://nlp.stanford.edu/software/segmenter.shtml
[2] http://nlp.stanford.edu/software/CRF-NER.shtml

[3] http://www.statmt.org/moses/

Sino-Vietnamese words. Thus, corresponding to one PN-UKW, there will usually be a set of corresponding Vietnamese names. For example, Chinese name 张丽花, there will be four Sino-Vietnamese translations respectively: “Trương Lệ Hoa”, “Trương Ly Hoa”, “Trướng Lệ Hoa”, “Trướng Ly Hoa”.

- Selecting a suitable translation: for PN-UKW having only one Vietnamese name, we will choose the Vietnamese name to replace PN-UKW in the final Vietnamese translation. Remaining cases, one PN-UKW has many Vietnamese names, we will select Vietnamese name with the highest probability. We calculate individual probabilities for family and given name, as follows:

Assuming FC is a Chinese family, FV is a Vietnamese family, GC is a Chinese given name and GV is a Vietnamese given name. We calculate these probabilities according to Bayes formula as follows:

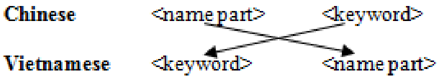
$$F^* = \arg\max_{FV} p(FV \mid FC) = \arg\max_{FV} p(FV) * p(FC \mid FV) \qquad (1)$$

$$G^* = \arg\max_{GV} p(GV \mid GC) = \arg\max_{GV} p(GV) * p(GC \mid GV) \qquad (2)$$

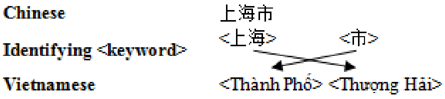
In which, *p(FV|FC)* and *p(GV|GC)* are Chinese and Vietnamese translation models respectively, *p(FV)* and *p(GV)* are respectively Vietnamese family and given name language models, *F** and *G** are the best Vietnamese family and given name corresponding to Chinese ones. Because *FC* and *GC* have only from one or two characters, specially, most of Chinese family names have only one character. Therefore, we choose unigram language model for Vietnamese language. Final translation *F*G** will replace PN-UKW in the final translation. As above example, “Trương Lệ Hoa” is the translation with the highest probability of 张丽花, we choose the translation to be the final translation.

B. Re-translating LC-UKW sub-word

LC-UKW is structured: <name part> <keyword>. In some cases, there is no <keyword> in LC. A <name part> is also translated into Vietnamese to be Sino-Vietnamese meaning, its translation method is similar to PN-UKW sub-word (in section *A*). We identify <keyword> based on a list of 120 <keywords>, their meanings are usually Pure-Vietnamese. One difference between LC and PN is that Vietnamese translation of LC-UKW is reordered, the reordering is conducted as follows:



For example: LC-UKW 上海市, <key word> will be identified and translated as follows:



C. Re-translating NNE sub-word

The NNE sub-words in “full OG” are either translated or un-translated (NNE UKW) by SMT. We preserve the meaning of NNEs translated by SMT and only re-translate NNE UKWs. The re-translation process is done sequentially as follows:

- Find meaning of the NNE UKWs based on Chinese-Vietnamese dictionary.
- If the NNE UKWs do not exist in the dictionary, we will decompose them into smaller units and find the meaning of those units. We use the Maximum Matching algorithm based on Chinese-Vietnamese dictionary to perform this decomposition. The NNE UKWs will be replaced by the meaning of the smaller units.

For example, the “full-OG” [北京] [语言] [学院] has one NNE “语言” and “语言” is translated as “ngôn ngữ” (language) by SMT.

V. EXPERIMENTS

Our bilingual corpus includes 12,000 Chinese-Vietnamese sentence pairs, which are collected from Chinese conversation textbooks and Chinese online forums. Documents in the corpus are mostly communication text, the length of the sentences is relatively short, an average of about 10 words in a sentence. The corpus’ quality is fairly clean, its content is uniformed and spread over 12,000 sentences. We use 90% of the total number of sentences for training, 5% of sentences for testing and the remaining 5% of the sentences for developing. Training corpus (the sentences for training and developing) is trained by MOSES tool with the default parameters (SMT Baseline). We use these corpora to perform four experiments such as baseline translation (1), word segmentation translation (2), re-translating OG-UKW in non-tagging “full OG” case (3) and re-translating OG-UKW in pre-tagging “full OG” case (4).

In (1), we consider the Chinese characters and the Vietnamese spelling words as the meaningful independent units. We insert one space between Chinese characters and insert one space between spelling words with the punctuations.

In (2), we segment Chinese word by Stanford Chinese Segmenter. For Vietnamese, we segment word by our group’s word segmentation tool. The segmenter was implemented by Dinh Dien *et al*, according to Maximum Entropy approach.

We segment word for (3) to be similar to (2). Then, we conduct to translate Chinese test corpus by MOSES tool. The translation result continues to be identified and re-translated OG-UKW by our system. (4) is similar to (3) but before translating Chinese test corpus by MOSES tool, we conduct to pre-tag “full OG” label for the corpus.

