The primary objective of our work is to build a large-scale English-Thai dataset for machine translation. We construct an English-Thai machine translation dataset with over 1 million segment pairs, curated from various sources, namely news, Wikipedia articles, SMS messages, task-based dialogs, web-crawled data and government documents. Methodology for gathering data, building parallel texts and removing noisy sentence pairs are presented in a reproducible manner. We train machine translation models based on this dataset. Our models’ performance are comparable to that of Google Translation API (as of May 2020) for Thai-English and outperform Google when the Open Parallel Corpus (OPUS) is included in the training data for both Thai-English and English-Thai translation. The dataset, pre-trained models, and source code to reproduce our work are available for public use.

Keywords Machine Translation · Parallel Corpus · Pretraining · Transformer · Thai Language

1 Introduction

Machine translation (MT) techniques have advanced rapidly in the last decade with many practical applications, especially for high-resource language pairs, for instance, English-German, English-French [Ott et al., 2018] and Chinese-English [Hassan et al., 2018]. While the translation quality of these machine translation systems is close to that of average bilingual human translators [Wu et al., 2016], they require a relatively large number of parallel segments to train and benchmark on. Examples of these parallel datasets include News Commentary Parallel Corpus [1] Europarl Parallel Corpus, UN Parallel Corpus [Ziems et al., 2016], Europarl [Koehn, 2005] and ParaCrawl Corpus [Esplà et al., 2019]. However, English-Thai is a low-resource language pair. Insufficient number of training examples is found to directly deteriorate translation quality [Koehn and Knowles, 2017] as current state-of-the-art
models ([Bahdanau et al., 2014; Gehring et al., 2017; Vaswani et al., 2017]) require substantial amount of training data to perform well. Therefore, we curate this dataset of approximately 1M English-Thai sentence pairs to solve the challenge of both quantity and diversity of English-Thai machine translation data.

The difficulties in constructing an English-Thai machine translation dataset include costs for acquiring high-quality translated segment pairs, complexity involved in segment alignment due to the ambiguity of Thai sentence boundaries, and limited number of web pages and documents with English-Thai bilingual content. Currently, the largest source of English-Thai segment pairs is the Open Parallel Corpus (OPUS) [Tiedemann, 2012]. It comprises of parallel segments for many language pairs including English-Thai. However, the contexts of those segment pairs are limited to subtitles (OpenSubtitles [Lison and Tiedemann, 2016], QED [Abdelali et al., 2014]), religious texts (Bible [Christodouloupoulos and Steedman, 2015], JW300 [Agić and Vulić, 2019], Tanzil), and open-source software documentation (Ubuntu, KDE, GNOME).

In order to build an English-Thai machine translation dataset with sufficient number of training examples from a variety of domains, we curate a total of 1,001,752 segment pairs from web-crawled data, government documents, model-generated texts and publicly available datasets for NLP tasks in English. For each data source, approaches to obtain and filter English-Thai segment pairs are described in details. Using OPUS and our dataset, we train machine translation models based on Transformer [Vaswani et al., 2017] and compare the model performance with Google and AI-for-Thai translation services. We used Thai-English IWSLT 2015 [Cettolo et al., 2015] as a benchmark dataset and BLEU [Papineni et al., 2002] as the evaluation metric. BLEU is widely used to evaluate translation quality by comparing translated segments with ground-truth segments. Higher BLEU score indicates better correspondence between the results and ground-truth translation. Our models are comparable to Google Translation API (as of May 2020) for Thai → English and outperform for both direction when OPUS is included in the training data.

The rest of the paper is organized as follows. In Section 2, we first describe the sources from which segment pairs are retrieved for our dataset. After that, we detail the methods to obtain segment pairs, verify translation quality, and filter out noisy segment pairs. In Section 3, we exhibit the statistics of our resulting dataset namely number of segment, number of tokens, and the distribution of segment pair similarity scores. Section 4 presents the results of our experiments training machine translation models on OPUS and our dataset, and evaluating the performance on IWSLT 2015, OPUS and our dataset. In the next section, we discuss the challenges in building the English-Thai machine translation dataset and explore the opportunities to further improve the methodology to obtain a dataset with larger size and higher quality. Our work is then concluded in Section 6.

Last but not least, our English-Thai machine translation dataset and pre-trained machine translation models are publicly available on our GitHub repositories. We also present additional datasets for other Thai NLP tasks such as review classification and sentence segmentation, which are created as a result of building the machine translation dataset, in Appendix 1.

## 2 Methodology

We collect and generate over one million English-Thai segment pairs from five data sources and preprocess them for English-Thai and Thai-English machine translation tasks. Since there is no formal definition of sentence boundaries in Thai [Aroonmanakun et al., 2007], we use English sentence boundaries as segment boundaries for parallel Thai segments. In some cases where the sentence boundaries are not clear even in English (for instance, product descriptions), we do not perform sentence segmentation and treat the entire texts as segments.

---

2 http://opus.nlpl.eu/Tanzil.php

3 http://opus.nlpl.eu/Ubuntu.php

4 http://opus.nlpl.eu/KDE4.php

5 http://opus.nlpl.eu/GNOME.php

6 https://github.com/vistec-AI/dataset-releases/releases/tag/scb-mt-en-th-2020_v1.0

7 https://github.com/vistec-AI/model-releases/releases/tag/SCB_1M+TBASE_v1.0
Figure 1: Preprocessing flow for each data source
2.1 Data Sources

2.1.1 Publicly Available Datasets

We use English segments from following public datasets for natural language processing (NLP) and natural language understanding (NLU) tasks as source segments. These datasets are translated to Thai by professional and crowdsourced translators.

- Taskmaster-1 [Byrne et al., 2019] is a dataset of 13,215 task-based dialogs in 6 domains: ordering pizza, making auto repair appointments, scheduling rides, ordering movie tickets, ordering coffee drinks and making restaurant reservations. The dialogs created in both written and spoken English.

- The National University of Singapore (NUS) SMS Corpus [Chen and Kan, 2011] is a collection of 67,093 SMS messages written by Singaporeans, mostly NUS students. The style of writing is informal and contains so-called Singlish dialect of English.

- Mozilla Common Voice is a crowdsourced collection of 61,584 voice recordings in various languages. We use the English transcriptions as the source segments. The dataset has segments both written and spoken English.

- Microsoft Research Paraphrase Identification Corpus [Dolan and Brockett, 2005] contains 5,801 English segment pairs from news sources. Each segment pair has a binary label of whether they are paraphrasing of each other (that is, semantically equivalent) or not.

2.1.2 Generated Product Reviews

We generate 372,534 product reviews in English using Conditional Transformer Language Model (CTRL) [Keskar et al., 2019] and use them as the source segments. The conditional transformer language model was trained on multiple domains such as Amazon reviews, Wikipedia, Project Gutenberg and Reddit. CTRL can generate texts with content and style specified by the control codes. For our dataset, we specified the following conditions:

- The content generated must be in the product review domain.

- The generated reviews must represent sentiments ranging from mostly dissatisfied to mostly satisfied (1-5 scale).

- The length of each generated review is limited to less than 150 tokens. Incomplete segments as a result of the generation process are filtered out.

2.1.3 Wikipedia

Wikipedia consists of articles about various topics such as biographies, events, organizations and places. Articles are written and edited by crowdsourced contributors. At the time of writing, we obtain 6,047,512 articles in English Wikipedia and 136,452 articles in Thai Wikipedia. We hypothesize that there are a number of articles among them that can be treated as parallel documents.

2.1.4 Web Crawling

Large machine translation datasets such as Paracrawl [?] are created from scraping websites with parallel texts. We gather domains of possible parallel websites from three sources:

- Paracrawl: Out of 208,349 domains from 23 language pairs of Paracrawl, we found that 1,047 domains have both English and Thai content.

---

8https://voice.mozilla.org/en
• Top 500 Thai Websites according to Alexa.com: We hypothesize that websites with high traffic volume are more likely to have pages both in Thai and English.
• Other specific bilingual websites such as Asia Pacific Defense Forum, Ministry of Foreign Affairs, and websites of various embassies in Thailand that provide sizeable amount of English-Thai content.

2.1.5 Thai Government Documents

Official government documents in Thai and English in PDF format are obtained from their respective organizations. The documents include but are not limited to:

• The Constitution of the Kingdom of Thailand 2017 (B.E. 2560)
• The Thailand Penal Code
• The Thailand Civil and Commercial Code
• Thailand’s Labour Relations Act 1975 (B.E. 2518)
• Thailand’s First - Twelfth National Economic and Social Development Plan
• Economic Outlook and Performance Report
• Social Outlook Report
• Gross Domestic Product report
• National Income of Thailand report
• Oil plan 2015 – 2036 (B.E. 2558 - 2579)
• Thailand 20-Year Energy Efficiency Development Plan 2011-2030 (B.E. 2554 - 2573)
• Alternative Energy Development Plan 2015-2036 (B.E. 2558 - 2579)
• Thailand Power Development Plan 2015-2036 (B.E. 2558 - 2579)
• Sustainable Future City Initiative Guideline for SFCI Cities

2.2 Translation of English Segments

One way to create segment pairs is to employ various translation methods. We employ 3 approaches to get the translation including professional translation, crowdsourced translation and Google Translation API.

Regarding professional translation, we employ 25 professional translators to translate 13,215 conversations of the Taskmaster-1 dataset and 43,374 generated product reviews from English to Thai. Secondly, we use a crowdsourcing platform to disseminate English-to-Thai translation tasks for NUS SMS, Mozilla Common Voice, and Microsoft Research Paraphrase Identification, and 21,590 generated product reviews.

