Songklanakarin J. Sci. Technol.
44 (3), 852–860, May – Jun. 2022

http://www.sjst.psu.ac.th

Original Article

TranSentCut – transformer based Thai sentence segmentation

Sumeth Yuenyong[1*], and Virach Sornlertlamvanich[2, 3]

[1] *Department of Computer Engineering, Faculty of Engineering, Mahidol University, Phutthamonthon, Nakhon Pathom, 73170 Thailand*

[2] *Asia AI Institute, Faculty of Data Science, Musashino University, Tokyo, 202-8585, Japan*

[3] *Faculty of Engineering, Thammasat University, Khlong Luang, Pathum Thani, 12121 Thailand*

Received: 22 Auigust 2021; Revised: 26 January 2022; Accepted: 23 February 2022

Abstract

We propose TranSentCut, a sentence segmentation model for Thai based on the transformer architecture. Sentence segmentation for Thai is a problem because there is no end of sentence marker like in other languages. Existing methods make use of POS tags, which is not easy to label and must be done for every word in the data. This limits the the applicability and performance of sentence segmentation on open-domain text, because the only high-quality Thai corpus that has sentence boundary and POS labels was constructed mostly from academic articles. Our approach only uses raw text for training and the only labelling required is to separate each sentence into its own line in a text file. This makes new datasets much easier to construct. Comparison with existing methods show that our proposed model is competitive with the most recent state-of-the-art when evaluated on in-domain texts, and improved significantly over existing publicly available libraries when applied to out-of-domain input texts.

Keywords: sentence segmentation, natural language processing, neural network, transformer model

1. Introduction

The sentence unit is an important information to process a language text as an initial unit. Many tasks in Natural Language Processing (NLP) such as information extraction (Cowie & Lehnert, 1996) rely on being able to extract complete sentences accurately. For most languages extracting sentences from text is a trivial task due to the use of end of sentence marker. Even languages that do not have space between words such as Chinese or Japanese use end of sentence marker. However, Thai does not use any sentence marker, but instead put a space between the end of one sentence and the start of the next one. This makes sentence segmentation in Thai very ambiguous, as the space character is used for many other purposes: separating items in a list, separating clauses in the same sentences (Thai does not use comma to separate clauses), and separating ordinal number from the unit such as "1 person", for example.

The Thai NLP community has tackled the sentence segmentation problem over the years. In the early 2000's there were (Charoenpornsawat & Sornlertlamvanich, 2001; Mittrapiyanuruk & Sornlertlamvanich, 2000) that used part-of-speech (POS) tags (Voutilainen, 2003) by forming bi/tri-gram of the POS tags leading up to a space or on both sides of a space as features, which were then used to train a machine learning model whose job was to classify a space as nsb (non-sentence boundary) or sb (sentence boundary). More recently (Nararatwong, Kertkeidkachorn, Cooharojananone, & Okada, 2018; Zhou, Aw, Lertcheva, & Wang, 2016) incorporated conditional random field (CRF), a technique invented for sequence labelling (Lafferty, McCallum, & Pereira, 2001). Using CRF allowed one to model the probabilistic transition between the current POS tag and the next one. This recursion then enabled the context (POS tags on either side) of a space in question to extend further than a few words on both sides. CRF also allowed for the possibility of inserting explicit rules,

*Corresponding author
Email address: sumeth.yue@mahidol.edu

such as "do not break the sentence between a number and a unit", into the model by defining these rules as feature functions. The most popular Thai NLP library PyThaiNLP uses CRF as the default engine for sentence segmentation. In (Zhou *et al.*, 2018) the authors proposed solving both POS tagging and sentence segmentation as the same problem by considering the space character as just a normal character that can be assigned the <SB> or <NSB> POS tags. They also used Factorial CRF (Wu, Lian, & Hsu, 2007) which models the connection between different layers in a multi-layered CRF chain in addition to the temporal connections found in standard (linear-chain) CRF. In (Nararatwong *et al*., 2018) the authors focused on improving the performance of word and sentence segmentation where compound words are involved. Compound words can be incorrectly POS tagged, causing problems for any models that use POS tags. They addressed this problem by proposing a word merging dictionary through which compound words can be separated into their individual parts and tagged correctly.

In recent years, due to the success of Deep Learning (LeCun, Bengio, & Hinton, 2015), many researchers proposed improvements over existing methods by applying deep learning models. In (Saetia, Chuangsuwanich, Chalothorn, & Vateekul, 2019) authors proposed adding n-gram embedding, an idea made possible by word2vec (Mikolov, Chen, Corrado, & Dean, 2013), to the Bidirectional LSTM-CRF model (Huang, Xu, & Yu, 2015), and incorporating attention mechanism (Vaswani *et al*., 2017) in order to model the long term dependency for words far away from the space under consideration.

