Improvement of Word Alignment Models for Vietnamese-to-English Translation

Takahiro Nomura, Hajime Tsukada, Tomoyoshi Akiba

Toyohashi University of Technology,
Aichi, Japan

nomura@nlp.cs.tut.ac.jp,tsukada@brain.tut.ac.jp,akiba@cs.tut.ac.jp

Abstract

Aiming at better SMT systems, two approaches for improving word alignment between Vietnamese and English are proposed and evaluated. One is to delete English words that never appear in Vietnamese; the other is to retokenize Vietnamese tokens so that each token of Vietnamese matches an English word. Although the baseline systems could not be improved by these methods at this moment, the results of the analysis show that these approaches are promising.

1. Introduction

Nowadays, a large number of bilingual corpora between popular languages such as English, Chinese, Arabic, and European languages (as listed in the "permissible training data" in this evaluation campaign) are available. In contrast, few corpora for many Asian languages are available. Although Vietnamese was one of the low resource languages, the TED task provided a fair amount of bilingual corpora between English and Vietnamese. Accordingly, Vietnamese has becomes a new target of statistical machine translation.

Since tokenization and grammatical constituents of Vietnamese are different from those of English, each token or word does not always correspond to an English word. This nature of Vietnamese leads a poor word alignment model between Vietnamese and English that will be a base of phrase alignment. To overcome this problem, two methods are proposed: (a) deleting English words that never appear in Vietnamese and inserting them afterward and (b) retokenize Vietnamese so that each token corresponds to an English word. To the authors' knowledge, this is the first application of these methods to Vietnamese translation. Although the baseline system could not be improved by these methods at this moment, we believe these methods will be helpful with further improvement.

The rest of the paper is organized as follows. Section 2 reviews the Vietnamese language. Section 3 explains the method used for retokenization and Section 4 explains our system configuration. Section 5 presents the results of an experimental evaluation of the proposed system, and Section 6 discusses the results. Section 7 concludes the paper.

2. Vietnamese language

Key features of the Vietnamese language are summarized as follows. Some Vietnamese sentences and their English translations are shown in Figure 1. The following features of the Vietnamese language can be seen in this example:

1. Vietnamese is tokenized into units that correspond approximately to syllables.

2. Vietnamese does not have words equivalent to English articles.

For example, in Figure 1, "kết quả" corresponds to "result". Also, there is no Vietnamese word corresponding to the English article "the."

The English side of the training data of the experiment has 2,492,239 words and the number of the articles is 213,710. Therefore, approximately 9% of the English words do not correspond to Vietnamese words. Since the Vietnamese side of the training data has 3,030,127 words, the Vietnamese sentence is approximately 1.3 times longer than that of English ignoring English articles that do not correspond to Vietnamese words.

To improve word alignment models, it is preferable to retokenize Vietnamese words into a unit that corresponds to an English word. Also, an English article must be deleted from the viewpoint of word alignment models, if possible. Considering the former point, we apply two tokenization methods (explained in Section 3). Considering the latter point, we apply two-step translation (explained in Section 4).

3. Tokenizer

Hereafter, the term "tokenization" is simply used for "retokenization" of Vietnamese words where some original tokens are consolidated.

Two tokenization systems are used. One is an existing tokenization tool for Vietnamese, namely, vnTokenizer[1]. It utilizes a word dictionary. Therefore, we refer this as a supervised method. The other is an unsupervised tokenizer proposed by Tagyoung et al. [2] and does not utilize any word dictionaries.

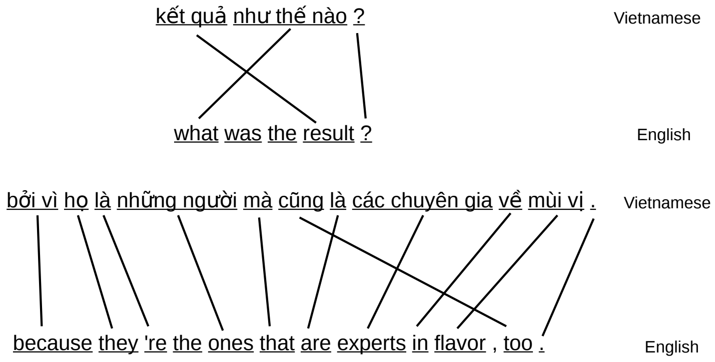
kết quả như thế nào ?
Vietnamese
what was the result ?
English
bởi vì họ là những người mà cũng là các chuyên gia về mùi vị .
Vietnamese
because they 're the ones that are experts in flavor , too .
English

Figure 1: Phrase alignment of Vietnamese and English

3.1. Unsupervised bilingual tokenizer

This tokenization method based on a word-level alignment model trained by using a parallel corpus was originally proposed by Tagyoung et al. for languages that are not tokenized by spaces (such as Chinese and Korean). It is used here for consolidating original Vietnamese tokens.