Aforementioned approaches are relatively expensive and time-consuming, therefore, we opt in Google Translation API to translate 307,570 generated English product reviews to Thai. After that, we employ annotators to assess the quality of each product review. We ask the annotators to classify whether the product reviews translation should be accepted or rejected. The criteria are fluency and adequacy of the translation. One product review may have several segments but we only include segments from product reviews that are labeled as acceptable.

2.3 Alignment of Existing English-Thai Segments

Apart from translation from English to Thai, we also perform segment alignment of existing English-Thai segment pairs parallel documents.
2.3.1 Sentence Segmentation

We use NLTK [Loper and Bird, 2002] for English sentence segmentation. For Thai texts, we train a conditional random field model to predict sentence boundary tokens based on the following datasets:

- Generated Product Reviews: 67,387 reviews and a total of 259,867 segments that are translated by Google Translate API and annotated by humans are used to train the model since we know the sentence boundaries marked by English texts.
- TED Transcripts: We obtain transcripts in Thai of TED talks containing 136,463 utterances. We treat each utterance as a segment.
- ORCHID Corpus: The corpus was originally meant for POS tagging but it contains 23,125 marked segment boundaries and are used as benchmark for Thai sentence segmentation.

We tokenize them into Thai words using newmm tokenizer of pyThaiNLP [Phatthiyaphaibun et al., 2020], then create unigram, bigram and trigram features with a sliding window of 2 steps before and after the token to predict if it is a sentence boundary or not. We also mark words that are often found to be sentence starters or sentence enders and apply the same feature extraction.

Our baseline model CRFCut achieves the following performance[9]

| Training set         | Validation set     | Non-boundary token | Sentence boundary token | space-correct |
|----------------------|--------------------|---------------------|-------------------------|--------------|
|                      |                    | Precision          | Recall                  | F1 score     | Precision | Recall | F1 score | space-correct |
| TED                  | TED                | 0.99               | 0.99                    | 0.99         | 0.74      | 0.70   | 0.72     | 0.82          |
| TED                  | Orchid             | 0.95               | 0.99                    | 0.97         | 0.73      | 0.24   | 0.36     | 0.73          |
| TED                  | Product Review     | 0.98               | 0.99                    | 0.98         | 0.86      | 0.70   | 0.77     | 0.78          |
| Orchid               | TED                | 0.98               | 0.98                    | 0.98         | 0.56      | 0.59   | 0.58     | 0.71          |
| Orchid               | Orchid             | 0.98               | 0.99                    | 0.99         | 0.85      | 0.71   | 0.77     | 0.87          |
| Orchid               | Product Review     | 0.97               | 0.99                    | 0.98         | 0.77      | 0.63   | 0.69     | 0.70          |
| Product Review       | TED                | 0.99               | 0.95                    | 0.97         | 0.42      | 0.85   | 0.56     | 0.56          |
| Product Review       | Orchid             | 0.97               | 0.96                    | 0.96         | 0.48      | 0.59   | 0.53     | 0.67          |
| Product Review       | Product Review     | 1                  | 1                       | 1            | 0.98      | 0.96   | 0.97     | 0.97          |
| TED + Orchid + Product Review | TED     | 0.99               | 0.98                    | 0.99         | 0.66      | 0.77   | 0.71     | 0.78          |
| TED + Orchid + Product Review | Orchid  | 0.98               | 0.98                    | 0.98         | 0.73      | 0.66   | 0.69     | 0.82          |
| TED + Orchid + Product Review | Product Review  | 1                  | 1                       | 1            | 0.98      | 0.95   | 0.96     | 0.96          |

Table 1: The precision, recall and F1 score for non-boundary and sentence boundary token of CRF-based sentence segmentor models trained and validated on different datasets. space-correct is accuracy of predicting if spaces are sentence boundaries or not.

[9]Training codes at https://github.com/vistec-AI/crfcut
2.3.2 Segment Extraction

Once we have a means to segment all texts, we proceed to extract all segments from each data source.

**Paracrawl Corpus Release v5.0 (September 2019)**

First, we aggregate the TMX files from 23 language pairs. The total number of domains listed is 208,349. The total number of URLs is approximately 12.8 M URLs. We directly substitute ISO 639-1, 639-2T, 639-2B language codes appeared in the URLs of non-English language code (e.g. /de/, /ger/, /es/, /spa/) to Thai language code (e.g. /th/, /tha/), and send HTTP request to verify whether the HTTP request of modified URL with Thai language code response with HTTP status 200.

With this approach, we obtain a total number of 1,047 domains that comprised of content in both English and Thai. We use the web crawling module from bitextor [Espl and Transducens, 2009] to crawl the websites and perform language detection to filtered out the pages whose contents are in neither English nor Thai. We then perform document alignment on crawled data of each domain name based on edit distance of tokens in URLs. A token in this case is defined by a group of characters separated by / except for the protocols (http:, https: and so on). URLs pairs with edit distance equal to one token were paired up, for instance, two URLs that are different only in the language code tokens. We successfully aligned 23,528 document pairs.

**Top-500 Thai Websites**

We obtain the list of top-500 websites in Thailand from the ranking website Alexa.com [alec]. We retrieved the sitemaps in XML format from those websites and read all the URLs listed. We wrote a web crawling script to crawl bilingual web pages based on these URLs. Similar to what we do with Paracrawl, if a URL contains English or Thai language code, we substitute the language code with /en/ or /th/ and verify if the document pair contains content both in English and Thai. The total number of aligned documents we crawled is 246,868 page pairs that have content both in English and Thai.

**Wikipedia**

To create parallel documents from Wikipedia pages, we align English and Thai articles based on their titles by transforming them into dense vectors using multilingual universal sentence encoder [Yang et al., 2019] and find cosine similarity. Out of all English and Thai articles, we find 13,853 articles that we consider parallel documents.

**Government Documents in PDF Format**

We extract segments from aligned government documents in PDF format with Apache Tika [10] Character errors in extracted Thai texts are fixed with handcrafted rules [11].

**Thai Translation of Generated Product Reviews**

We obtained the translation in Thai of 43,374 generated product reviews by professional translation. Since the translation is in document-level, we need to extract segments from the source reviews and translated reviews in order to obtain the alignment at segment-level.

2.3.3 Segment Alignment

For each pair of aligned documents, we have two approaches in aligning segments. The first approach is applicable for documents crawled from the web. We segment the content in the documents by HTML tags (e.g. <p>, <li>, <h>). All content within a tag is treated as one segment. We then choose only document pairs that have the same number of equivalent tags and align the segments in order. The downside of this approach is that we might end up with multiple segments per tag.

---

[10] https://tika.apache.org/
[11] See https://github.com/vistec-AI/pdf2parallel
The second approach is to use sentence segmenter in the previous section to segment Thai texts and NLTK sentence segmenter [Loper and Bird, 2002] to segment English texts then align them based on semantic similarity. We found that after sentence segmentation there are more Thai segments than their English counterparts. In order to correctly align the segments, multiple segments in Thai language have to align with one segment in English in a many-to-one manner. For each English segment, we align them with a concatenation of one to three consecutive Thai segments. To extract the semantic features, we use multilingual universal sentence encoder [Yang et al., 2019] trained on 13 languages including English and Thai to transform each segment into a 512-dimension dense vector. After that, for each segment pair, we compute cosine similarity of their respective vectors. Therefore, one English segment can have up to three versions of alignment with one, two or three concatenated consecutive Thai segments. For each English segment, we select the version that has the highest cosine similarity score.

2.4 Preprocessing for Machine Translation

We apply rule-based text cleaning to all texts obtained. After that, we filter out segments that are incorrectly aligned using handcrafted rules and multilingual universal sentence encoder [Yang et al., 2019].

2.4.1 Text Cleaning

We perform text cleaning on each sub-dataset with text-cleaning rules including NFKC Unicode text normalization, replacing HTML entity and number code (e.g. &quot;, &amp;) with corresponding ASCII characters, Removing redundant spaces, and standardizing quote characters. Note that emojis and emoticons are not filtered out from the texts.

2.4.2 Segment Pair Filtering

Since we obtain our segment pairs by different sources and approaches with varying degree of quality, we have to filter out some segment pairs that are not parallel to each other by handcrafted rules and text similarity based on multilingual universal sentence encoder.

Handcrafted Rules

For each dataset, we define a set of thresholds for the following handcrafted rules to filter out low-quality segment pairs:

- Percentage of English or Thai characters in each English or Thai segment; for instance, Thai segments with lower percentage of Thai characters are most likely not actually Thai segments but segments from other languages that have been mistakenly crawled

- Minimum and maximum number of word tokens for Thai and English segment. We use newmm tokenizer from pyThaiNLP [Phatthiyaphaibun et al., 2020] to tokenize Thai words, and NLTK [Loper and Bird, 2002] to tokenize English words. Spaces are excluded from the token counts.

- Ratio of word tokens between English and Thai segments; for example, a pair of segment with 100 tokens for English and 5 tokens for Thai will be filtered out from the resulting dataset.

We also remove all duplicated segment pairs both by exact match and by text similarity based on multilingual universal sentence encoder.

Text Similarity based on Multilingual Universal Sentence Encoder

We transform all segments into 512-dimension dense vectors using multilingual universal sentence encoder, trained on 13 languages including English and Thai [Yang et al., 2019]. We then calculate the cosine similarity between English and Thai segments of each segment pair. The rationale is that segments that are translation of each other should be semantically similar and thus have high cosine similarity score.