While the performance of the latest Thai sentence segmentation algorithms are already outstanding, every one of them rely on training data with POS tags. The ORCHID corpus (Charoenporn, Sornlertlamvanich, & Isahara, 1997; Sornlertlamvanich, Charoenporn, & Isahara, 1997) is an excellent Thai text corpus that have labels both for POS tags as well as word/sentence boundaries. However, constructing such as corpus was very time-consuming and required special expertise. ORCHID uses a system of over 20 different POS tags, as such, labeling text in such system is a difficult task in itself. Moreover, every single word in the corpus must be labelled, not just the sentence boundaries. This is a disadvantage because ORCHID consists of mostly technical/ academic articles, where the language is very specific. Any model trained on it will face out-of-domain inputs when applied to open-domain texts, and not being able to easily construct new training data for other domains of text, due the difficulty in labelling, limits the applicability of any sentence segmentation methods "in the wild".

In order to overcome this limitation and inspired by the recent success of the transformer architecture (Vaswani *et al*., 2017) in NLP, in this paper we proposed a Thai sentence segmentation method based on a derivative of BERT (Devlin, Chang, Lee, & Toutanova, 2018) called RoBERTa (Liu *et al.*, 2019). The idea is simple: the model receives a pair of sequences as input. Sequence A is everything to the left of a space to be decided as sb/nsb, and similarity sequence B is everything to the right, up to the maximum length of the model (512 tokens), or a lower prescribed limit, or the beginning/end of a paragraph. The sequences are in raw text without the need for any word tokenization. POS tags are also not needed. The task of the model is binary classification between sb/nsb, which is repeated for each space character is the text. We release our code on GitHub (https://github.com/ sumethy/TranSentCut) In Section 2, we describe our proposed method for sentence segmentation of the Thai text. We discuss on the experiment results in Section 3 by evaluating against the existing approaches, and show the results of the class weight adjustment for precise evaluation and fine-tuning of the context length. Finally, we come up with the Section of conclusion and some samples of the sentence segmentation.

2. Proposed Method

Transformers models are usually "pretrained" in a self-supervised manner on a large text corpus and then finetuned for a specific problem. The pretraining task is usually a language modelling task, here the model is asked to predict the next word for the GPT (Brown *et al*., 2020) family of models, or to predict the masked words in what is called the masked language model (MLM, Figure 6) task for the BERT family. Additionally, the pretraining task may include some sort of sentence-level task such as predicting whether sentence B should follow sentence A, called the next sentence prediction in BERT. This is not ideal for Thai since we are trying to solve sentence segmentation in the first place. However the RoBERTa model uses only the MLM task and no sentence-level task for pretraining, making it ideal for use with Thai. Recently a model called WangchanBERTa was released by (Lowphansirikul, Polpanumas, Jantrakulchai, & Nutanong, 2021), pretrained on approximately 70 GB of text, the largest publicly available pretrained transformer model for Thai. WangchanBERTa is identical in structure to the RoBERTa model, with the difference being the training data. RoBERTa itself is identical in structure to BERT, with the difference being the training lost. BERT uses next sentence prediction task as part of the lost, while RoBERTa only uses the MLM lost. This means that WanchanBERTa is basically BERT trained on Thai data without next sentence prediction lost. In particular, its structure is BERT-base with 12 layers, 768 hidden size, 12 attention heads, and a vocabulary size of 25,002. The number of weights is approximately 110 million. The maximum input length is 512 tokens. An input string can be separated into input A and input B by inserting the special <sep> token between the two inputs.

We parsed the ORCHID corpus, which is given in XML file, into a text file which has the following structure: each line is a complete sentence, and paragraphs/documents are separated by one blank line. We did not consider the pairs between a last sentence in a paragraph and the first sentence in the next paragraph. That is, we assume that the model will only work on one paragraph at a time. Paragraphs segmentation is a trivial matter with the newline character.

We implemented the training of the model in Pytorch (Paszke *et al*., 2019) and the Huggingface library (Wolf *et al*., 2019). The released pretrained WangchanBERTa model is available on the Huggingface Model Hub. An input training example to the model looks like the following: *<s>sequenceA</s>sequenceB</s>* where <s> and </s> are special token used by the model. <s> denote the beginning of input and </s> acts as both the separator between two sequences and to denote the end of input. Figure 1 and 2 show the flowcharts of our proposed method. As an example of the input that the model sees, see Figure 3 and 4, where the

