Thai-to-Any-Language Parallel Corpora from Wikipedia Dumps with CRF-based Sentence Segmentation and Multilingual Sentence Encoder

Charin Polpanumas
PyThaiNLP / Bangkok, Thailand
charin.polpanumas@datatouille.org

Problem Parallel sentence pairs that include Thai are a scarce resource, especially in open source settings. For instance, whereas there are 234.3M sentence pairs in English-German and 1.1M sentence pairs in Chinese-German, there are only 5.5M English-Thai sentence pairs and 700k Chinese-Thai sentence pairs in OPUS (Tiedemann, 2012). Wikipedia dumps provide non-parallel, monolingual datasets that can be mined to create parallel corpora. WikiMatrix (Schwenk et al., 2019) took this apporach but excluded Thai from their 1,620 language pairs extracted from Wikipedia dumps due to lack of a reliable sentence segmentor.

In this work, we propose a method to mine Thai-to-any-language sentence pairs, starting with English-Thai, using open source data from Wikipedia dumps.

Thai Sentence Segmentation For our conditional random fields model CRFCut, we use training and validation data from ORCHID (Sornlertlamvanich et al., 1997, 23,125 sentences), TED transcripts (Lowphansirikul et al., 2020, 136,463 sentences) and generated product reviews (Lowphansirikul et al., 2020, 217,482 sentences), all of which are translated to Thai from English. Since there is no linguistically defined sentence boundary in Thai (Aroonmanakun et al., 2007), we use English sentence boundaries segmented by NLTK (Loper and Bird, 2002) as our sentence boundary labels. We tokenize words using PyThaiNLP's maximal matching 'newmm' tokenizer (Phatthiyaphaibun et al., 2020). For input features, we use unigrams, bigrams and trigrams within the sliding window of two steps before and after the space token we are predicting as end of sentence or not. We also mark words that are frequent sentence starters and enders such as honorifics, demonstrative pronouns, and discourse connectors as additional features. The model predicts which spaces are sentence boundaries.

train set	validation set	F1	accuracy
Ted	Ted	0.72	0.82
Ted	Orchid	0.36	0.73
Ted	Product review	0.77	0.78
Orchid	Ted	0.58	0.71
Orchid	Orchid	0.77	0.87
Orchid	Product review	0.69	0.70
Product review	Ted	0.56	0.56
Product review	Orchid	0.53	0.67
Product review	Product review	0.97	0.97
All	Ted	0.71	0.78
All	Orchid	0.69	0.82
All	Product review	0.96	0.96

Table 1: CRFCut performance at 80/20 train-validation split, counting only spaces

Sentence Alignment There are 6,047,512 articles in English Wikipedia and 136,452 articles in Thai Wikipedia. We transform the titles into sentence vectors using multilingual universal sentence encoder (Yang et al., 2019). Then, we select 13,853 articles whose titles have a cosine similarity score above 0.7 as our parallel articles. After that, we match a group of one to three adjacent sentences in both languages with their counterparts according to the highest cosine similarity score. We choose this approach instead of matching the sentences one-to-one to avoid incomplete sentences due to errors in sentence segmentation as well as the fact that one sentence in one language might match to two in another. Lastly, we filter out those sentence pairs that have a lower cosine similarity score than the 0.7 threshold. As a result, we retrieved 33,756 sentences from 13,853 parallel articles. The sentence pairs are part of the scb-mt-en-th-2020 datasets (Lowphansirikul et al., 2020).

References

Aroonmanakun, W. et al. (2007). Thoughts on word and sentence segmentation in thai. In *Proceedings of the Seventh Symposium on Natural language Processing, Pattaya, Thailand, December 13–15*, pages 85–90.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, page 63–70, USA. Association for Computational Linguistics.

Lowphansirikul, L., Polpanumas, C., Rutherford, A. T., and Nutanong, S. (2020). scb-mt-en-th-2020: A large english-thai parallel corpus. *arXiv preprint arXiv:2007.03541*.

Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., and Chormai, P. (2020). Pythainlp/pythainlp: Pythainlp 2.1.4.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Sornlertlamvanich, V., Charoenporn, T., and Isahara, H. (1997). Orchid: Thai part-of-speech tagged corpus. *National Electronics and Computer Technology Center Technical Report*, pages 5–19.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Yang, Y., Cer, D. M., Ahmad, A., Guo, M., Law, J., Constant, N., Ábrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2019). Multilingual universal sentence encoder for semantic retrieval. *ArXiv*, abs/1907.04307.