Computational Linguistics and Chinese Language Processing
Vol. 19, No. 1, March 2014, pp. 1-10

A Novel Approach for Handling Unknown Word Problem in Chinese-Vietnamese Machine Translation

Phuoc Tran*, and Dien Dinh+

Abstract

For languages where space cannot be a boundary of a word, such as Chinese and Vietnamese, word segmentation is always the task to be done first in a statistical machine translation system (SMT). The word segmentation increases the translation quality, but it causes many unknown words (UKW) in the target translation. In this paper, we will present a novel approach to translate UKW. Based on the meaning relationship between Chinese and Vietnamese, we built a model which based on the meaning of the characters forming the UKW before translating the UKW through the model. Experiments show that our method significantly improved the performance of SMT.

Keywords: Chinese-Vietnamese SMT, Unknown Word, Sino-Vietnamese, Pure-Vietnamese, SVBUT Model, PVBUT Model.

1. Introduction

Unlike Western languages (typically English), Chinese and Vietnamese words are not separated by a space. A Chinese sentence consists of a series of characters, including punctuation, and no spaces between the characters. In Vietnamese, the spelled words (one-syllabled word) are separated by only one space, and the punctuation is located after the spelled words. Therefore, word segmentation is always solved first in Chinese or Vietnamese statistical machine translation (SMT) into other languages. The word segmentation increases the translation quality but generates many unknown words (UKW).

A Chinese word usually includes many meaningful characters; when translating it into Vietnamese, its meaning is usually divided into three cases. The first case is where the meanings of Chinese characters are their Sino-Vietnamese meanings, usually a 1-1 correspondence. The second case is where the meanings of the Chinese characters are similar

* Faculty of Information Technology, University of Food Industry, Ho Chi Minh City, Vietnam
E-mail: phuoctt@cntp.edu.vn

+ Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
E-mail: ddien@fit.hcmus.edu.vn

or related to the meaning 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 largely are borrowed from Chinese words (often called Sino-Vietnamese, which make up about 65% of the total number of Vietnamese words). Thus, the Sino-Vietnamese words generally appear in Vietnamese text. This very important feature is the basis for our handling UKW approach. In the second case, the meaning of the Chinese word is a combination of Pure-Vietnamese meanings of Chinese characters that form the Chinese word. For these two cases, we re-split a Chinese UKW into characters and translate the characters into Sino-Vietnamese or Pure-Vietnamese. Then, we proceed to incorporate the meanings of the characters and filter their meanings to be suitable to Vietnamese meaning.

In the final case, the meaning of Chinese word is not related to the meanings of the characters forming them. Named entity is a fairly common type of this case. In Chinese-Vietnamese SMT, a Chinese named entity is usually translated into its Sino-Vietnamese. Therefore, for these UKW, we will translate them into Sino-Vietnamese. Maybe the translation result is still not correct, but the quality is better than the previous translation, because the UKW are likely named entities.

This paper is presented as follows: in Section 2, we present related work. Our approach for handling UKW will be presented in Section 3. Meanwhile, in Section 4, we will present experiments and some discussion. Our conclusion will be presented in Section 5.

2. Related Work

Currently, there are many studies with different approaches to handle UKW to improve machine translation performance. Based on word's cognates and logical analogy, Joao *et al.* (2012) proposed two methods (cognates' detection and logical analogy) to translate UKW.

Another handling UKW approach was conducted by Matthias *et al.* (2008). The authors looked for the definition of the UKW in the source language and translated the definition (instead of translating the UKW). The definitions of UKW were automatically extracted from online dictionaries and encyclopedias and they were translated through the SMT system. The translation result would replace the UKW in the previous translation.

On the other hand, Zhang *et al.* (2008) translated Chinese UKW by re-splitting UKW into sub-words and translating the sub-words (sub-word based translation). Sub-word is a unit in the middle of a character and word. In addition, the authors also found that the quality of translation would increase significantly if applying NER to translate the UKW before using the sub-word based translation. Our approach is similar to this approach. Nevertheless, instead of re-splitting UKW into sub-words (greater than character), we re-split UKW into single characters and find their Sino-Vietnamese or Pure-Vietnamese meanings.

