Similar Southeast Asian Languages: Corpus-Based Case Study on Thai-Laotian and Malay-Indonesian

Chenchen Ding, Masao Utiyama, Eiichiro Sumita
Advanced Translation Technology Laboratory, ASTREC
National Institute of Information and Communications Technology
3-5 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0289, Japan
{chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

This paper illustrates the similarity between Thai and Laotian, and between Malay and Indonesian, based on an investigation on raw parallel data from Asian Language Treebank. The cross-lingual similarity is investigated and demonstrated on metrics of correspondence and order of tokens, based on several standard statistical machine translation techniques. The similarity shown in this study suggests a possibility on harmonious annotation and processing of the language pairs in future development.

1 Introduction

Research and technique development of natural languages processing (NLP) on many understudied and low-resource Southeast Asian languages are launched in recent years. Some attempts on transferring available techniques on a well-developed language to an understudied language have been proved efficient (Ding et al., 2014). However, such a research direction must be established on an *a priori* observation on the similarity between languages. Generally, linguistically oriented issues, i.e., etymology, vocabulary, or syntactic structure, are discussed when two (or more) languages are referred to as "similar to each other". In engineering practice, however, statistical metrics, i.e., word alignment precision, or word order differences, are more addressed in NLP applications.

In this study, we focus on two Southeast Asian languages pairs, Thai-Laotian and Malay-Indonesian. Both language pairs have mutual intelligibility to a certain extend in spoken form. We conducted a data-driven investigation of the language pairs, trying to figure out the similarity from an engineering viewpoint, which can provide a basis of further NLP techniques development on these languages. The Asian Language Treebank (ALT)[1] (Utiyama and Sumita, 2015; Riza et al., 2016), containing approximately $20,000$ parallel sentence pairs on several Asian languages, facilitates statistical approaches. Specifically, we conducted word aligning on the two language pairs to find out the accordance on token-level, and translation experiments to find out the accordance on sentence-level. The experimental results reveal that the similarity on Thai-Laotian pair is nearly as same as that of Japanese-Korean, i.e., an extremely similar East Asian language pair; for the case of Malay-Indonesian, they are extremely similar to each other that basically they can be considered as two registers of one language. Based on the observation, we think the Thai-Laotian and Malay-Indonesian pairs can be handled simultaneously and harmoniously in further research, including corpus annotation, technique development, and NLP applications.

The remaining of the paper is arranged as following. In section 2, we introduce the background of the languages discussed in this paper. In section 3, we describe the experiment settings used and discuss the numerical results obtained. Section 4 concludes the paper and provides our future work.

2 Background

Specific approaches to process similar languages is an interesting topic in NLP (Vilar et al., 2007; Ding et al., 2015). In this research direction, *a priori* knowledge of the languages is required and specific approaches can thus be designed by taking advantage of the similarities to outperform a general approach. Here we provide outlines of linguistic features of the four languages mentioned in this paper.

[1] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html

Proceedings of the 3rd Workshop on Asian Translation,
pages 149–156, Osaka, Japan, December 11-17 2016.

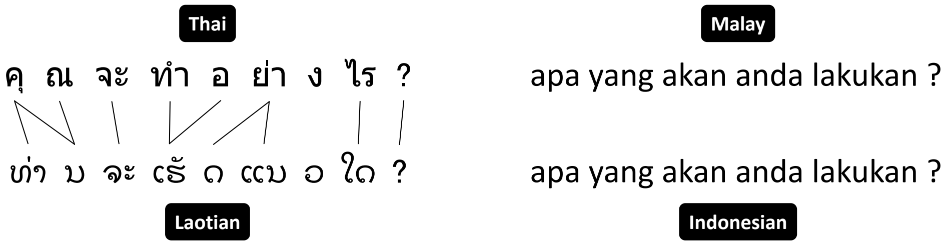

Figure 1: Parallel translations on different languages from SNT.82657.8332-1 in ALT data. The original English is "*What would you do about it?*"

Thai and Laotian are both tonal languages from the Tai-Kadai language family. Spoken Thai and Laotian are mutually intelligible. The two languages share a large amount of etymologically related words and have similar head-initial syntactic structures. Both the languages are written with abugida scripts, slightly different from each other but linguistically similar. A Thai-Laotian example is shown in the left part of Fig. 1. The alignment of tokens is generated by the approach mentioned in the next section. It can be observed the similarity in the shape of certain tokens.