Depending on selecting sentence in the test corpus that the BLEU score has difference depending on the selection. The choice giving result with many OG-UKWs is sure that the BLEU score of baseline translation is much lower than the re-translating OG-UKW system and vice versa. Here is the translation result of the test selected following the format:

each 20 sentences in the corpus, the first 18 sentences for training, the 19th sentence for developing and the 20th for testing. The translation result is presented in Figure 5

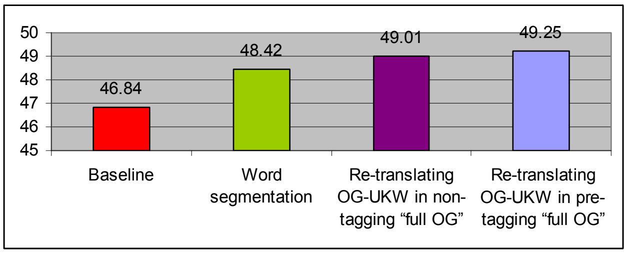


Fig. 4. Experiment result

VI. DISCUSSION

From the experimental results, we found that the translation result in the word segmentation case usually gives the result better than the baseline system. This problem has been presented in [1]. However, the word segmentation translation system gives result with more UKWs. Besides, the re-translating quality of our system is much better than the word segmentation translation case, especially in pre-tagging "full OG" case. This is understandable because the quality of the re-translation system included translation quality of the word segmentation plus translation quality of OG-UKW re-translation system. Assuming the re-translation result is completely wrong, the quality of re-translation will not be also lower than the word segmentation. Here, we would like to present more clearly the difference between non-tagging and pre-tagging "full-OG". The UKWs in the two cases are the same Vietnamese meaning. The unique difference between them is the quality of word order in Vietnamese translation. Word order in the non-tagging case depends entirely on MOSES tool which often gives an incorrect order result when translating "full OG". In the case of pre-tagging, based on the meaning relationship between Chinese and Vietnamese in OG translation, we reorder the sub-words within "full OG" for the corresponding Vietnamese word order and implement to translate. This re-translation result has word order to be more suitable to Vietnamese language. Here are four specific cases in the test corpus (table I):

TABLE I. TRANSLATION RESULT OF FOUR SYSTEMS

Chinese sentence **Translation System**	1. 我 在 清华 大学 学习	2. 我 爸爸 在 中央 语言 研究院 工作	3. 王兰 去 到 北京 语言 学院	4. 他 在 上海 证券 交易所 工作
Vietnamese meaning	Tôi học tại Đại học Thanh Hoa	Ba tôi làm việc tại Viện nghiên cứu ngôn ngữ trung ương	Vương Lan đi đến Học viện ngôn ngữ Bắc Kinh	Anh ta làm việc tại Sở giao dịch chứng khoán Thượng Hải
English meaning	I study at Thanh Hoa University	My father works at the Central Language Research Institute	Vuong Lan went to the Bac Kinh Language Institute	He works at the Thuong Hai Stock Transaction Office
Baseline system	Tôi ở đại học thanh hoa học tập	Ba tôi ở trong ương ngôn ngữ nghiên cứu viện làm việc	Vương Lan đến học viện ngôn ngữ Bắc Kinh	Anh ta ở Thượng Hải chứng khoán giao dịch sở làm việc
Word segmentation	Tôi ở 清华 đại học học tập	Ba tôi ở 中央 ngôn ngữ 研究院 làm việc	Vương Lan đến học viên ngôn ngữ Bắc Kinh	Anh ta ở Thượng Hải chứng khoán 交易所 làm việc
Non-tagging "full OG"	Tôi ở Thanh Hoa Đại học học tập	Ba tôi ở trung ương ngôn ngữ viện nghiên cứu làm việc	---------------------------	Anh ta ở Thượng Hải chứng khoán Sở giao dịch làm việc
Pre-tagging "full OG"	Tôi ở Đại học Thanh Hoa học tập	Ba tôi ở viện nghiên cứu ngôn ngữ trung ương làm việc	---------------------------	Anh ta ở Sở giao dịch chứng khoán Thượng Hải làm việc

All systems have errors of meaning, word order, etc. However, we do not discuss about these errors. Here, we only focus on the errors of the systems when they translate sentences containing OGs. The baseline system gives some results (not contain UKW) in all four cases, but the results are mostly inaccurate except sentence 3 (translating correctly OG "北京语言学院" because of the OG existed in training corpus with the highest probability). In the remaining cases, the characters in OGs also exist in the training corpus, so that the baseline system has chosen Vietnamese meanings having the highest probability to be the outcome. However, these meanings are not true for the OG case, the common errors are meaning of word and word order. Typically, in case 2, "中央语言研究院", correct translation as "viện nghiên cứu ngôn ngữ trung ương", though, the baseline system translated the word to be "trong ương ngôn ngữ nghiên cứu viện". The translation is incorrect about Vietnamese word order. These are similar errors for remaining cases.