3. Chinese Character Meaning based UKW Translation Model

A Chinese UKW is re-translated by our model as follows.

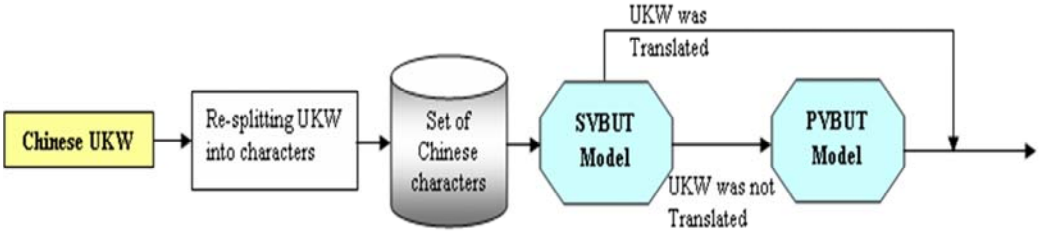


Figure 1. Chinese character meaning-based UKW translation model.

First, a Chinese UKW is disintegrated into Chinese characters before these characters are handled by the SVBUT model. Through this model, the UKW may be translated or not. If the UKW still has not been translated, it will continue to be translated by the PVBUT model. The two models will be presented in Section 3.1 and Section 3.2, respectively.

3.1 SVBUT Model (Sino-Vietnamese based Unknown Word Translation Model)

3.1.1 About Sino-Vietnamese

Chinese, even in China, is pronounced differently, depending on the area, because there are many different voices or pronunciations, such as Cantonese, Hokkien and Beijing (Mandarin). Neighboring countries also have their own reading of Chinese, such as Korea having Sino-Korean (汉朝), Japanese having Sino-Japanese (汉和), and the Vietnamese having Sino-Vietnamese (汉越). Thus, Sino-Vietnamese is the reading way of Vietnamese people. For example, the Chinese word 银行 (bank) is pronounced “yín háng” (rendered using Pinyin), with the 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 the unique word, 银行, we only pronounce it “ngân hàng”.

3.1.2 SVBUT Model

Based on the meaning relationship between Chinese and Sino-Vietnamese, we built a novel model to translate UKW as follows.

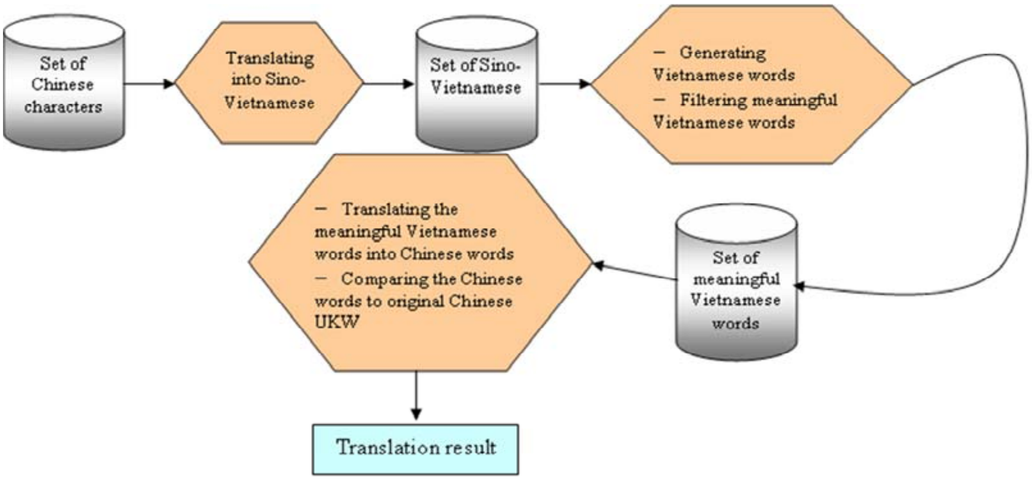
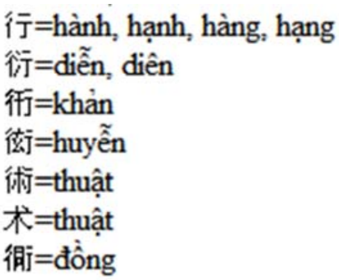