12The source code and thresholds used for the preprocessing can be found at: https://github.com/vistec-AI/thai2nmt_preprocess
We found that after sentence segmentation there are more Thai segments than their English counterparts. This is to be expected. In order to correctly align the segments, multiple segments in Thai language have to align with one segment in English (many-to-one). Thus, we compute cosine similarity between a pair of English segment and Thai concatenated segments.

We use a different cosine similarity threshold for segments from each domain. For example, texts retrieved from web crawling have a relatively higher threshold of 0.7 as we see higher rate of misalignment, whereas the segment pairs from Thai government documents have the threshold of 0.5 as they follow set patterns and are easier to align.

### 3 Resulting Datasets

#### 3.1 English-Thai Machine Translation Dataset

We collected segment pairs from 12 sources and performed the text processing procedures described in Methodology. Table 2 and 3 present the statistics of the resulting datasets after text processing. The total number of segment pairs is 1,001,752. We tokenize Thai segments with pyThaiNLP’s `newmm` dictionary-based tokenizer where space token is excluded and Moses tokenizer for English segments.

| Method                              | Sub-dataset                | Number of segment pairs |
|-------------------------------------|----------------------------|-------------------------|
| Professional Translators            | task_master_1              | 222,733                 |
|                                     | product_review_translator  | 133,330                 |
| Crowd-sourced Translators           | nus_sms                    | 43,750                  |
|                                     | msr_paraphrase             | 10,371                  |
|                                     | mozilla_common_voice       | 33,797                  |
|                                     | product_review_crowd       | 24,587                  |
| Annotation by Translators           | product_review yn          | 280,208                 |
| Segment Alignment on PDF Documents  | assorted_government        | 25,398                  |
|                                     | thai_websites               | 120,280                 |
| Segment Alignment on Web-crawled Data | paracrawl                 | 60,039                  |
|                                     | wikipedia                  | 33,756                  |
|                                     | apdf                       | 13,503                  |
|                                     |                            | 1,001,752               |

Table 2: Number of segment pairs categorized by data source and method to obtain parallel segment pairs.
| Sub-dataset name          | Tokens | Unique tokens | Token Distribution          |
|--------------------------|--------|---------------|----------------------------|
|                          |        |               | mean | median | (min, max) |
| task_master_1            | en     | 2,615,760     | 32,888 | 11.74  | 10        | (1, 211) |
|                          | th     | 2,349,135     | 20,406 | 10.55  | 8         | (3, 203) |
| generated_reviews_translator | en     | 2,128,286     | 32,025 | 15.96  | 14        | (1, 102) |
|                          | th     | 1,974,424     | 22,109 | 14.81  | 13        | (2, 117) |
| nus_sms                  | en     | 538,584       | 33,816 | 12.31  | 10        | (1, 171) |
|                          | th     | 561,907       | 13,329 | 12.84  | 10        | (1, 172) |
| msr_paraphrase           | en     | 231,897       | 18,191 | 22.36  | 22        | (3, 46)  |
|                          | th     | 219,682       | 15,776 | 21.18  | 21        | (3, 52)  |
| mozilla_common_voice     | en     | 325,856       | 17,377 | 9.64   | 9         | (2, 28)  |
|                          | th     | 288,066       | 15,578 | 8.52   | 8         | (1, 54)  |
| generated_reviews_crowd  | en     | 441,804       | 13,246 | 17.97  | 16        | (3, 89)  |
|                          | th     | 391,505       | 12,169 | 15.92  | 14        | (2, 91)  |
| generated_reviews_yn     | en     | 4,429,469     | 37,202 | 15.81  | 14        | (2, 104) |
|                          | th     | 3,909,029     | 26,261 | 13.95  | 12        | (3, 96)  |
| assorted_government      | en     | 1,711,174     | 25,139 | 67.37  | 63        | (5, 500) |
|                          | th     | 1,931,200     | 25,802 | 76.04  | 64        | (4, 441) |
| thai_websites             | en     | 9,934,983     | 117,267 | 82.60  | 70        | (3, 543) |
|                          | th     | 11,105,989    | 85,096 | 92.33  | 80        | (1, 455) |
| wikipedia                | en     | 1,655,315     | 54,173 | 49.04  | 47        | (6, 226) |
|                          | th     | 1,839,488     | 40,570 | 54.49  | 40        | (5, 272) |
| paracrawl                | en     | 1,688,408     | 56,196 | 28.12  | 19.0      | (5, 316) |
|                          | th     | 1,691,030     | 39,035 | 28.17  | 19.0      | (3, 322) |
| apdf                     | en     | 685,864       | 25,516 | 50.79  | 46        | (6, 303) |
|                          | th     | 736,931       | 15,301 | 54.58  | 49        | (5, 331) |

Table 3: Number of segment pairs, Thai/English word tokens, unique word tokens and distribution of English and Thai word tokens in segments for each sub-dataset.
| Sub-dataset name                  | Average | Min  | Max  |
|----------------------------------|---------|------|------|
| generated_reviews_yen            | 0.81    | 0.40 | 0.40 |
| task_master_1                    | 0.59    | 0.20 | 0.20 |
| generated_reviews_translator     | 0.74    | 0.51 | 0.51 |
| thai_websites                    | 0.78    | 0.09 | 0.09 |
| paracrawl                        | 0.80    | 0.50 | 0.50 |
| nus_sms                          | 0.58    | 0.10 | 0.10 |
| mozilla_common_voice             | 0.71    | 0.30 | 0.30 |
| wikipedia                        | 0.80    | 0.70 | 0.70 |
| assorted_government              | 0.80    | 0.31 | 0.31 |
| generated_reviews_crowd          | 0.75    | 0.35 | 0.35 |
| apdf                             | 0.79    | 0.40 | 0.40 |
| msr_paraphrase                   | 0.82    | 0.28 | 0.28 |

Table 4: Minimum, maximum and average segment pairs cosine similarity for each sub-dataset

Table 4 presents the distribution of segment similarity score for each sub-dataset. Examples of segment pairs and their similarity score are shown in Appendix 3.

4 Experiments

4.1 Training data

We use the preprocessed and filtered segments pairs summing up to 1,001,752 pairs for the experiments. The ratio for training/validation/test sets is 80/10/10. The validation set and test set are sampled in a stratified manner in respect to their sources. We also ensure that their are no duplicate segments within the same language shared between validation and test sets.

Additionally, we use approximately 5M parallel English-Thai segments from OPUS [Tiedemann, 2012], an open source parallel corpus. Out of 9 English-Thai parallel datasets currently listed in OPUS, we use the following 6 datasets: OpenSubtitles [Lison and Tiedemann, 2016], Tatoeba [13] Tanzil [14] QED [Abdelali et al., 2014], Ubuntu and GNOME. The total number of segment pairs is 3,715,179. Then, we perform hand-crafted text cleaning as defined in the section 2.4.1 and segment filtering rules including setting Thai/English character ratio limit up to 0.1, number tokens up to 500 for each segment, removing segments meant for English translation with Thai characters and removing duplicated segment pairs. The resulting datasets contain 3,318,153 segment pair in total. The ratio for training/validation/test sets is 80/10/10.

[13] tatoeba.org
[14] tanzil.net
4.2 Models & Architectures

We use the Transformer [Vaswani et al., 2017], a supervised neural machine translation model, implemented in the Fairseq toolkit [Ott et al., 2019] as our NMT models in both English → Thai and Thai → English direction. We train Transformer models with the number of 6 encoder and 6 decoder blocks, 512 embedding dimensions, and 2,048 feed forward hidden units. The dropout rate is set to 0.1 only for the encoder and decoder input layer. The embedding of decoder input and output are shared. Maximum number of tokens per mini-batch is 9,750. The optimizer is Adam with initial learning rate of 1e-7 and weight decay rate of 0.0. The learning rate has an inverse squared schedule with warmup for the first 4,000 updates. Label smoothing of 0.1 is applied during training. The criteria for selecting the best model checkpoint is label-smoothed cross entropy loss.

There are 3 types of tokens used in the experiment namely word-level token tokenized by pyThaiNLP’s dictionary based tokenizer (newmm) for Thai, word-level token with Moses tokenizer for English (moses), and subword-level tokenized by SentencePiece [Kudo and Richardson, 2018] trained on the training set for both English and Thai (spm). The translation directions for MT model are both th → en, and en → th. The token type for each direction consists of word → word, word → subword, subword → word, and subword → subword (joined dictionary).

In addition, for the word-level tokens where Thai is the target language, space tokens are included during the word tokenization process with pyThaiNLP. When training transformer base and large, the maximum number of tokens for each batch is set to 9,750 and 6,750 respectively. The number of epoch for transformer base and large is set to 150 and 75 respectively. All the models in this experiment are trained on NVIDIA V100 GPU with mixed-precision training (fp16) and gradient accumulation for 16 steps. \(^{15}\)

4.3 Evaluation Methods

SacreBLEU [Post, 2018] is used to evaluate translation quality in both directions. For th → en translation, word-level outputs are detokenized with Moses detokenizer and subword outputs for both Thai and English are detokenized Sentencepiece [Kudo and Richardson, 2018]. The version string used for computing BLEU score for case-sensitive and case-insensitive are BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.2.10 and BLEU+case.lc+numrefs.1+smooth.exp+tok.13a+version.1.2.12 respectively.

For the en → th translation, the word-level outputs are detokenized by joining all the output tokens including space tokens as specified when preparing word-level tokens. The detokenized texts are tokenized again by the pyThaiNLP word tokenizer. We then compute BLEU score with the tokenized texts.