3.1.1. Bilingual model

The bilingual model is denoted by the following equation. The input data are a tokenized English string e^n and an untokenized Vietnamese string s^m, where "untokenized" means the original tokens are left.

$$P(f, a = k|e) \quad = \quad \frac{\alpha(i)P(f_i|e_k)P(a = k)\beta(j)}{P(\mathbf{c}|\mathbf{e})}$$

where $f = \{s_i s_{i+1} ... s_j\}$ is a new token formed by concatenating from the i-th to the j-th Vietnamese tokens, and a is a variable indicating the position of the English word that generates f. α and β are given by the following equations:

$$\begin{aligned}\alpha(i) &= \sum_{l=1}^{L} \alpha(i-l) \sum_{a} P(a)P(s_{i-l}^{i}|e_a) \\ \beta(j) &= \sum_{l=1}^{L} \sum_{a} P(a)P(s_{j}^{j+l}|e_a)\beta(j+l)\end{aligned}$$

where L is the maximum syllable length for a word.

This model is trained by using an EM algorithm. First, it calculates the expected counts of individual word pairs:

$$ec(s_i^j, e_k) \quad = \quad \frac{\alpha(i)P(a)P(s_i^j|e_k)\beta(j)}{\alpha(m)}$$

Second, an M step simply normalizes the counts:

$$P(f|e) \quad = \quad \frac{ec(f,e)}{\sum_f ec(f,e)}$$

Given two sentences, **e** and **f**, the optimal segmentation of a new source-language sentence can be obtained by using the Viterbi algorithm.

$$segments = argmin_{\mathbf{s}} \sum_{i}^{n} \left(-log \sum_{\mathbf{a}} P(s_i|e_a) + \theta \right)$$

where $\mathbf{s} = \{s_1 s_2 ... s_n\}$ is a segment set of source sentences **f**, and **a** is the alignment of the source segments to the target words. This model can be applied only when a target sentence is available.

3.2. Monolingual model

The monolingual model is denoted by the following equation.

$$P(f) \quad = \quad \sum_{e} P(f|e)P(e)$$

where $P(f|e)$ is the probability of the bilingual model explained Section 3.1.1. $P(e)$ is a monolingual model calculated by the following equation.

$$P(e_i) \quad = \quad \frac{count(e_i)}{\sum_k^K count(e_k)}$$

where $count()$ is the number of occurrences on the English side of the training corpus, and K is the size of the vocabulary.

4. System configuration

The configuration of the proposed system is shown in Figure 2. Each SMT system, namely, (a) baseline system,(b) two-step translation system, and (c) retokenized system, preform the translation for the test set. Multi-Engine Machine Translation (MEMT) then performs the system combination. It receives the results of the combined systems as inputs.

4.1. Baseline system

A phrase-based SMT system and a hierarchical phrase-based SMT system were adopted as baseline systems. These systems are trained by Moses scripts from parallel corpus that is tokenized. The phrase table is trained by the grow-diag-final method and the reordering model is msd-bidirectional-fe.

4.2. Two-step translation system for inserting articles

Two-step translation was performed to deal with English articles properly. The first step is a translation from Vietnamese to English that erases the article. The second step is a translation from English without articles to original English.

First, this two-step approach makes a corpus in which articles of the English side of the parallel corpus are removed and to makes a trilingual parallel corpus: both languages of the original parallel corpus and the newly made English corpus without articles. The two systems are trained by using the trilingual corpus. The first system is a phrase-based SMT system or hierarchical phrase-based SMT system trained by Vietnamese and English without articles. This system receives Vietnamese as an input and outputs English without articles. The second system is a phrase-based system trained by English without articles and original English. It receives the first system's output as an input and complements the removed articles.

4.3. Retokenized system

The training set is tokenized by using vnTokenizer and unsupervised Tokenizer. A phrase-based SMT system is trained by using these tokenized corpora.

The phrase-based systems are trained by a corpus tokenized by vnTokenizer and unsupervised Tokenizer. The systems may have more unknown words than the baseline system because retokenization may not be consistent and produce unknown combined tokens. To solve this problem, the tokens in the phrase table are divided, and the original notation is recovered after the phrase table is built. This system does not perform two-step translation in the experiment.

4.4. System combination

To improve of the translation quality, the outputs of each system are combined by using MEMT[3] (developed by Kenneth Heafield et al).

5. Experiment

The effectiveness of our proposed methods was experimentally evaluated by using a Vietnamese-to-English translation task in IWSLT2015.

5.1. Submitted results

Our submitted results used all of the development sets and test sets provided by IWSLT2015 as a development set. Contrastive1 is the result given by the hierarchical phrase-based baseline system. Contrastive2 is the result given by the phrase-base baseline system. Contrastive3 is the result given by the hierarchical two-step translation system. Contrastive4 is the result given by the phrase-based two-step translation system. Contrastive5 is the result given by the phrase-based retokenized system. The primary is the results obtained by MEMT combining all results listed above. However, the systems had some bugs. The following shows the results where the bugs were fixed.

5.2. Conditions

Only in-domain training and development data of TED talks provided for the IWSLT2015 evaluation campaign were used in the experiments.

Both languages in the training set were tokenized, and the first letter of sentences was recased. The case of the original form is determined by the majority in the training data. The training data was cleaned so that the length of a sentence is 80 words at most.