Malay and Indonesian are both from the Austronesian languages family, applying Latin alphabet in their writing systems. Actually, Indonesian can be considered as a standardized register of Malay. The two languages are also mutually intelligible, with several difference in orthography, pronunciation, and vocabulary. The right part in Fig. 1 is an example on the Malay-Indonesian pair, where the expressions are actually identical.

3 Investigation

3.1 Data

We used the parallel raw sentences of corresponding languages from the ALT data. There are $20,106$ sentences in total, which is not a huge size, but a valuable parallel data set. As the Malay and Indonesian use Latin alphabet in their writing systems, the pre-processing of them is relatively simple, we applied naïve tokenization to detach the punctuation marks and lowercasing all the letters. The abugida writing systems of Thai and Laotian are more problematic. As we did not have available tokenization tools for the two languages, we segmented the sentences of the two languages into *unbreakable writing units* for experiments. Specifically, standalone consonant letters with dependent diacritics[2] attached to them are segmented and taken as tokens. The western name entities in sentences were also lowercased.

For the following alignment-based analysis, all the $20,106$ sentences were used as training data for unsupervised word aligning. For the statistical machine translation experiments, the last $2,000$ to $1,000$ sentences were left out as the development set and the last $1,000$ sentences were reserved for test. As the corpus is not large, the aim of the translation experiments is not a pursuit of high performance, but to provide evidence for the similarity of the languages. Statistics of the data we used is list in Table 1.

3.2 Alignment-Based Analysis

We used **GIZA++**[3] (Brown et al., 1993; Och and Ney, 2003) to align all the $20,106$ tokenized sentences for the two language pairs. Based on the aligned results, we investigate (1) token orders by *Kendall's* τ and (2) varieties in token accordance by entropies.

The Kendall's τ was calculated according to several previous work (Isozaki et al., 2012), which mainly focused on the difficulties in word reordering in SMT. The distribution of Kendall's τ is illustrated in Figs. 2 and 3 on the two languages pairs. The Thai-Laotian pair shows a relative similar order with an average

[2] Both post-positioned and pre-positioned

[3] http://www.statmt.org/moses/giza/GIZA++.html

τ around 0.71, and the Malay-Indonesian pair shows an extremely identical order that the average τ is as high as 0.98. These evidences demonstrated the similarity in token orders on the two language pairs.

The statistics on token-level entropy are shown from Fig. 4 to Fig. 9, where Figs. 6 and 7 are based on the patent data from WAT2015's Japanese-Korean task (Nakazawa et al., 2015), shown here for comparison. The entropies were calculated by the lexical translation probabilities from *grow-diag-final-and* symmetrized word alignment (Och and Ney, 2003). Tokens of punctuation marks and numbers are not included in these figures. Generally, the entropies observed in Thai-Laotian and Malay-Indonesian are not large, which suggests the varieties are not large in token corresponding.[4] The scatter plots on Japanese and Korean seem more similar to Thai and Laotian rather than Malay and Indonesian, because the statistics of Japanese and Korean are based on characters, which are smaller units than words as the units used of Thai and Laotian. From the Thai-Laotian pairs, a relative clear tendency can be observed that tokens with very high and very low probabilities have lower entropies in translation. This phenomena is reasonable, because a large portion of vocabulary of the two languages are etymologically related, as well as their syntactic structures. So, common tokens may be aligned well by the similarity in syntax and rare tokens may be aligned well by the similarity in vocabulary. The tendency on Malay-Indonesian is not as obvious as that on Thai-Laotian. A reason is that the vocabulary size is much larger on the Malay-Indonesian pair than the number of unbreakable unit types on the Thai-Laotian pair, which may decrease the precision of alignment on the small training set.

3.3 Translation-Based Analysis

We used the phrase-based (PB) SMT in **MOSES**[5] (Koehn et al., 2007) for the translation experiments. Default setting were applied except for the Thai-Laotian pair the maximum phrase length was set to nine due to the tokens are over-segmented. **SRILM**[6] was use to train a 5-gram language model (LM) for the Malay-Indonesian pair and a 9-gram LM for the Thai-Laotian pair. Modified Kneser-Ney interpolation (Chen and Goodman, 1996) was applied for the LMs. We tested different distortion-limit (DL) in experiments to check the requirement of reordering process in translation. The test set BLEU (Papineni et al., 2002) and RIBES (Isozaki et al., 2010) are listed in Table 2.