Set of Chinese characters
Translating into Sino-Vietnamese
Set of Sino-Vietnamese
– Generating Vietnamese words
– Filtering meaningful Vietnamese words
Set of meaningful Vietnamese words
– Translating the meaningful Vietnamese words into Chinese words
– Comparing the Chinese words to original Chinese UKW
Translation result

Figure 2. SVBUT model

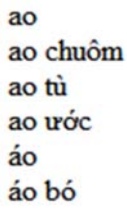
- Step 1: Translating the Chinese characters into Sino-Vietnamese. Based on a Sino-Vietnamese lexicon (Figure 3), we list all Sino-Vietnamese words of Chinese characters. A Chinese character may have many different Sino-Vietnamese words, but in a specific context, one Chinese character corresponds to one Sino-Vietnamese.



行=hành, hạnh, hàng, hạng
衍=diễn, diên
衎=khản
衒=huyễn
術=thuật
术=thuật
衕=đồng

Figure 3. Sino-Vietnamese lexicon format

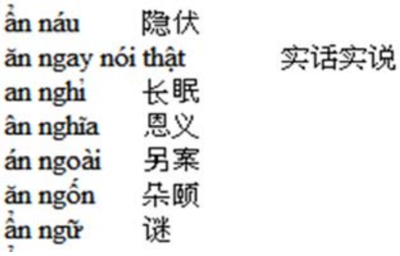
- Step 2: Generate a set of Vietnamese words from the Sino-Vietnamese words in Step 1. The generated Vietnamese words are formed by combining Sino-Vietnamese words together in the correct order in the source language. Then, based on a monolingual Vietnamese dictionary, we carry out filtering of the Vietnamese words, just using the meaningful Vietnamese words. The monolingual Vietnamese dictionary includes Pure-Vietnamese words and loanwords (mainly Sino-Vietnamese words). The format of the dictionary is presented in Figure 4.



ao
ao chuôm
ao tù
ao ước
áo
áo bó

Figure 4. Monolingual Vietnamese dictionary format

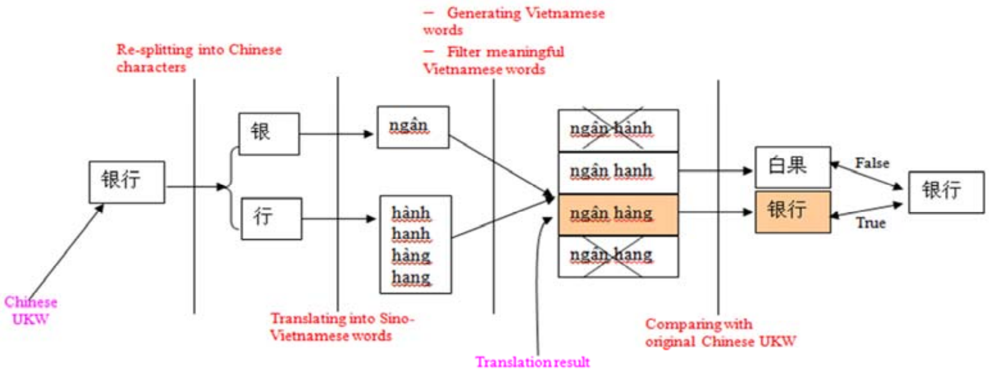
- Step 3: One Chinese word usually has one meaningful Sino-Vietnamese word and that is the meaning of the Chinese UKW. In case there are many meaningful generated Vietnamese words from one Chinese UKW, based on the Vietnamese-Chinese dictionary (Figure 5), we will look up the Chinese words corresponding to those Vietnamese words and compare them with the original Chinese UKW. If the Vietnamese word has a Chinese word that is the same as the Chinese UKW, it is the meaning of the UKW and it replaces the UKW in the translation results. If there are many meaningful Vietnamese words without any corresponding Chinese words, we will select the first word in a set of meaningful Vietnamese words to be meaning of Chinese UKW. Finally, if all generated Vietnamese words are meaningless, we will translate this Chinese UKW by the PVBUT model (Section 3.2).