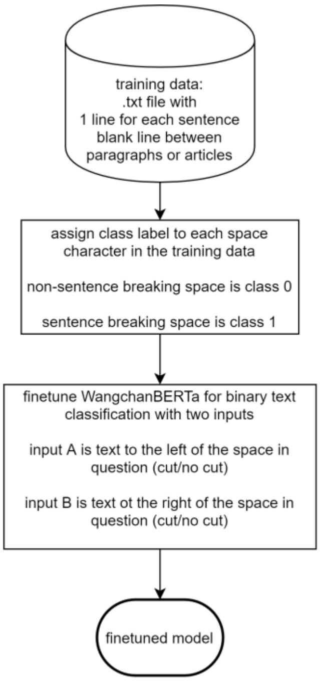

Figure 1. The flowchart of our proposed method, during training phase

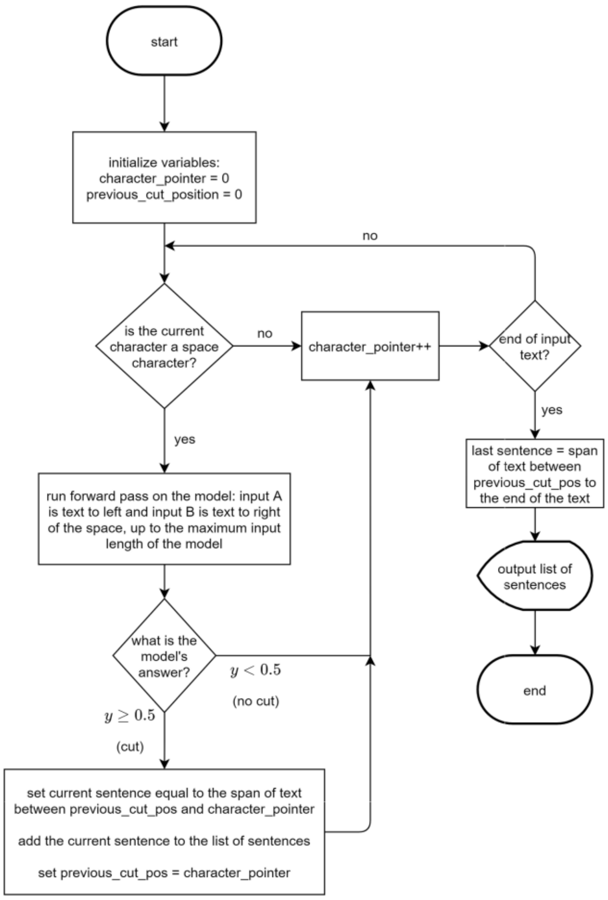

Figure 2. The flowchart of our proposed method, during inference

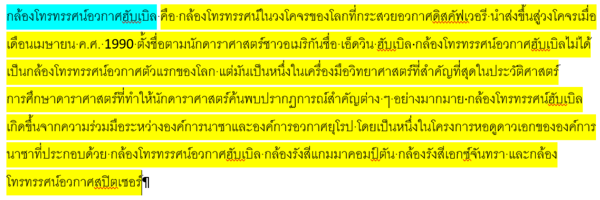
กล้องโทรทรรศน์อวกาศฮับเบิล คือ กล้องโทรทรรศน์ในวงโคจรของโลกที่กระสวยอวกาศดิสคัฟเวอรี นำส่งขึ้นสู่วงโคจรเมื่อ
เดือนเมษายน ค.ศ. 1990 ตั้งชื่อตามนักดาราศาสตร์ชาวอเมริกันชื่อ เอ็ดวิน ฮับเบิล กล้องโทรทรรศน์อวกาศฮับเบิลไม่ได้
เป็นกล้องโทรทรรศน์อวกาศตัวแรกของโลก แต่มันเป็นหนึ่งในเครื่องมือวิทยาศาสตร์ที่สำคัญที่สุดในประวัติศาสตร์
การศึกษาดาราศาสตร์ที่ทำให้นักดาราศาสตร์ค้นพบปรากฏการณ์สำคัญต่าง ๆ อย่างมากมาย กล้องโทรทรรศน์ฮับเบิล
เกิดขึ้นจากความร่วมมือระหว่างองค์การนาซาและองค์การอวกาศยุโรป โดยเป็นหนึ่งในโครงการหอดูดาวเอกขององค์การ
นาซาที่ประกอบด้วย กล้องโทรทรรศน์อวกาศฮับเบิล กล้องรังสีแกมมาคอมป์ตัน กล้องรังสีเอกซ์จันทรา และกล้อง
โทรทรรศน์อวกาศสปิตเซอร์¶

Figure 3. An example of text presented to the model, the cyan hi-lighted (darker) part is sequence A and the yellow hi-lighted (lighter) part is sequence B. Note that there is a space between the cyan part and the yellow part. This space is an nsb (non-sentence boundary). The visible dot between the cyan part and the yellow part is from MS word, not the text itself. Translation of this paragraph is in Appendix B.

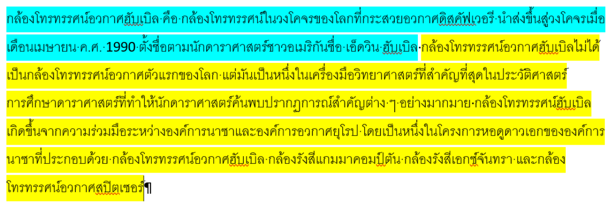
กล้องโทรทรรศน์อวกาศฮับเบิล คือ กล้องโทรทรรศน์ในวงโคจรของโลกที่กระสวยอวกาศดิสคัฟเวอรี นำส่งขึ้นสู่วงโคจรเมื่อ
เดือนเมษายน ค.ศ. 1990 ตั้งชื่อตามนักดาราศาสตร์ชาวอเมริกันชื่อ เอ็ดวิน ฮับเบิล กล้องโทรทรรศน์อวกาศฮับเบิลไม่ได้
เป็นกล้องโทรทรรศน์อวกาศตัวแรกของโลก แต่มันเป็นหนึ่งในเครื่องมือวิทยาศาสตร์ที่สำคัญที่สุดในประวัติศาสตร์
การศึกษาดาราศาสตร์ที่ทำให้นักดาราศาสตร์ค้นพบปรากฏการณ์สำคัญต่าง ๆ อย่างมากมาย กล้องโทรทรรศน์ฮับเบิล
เกิดขึ้นจากความร่วมมือระหว่างองค์การนาซาและองค์การอวกาศยุโรป โดยเป็นหนึ่งในโครงการหอดูดาวเอกขององค์การ
นาซาที่ประกอบด้วย กล้องโทรทรรศน์อวกาศฮับเบิล กล้องรังสีแกมมาคอมป์ตัน กล้องรังสีเอกซ์จันทรา และกล้อง
โทรทรรศน์อวกาศสปิตเซอร์¶