From the evaluation, it can be observed the absolute BLEU scores are not quite high, i.e., between 30 and 40, compared with the performance on Japanese-Korean task in WAT2015, which achieved over 70 in terms of BLEU. Generally, the data we used for the experiment is quite limited for statistical model training. Furthermore, as the sentences in different languages are translated from original English articles, the quality between specific language pairs may affected. On the other hand, we observed two phenomena from the translation evaluation. One is the RIBES meets the Kendall'd τ quite well, to show that reordering is not a serious problem in the translation. A further evidence is that the distortion limit did not affect the performance much. This feature is quite like those observed in Japanese-Korean pair. Based on the observation, we consider the Thai-Laotian and Malay-Indonesian have considerable similarities, even from the observation on the relatively small data set.

4 Conclude and Future Work

This paper illustrates the similarity between Thai and Laotian, and between Malay and Indonesian, based on the ALT data. The similarity shown in this study suggests a possibility on harmonious annotation and processing of the language pairs in our further annotated corpus construction based on the ALT data.

References

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proc. of ACL*, pages 310–318.

[4] The logarithm used in this paper is based on 10.

[5] `http://www.statmt.org/moses/`

[6] `http://www.speech.sri.com/projects/srilm/`

	Thai (th)	**Laotian** (lo)	**Malay** (ms)	**Indonesian** (id)
training	1,291,784	1,245,748	435,705	432,456
development	65,387	64,538	23,143	22,978
test	65.014	64,420	23,880	23,392
total	1,422,185	1,374,706	482,728	478,826

Table 1: Number of tokens in the data used in experiment.

DL.	th-lo	lo-th	ms-id	id-ms
0	32.2 / .745	37.0 / .753	31.5 / .867	31.0 / .869
3	32.2 / .743	36.8 / .753	31.3 / .867	31.2 / .869
6	31.4 / .737	37.1 / .754	31.4 / .866	31.2 / .869
9	32.2 / .744	37.0 / .753	31.3 / .866	31.1 / .869

Table 2: BLEU / RIBES for source-target language pairs.

Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2014. Empirical dependency-based head finalization for statistical chinese-, english-, and french-to-myanmar (burmese) machine translation. In *Proc. of IWSLT*, pages 184–191.

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2015. NICT at WAT 2015. In *Proc. of WAT*, pages 42–47.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proc. of EMNLP*, pages 944–952.

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2012. Hpsg-based preprocessing for english-to-japanese translation. *ACM TALIP*, 11(3):8.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL (Demo and Poster Sessions)*, pages 177–180.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2015. Overview of the 2nd Workshop on Asian Translation. In *Proc. of WAT2015*, pages 1–28.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318.

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the asian language treebank. In *Proc. of Oriental COCOSDA (to apprear)*.

Masao Utiyama and Eiichiro Sumita. 2015. Open collaboration for developing and using asian language treebank (ALT). ASEAN IVO Forum.

David Vilar, Jan-T Peter, and Hermann Ney. 2007. Can we translate letters? In *Proc. of WMT*, pages 33–39.

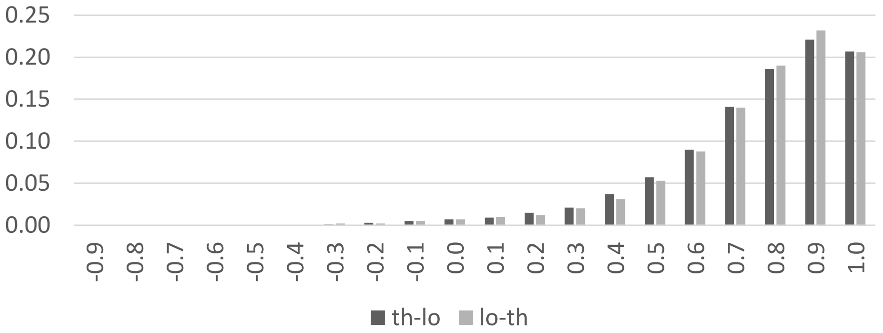

Figure 2: Distribution of Kendall's τ on Thai-to-Laotian (th-lo) and Laotian-to-Thai (lo-th).

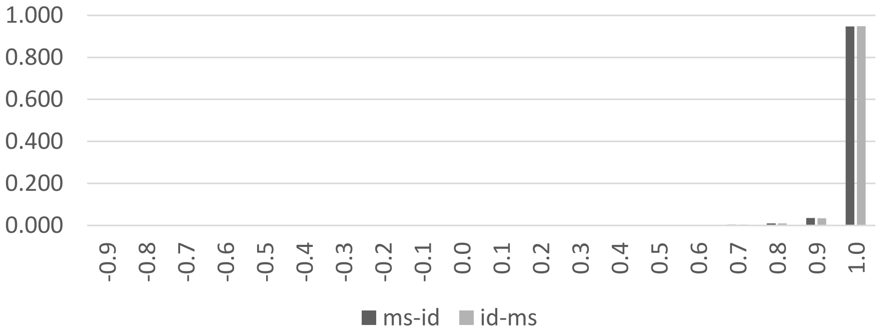

Figure 3: Distribution of Kendall's τ on Malay-to-Indonesian (ms-id) and Indonesian-to-Malay (id-ms).

