Le-Minh Nguyen
School of Information Science,
Japan Advanced Institute of Science and Technology.
1-1 Asahidai, Nomi-city, Ishikawa, 923-1292, JAPAN.
☎ *+81 (0761) 51 1221*
FAX *+81 (0761) 51 1149*
✉ *nguyenml@jaist.ac.jp*

Dr. Nianwen Xue
Brandeis University
Waltham, MA, United States

March 27, 2017

Dear Dr. Nianwen Xue,

I am writing to submit our manuscript entitled: Enhancing Statistical Machine Translation For Low-Resource Languages Using Semantic Similarity, which is an improved and extended version of the two papers: Dealing with Out-Of-Vocabulary Problem in Sentence Alignment Using Word Similarity (in Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation, PACLIC 2016), and Applying Semantic Similarity to Phrase Pivot Translation (in Proceedings of the 28th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2016), for the consideration of publication in the ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP).[1]

Statistical machine translation (SMT) models require large bilingual corpora to produce high quality results. Nevertheless, such large bilingual corpora are unavailable for almost language pairs. We focus on two strategies to improve SMT for low-resource languages: sentence alignment and pivot translation. For sentence alignment, we use the representative method that based on sentence length and word alignment as a baseline method. We utilize word2vec to extract word similarity from monolingual data to improve the word alignment phase in the baseline method. The proposed sentence alignment algorithm is used to build bilingual corpora from Wikipedia. In pivot translation, the representative method called triangulation connects source to target phrases via common pivot phrases in source-pivot and pivot-target phrase tables. Nevertheless, it may lack information when some pivot phrases contain the same meaning, but they are not matched to each other. Therefore, we use similarity between pivot phrases to improve the triangulation method. Finally, we introduce a framework that combines the two proposed algorithms to improve SMT for low-resource languages.

Experimental results show two important points: (1) our proposed methods of sentence alignment and pivot translation based on semantic similarity improve the baseline methods and (2) the proposed framework significantly improves baseline SMT models trained on small bilingual corpora. We believe that our findings are going to be the great interests to scientists who work on SMT for low-resource languages, and are relevant to the focus of the ACM Transactions on Asian and Low-Resource Language Information Processing.

Comparing to the original papers, there are new points and significant improvements in this manuscript as follows.

- We apply the proposed sentence alignment algorithm to build bilingual corpora from Wikipedia for the language pairs: Japanese, Indonesian, Malay, and Filipino paired with Vietnamese. In the original paper, experiments are conducted on English-Vietnamese without building bilingual corpora.
- We introduce a new point in this manuscript in which a new framework is proposed that combines the proposed sentence alignment and pivot methods. Experiments are conducted on a set of language pairs: Japanese-Vietnamese, Indonesian-Vietnamese, Malay-Vietnamese, and Filipino-Vietnamese to show the contribution of the proposed framework.

All authors approved the manuscript and this submission. Thank you very much for receiving our manuscript and considering it for review. We appreciate your time and look forward to your response.

Sincerely,

Le-Minh Nguyen

[1] All the necessary documents can be accessed at: https://github.com/nguyenlab/SMT-LowRec-TALLIP-Submission