Figure 4. The same paragraph as in Figure 1 but now the space under consideration is a different one. The sequences A and B with respect to this space is hi-lighted using the same color code as in the previous figure. This space is an sb (sentence boundary). Translation of this paragraph is in Appendix B.

paragraph was taken from a Thai Wikipedia article about the Hubble Space Telescope. Figure 5 illustrates how the input is fed into the TranSentCut model. Each space character in the input string yields one input to the model.

It can be seen from Figures 3 and 4 that using a transformer model with a maximum input length of 512 tokens allows for the context to become very long, spanning an entire paragraph. One could argue that it can even be too long, a word very far away from the space under consideration probably does not influence whether it is sb or nsb. As will be shown in the ablation study, above a certain length making the context longer does not help. However, the optimal context length is still well over 100 tokens long, demonstrating that deciding between sb/nsb does benefit from having longer context information. This is a strong argument for the use of the transformer architecture.

3. Experiments

Going through the entire ORCHID corpus in a manner described in the previous section, there were 79137 examples of nsb spaces and 13384 examples of sb space. The imbalance is by the nature of the problem. In the ablation study we show the results of different ways of dealing with the imbalance. Here we state the best result which was obtained using the following set of hyper-parameters: context length = 256 tokens, number of epochs = 20, seed = 12345, batch size = 64, weight decay = 9.51207×10^{-5}, learning rate = 4.05813×10^{-5} and class weight strategy 2. The different strategies for assigning weight to each class will be discussed in the ablation study below. The weight decay and learning rate were taken from hyper-parameter optimization on another

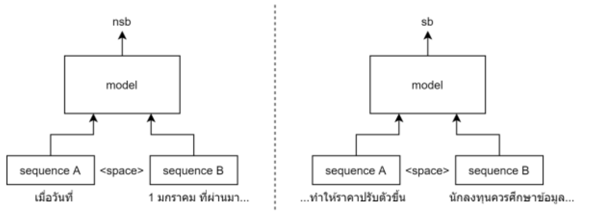

Figure 5. Illustration of how we apply the transformer model to solve sentence segmentation. Transformer model can accept one or two sequences as an input. The two-sequences input is used for the tasks such as next sentence prediction or questions answering, and can be applied to sentence segmentation.

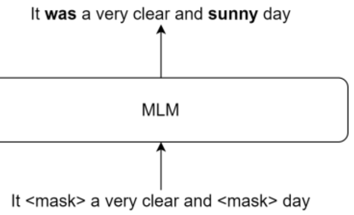

Figure 6. Illustration of the MLM task. The tokens "<mask>" are hidden from the model during pretraining, the model job is to predict them from a set of all possible tokens in the vocabulary.

Thai text classification problem using the same model architecture. The same seed was used for both splitting the data into train/test, shuffling the data and initializing the model, ensuring that the training is perfectly repeatable given the same hyper-parameters. Comparison between our results with the numbers stated in crfcut (the sentence segmentation engine for PyThaiNLP), on the ORCHID dataset, we have the result in Table 1.

The prefix I and E in Table 1 denote "inside sentence" and "end of sentence" respectively, corresponding to our notation of nsb and sb, respectively. The metric space-correct (sc) is just the overall classification accuracy, which is given by sc = (#correct sb+#correct nsb)/(total # of space tokens). These metrics were introduced in (Mittrapiyanuruk & Sornlertlamvanich, 2000). While we did not achieve higher number for every single metric, we made large gains on E-recall, E-fscore and space-correct, while maintaining within around 2% of the other metrics. Taking the macro average of I-fscore and E-fscore, we got 0.9296 vs. 0.8800 for crfcut. And comparing our results with the ORCHID part of Table 3 in (Saetia *et al*., 2019), which is the most recent and similar to this work, their macro average fscore as reported was 0.9250.

3.1 Performance on out of domain data

In order to test the performance of sentence segmentation on out-of-domain data, we constructed a small test set consisting of paragraphs from news articles. We choose only recent articles to make sure that they were not part of the training data of any model. The articles were about Covid-19 and the 2021 Olympics, so it is certain that they did not not exist in, or were similar to ORCHID in any way. When constructing the test set, if the taggers cannot reach an agreement whether a space is nsb or sb, one possible way to reach a decision was to translated the text surrounding the space under consideration in Google Translate and put the sb in the same place as in the English translation. We acknowledge that this is not theoretically rigorous, however it was used very sparingly since the taggers usually were able to discuss and reach a decision. Figures 7 and 8 compare an excerpt of this new test data vs. an excerpt from ORCHID, respectively. It can be seen that, at least for the purpose of sentence segmentation, the data for our model which does not require POS tags is much easier to label than having to label POS tag for each word.