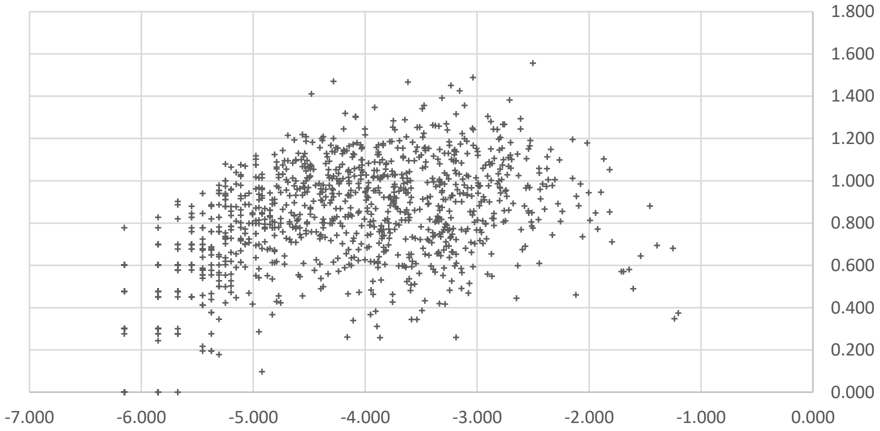

Figure 4: Scatter plot of Thai tokens. *X*-axis is the logarithmic probability of tokens; *Y*-axis is the entropy on corresponding Laotian tokens.

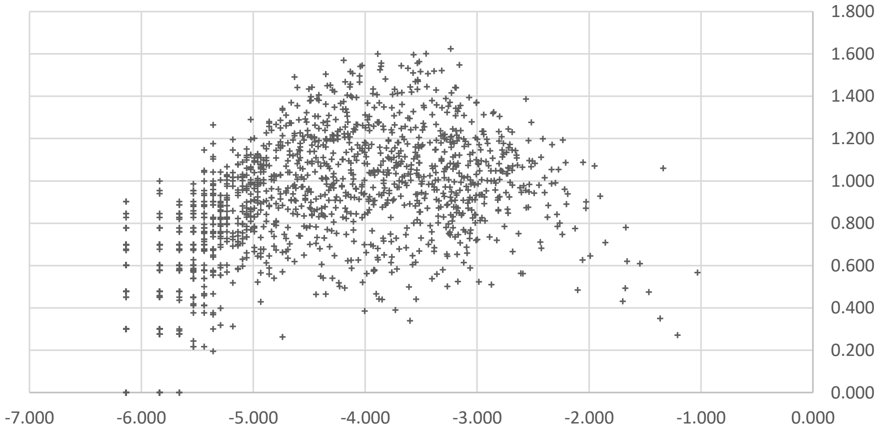

Figure 5: Scatter plot of Laotian tokens. *X*-axis is the logarithmic probability of tokens; *Y*-axis is the entropy on corresponding Thai tokens.

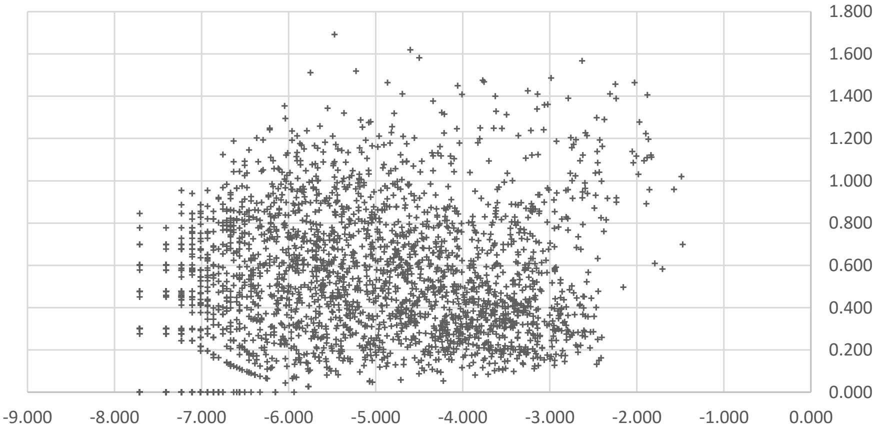

Figure 6: Scatter plot of Japanese tokens. *X*-axis is the logarithmic probability of tokens; *Y*-axis is the entropy on corresponding Korean tokens.