For model decoding, the model checkpoint selected is the epoch with minimum label-smoothed cross entropy loss. The beam width used is 4.

4.4 Experiment Results

4.4.1 Our Dataset and Parallel English-Thai Segments from OPUS

We report the evaluation results on the test set of our dataset, denoted as SCB_1M, and parallel English-Thai segments from OPUS, denoted as MT.OPUS. The the total number of segment pairs from SCB_1M and MT.OPUS test set are 100,177 and 297,874 respectively.

We trained models on each train set and cross validate on the test sets from 2 sources.

\(^{15}\)The source code used for the experiments can be found at: https://github.com/vistec-AI/thai2nmt
Table 5: Results on SCB_1M and MT_OPUS test set for th → en and en → th of the Transformer BASE models trained on either SCB_1M or MT_OPUS train set.

| Language pair | Token type | BLEU score (train set → test set) |
|---------------|------------|-----------------------------------|
| th → en       | newmm → moses | 14.32 20.88 25.48 |
|               | newmm → spm   | 14.36 23.57 25.21 |
|               | spm → moses   | 16.42 27.51 28.33 |
|               | spm → spm     | 17.15 28.09 26.37 |
| en → th       | moses → newmm | 12.68 16.56 17.77 |
|               | moses → spm   | 12.45 16.09 17.02 |
|               | spm → newmm   | 12.95 17.24 16.61 |
|               | spm → spm     | 12.54 15.35 15.27 |

Table 6: Results on Thai-English IWSLT 2015 test sets (tst2010-2013) for th → en and en → th of the Transformer BASE model trained on SCB_1M, MT_OPUS, and both.

| Language pair | Token type | BLEU score |
|---------------|------------|------------|
| th → en       | newmm → moses | 14.32       |
|               | newmm → spm   | 14.36       |
|               | spm → moses   | 16.42       |
|               | spm → spm     | 17.15       |
| en → th       | moses → newmm | 12.68       |
|               | moses → spm   | 12.45       |
|               | spm → newmm   | 12.95       |
|               | spm → spm     | 12.54       |
In Table 6, we compare the performance of our baseline models trained on SCB_1M, MT_OPUS, and both. We report detokenized SacreBLEU (case-sensitive) for th → en direction and BLEU4 (case-sensitive) for en → th direction.

| Language pair | Type  | BLEU score |
|---------------|-------|------------|
|               |       | Google     | AI-for-Thai | SCB_1M | MT_OPUS | SCB_1M + MT_OPUS |
| th → en       | cased | 14.19      | -           | 17.15  | 28.09   | 28.33            |
|               | uncased | 17.64     | -           | 17.90  | 28.72   | 29.0             |
| en → th       | cased | 15.36      | 6.14        | 12.95  | 17.24   | 17.77            |

Table 7: Results on Thai-English test sets (tst2010-2013). We submitted detokenized source segments in Thai language to Google Translation API to obtain translation in English. Our baseline model is Transformer (BASE) where the source and target token is BPE token built with SentencePiece library.

In Table 7, we compare the performance of our models with Google Translation API. We submitted the pre-processed Thai segments to Google Translation API (Neural Translation Model Predictions In Translation V3) on May 12 2020 to obtain translated segments in English and English segments from IWSLT 2015 to obtain translated segments in Thai. We submitted English segments to the Translation API provided by AI-for-Thai to obtain translated segments in Thai on May 16 2020. We evaluated only in English → Thai direction as that moment AI-for-Thai provided only English → Thai translation. We report detokenized SacreBLEU (case-sensitive) for th → en direction, and BLEU4 (case-sensitive) for en → th direction.

5 Discussion

Segment Alignment between Languages With and Without Boundaries

Unlike English, there is no segment boundary marking in Thai. One segment in Thai may or may not cover all the content of an English segment. Currently, we mitigate this problem by grouping Thai segments together before computing the text similarity scores. We then choose the combination with the highest text similarity score. It can be said that adequacy is the main issue in building this dataset.

Quality of Translation from Crawled Websites

Some websites use machine translation models such as Google Translate to localize their content. As a result, Thai segments retrieved from web crawling might face issues of fluency since we do not use human annotators to perform quality control.

Quality Control of Crowdsourced Translators

When we use a crowdsourcing platform to translate the content, we can not fully control the quality of the translation. To combat this, we filter out low-quality segments by using a text similarity threshold, based on cosine similarity of universal sentence encoder vectors. Moreover, some crowdsourced translators might copy and paste source segments to a translation engine and take the results as answers to the platform. To further improve, we can apply techniques such as described in [Zaidan, 2012] to control the quality and avoid fraud on the platform.

Domain Dependence of Machine Translation Models

We test domain dependence of machine translation models by comparing models trained and tested on the same dataset, using 80/10/10 train-validation-test split, and models trained on one dataset and tested on the other.

16https://www.aiforthai.in.th
For SCB_1M test set, models trained on SCB_1M training set have consistently 4-8 times higher BLEU score than those trained on MT_OPUS. In similar manner, for MT_OPUS test set, models trained on MT_OPUS have 2-4 times higher BLEU score than those trained on SCB_1M. This suggests that diversity of domains in the training set greatly impacts the performance of the models.

**Performance Uplifts from Models Trained on Existing Datasets**

For the IWSLT 2015 test set, the model trained on both OPUS [Tiedemann, 2012] and our dataset achieve 0.24 uplift in SacreBLEU for Thai to English translation and 0.53 uplift in SacreBLEU for English to Thai translation. The uplifts might be smaller due to the fact that IWSLT 2015 is a collection of TED Talk transcripts which are in the same domain as OpenSubtitles [Lison and Tiedemann, 2016], the majority of OPUS dataset.

In this section, we discuss the challenges in building a large-scale English-Thai machine translation and corresponding machine translation models.

**6 Conclusions**

We release English-Thai parallel corpus comprising of over 1 million segment pairs including both written and spoken language. The segment pairs in the corpus comprise text from various domains such as product reviews, laws, reports, news, spoken dialogues, and SMS messages. We also release 4 additional datasets for Thai text classification tasks and Thai sentence segmentation task.

We present an approach to filtering segment pairs with universal sentence encoder to remove misaligned segments. This approach can be used to only filtered out unrelated segments but it still prone to target segment adequacy error. Our further improvement is to develop a sophisticated method in order to obtain less noisy parallel corpus.

We conduct experiments on English → Thai and Thai → English machine translation systems trained on our dataset and the Open Parallel Corpus (OPUS) with different types of source and target token (i.e. word-level and subword-level). The evaluation results on Thai-English IWSLT 2015 test sets show that performance of our baseline models is on par with Google Translation API for Thai→English and outperform for both direction when OPUS is included in the training data.

**Acknowledgement**

This investigation is partially supported by the Digital Economy Promotion Agency Thailand under the infrastructure project code MP-62-003 and Siam Commercial Bank. We thank our data annotation partners Hope Data Annotations and Wang: Data Market; Office of the National Economic and Social Development Council (NESDC) through Phannisa Nirattiwongsakorn for providing government documents; Chonlapat Patanajirasit for training CRFCut sentence segmentation models on new datasets; Witchapong Daroontham for product review classification baselines; Pined Laohapiengsak for helping with sentence alignment using universal sentence encoder.

**References**

[ale, ] Top sites in thailand the sites in the top sites lists are ordered by their 1 month alexa traffic rank.the 1 month rank is calculated using a combination of average daily visitors and pageviews over the past month. the site with the highest combination of visitors and pageviews is ranked #1.

[Abdelali et al., 2014] Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC ’14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).
[Agić and Vulić, 2019] Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

[Aroonmanakun et al., 2007] 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.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. ArXiv, 1409.

[Byrne et al., 2019] Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Duckworth, D., Yavuz, S., Goodrich, B., Dubey, A., Cedilnik, A., and Kim, K.-Y. (2019). Taskmaster-I: Toward a realistic and diverse dialog dataset. arXiv preprint arXiv:1909.05358.

[Cettolo et al., 2015] Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., and Federico, M. (2015). The iwsit 2015 evaluation campaign.

[Chen and Kan, 2011] Chen, T. and Kan, M.-Y. (2011). Creating a live, public short message service corpus: The nus sms corpus. Language Resources and Evaluation, 47.

[Christodouloupoulos and Steedman, 2015] Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: The bible in 100 languages. Lang. Resour. Eval., 49(2):375–395.

[Dolan and Brockett, 2005] Dolan, W. B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005).

[Espl and Transducens, 2009] Espl, M. and Transducens, G. (2009). Bitextor, a free/open-source software to harvest translation memories from multilingual websites.

[Esplà et al., 2019] Esplà, M., Forcada, M., Ramírez-Sánchez, G., and Hoang, H. (2019). ParaCrawl: Web-scale parallel corpora for the languages of the EU. In Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

[Gehring et al., 2017] Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. CoRR, abs/1705.03122.

[Hassan et al., 2018] Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation. ArXiv, abs/1803.05567.

[Keskar et al., 2019] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation.

[Koehn, 2005] Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In MT summit, volume 5, pages 79–86. Citeseer.

[Koehn and Knowles, 2017] Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver. Association for Computational Linguistics.

[Kudo and Richardson, 2018] Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

[Lison and Tiedemann, 2016] Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).
[Loper and Bird, 2002] 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.

[Ott et al., 2019] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). Fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of NAACL-HLT 2019: Demonstrations.