In total, our new test data consists of 104 sentences, with 782 nsb and 84 sb spaces. The number of sb spaces is less than the number of sentences because we look at only one paragraph at a time. Running our trained model on this data, we got macro average fscore of 0.6903, while crfcut and thai-segmentor got 0.6271 and 0.6283 respectively. These are the only two methods that we can actually run our own comparison against, since they are the only ones with openly available libraries. The results demonstrate that our model can generalize better to out-of-domain input. Examples of segmentation results are given in appendix A. Table 2 shows the classification performance of crfcut, thai-segmentor and TranSentCut on our new test dataset.

```
1  กรุงโตเกียวได้รับเกียรติเป็นเจ้าภาพกีฬาโอลิมปิก เมื่อวันที่ 7 กันยายน พ.ศ. 2556
   ในประชุมคณะกรรมการโอลิมปิกสากล สมัยที่ 123 ณ กรุงบัวโนสไอเรส ประเทศอาร์เจนตินา
2  นับเป็นครั้งที่ 3 ที่กรุงโตเกียวได้รับสิทธิ์เป็นเจ้าภาพโอลิมปิก ครั้งแรกเมื่อ ค.ศ.
   1940 ได้รับสิทธิ์เป็นเจ้าภาพโอลิมปิกฤดูร้อนครั้งแรกของทวีปเอเชีย
   และเมืองซัปโปโรสำหรับโอลิมปิกฤดูหนาว
   แต่ได้ถอนตัวจากการแข่งขันเนื่องจากสงครามระหว่างจีนและญี่ปุ่น
   และกลับมาเป็นเจ้าภาพอีกครั้งในกีฬาโอลิมปิกฤดูร้อน 1964 (พ.ศ. 2507)
3  ซึ่งครั้งนี้ กรุงโตเกียวเป็นเมืองที่ 5 (และเมืองที่ 1 ในทวีปเอเชีย)
   ที่ได้จัดการแข่งขันกีฬาโอลิมปิกฤดูร้อนมากกว่า 1 ครั้ง
   รวมถึงกรุงโตเกียวก็ได้รับเกียรติเป็นเจ้าภาพกีฬาพาราลิมปิกฤดูร้อน 2020
   สำหรับนักกีฬาคนพิการเช่นกัน

4
```

Figure 7. One paragraph excerpt from the new sentence segmentation test data that we constructed. Each sentence is one line, note the line numbers on the left margin. Paragraphs are separated by a blank line (line 4).


```
<corpus>
<document TPublisher="ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, กระทรวงวิทยาศาสตร์ เทคโนโลยีและการพ
<paragraph id="1" line_num="12">
<sentence id="1" line_num = "13" raw_txt = "การประชุมทางวิชาการ ครั้งที่ 1">
<word surface="การ" pos="FIXN"/>
<word surface="ประชุม" pos="VACT"/>
<word surface="ทาง" pos="NCMN"/>
<word surface="วิชาการ" pos="NCMN"/>
<word surface="&lt;space&gt;" pos="PUNC"/>
<word surface="ครั้ง" pos="CFQC"/>
<word surface="ที่ 1" pos="DONM"/>
</sentence>
<sentence id="2" line_num = "23" raw_txt = "โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์">
<word surface="โครงการวิจัยและพัฒนา" pos="NCMN"/>
<word surface="อิเล็กทรอนิกส์" pos="NCMN"/>
<word surface="และ" pos="JCRG"/>
<word surface="คอมพิวเตอร์" pos="NCMN"/>
</sentence>
```


Figure 8. The first paragraph and the first two sentences of ORCHID. Note that every word has a POS tag.

Table 1. Comparison between TranSentCut and crfcut on ORCHID data

	I-precision	I-recall	I-fscore	E-precision	E-recall	E-fscore	Space-correct
crfcut	0.9800	0.9900	0.9900	0.8500	0.7100	0.7700	0.8700
TranSentCut	0.9860	0.9697	0.9778	0.8354	0.9175	0.8746	0.9622

Table 2. The classification performance of crfcut, thai-segmentor, and TranSentCut on the new, out-of-domain test dataset

		Precision	Recall	Fscore	Support
crfcut	sb	0.2727	0.5357	0.3614	84
	nsb	0.9444	0.8465	0.8928	782
	macro avg.	0.6085	0.6911	**0.6271**	
thai-segmentor	sb	0.3400	0.3148	0.3269	84
	nsb	0.9260	0.9335	0.9297	782
	macro avg.	0.6330	0.6241	**0.6283**	
TranSentCut	sb	0.3362	0.9285	0.4937	84
	nsb	0.9905	0.8031	0.8870	782
	macro avg.	0.6634	0.8658	**0.6903**	

3.2 Ablation study

Like most machine learning problems, sentence segmentation suffers from imbalance data. There are many nsb than there are sb in any piece of text. The article (Chawla, Japkowicz, & Kotcz, 2004) outlines different approaches to deal with imbalance data, such as class weight, under-sampling, using ensembles, and one-class classification. Since nb vs. nsb is not highly imbalanced (the class ratio is only about 6:1), we investigated two approaches in this study: making the data balanced by discarding examples from the majority class until the data is balanced. This is the under-sampling approach. The other approach was adding class weights to the loss function during training, which is the class weight approach.