In the word segmentation translation system, the number of words in the corpus of this case will be less than a baseline system, so its "alignment word" dictionary as well as word recognition ability is also less than baseline system. As a result, this translation system arises many UKWs. On the other hand, with the abundance of the OG-UKW, although the corpus is extremely large, it is very difficult to cover all words in a language, so the system does not translate them is unavoidable. The word segmentation translation result is re-translated by our identifying and re-translating OG-UKW model. The re-translation performance significantly increased.

We continue to consider the two OG-UKW re-translation models, which are non-tagging "full OG" (a) and "pre-tagging" full OG (b). A "full OG" often consists of many other components such as NEs and NNEs. The NNE components are common words and SMT usually translate these words. Therefore, in case (a), the model only recognize and re-translate NE-UKWs in "full OG" and not interested in the non-NE components which are translated by SMT. Therefore, the final outcome of the "full OG" are often wrong word order. In contrast, in case (b), because we have labeled "full OG" so we easily identify them after they are translated by SMT. Next, we conduct to reorder sub-words in "full OG" in accordance with Vietnamese OG and re-translate these UKW sub-words. The sentence 4 is a typical case about improvement of case (b), OG-UKW "交易所" is re-translated successfully in both cases (a) and (b) but the word order of "full OG" of the case (a) is false ("Thượng Hải chứng khoán sở giao dịch"). Meanwhile, in case (b), because the system has marked "full OG" (<ORG>上海 证券 交易所</ORG>), at the re-translation stage, the system reordered the sub-words in "full OG" into "交易所 证券 上海" and re-translate according to this new order. And this is the right order of Vietnamese OG.

Besides the obvious improvement of the model, through the experiments, we have also found some re-translating cases un-accurately. These incorrect cases fall into irregular NEs in "full OG", typically the PN sub-words. Our system only re-translates right the full Chinese PN. In addition, the NE recognition quality of Stanford NER tool also affects the translation quality of our system. In some cases, the tool does not recognize or recognize OG incorrectly.

VII. CONCLUSION

In this paper, we propose a method to handle the OG-UKW in Chinese-Vietnamese SMT based on their meaning relation. The experimental result shows that our system re-translated OG-UKW with good quality to contribute how significantly improve Chinese-Vietnamese SMT performance. Besides, we have also found that our system was ambiguous meaning when it encounters irregular NE sub-words in "full OG", the quality of OG recognition depends on the performance of the Stanford NER tool.

In the future, we continue to study and learn the other approaches to identify and translate OG-UKW more accurately, improving machine translation with the best quality.

ACKNOWLEDGMENT

This paper was performed under the sponsorship of Kim Tu Dien Multilingual Data Center.

REFERENCES

[1] Trần Thanh Phước – Đinh Điền, Khảo sát yếu tố ranh giới từ trong dịch thống kê Hoa-Việt, Hội nghị Khoa học lần XIII Đại Học Khoa Học Tự Nhiên TP.HCM, 2012 (được chấp nhận đăng trong kỷ yếu của Hội nghị trên tạp chí Khoa học và Công nghệ).

[2] Joao Silva, Luisa Coheur, Angela Costa, Isabel Trancoso, Dealing with unknown words in statistical machine translation, in proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12, 2012.

[3] Khan Md. Anwarus Salam, Setsuo Yamada and Setsuo Yamada, How to Translate Unknown Words for English to Bangla Machine Translation Using Transliteration, Journal of computers, vol. 8, no. 5, may 2013.

[4] Philippe Langais and Alexandre Patry, *Translating Unknown Words by Analogical Learning*, Conference on Empirical Methods in Natural Language Processing, 2007.

[5] Matthias Eck, Stephan Vogel, Alex Waibel, Communicating Unknown words in machine translation, in International Conference on Language Resources and Evaluation, 2008.

[6] Ruiqiang Zhang, Eiichiro Sumita, Chinese Unknown word Translation by Subword Re-segmentation, in International Joint Conference on Natural Language Processing, 2008.

[7] Keh-Jiann & Chao-jan Chen, Knowledge Extraction for Indentification of Chinese Organization Names, In Second Chinese Language Processing Workshop, Hong Kong, 2000.

[8] Jianfeng Gao, Mu Li, and Chang-Ning Huang, Improved Source-Channel Models for Chinese Word Segmentation, in ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, 2003.

[9] Youzheng Wu, Jun Zhao and Bo Xu, Chinese Named Entity Recognition Combining a Statistical Model with Human Knowledge, in MultiNER '03 Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition - Volume 15, 2003

[10] Liu Hongjian, Guo Defang, Zhou Quan, Nagamatsu Kenji, Sun Qinghua, A pre-identification method for Chinese Named Entity Recognition, 2010.

[11] Dinh Dien and Vu Thuy (2006), A maximum entropy approach for Vietnamese word segmentation, in Research, Innovation and Vision for the Future, 2006 International Conference on.