A language model was created by using test set IWSLT2015 to select the development set based on the perplexity, and test set IWSLT2010 was adopted as the development set.

Moses[4] was used for translation tools, and GIZA++[5] for word alignment tools. The language model was trained by using kenLM[6], and MEMT was used for combining the systems.

5.3. Results

The results obtained by the proposed system are listed in Table 1. In this table, the baseline denotes the systems explained in Section 4.1, and "two-step translation" is the system explained in Section 4.2. The method "vnTokenizer" utilized the corpus tokenized by vnTokenizer. The method "unspTokenizer(bi)" utilized the corpus tokenized by the bilingual model described in Section 3.1.1, and the method "unspTokenizer(mono)" utilized the corpus tokenized by the monolingual model described in Section 3.1.2. And "system comb" is the SMT system that combines all systems listed above.

The two-step translation for articles and retokenization of Vietnamese could not improve the performance of the baseline systems. Also, the result of the system combination fell below the baselines.

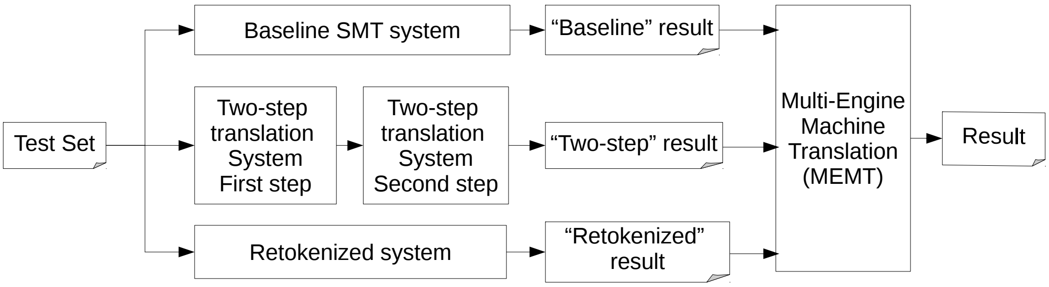
Test Set
Baseline SMT system
"Baseline" result
Two-step translation System First step
Two-step translation System Second step
"Two-step" result
Retokenized system
"Retokenized" result
Multi-Engine Machine Translation (MEMT)
Result

Figure 2: Outline of proposed system

method	model	BLEU %
baseline	phrase base	24.41
	hierarchical	25.00
two-step translation	phrase base	19.06
	hierarchical	19.22
retokenized	vnTokenizer+phrase base	20.38
	unspTokenizer(bi)+phrase base	19.11
	unspTokenizer(mono)+phrase base	19.97
system comb		20.78

Table 1: Experiment results

6. Discussion

As for the two-step translation, the performance improvement of the first step is worse than we expected. The BLEU score of the first step in the development set is 23.47 and that of the baseline system ignoring English articles is 23.34. We have no idea on the very small improvement at this moment. Clearly, this problem must be further investigated.

As for retokenization, vnTokenize may cause mismatch between the training data and the TED task, and its tokenization performance may not be good enough. The unsupervised tokenizer does not cause task mismatch between training data and test data. However, the model does not guarantee each tokenized unit corresponds to an English word, although the model considers bilingual natures. This is a weakness of the current model of the unsupervised tokenizer.

In addition, the result given by the unsupervised tokenizer is not consistent. Therefore, it causes a large number of out-of-vocabulary words if the phrase table is used without reforming the original tokens.

7. Conclusion

Two methods for improving baseline translation were applied. One is deleting English articles that never appear in Vietnamese and inserting them afterward. The other is to retokenize Vietnamese so that each Vietnamese word corresponds to an English word by applying both supervised and unsupervised tokenizers. Although these methods were not helpful at the moment, our analysis shows that the approaches themselves are promising.

8. Acknowledgment

The advice and comments given by Doan Thi Thuy Trinh greatly helped us to understand the Vietnamese language.

9. References

[1] L. H. Phuong, N. Thi Minh Huyên, A. Roussanaly, and H. T. Vinh, "Language and automata theory and applications," C. Martín-Vide, F. Otto, and H. Fernau, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. A Hybrid Approach to Word Segmentation of Vietnamese Texts, pp. 240–249. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-88282-4_23

[2] T. Chung and D. Gildea, "Unsupervised tokenization for machine translation," in *In Proc. EMNLP 2009*, 2009.

[3] K. Heafield and A. Lavie, "Combining machine translation output with open source: The Carnegie Mellon multi-engine machine translation scheme," *The Prague Bulletin of Mathematical Linguistics*, vol. 93, pp. 27–36, January 2010. [Online]. Available: http://kheafield.com/professional/avenue/marathon2010.pdf

[4] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen,

C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open source toolkit for statistical machine translation,” in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ser. ACL ’07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 177–180. [Online]. Available: http://dl.acm.org/citation.cfm?id=1557769.1557821

[5] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[6] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, “Scalable modified Kneser-Ney language model estimation,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013, pp. 690–696. [Online]. Available: http://kheafield.com/professional/edinburgh/estimate_paper.pdf