In the under-sampling approach, we put all the examples of the nsb class in a list, shuffled that list (after the seed had been set, so each run got exactly the same data), and then keeping only the first *n* elements of the list, where *n* is the number of sb examples. This was done before the usual train/test split, so both the training and test data were balanced. The model was then trained with the standard cross-entropy loss.

For the class weight approach, we investigated three strategies for assigning the class weights. To illustrate them, note that class nsb has 79,137 examples and class sb has 13,384 examples. Strategy 0 (the naive strategy) was to simply assign the majority class a weight of 1, and the weight of the minority class was the ratio between the two classes. That is, class sb (minority) gets a weight of *79137/13384 = 5.9128*, while class nsb (majority) gets a weight of 1. Strategy 1 was to use the scikit-learn (Pedregosa *et al*., 2011) library's *compute_class_weight* function, which assign class weights according to the following formula for each class *i*

$$w_i = n_{samples} / (n_{classes} * count(i))$$

where *count* is the function that counts the number of examples of class *i*. Using this formula, the weights for nsb and sb classes were as follows

$$w_{nsb} = 92521 / (2 * 79137) = 0.5845$$

and

$$w_{sb} = 92521 / (2 * 13384) = 3.456$$

Finally, strategy 2 was to ensure that the maximum weight is 1, and to assign the smaller weight to preserve the class ratio. In this strategy, class sb (minority) gets the weight value of 1, while class nsb (majority) get the weight of 0.1691. Another way to think about strategy 2, is that it's simply a normalized version of strategy 0, as in *[1,5.9128]/5.9128 = [0.1691,1]*. Note that the ratio between the two weights remains the same, the main difference from the strategy 0 is that the maximum weight is 1, ensuring that the magnitude of the loss function is not amplified. This strategy can be extended to number of classes ≥ 3 by assigning the smallest class a weight of 1, give each of the other classes weight according to its ratio to the smallest class, then dividing all the weights by the largest weight.

For this round of experiments, we trained the model on ORCHID using the same configuration as reported in the beginning of section 3. As is common practice in training deep neural networks, an early stopping policy was enforced. If the model did not improve on the validation fscore after 5 consecutive validation rounds, the training was stopped. Validation was performed every 200 iterations. Figure 9 shows the validation macro average fscore curve for the balanced case, and the difference class weight strategies. While the figure suggest that balanced training is the best, we evaluate the trained models on the out-of-domain test data and show that this was not the case. The result is shown in Table 3. It can be seen that while balanced training seems to have the best performance on the ORCHID data, it was not able to adapt to out-of-domain data as well as class weight training strategy 1 and 2. This is because the actual data distribution when the model is deployed is imbalanced, and having been exposed to a distribution with the same characteristic during training helps the model to better adapt.

3.2.1 The effect of context length and batch size

In this section, we studied the effects of context length and the batch size. We used the exact same data split as in the previous section. All parameters were kept fixed as the ones in the beginning of Section 3, except for the one that was being tested.

The context length plays a key role in the performance of the model. If the length is too short, the model might not have enough information to make a good decision. On the other hand, if the context is too long, the extra tokens that do not help are basically noise that the model must learn to assign low attention weights to. Even if the model can do this, having a context length that is too long means a bigger model that takes longer for both training and inference. Therefore, it is important to find the right context length. For this purpose, we compared different context lengths: 32, 64, 96, 128, 256, and 504 (The maximum length of the model is 512, but some tokens must be reserved for the special tokens, so we took the next lower multiple of 8.). The other parameters of the model were fixed as the same as those in

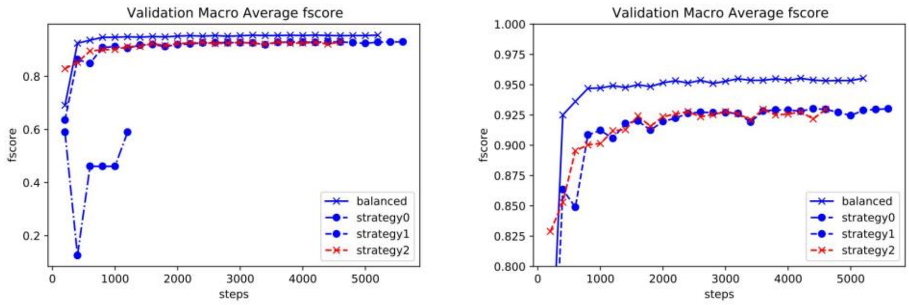

Figure 9. The macro average fscore validation curve for balanced data training, and the three class weight strategies. The right panel is the zoomed in version of the left panel. Validation was performed every 200 iterations, not including the beginning of training, so the curves do not start from 0 on the x-axis. The curve for strategy 0 shows that training was not very successful and was terminated early by the early stopping policy. Note that the "did not improve anymore" portion of the curves was not recorded by the training loop. Had it been included, the bottom curve would not look like it was still going up. The curve for balanced training seems to be the best, but it did not perform very well when the trained model was applied to out-of-domain test data.