ẩn náu	隐伏
ăn ngay nói thật	实话实说
an nghỉ	长眠
ân nghĩa	恩义
án ngoài	另案
ăn ngốn	朵颐
ẩn ngữ	谜

Figure 5. Vietnamese-Chinese dictionary format

For example, consider 银行 as a Chinese UKW; it will be translated through the SVBUT model as follows.



– Generating Vietnamese words
– Filter meaningful Vietnamese words
Re-splitting into Chinese characters
银行
银
行
ngân
hành
hanh
hàng
hang
ngân hành
ngân hanh
ngân hàng
ngân hang
白果
银行
False
True
银行
Chinese UKW
Translating into Sino-Vietnamese words
Translation result
Comparing with original Chinese UKW

Figure 6. Chinese UKW 银行 is translated through SVBUT model.

The Chinese UKW 银行 includes two characters, 银 and 行. 银 has a corresponding Sino-Vietnamese word “ngân” and 行 has four Sino-Vietnamese words, these are “hành” “hạnh” “hàng” “hạng”. Combining them together, we have four corresponding generated Vietnamese words. In these words, there are only two words that are meaningful, which are “ngân hàng” and “ngân hạnh”. Since “ngân hạnh” is a fruit type that is translated into Chinese to be “白果” we exclude the Vietnamese word because its Chinese word does not suit the original UKW. The remaining word “ngân hàng” (bank) has a corresponding Chinese word that is also Chinese UKW, so “ngân hàng” is chosen to be the meaning of the UKW 银行.

3.2 PVBUT Model (Pure-Vietnamese based Unknown Word Translation Model)

3.2.1 About Pure-Vietnamese

Vietnamese vocabulary, apart from words borrowed from other languages (mainly from Sino-Vietnamese words), 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 the few remaining can be translated into a Vietnamese word with more syllables. Some examples are 天/trời (heaven), 地/đất (land), 市/thành_phố (city). Another feature of the translation from Chinese to Pure-Vietnamese is that the meaning of the Chinese characters can be reorder in the Pure-Vietnamese translation. For example, the Chinese word 零钱 with 零/lẻ (loose) and 钱/tiền (cash, money), it is translated into Vietnamese as “tiền lẻ” (loose cash) (instead of “lẻ tiền”).

3.2.2 PVBUT Model

Based on the relationship of meaning between the Chinese and their Pure-Vietnamese, we built a UKW translation model as follows:

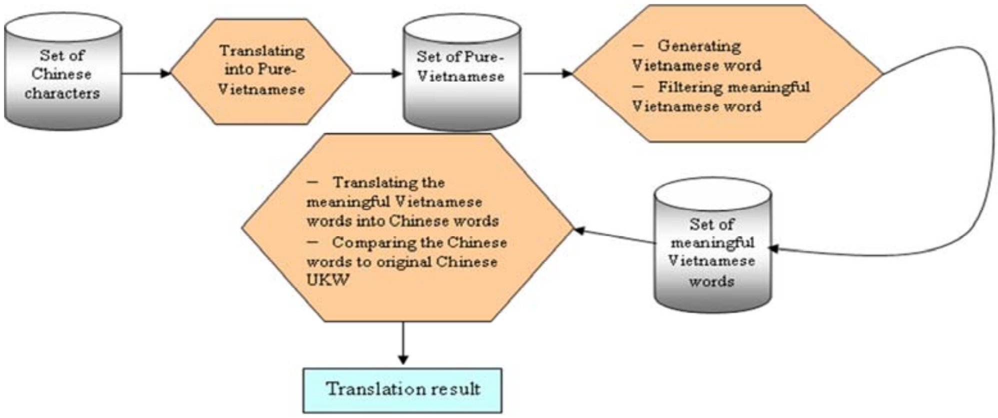

Set of Chinese characters
Translating into Pure-Vietnamese
Set of Pure-Vietnamese
– Generating Vietnamese word
– Filtering meaningful Vietnamese word
– Translating the meaningful Vietnamese words into Chinese words
– Comparing the Chinese words to original Chinese UKW
Set of meaningful Vietnamese words
Translation result

Figure 7. PVBUT model.

The PVBUT model is similar to SVBUT but there are some expansions. In Step 1, the meaning of a Chinese character can be a multi-syllabic word. In Step 2, the generated Vietnamese words, apart from the words being formed according to the order in the source language, must also include the words being established by reordering Vietnamese words that translated from the Chinese characters. The generated words will be filtered like in the SVBUT model.

After this period, the collection of meaningful Vietnamese words may not have any elements, may also have one element, or may have two elements or more. In the case where there is no element, we will translate the UKW as the Sino-Vietnamese (assuming the UKW to be a named entity). For the case of one element, the generated Vietnamese word is the meaning of the UKW. In the other case, where there is more than one meaningful element, we will select the first element in this collection to be the UKW's meaning. For example, Chinese UKW 零钱 will be translated by PVBUT model as follows.

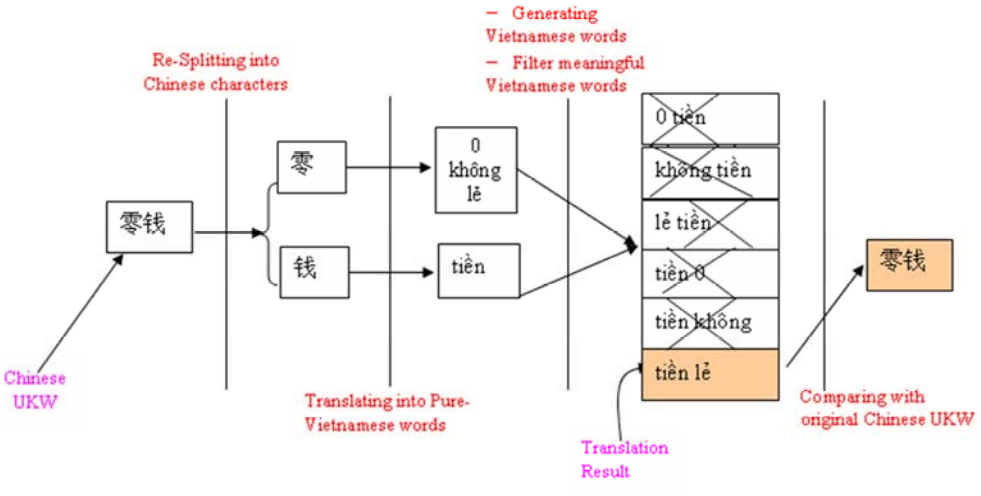

– Generating Vietnamese words
– Filter meaningful Vietnamese words
Re-Splitting into Chinese characters
零
0
không
lẻ
零钱
钱
tiền
0 tiền
không tiền
lẻ tiền
tiền 0
tiền không
tiền lẻ
零钱
Chinese UKW
Translating into Pure-Vietnamese words
Translation Result
Comparing with original Chinese UKW

Figure 8. Chinese UKW 零钱 is translated through PVBUT model.