[Ott et al., 2018] Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. ArXiv, abs/1806.00187.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

[Post, 2018] Post, M. (2018). A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

[Tiedemann, 2012] Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. CoRR, abs/1706.03762.

[Wu et al., 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G. S., Hughes, M., and Dean, J. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. ArXiv, abs/1609.08144.

[Yang et al., 2019] 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.

[Zaidan, 2012] Zaidan, O. (2012). Crowdsourcing annotation for machine learning in natural language processing tasks (non-final version! proofread version will be uploaded april 30, 2012.).

[Ziemski et al., 2016] Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).
Appendix 1: Datasets for Other Tasks

In addition to the machine translation tasks, we can also use some datasets for other natural language processing tasks in Thai.

1.1 Paraphrase Identification

For the paraphrase identification task, we take the crowdsourced translations from English to Thai based on Microsoft Research Paraphrase Identification corpus [Dolan and Brockett, 2005]. The current version of msr_paraphrase has 10,122 translated sentences. As a result, the dataset includes 3,513 and 1,485 sentence pairs for training and test set respectively (reduced from the original dataset by 563 pairs for training set and 240 pairs for test set).

| Dataset  | Sentence pairs | # Paraphrased | # Non-paraphrased |
|----------|----------------|---------------|-------------------|
| Train set| 3,513          | 2,349         | 1,164             |
| Test set | 1,485          | 516           | 969               |

Table 8: Number of sentences pairs along with paraphrased and non-paraphrased sentences from Microsoft Research Paraphrase Identification corpus that we have translated into Thai.

1.2 Sentence Segmentation

We can build sentence segmentation models with the generated product review dataset as described in Section 2.3.1.

1.3 Translation Quality Estimation

The fact that generated_reviews_yen use human annotators to label the Google-Translated reviews allows us to have another dataset for translation quality estimation. The total number of reviews in this dataset is 302,066.

![Figure 2: Distribution of sentences per review of the correctly translated reviews (a) and correctly translated reviews (b) in the Sentence Segmentation dataset](image-url)
### 1.4 Product Review Classification

We combine `generated_reviews_translator` and `generated_reviews_yn` to create a product review classification dataset with 64,760 reviews. The distribution of label is shown below. Note that we might want to exclude those reviews in `generated_reviews_yn` that are labelled as not human-readable from validation set when evaluating a text classification model.

| Review star | Total number of reviews | Percentage |
|-------------|-------------------------|------------|
| 1           | 11,602                  | 26.75      |
| 2           | 934                     | 2.15       |
| 3           | 9,976                   | 23.00      |
| 4           | 11,654                  | 26.87      |
| 5           | 9,207                   | 21.23      |

Table 10: Label distribution of the `generated_reviews_translator`

| Review star | Total number of reviews | Percentage |
|-------------|-------------------------|------------|
| 1           | 4,263                   | 19.93      |
| 2           | 4,245                   | 19.85      |
| 3           | 4,504                   | 21.06      |
| 4           | 5,176                   | 24.20      |
| 5           | 3,199                   | 14.96      |

Table 11: Label distribution of the `generated_reviews_yn`

| Review star | Total number of reviews | Percentage |
|-------------|-------------------------|------------|
| 1           | 15,865                  | 24.50      |
| 2           | 5,179                   | 8.00       |
| 3           | 14,480                  | 22.36      |
| 4           | 16,830                  | 25.99      |
| 5           | 12,406                  | 19.16      |

Table 12: Label distribution of the resulting product review classification dataset
Appendix 2: Example Sentence Pairs

The sentences pairs examples from our English-Thai machine translation dataset are listed below:

2.1 Manual translation by hired and crowd-sourced translators

1) Dialogues in spoken language from Taskmaster-1

| Source (en): Hakka, and uptown restaurant Philippe Chow are top rated | Target (th): ฮั้กการะ ที่บริการในอัพทาวน์ร้านเชอ ฟิลิปเป่ เข้าได้ เวลาตั้งอยู่นะ |
| Source (en): What showtimes do they have at night? | Target (th): ตอนกลางคืนมีรอบถึงเที่ยงยังไหม?
| Source (en): Who doesn’t deliver these days? Alright, so a White Wonder with chicken & onions? | Target (th): เอ๋ยนี้ใครเข้าไม่มีผละวัชชีแจ่แล้วถ้า? เอาเถอะ เอาไว้ที่ร้านเลยไม่ได้กับตัวผมแน่?

2) SMS Messages from NUS SMS corpus

| Source (en): They said ถ้าคุณไม่ชำมอาร์ท แล้วจุส์เซ็นที่รีเมйลของฉัน | Target (th): เข้าขอบอกว่าตอนนี้คุณทำสปอร์ตให้ผู้เรียบอร่อยแล้วใช้ไหม หรือคุณแต่งส่างค่าหีบเหล่านั้น |
| Source (en): Watch lor. I saw a few swatch one i thk quite ok. Ard 116 but i need 2nd opinion leh... | Target (th): นาฬิกาเร็ว เรื่องนี้เห็นว่ามีแบบอยู่สองแบบก็คินะ ประมาณ 116 แต่ก็ต้องเลือกถึงเรื่องนี้จะ |
| Source (en): It’s true already. I thk she muz c us tog then she believe. | Target (th): หมาจริงนะ แล้วว่าเธอต้องฟังเราเพื่อก่อนลังจะเขย |

3) Generated product reviews

| Source (en): I actually just finished it because i thought maybe i’d beat every level. Nope. | Target (th): ตัดเส้นเสียงผ่านเพราะคิดว่าฝันจะชนะได้จริงด้วยเกมส์ แต่ไม่เลยค่ะ |
| Source (en): My husband wanted to try this on his black and yellow tabby, who has very mild digestive problems. | Target (th): สามีต้องลองบอกลองของที่นี้กับแมวสีดำและเหลืองของเอช และเป็นคนที่ใจดีเกินกว่าที่ถ่ายที่เกี่ยวกับห้องสีมาก |

| Source (en): The connector on it is different, so I’m hesitant whether or not it’s an actual OEM one. | Target (th): ตัวเชื่อมต่อหน้านั้นแตกต่างกันดั้นหน้านั้นจึงล่าสุดว่าจะเป็น ที่จริงหรือไม่ |

20
5) Mozilla Common Voice

Source (en): The fool wanders, the wise man travels.
Target (th): คนโง่มาเที่ยว คนฉลาดเที่ยวที่อื่น

Source (en): Would you like a game of noughts and crosses?
Target (th): คุณอยากเล่นเกมเสือดำหรือเปล่า

Source (en): Paul moved to Oxford for his D Phil.
Target (th): พอแล้วไปออกฝรั่งเพื่อเรียนต่อปริญญาเอก

6) Microsoft Research Paraphrase Identification corpus

Source (en): She started taking supplements two years ago - partly to stave off mild dementia that affects her elderly parents.
Target (th): เธอเริ่มทานยาเพื่อป้องกันภาวะสมองเสื่อมที่ไม่รู้จะชะลอไม่ได้ ที่กับอายุเกิดสูญหายก่อนประสม

Source (en): The vulnerability affects Windows NT 4.0, NT 4.0 Terminal Services Edition, XP and 2000, as well as Windows Server 2003.
Target (th): ที่อยู่ในกลุ่มระบบปฏิบัติการ Windows NT 4.0, NT 4.0 Terminal Services Edition, XP และ 2000 รวมถึง Windows Server 2003

Source (en): In July, EMC agreed to acquire Legato Systems (Nasdaq: LGTO) for about $1.2 billion.
Target (th): ในเดือนกรกฎาคม EMC ตกลงจะซื้อธุรกิจ Legato (แอดแส: LGTO) ประมาณ 1.2 พันล้านดอลลาร์

2.2 Translated segment pairs via Google Translation API verified by translators

1) Generated product reviews

Source (en): I read this book on the advice of an acquaintance.
Target (th): ฉันอ่านหนังสือเล่มนี้ตามคำแนะนำของคนรู้จัก

Source (en): Bought the Cuisinart DCC-2700 coffeemaker from Amazon based on other people’s reviews.
Target (th): ซื้อเครื่องทำกาแฟ เครื่องปั้นกาแฟ 2700 จาก อะเมซอน ตามข้อความในรีวิว

Source (en): I’ve been through a number of screen protectors in my life and all were from ZAGG – until these.
Target (th): ฉันเคยใช้ฟิล์มกรองหน้าจอมากมายในชีวิตของฉันและทั้งหมดมาจาก ZAGG - จนกระทั่งสิ้นที่
### 2.3 Aligned segment pairs from web-crawled data and PDF documents

#### 1) Assorted government

| en: | Furthermore, the car sale volume reached 1.25 million cars comparing to an average of 500,000 -700,000 units per year. |
| th: | ทั้งนี้ การจำหน่ายรถยนต์ในประเทศทั้งปีสูงถึง 1.25 ล้านคันเทียบกับเฉลี่ยประมาณ 500,000 – 700,000 ตันต่อปี |

| en: | Meanwhile, NPLs1 rose from 0.96 percent in the first quarter to 1.0 percent. Excess liquidity of commercial bank system considerably tightened. |
| th: | ในขณะที่สัดส่วนหนี้ที่ไม่ก่อให้เกิดรายได้ (NPLs1) ต่อสินเชื่อดังกล่าวขึ้นจากร้อยละ 0.96 ในไตรมาสก่อนหน้านี้เป็นร้อยละ 1 สภาพคล่องในระบบธนาคารเหลื่อมล้ำตัวเขิน |

| en: | Private consumption in this quarter dropped by 0.1 percent (qoq). |
| th: | โดยในไตรมาสนี้การบริโภคของเอกชนลดลงร้อยละ 0.1 (qoq) |