Table 3. Classification performance comparison on the out-of-domain data between balanced training and different class weight strategies. Note that the performance of the balanced training did not beat crfcut and thai-segmentor from table 2 and that naive (strategy 0) class weighting performed very poorly. Our strategy 2 was able to beat strategy 1 from the scikit-learn library.

		Precision	Recall	Fscore	Support
	sb	0.2606	0.9524	0.4092	0.2606
balanced	nsb	0.9928	0.7097	0.8277	0.9928
	macro avg.	0.6267	0.8310	0.6185	0.6267
	sb	0.1005	0.2619	0.1452	0.1005
strategy 0	nsb	0.9042	0.7481	0.8188	0.9042
	macro avg.	0.5023	0.5050	0.4820	0.5023
	sb	0.3290	0.9048	0.4825	0.3290
strategy 1	nsb	0.9874	0.8018	0.8850	0.9874
	macro avg.	0.6582	0.8533	0.6838	0.6582
	sb	0.3362	0.9285	0.4937	0.3362
strategy 2	nsb	0.9905	0.8031	0.8870	0.9905
	macro avg.	0.6634	0.8658	0.6903	0.6634

strategy 2 in the previous section. Figure 10 shows the result of this experiment. The best context length was 256. Furthermore, in order to find the best context length in more detail, we tested several values for context length from 220 to 300, the result is shown in Table 4. It is shown in table from because the values are very close to each other. Context length of 280 gives the best fscore result, and there was no shorter context length that had better performance than 256. For longer context length, the improvement over the standard (a power of 2) length of 256 is not very large, and increasing the length beyond 280 seems to offer no further improvement. In practice, one might choose to use length 256 during deployment due to the computational advantage on the GPU by using a length that is a power of 2.

For the batch size, we tested batch sizes of 16, 32, and 64. Batch size of 64 was the maximum batch size possible for the context length of 256 and for the GPU that we have. It can be seen in Figure 11 that batch sizes of 32 and 64 performed about the same, with 64 being slightly better. The batch size of 16 was too low and was stopped very early. This confirms that one should use the largest batch size possible without exceeding the GPU memory.

4. Conclusions

We presented a new sentence segmentation model for Thai. The main advantage of our models compared to existing methods is that the training data does not need to be POS tagged, allowing new datasets to be constructed easily without needing special expertise. The model performance is competitive. Comparison with existing libraries shows that our model has higher macro average fscore of about 0.04 and 0.06 on ORCHID corpus and on out-of-domain texts, respectively. Comparing with the most recent research that also uses the transformer architecture. We got approximately

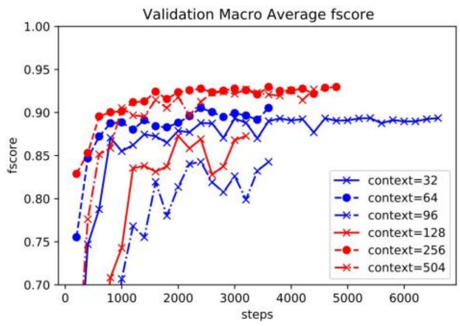

Figure 10. The validation fscore of different context lengths. Models with length-32 and length-64 performed better than both length-96 and length-128. However, length-256 and length-504 were both clearly better than all the lower length ones. There was very slight difference between length-256 (fscore=0.9296) and length-504 (fscore= 0.9268). Overall, length-256 was the best context length.

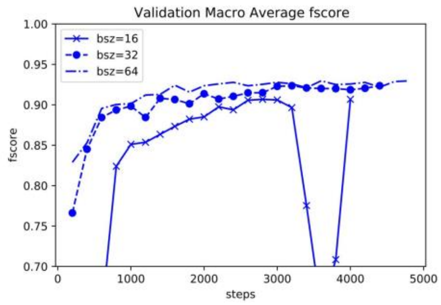

Figure 11. The validation fscore of different batch sizes. The best batch size was 64.

Context length	Maximum validation fscore
220	0.8944
240	0.9290
256	0.9296
260	0.9304
280	0.9331
300	0.9303

Table 4. Classification Additional experiments for determining the optimal context length. Context length of 280 gives the best fscore result. However, the improvement over the standard (a power of 2) length of 256 is not very large. Increasing the length beyond 280 seems to offer no further improvement. In practice, one might choose to use length 256 during deployment due to the computational advantage on the GPU by using a length that is a power of 2.

the same fscore as reported in the paper, but without needing POS tags for training. We release the code and the trained model.