The Chinese UKW 零钱 has two characters 零 and 钱. 零 has three Pure-Vietnamese meanings, which are "0" (zero), "không" (not) and "lẻ" (loose); 钱 has a common meaning of "tiền" (cash, money). Combining the Pure-Vietnamese meanings together, including reordering them, we get six generated Vietnamese words. In these six words, there is only "tiền lẻ" (loose cash) that is a meaningful Vietnamese word, the generated word "không tiền" (no money) is meaningful but it is not a Vietnamese word (it is a Vietnamese phrase). Fortunately, the word "tiền lẻ" has a corresponding Chinese word that is also an original UKW, so it replaces for the UKW in the final translation.

4. Experiments

Our experiment bilingual corpus consists of 20,000 Chinese-Vietnamese sentence pairs, which were extracted from Chinese conversational textbooks and online Chinese-Vietnamese forums, such as: "Textbook of 301 sentences in Chinese Conversation, Beijing Language Institute" and "Learning Chinese online, www.dantiengtrung.com.vn". Documents in the corpus are mostly communication text, so the length of the sentences is relatively short, with an average of about 10 words in a sentence. We use 90% of the sentences to train, 5% of sentences to test, and the remaining 5% of the sentences to develop. The training corpus (sentences to rain and developing) was trained by Moses[1] tool with the default parameters (SMT Baseline). We performed three experiments, Baseline translation, word segmentation translation, and translating UKW, by our model.

In the Baseline system, we considered the Chinese characters and the Vietnamese spelling words as the meaningful independent units. We inserted one space between Chinese characters and inserted one space between spelled words with the punctuation.

In the word segmentation system, we segmented Chinese words by the Stanford Chinese Segmenter tool[2]. This tool was installed by the CRF method (Conditional Random Field). For Vietnamese, we segmented words by our group's word segmentation tool. The segmenter was implemented by Dinh Dien *et al*. (2006), according to the Maximum Entropy approach.

Based on the results in the segmentation translation, we translated the sentences containing the UKW by our model. The BLEU score for each cases as follows.

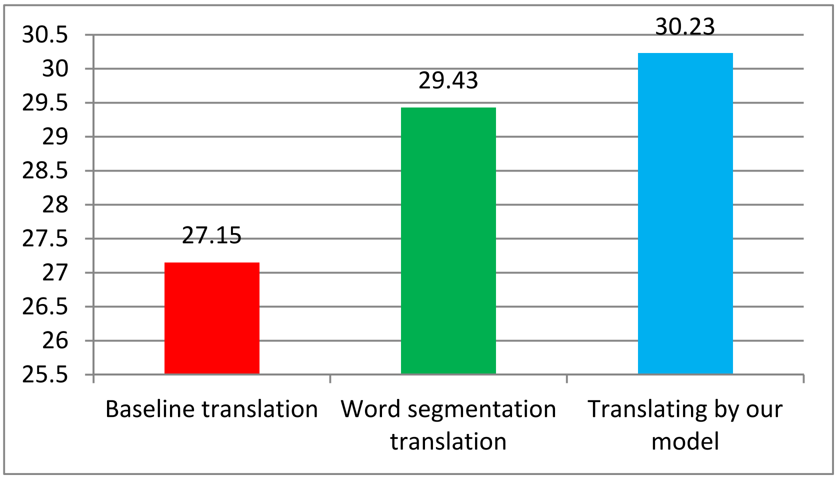


Figure 9. Experiment results.