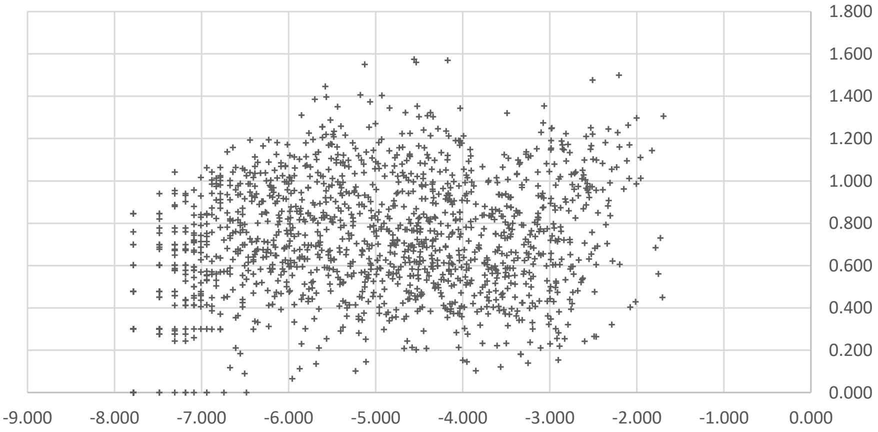

Figure 7: Scatter plot of Korean tokens. *X*-axis is the logarithmic probability of tokens; *Y*-axis is the entropy on corresponding Japanese tokens.

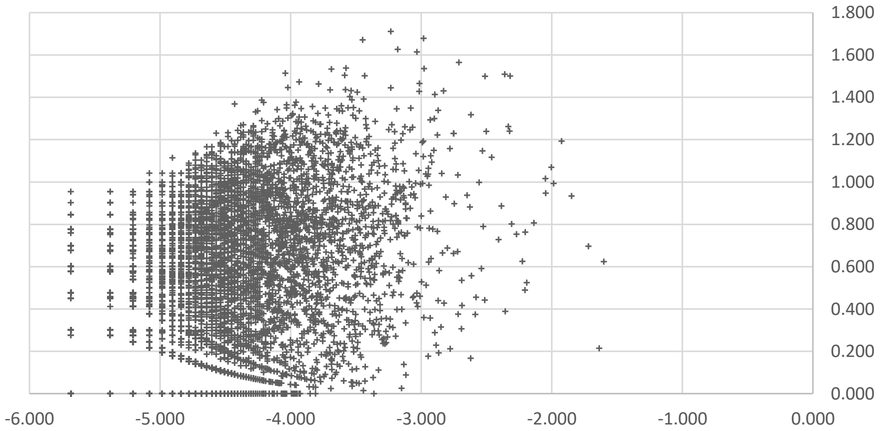

Figure 8: Scatter plot of Malay tokens. *X*-axis is the logarithmic probability of tokens; *Y*-axis is the entropy on corresponding Indonesian tokens.

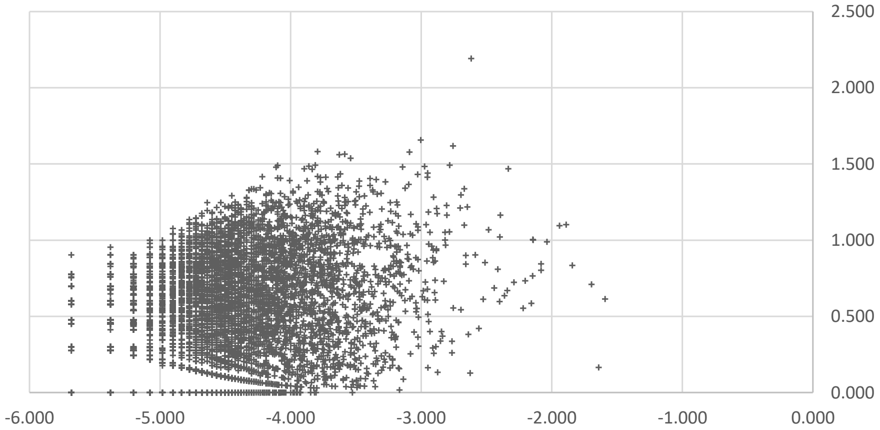

Figure 9: Scatter plot of Indonesian tokens. *X*-axis is the logarithmic probability of tokens; *Y*-axis is the entropy on corresponding Malay tokens.