#### 2) English-Thai parallel Wikipedia corpus

| en: | Polish forces then withdrew to the southeast where they prepared for a long defence of the Romanian Bridgehead and awaited expected support and relief from France and the United Kingdom. |
| th: | ทหารโปแลนด์จึงถอนออกไปใต้ทิศใต้ที่เตรียมตั้งกองทัพในการรบกับโรมานีาและรอคอยการสนับสนุนจากฝรั่งเศสและอังกฤษ |

| en: | Railway lines of JR East primarily serve the Kanto and Tohoku regions, along with adjacent areas in Kōshin’etsu region (Niigata, Nagano, Yamanashi) and Shizuoka prefectures. Section:::Shinkansen. |
| th: | ทางรถไฟของโดยสารออกนั้นเริ่มก่อสร้างตั้งแต่รัฐทางเหนือ ทั้งจังหวัดที่อยู่ในบริเวณแหล่งน้ำและรัฐของจังหวัดที่อยู่ทางตะวันตกและจังหวัดที่อยู่ทางตะวันออก Section:::สินค้านี้ |

| en: | Section:::Computer simulation. A computer simulation (or "sim") is an attempt to model a real-life or hypothetical situation on a computer so that it can be studied to see how the system works. |
| th: | นักวิทยาศาสตร์คอมพิวเตอร์ ที่มีส่วนในการพัฒนาคอมพิวเตอร์ที่มี "sim" เป็นการสร้างแบบจำลองของวัตถุจริง หรือเหตุการณ์ตามธรรมชาติ ด้วยคอมพิวเตอร์เพื่อใช้ในการศึกษากระบวนการใดๆอย่างไร |

#### 3) News sites (Asia Pacific Defense Forum)

| en: | Fiji’s Defense Ministry said it paid U.S. $8.8 million for the shipment and declined to give specifics about what it entailed, other to say that a second shipment was forthcoming, the Nikkei Asian Review reported in February 2016. Russian military advisors were also expected to arrive in Fiji to teach Soldiers there how to use the equipment. |
| th: | กระทรวงกลาโหมฟิจิได้ประกาศว่าได้ซื้อวัสดุที่มีหน่วยงานเป็นจำนวน 8.8 ล้านดอลลาร์สหรัฐฯ (ประมาณ 308 ล้านบาท) และปฏิเสธที่จะให้ข้อมูลเพิ่มเติมเกี่ยวกับโครงการที่ได้รับ ผู้สื่อข่าวจากมิติการช่วยเหลือทางการทหารของรัสเซียที่ที่มาทรงพลัง นิคีกิ เอเชีย ได้เรียกระดับนิคีกิ เอเชีย ให้ร่วมงานของรัฐบาลฟิจิ พ.ศ. 2559 นอกจากนี้ หัวหน้าการทหารของรัสเซียจะมาเยี่ยมชมห้องปฏิบัติการซึ่งมีการกระจายของสถานีวิทยุการใช้ปืนกลใหญ่ให้กับทหารที่ฟิจิ |

---

22
Cambodia, China, Laos, Pakistan, Papua New Guinea and Thailand passed new cyber laws in 2015 and 2016. Cambodia’s new telecommunications law and other e-commerce and cyber crime legislation are “promising examples of growth in cyber maturity in one of the region’s cyber underperformers,” the report said. Laos also passed new cyber crime legislation that included definitions from the Council of Europe’s Convention on Cybercrime. The ASEAN Economic Community, which was established in late December 2015, will propel new cyber crime legislation in Southeast Asia, the report predicted.

4) Crawled pages from top-500 websites

Chomchuen said that in recent times, young Thai grooms give dowries as a simple symbolic gesture, and then have the money returned to them by the bride’s family after the wedding is over.

The Bangkok Metropolitan Administration has launched a three-day celebration of the new Giant Swing located in front of the Bangkok City Hall.

5) Crawled pages from websites listed in ParaCrawl v5

Inhabitants London has approximately 8,673,713 inhabitants.

Women’s Pink Three-Quarter Sleeved T-Shirt Plus Size Style Pocket Trimmed Top

Regardless of Bar Forming Machine, meat processing machine, vegetable processing machine, bread making equipment or commercial deep fryer, every commercial kitchen equipment designed by Ding-Han is to meet your requirement of high productivity, and low cost.
Appendix 3: Sentence Pairs Similarity with USE

Figure 3: Distribution of sentence pairs similarity for each source before applying text cleaning and filtering rules

Figure 4: Distribution of sentence pairs similarity for each source after applying text cleaning and filtering rules
3.1 Example of correctly aligned sentence pairs with high similarity score

**sub-dataset: wikipedia**

en: The first portable nuclear reactor "Alco PM-2A" was used to generate electrical power (2 MW) for Camp Century from 1960.

th: เครื่องปฏิกรณ์นิวเคลียร์แบบพกพาเร็วแรกคือ "Alco PM-2A" ใช้ในการสร้างพลังงานไฟฟ้า (2 เมกะวัตต์) ได้ในปี 1960

similarity: 0.928

**sub-dataset: assorted government**

en: Both side discussed and exchanged views on the topics of mutual interests both at bilateral and regional levels, including, Thai - European Union relations, Thailand’s political developments, ASEAN - European Union Relations, Thailand’s ASEAN Chairmanship 2019, and various regional security issues.

th: ทั้งสองฝ่ายได้หารือและแลกเปลี่ยนความคิดเห็นในประเด็นที่อยู่ในความสนใจต่าง ๆ ทั้งในระดับพาดิศและในระดับภูมิภาค อาทิ ความสัมพันธ์ไทย - ยุโรป, ทุกข์การทางการเมืองไทย, ความสัมพันธ์อาเซียน - ยุโรป, การเป็นประธานอาเซียนของไทยในปี 2546 และท้าทายทั้งสองฝ่ายต่อประเด็นความภัยคุกต่าง ๆ เป็นต้น

similarity: 0.910

**sub-dataset: assorted government**

en: Thus, import of goods and services at constant price in 2004 is expected to expand by 9.2 percent, higher than 7.4 percent in 2003.

th: ดังนั้นการนำเข้าสินค้าและบริการ ณ ราคาคงที่ ในปี 2547 จึงเพิ่มขึ้นร้อยละ 9.2 สูงกว่าร้อยละ 7.4 ในปี 2546

similarity: 0.906

**sub-dataset: apdf**

en: Satellite images taken in November 2016 show that Vietnam lengthened its runway on Spratly Island from less than 760 meters to more than 1 kilometer, the Asia Maritime Transparency Initiative (AMTI) said.

th: ภาพจากดาวเทียมที่ถ่ายเมื่อต้นพฤศจิกายน พ.ศ. 2559 แสดงให้เห็นว่าเวียดนามเพิ่มความยาวของทางช้างเผือกใน เครื่องบินเหยื่อเครื่องบินของมิชชันนิวส์เพิ่มขึ้นกว่า 760 เมตรเป็นมากกว่า 1 กิโลเมตร โครงการด้านไปรษณีย์ทางทะเลในเอเชียกล่าว

similarity: 0.902

**sub-dataset: paracrawl**

en: Abundant vegetable proteins and dietary minerals are the best nutrients for shiny coat and smooth skin for pet .

th: โปรตีนจากพืชและสารอาหารที่ดีที่สุดสำหรับการบริโภคและสุขภาพจากเนื้อสัตว์ที่ดีที่สุด สำหรับการเคลื่อนย้ายและผิวที่เรียบเนียนสัตว์
3.2 Example of correctly aligned sentence pairs with low similarity score

sub-dataset: task_master_1

en: Sure thing, and what would you like to drink?
th: ได้แน่นอนค่ะ ไม่ทราบว่าสิ่งอะไรหรือคะ

similarity: 0.255

sub-dataset: task_master_1

en: great, and you said for pick-up is that right?
th: เบี้ยมคะ พี่ชูใจจะรับกลับบ้านถูกไหมคะ

similarity: 0.224

sub-dataset: mozilla_common_voice

en: A penny wise and a pound foolish.
th: เสียเงินเสียชีวิต

similarity: 0.222

sub-dataset: mozilla_common_voice

en: Not yet, madam.
th: ยังไม่ได้

similarity: 0.192

sub-dataset: nus_sms

en: Take your time.
th: ดูสมาทบเลยไม่ต้องรีบ

similarity: 0.246
### 3.3 Example of incorrectly aligned sentence pairs with low similarity score

| Sub-dataset | Sentence (English) | Sentence (Thai) | Similarity |
|-------------|--------------------|-----------------|------------|
| nus_sms     | Search your mailbox now. | ส่งมาแล้ว ตรวจสอบบัตรคีย์ของตนเองแล้วหน้า | 0.291 |
| apdf        | If I were to characterize the border environment in one word, it would be in ‘volumes.’ The volumes of people and goods crossing our border continues to grow exponentially. | ถ้าจะเดาถึงสภาพแวดล้อมรัฐธรรมนูญในเรื่องว่าจะมีอะไรอยู่ใน mux ในผสมผสานมากขึ้น และเพิ่มความเร็วและประสิทธิภาพในการดำเนินการปฏิบัติที่สะอาด นั่นก็คือทักษะรูปแบบของอัตราส่วนลดกา-ลังทุ่งในดำเนินการตามเกณฑ์หลักและเทคโนโลยีที่ดีเพื่อให้มีการปฏิบัติตามกฎหมายมากขึ้น และเพิ่มอัตราส่วนในคงที่ | 0.206 |
| assorted government | It is advised to follow these steps to avoid heat-related stress: | นี้เป็นวิธีการป้องกัน | 0.043 |
| assorted government | - 18 January 2019 from 07.00 – 16.00 hrs. | วันที่ 18 มกราคม พ.ศ. 2562 ในช่วงเวลา 07.00 – 16.00 น. | 0.008 |
| paracrawl   | This rubber seal blocks water and foreign materials from entering the drag system. | นิวเมติกนี้ป้องกันน้ำและสิ่งของไม่ที่มาจากที่ใช้ในการอุปกรณ์การลาก | 0.181 |
sub-dataset: paracrawl

en: Strawberries are available January through May, melons and grapes are available May through September and Mandarin Oranges are available October through December.