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

[2] Download: http://nlp.stanford.edu/software/segmenter.shtml

In the Baseline system, although it does not generate UKW, but it gives wrong result. For the case of word segmentation translation, its translation result is better than the Baseline's, but it generates many UKWs. The UKWs are translated through our system. The translation result shows that our system's translation quality is better than the Baseline system, as well as the word segmentation system. Here are two specific cases:

Table 1. Two specific cases

ID	Chinese	True Translation	Baseline Translation	Word Segmentation Translation	Our Model
1	假使	Giả sử, nếu (if)	Kỳ nghỉ làm cho (holiday make)	假使	Giả sử (if)
2	地点	Địa điểm (location)	Địa giờ (land hour)	地点	Địa điểm (location)

In both cases, the Baseline system did not generate UKWs but it gave wrong results. In the first case, the Chinese word 假使 includes character 假/ "kỳ nghỉ" (holiday) (in 放假 -> "nghỉ phép" (holiday)) and 使/ "làm cho" (make). Therefore, 假使 was translated "kỳ nghỉ làm cho" (holiday make). This result is completely wrong. A similar explanation can be seen for the second case.

For the word segmentation translation system, because the system did not recognize the Chinese words, it could not translate them and generated UKW. The UKW were translated by our model. In both cases, the meaning of UKW was also their Sino-Vietnamese meaning. Therefore, the UKWs were translated successfully by the SVBUT model.

In addition, to clarify the improvement of our model, we computed the Precision of the re-translation of UKW. Based on the word segmentation result, we selected 100 sentences containing UKW. Since the documents in the corpus are mostly communication texts, the length of each sentence is an average of about 10 words. Moreover, after segmenting words, the number of words in a sentence is less than 10 words. They were translated by MOSES; if there were UKW, each sentence often had only one UKW. Thus, in this paper, we only chose the sentences containing one UKW for precision calculation. We calculated the precision by the following formula:

$$\mathrm{Precision} = \frac{\sum \mathrm{Correct\ Pairs}}{\sum Total} \qquad (1), \mathrm{Total} = 100 \text{ in this case.}$$

The 100 sentences were re-translated through our system. The system translated exactly 83 UKWs, gaining 83%. The remaining UKWs were translated into Sino-Vietnamese words. These words have no meaning in Vietnamese and also are not person names, place names, or organization names (these names are usually translated into Sino-Vietnamese). UKW 好的 is a specific case. The Sino-Vietnamese of this UKW is "hảo" "đích" and its Pure-Vietnamese is "tốt" (good), "của" (of). Both of "hảo đích" as well as "tốt của" are not Vietnamese words, so that our system will choose Sino-Vietnamese "hảo đích" to be the translation of UKW 好的. This result is completely wrong. We accept this incorrectness with perspective: a mistranslated result is not worse than a UKW result.

5. Conclusion

In this paper, we propose a novel approach to handle UKW in Chinese-Vietnamese SMT. This approach bases on meaning relations between Chinese and Vietnamese, including the relations between Chinese and Sino-Vietnamese and between Chinese and Pure-Vietnamese. The experiments show that our approach has significantly improved Chinese-Vietnamese SMT performance.

Acknowledgement

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

References

Dinh, D., & Vu, T. (2006). A maximum entropy approach for Vietnamese word segmentation. In *Research, Innovation and Vision for the Future, 2006 International Conference on*, Ho Chi Minh, Vietnam, 248-253.

Eck, M., Vogel, S., & Waibel, A. (2008). Communicating Unknown words in machine translation. In *International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Marocco.

Silva, J., Coheur, L., Costa, A., & Trancoso, I. (2012). Dealing with unknown words in the Statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 3977-3981.

Tran, P., & Dinh, D. (2012). Surveying word boundary factor in Chinese-Vietnamese SMT. In *8th Science conference (HCMC University of Science, 2012)*, Ho Chi Minh, Vietnam.

Tran, P., & Dinh, D. (2012). Identifying and reordering prepositions in Chinese-Vietnamese machine translation. *First International Workshop on Vietnamese language and speech processing (VLSP), In conjunction with 9th IEEE-RIVF conference on Computing and Communication Technologies (RIVF 2012)*, Ho Chi Minh, Vietnam.

Zhang, R., & Sumita, E. (2008). Chinese Unknown word Translation by Sub-word Re-segmentation. In *International Joint Conference on Natural Language Processing*, Hyderabad, India.