th: วานาทิที่ 2 ของเดือนกุมภาพันธ์ถึงตุลาคมมีการเริ่มลงโทษในกฎหมายที่กำหนดวัตถุสิ่งแวดล้อม (การผลิตและคุณภาพ)

similarity: 0.128

3.4 Example of sentence pairs with high similarity score but lack adequacy in source or target sentence

| sub-dataset: generated_reviews_translator |
|------------------------------------------|

en: Battery life not what I’d hoped for, maybe 2-3 hours shooting continuous video and then have to recharge before you can fire again.

th: อาจใช้เวลาประมาณ 2-3 ชั่วโมงแล้วกีฬาจะใหม่จะใช้ได้อีก

similarity: 0.633

| sub-dataset: generated_reviews_translator |
|------------------------------------------|

en: This is a pretty good album and I’m glad I got it, however it just doesn’t have the classic vibe that his other albums or mixtapes seemed to have, plus there are several tracks from his mixtapes.

th: นี่เป็นอัลบั้มที่ดีและดีที่สุดที่ฉันได้มีมาก ยังไงก็ตาม ฉันไม่มีความเพิ่มเติมเหมือนที่อัลบั้มอื่น ๆ หรือมิกซ์เทปของฉัน

similarity: 0.792

| sub-dataset: generated_reviews_translator |
|------------------------------------------|

en: I don’t do the paranormal stuff as much so that doesn’t bother me.I’m not sure if I’ll read from this author again.It seemed at times more story rather than character.

th: ฉันไม่ได้สนใจเรื่องราบนี้หรือธรรมชาติที่มันเรื่องนั้นเลยไม่ผ่าข่าย

similarity: 0.517

| sub-dataset: generated_reviews_translator |
|------------------------------------------|

en: It will be going back immediately!

th: ย่าเสียใจจนเสียใจกับสินค้าเหล่านี้เยอะ เต็มตัวจะซื้อมันกลับไปให้เลยท่า!

similarity: 0.417
Appendix 4: Sample of Translation Results

The sampled translation results bellow are from the Transformer Base model trained on the train set (80%) from our 1 million segment pairs dataset where the source and target token for the MT model is subword (joined dictionary).

Direction: English → Thai

| Source: The centre was based at the Munich Fairgrounds, in what was formally Munich Airport. The building is now known as the Munich Exhibition Centre. |
|---|
| Reference: ศูนย์แสดงสินค้าสัมพันธ์ที่ “มิวนิเกอร์ฟาร์” (Munich Fair) ซึ่งก่อสร้างขึ้นในบริเวณของท่าอากาศยานมิวนิเกก้าจุ้นน์ ปัจจุบันอาคารแห่งนี้เป็นที่รู้จักในชื่อ "ศูนย์แสดงสินค้ามิวนิเกก้า" (Munich Exhibition Centre) |
| Hypothesis: ศูนย์แสดงสินค้ามิวนิเกก้า ปัจจุบันอาคารแห่งนี้เป็นที่รู้จักในชื่อ "ศูนย์แสดงสินค้ามิวนิเกก้า" |

| Source: I want the Almond Milk, and if they are out of that I would like the Coconut Milk. |
|---|
| Reference: ฉันต้องการนมอัลมอนด์ ถ้าไม่มีนมอัลมอนด์ ฉันจะต้องการนมมะพร้าว |
| Hypothesis: ฉันต้องการนมอัลมอนด์ แต่ถ้าไม่มีนมอัลมอนด์ ฉันจะต้องการนมมะพร้าว |

| Source: Traveling intercity by bus is generally cheaper than traveling by train. Buses vary widely in terms of comfort and onboard options depending on your budget. One big advantage of traveling by bus is that you can journey overnight, meaning that you save the money of a night’s accommodation. Expect to take around eight or nine hours from Tokyo to the western city of Osaka. The biggest transport hub for buses is the Shinjuku Expressway Bus Terminal, where you can board a bus headed for every corner of the country. |
|---|
| Reference: การเดินทางระหว่างเมืองโดยรถโดยสารเป็นทางเลือกที่ถูกกว่าการเดินทางโดยรถไฟ รถโดยสารมีราคาแตกต่างกันไปตามความสามารถในการพักผ่อนและความสะดวกสบายและความปลอดภัยของผู้โดยสาร ความเร็วของรถโดยสารมีอยู่ในกลุ่มการเดินทางโดยรถไฟ ความเร็วendafricaj ผู้โดยสารมีที่พักผ่อนที่เหมาะสมได้ โดยจะใช้เวลาประมาณ 8 หรือ 9 ชั่วโมงจากโตเกียวไปยังโอซาก้าจะวัดจากประมาณ 8-9 ชั่วโมง และเป็นสุนัขที่ระยองการเดินทางถึงที่สุดของรถโดยสารอัป inconvenientุส สถานีรถไฟอัป inconvenientุส สามารถขับรถโดยสารอัป inconvenientุส ผู้โดยสารมีที่พักผ่อนที่เหมาะสมได้ |
| Hypothesis: การเดินทางโดยรถโดยสารระหว่างเมืองโดยรถโดยสารเป็นทางเลือกที่ถูกกว่าการเดินทางโดยรถไฟ รถโดยสารมีราคาแตกต่างกันไปตามความสามารถในการพักผ่อนและความสะดวกสบายและความปลอดภัยของผู้โดยสาร ความเร็วของรถโดยสารมีอยู่ในกลุ่มการเดินทางโดยรถไฟ ความเร็วendafricaj ผู้โดยสารมีที่พักผ่อนที่เหมาะสมได้ โดยจะใช้เวลาประมาณ 8 หรือ 9 ชั่วโมงจากโตเกียวไปยังโอซาก้าจะวัดจากประมาณ 8-9 ชั่วโมง และเป็นสุนัขที่ระยองการเดินทางถึงที่สุดของรถโดยสารอัป inconvenientุส สถานีรถไฟอัป inconvenientุส สามารถขับรถโดยสารอัป inconvenientุส ผู้โดยสารมีที่พักผ่อนที่เหมาะสมได้ |
Source: Additionally, B cells present antigens (they are also classified as professional antigen-presenting cells (APCs)) and secrete cytokines. In mammals, B cells mature in the bone marrow, which is at the core of most bones. In birds, B cells mature in the bursa of Fabricius, a lymphoid organ where they were first discovered by Chang and Glick, (B for bursa) and not from bone marrow as commonly believed.

Reference: ปิแเด (,) เป็นเซลล์ที่มีเซลล์หน้าที่เป็นเซลล์ที่มีหน้าที่เป็นเซลล์หน้าที่มีหน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์หน้าที่เป็นเซลล์แ

Hypothesis: นอกจากนี้ เซลล์ B อิงเน้สนอแอดแอน (ซึ่งเป็นเซลล์ที่เป็นด่านในแอนติเจนของอิมมิชั่น) และ เลสเตอร์ cytokins ในสิ่งเรียบลูกตัวของ เซลล์ B เรียกในโครงสร้างเป็นแผนภูมิของรายส่วนใหญ่ ในนัก เซลล์ B เรียกใน Bursa of Fricius, lymphoid organ ที่ที่หลากหลายเห็นครั้งแรกโดย Chang and Glick (B for Bursa) และไม่ใช้จากไฮเพอร์ลูกที่พบทั่วไป

Direction: Thai → English

Source: ภาพยนตร์ที่สวยงามเรื่องนี้ถ่ายทำอย่างสวยงามโดยนักถ่ายทำภาพยนตร์ Jonathan Frakes ในเดือนกันยายนที่ผ่านมา 2545

Reference: This beautiful film is beautifully filmed by cinematographer Jonathan Frakes in the early spring of 2002.

Hypothesis: This beautiful film is beautifully filmed by the filmmaker Jonathan Frakes in early spring 2002.

Source: เรามีแนะนำอยู่ที่นี่ มีอะไรที่มีชื่อเสียงนี้ถูกจับกันมาอยู่ที่ทั่วไป อินเทอร์ร์เซลส์ แนะนำเด็กที่แอดแอนและมีมิชันกับอินเทอร์ร์เซลส์

Reference: Okay. I have two suggestions. How about Arrival, a drama sci-fi with themes of time travel and aliens? Or how about Interstellar, an action and adventure sci fi with themes of space and time travel?

Hypothesis: I’d recommend it. What’s Ful? Dramma Xyfi has time theme with aliens and Interstellars. Action events like Avengers Science have space themes and overseas.

Source: เพื่อให้ท่านก้าวข้ามเทคโนโลยีที่รวดเร็วในปัจจุบันและเพื่อให้แน่ใจว่า SOP ที่เหมาะสมทุก บริษัท และโรงงานของเราได้รับไปรับรอง ISO 9001: 2008, ISO 14001: 2004 และได้รับการยอมรับคุณภาพ EC รวมถึงมาตรฐาน 11B เรียบร้อยแล้ว

Reference: In order to keep pace with the fast technology development nowadays and to ensure proper SOP, all our company and factories have successfully obtained the certificates of ISO 9001:2008, ISO 14001:2004 and EC
Quality System Certificate including Article 11B.

Hypothesis: To keep up with current rapid technology development and ensure that all companies and our plants have received ISO 9001: 2008, ISO 14001: 2004 and EC quality system certificates, including Article 11B.

Source: รางวัลที่ดีที่สุดของเทศกาลหนังชื่อว่า "หนังทองคำ" (Goldener Bär) และ "หนังเงิน" (Silberner Bär) เป็นสัญลักษณ์ของเบอร์ลิน รางวัลหนังทองคำสื่อถึงรางวัลหนังทองคำสำหรับภาพยนตร์ยอดเยี่ยม (Best Motion Picture) และรางวัลหนังทองคำเกียรติยศสำหรับผู้ที่มีทุนทรัพย์ให้กับภาพยนตร์ (Lifetime Achievement)

Reference: Golden Bear ("Goldener Bär") Silver Bear ("Silberner Bär") The Silver Bear was introduced in 1956 as an award for individual achievements in direction and acting, and for best short film.

Hypothesis: The Golden Bear (Silver Bear) and the Golden Bear (Silver Bear) are two categories: the Golden Bear Award for Best Motion Picture and the Golden Bear Award for Lifetime Achievement.

The following sampled translation results shows the different in translated sentence for each pair of source and target token (word-level, subword-level) of the MT model.

**Direction: English  →  Thai**

Source: ถ้าคุณมีจุดไปใกล้จากเมืองเข็มแข็งไปถึงคลอง "เกษตรุ บาร์บาร่า คาเฟ่" ที่สามารถอยู่บนที่ตั้งระเบียงไม้ของร้าน .AppendFormatๆ กับการชมวิว "สะพานสายรุ้งแบร์บริคซ์"

Reference: Closer to central Tokyo is Dejikyu BBQ Café in Odaiba, where you can barbecue on a comfortable wooden deck overlooking Rainbow Bridge.

Hypotheses:

bpe → bpe : Another closer to downtown is Dejikyu’s BBQ Cafe, where you can grill BBQ on a woody balcony with a view of Rainbow Bridge.

word → word : Another closer to downtown is DejiQ’s BBQ Cafe, where you can barbecue on a cozy wooden porch with a view of Rainbow Bridge.

word → bpe : Another closer to the city center is DejiQ BBQ Cafe, where you can barbecue on a wooden balcony with a casual atmosphere while watching Rainbow Bridge.

bpe → word : Another closer location to downtown is DejiQ BBQ Cafe, where you can barbecue on a casual wooden balcony with a view of Rainbow Bridge.
Source: Shares of Mattel were down 13 cents to $19.72 on the New York Stock Exchange.

Reference: Shares of Mattel were down 13 cents to $19.72 on the New York Stock Exchange.

Hypotheses:

*bpe → bpe*: Mattel’s shares fell 13 cents to $19.22 on the New York Stock Exchange.

*word → word*: Shares of the junk have been down 13 cents to $25 in the New York Stock Exchange.

*word → bpe*: Shares of Mattel fashion fell 13 cents to dollar on the New York Stock Exchange.

*bpe → word*: Matte’s shares were down 13 cents to $72 on the New York Stock Exchange.

Direction: Thai → English

Source: Closer to central Tokyo is Dejikyu BBQ Café in Odaiba, where you can barbecue on a comfortable wooden deck overlooking Rainbow Bridge.

Reference: 

*bpe → bpe*: ใกล้ๆ ใกล้ๆ สถาปัตย์ สถาปัตย์ "บริการ บริการ คาเฟ่" ที่สามารถอย่างบริการบนเตาไม้ที่มองเห็นวิวสะพานสะขั้นเรียงในบริเวณใกล้ทางตรวจได้

*word → word*: ใกล้ๆ ใกล้ๆ ศาสน์เมือง คือ " โอ โอ ค่าหน้า " ที่สามารถ บริการได้ บน ลาดท้าย ใสนั้น ๆ สามารถมองเห็นสะพานสะขั้นเรียง บริเวณ บริเวณ ได้

*word → bpe*: ใกล้ๆ ใกล้ๆ สถาปัตย์ สถาปัตย์ "บริการบนเตาไม้ที่สามารถมองเห็นสะพานสะขั้นเรียงในบริเวณ

*bpe → word*: ใกล้ๆ ใกล้ๆ สถาปัตย์ สถาปัตย์ " โอ โอ ค่าหน้า " ที่สามารถ บน ลาดท้าย ใสนั้น ๆ สามารถมองเห็นสะพานสะขั้นเรียง บริเวณ บริเวณ ได้
bpe → bpe: หุ้นของ Mattel ลดลง 13 เลขเป็น 19.72 ต่อล้านในตลาดหุ้นนิวยอร์ก

word → word: หุ้นของแท้เทลลดลง 13 เลขเป็น 29.32 ต่อล้านในตลาดหุ้นนิวยอร์ก

word → bpe: หุ้นแมทเทลลดลง 13 เลขเป็น $87.54 ในตลาดหลักทรัพย์นิวยอร์ก

bpe → word: หุ้นของ Matte ลดลง 13 เลขเป็น 19.7 เลขในตลาดหุ้นนิวยอร์ก
Appendix 5: Descriptive Statistics of Resulting Dataset

5.1 English/Thai Character Ratio for Each Sub-dataset

| Sub-dataset name        | Character Ratio |          |          |          |
|-------------------------|----------------|----------|----------|----------|
|                         |                | mean     | median   | (min, max) |
| task_master_1           | en             | 0.78     | 0.78     | (0.51-1.00) |
|                         | th             | 0.96     | 0.96     | (0.51-1.00) |
| generated_reviewsTranslator | en          | 0.79     | 0.80     | (0.42-0.94) |
|                         | th             | 0.97     | 0.99     | (0.40-1.00) |
| nus_sms                 | en             | 0.76     | 0.76     | (0.00-1.00) |
|                         | th             | 0.94     | 0.95     | (0.00-1.00) |
| msr_paraphrase          | en             | 0.81     | 0.81     | (0.66-0.88) |
|                         | th             | 0.89     | 0.93     | (0.11-1.00) |
| mozilla_common_voice    | en             | 0.79     | 0.80     | (0.56-0.94) |
|                         | th             | 0.98     | 1.00     | (0.50-1.00) |
| generated_reviews_crowd | en             | 0.79     | 0.79     | (0.60-0.89) |
|                         | th             | 0.97     | 0.99     | (0.60-1.00) |
| generated_reviews yn    | en             | 0.79     | 0.80     | (0.52-0.94) |
|                         | th             | 0.97     | 1.00     | (0.50-1.00) |
| assorted_government     | en             | 0.81     | 0.82     | (0.52-0.92) |
|                         | th             | 0.93     | 0.94     | (0.25-1.00) |
| thai_websites            | en             | 0.81     | 0.81     | (0.55-0.94) |
|                         | th             | 0.83     | 0.85     | (0.46-1.00) |
| wikipedia                | en             | 0.82     | 0.82     | (0.54-0.91) |
|                         | th             | 0.90     | 0.93     | (0.50-1.00) |
| paracrawl                | en             | 0.81     | 0.81     | (0.51-0.91) |
|                         | th             | 0.89     | 0.93     | (0.50-1.00) |
| apdf                     | en             | 0.82     | 0.82     | (0.65-0.89) |
|                         | th             | 0.96     | 0.97     | (0.52-1.00) |

Table 13: Mean, median, minimum and maximum ratio of English and Thai characters in the segments for each sub-dataset. English characters include English alphabet. Thai characters include Thai consonants, vowels, tone diacritics, currency symbol (฿) and digits (0,1,2,3,4,5,6,7,8,9). Specifically, the range of Unicode is 0E01-0E29. Numbers are counted as characters for both English and Thai.
## 5.2 English-to-Thai Tokens Ratio for Each Sub-dataset

| Sub-dataset name           | English-to-Thai Tokens Ratio |
|----------------------------|-------------------------------|
|                            | mean  | median | (min, max)     |
| task_master_l              | 1.18  | 1.14   | (0.16-2.86)    |
| generated_reviews_translator| 1.15  | 1.10   | (0.20-11.75)   |
| nus_sms                    | 1.01  | 1.00   | (0.07-16.00)   |
| msr_paraphrase             | 1.09  | 1.07   | (0.41-2.38)    |
| mozilla_common_voice       | 1.21  | 1.17   | (0.18-8.00)    |
| generated_reviews_crowd    | 1.17  | 1.14   | (0.21-4.50)    |
| generated_reviews_yn       | 1.18  | 1.14   | (0.33-4.25)    |
| assorted_government        | 1.02  | 1.00   | (0.16-4.67)    |
| thai_websites               | 0.92  | 0.89   | (0.04-11.28)   |
| wikipedia                  | 0.97  | 0.97   | (0.40-2.11)    |
| paracrawl                  | 1.07  | 1.00   | (0.12-4.11)    |
| apdf                       | 0.95  | 0.94   | (0.18-2.79)    |

Table 14: Mean, median, minimum and maximum ratio of English-to-Thai tokens. We use `newmm` tokenizer from pyThaiNLP to tokenize Thai words, and NLTK to tokenize English words. Spaces are excluded from English-to-Thai ratio calculation.