Acknowledgements

The authors gratefully acknowledge the financial support provided by Thammasat University Research fund under the TSRI, Contract No. TUFF19/2564, and TUFF24/2565 for the project of “AI Ready City Networking in RUN”, based on the RUN Digital Cluster collaboration scheme.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. Retrieved from https://arxiv.org/abs/2005.14165

Charoenporn, T., Sornlertlamvanich, V., & Isahara, H. (1997). Building a large Thai text corpus-part-of-speech tagged corpus: Orchid. *Proceedings of the Natural Language Processing Pacific Rim Symposium*, 509-512.

Charoenpornsawat, P., & Sornlertlamvanich, V. (2001). Automatic sentence break disambiguation for Thai. *International Conference on Computer Processing of Oriental Languages* 33, 231-235.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, *6*(1), 1-6.

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM*, *39*(1), 80-91.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: 1810.04805*. Retrieved from https://arxiv.org/abs/1810.04805

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*. Retrieved from https://arxiv.org/abs/1508.01991?context=cs#:~:text=The%20BI%2DLSTM%2DCRF%20model%20can%20produce%20state%20of%20the,as%20compared%20to%20previous%20observations.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, 282–289.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. Retrieved from https://openreview.net/forum?id=SyxS0T4tvS

Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., & Nutanong, S. (2021). WangchanBERTa: Pretraining transformer-based Thai Language Models. *arXiv preprint arXiv:2101.09635*. Retrieved from https://airesearch.in.th/releases/wangchanberta-pre-trained-thai-language-model/

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. Retrieved from https://arxiv.org/abs/1301.3781

Mittrapiyanuruk, P., & Sornlertlamvanich, V. (2000). The automatic Thai sentence extraction. *Proceedings of the Fourth Symposium on Natural Language Processing*, 23-28.

Nararatwong, R., Kertkeidkachorn, N., Cooharojananone, N., & Okada, H. (2018). Improving Thai word and sentence segmentation using linguistic knowledge. *IEICE TRANSACTIONS on Information and Systems*, *101*(12), 3218-3225.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*, 8026-8037.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Saetia, C., Chuangsuwanich, E., Chalothorn, T., & Vateekul, P. (2019). Semi-supervised Thai Sentence segmentation using local and distant word representations. *arXiv preprint arXiv:1908.01294*. Retrieved from https://arxiv.org/abs/1908.01294

Sornlertlamvanich, V., Charoenporn, T., & Isahara, H. (1997). ORCHID: Thai part-of-speech tagged corpus.

National Electronics and Computer Technology Center Technical Report, 5-19.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998-6008.

Voutilainen, A. (2003). Part-of-speech tagging. *The Oxford Handbook of Computational Linguistics*, 219-232. doi:10.1093/oxfordhb/9780199276349.013.0011

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*. Retrieved from https://arxiv.org/abs/1910.03771

Wu, T. Y., Lian, C. C., & Hsu, J. Y. J. (2007). Joint recognition of multiple concurrent activities using factorial conditional random fields. *Twenty-Second AAAI Conference on Artificial Intelligence*, 82-87.

Zhou, N., Aw, A., Lertcheva, N., & Wang, X. (2016). A word labeling approach to Thai sentence boundary detection and POS tagging. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 319-327.

Appendix

Appendix A

In this section we show the result of sentence segmentation on some out-of-domain data. The input to the model is entire paragraphs as one long string. The output for each paragraph is a list of string, where each string is one sentence. The following URLs are the sources of the paragraphs.

- Figures A1 - A2: https://th.wikipedia.org/wiki/%E0%B9%82%E0%B8%AD%E0%B8%A5%E0%B8%B4%E0%B8%A1%E0%B8%9B%E0%B8%B4%E0%B8%81%E0%B8%A4%E0%B8%94%E0%B8%B9%E0%B8%A3%E0%B9%89%E0%B8%AD%E0%B8%99_2020
- Figure A3: https://www.khaosod.co.th/sports/news_6545105
- Figure A4: https://www.khaosod.co.th/special-stories/news_6546164
- Figure A5: https://www.voathai.com/a/us-covid19-delta-variant-fauci-mask-cdc-directives-republican-governors/5986866.html

```
Input Paragraph:

กรุงโตเกียวได้รับเกียรติเป็นเจ้าภาพกีฬาโอลิมปิก เมื่อวันที่ 7 กันยายน พ.ศ. 2556
ในประชุมคณะกรรมการโอลิมปิกสากล สมัยที่ 123 ณ กรุงบัวโนสไอเรส ประเทศอาร์เจนตินา นับเป็นครั้งที่ 3
ที่กรุงโตเกียวได้รับสิทธิ์เป็นเจ้าภาพโอลิมปิก ครั้งแรกเมื่อ ค.ศ. 1940
ได้รับสิทธิ์เป็นเจ้าภาพโอลิมปิกฤดูร้อนครั้งแรกของทวีปเอเชีย และเมืองซัปโปโรสำหรับโอลิมปิกฤดูหนาว
แต่ได้ถอนตัวจากการแข่งขันเนื่องจากสงครามระหว่างจีนและญี่ปุ่น
และกลับมาเป็นเจ้าภาพอีกครั้งในกีฬาโอลิมปิกฤดูร้อน 1964 (พ.ศ. 2507) ซึ่งครั้งนี้
กรุงโตเกียวเป็นเมืองที่ 5 (และเมืองที่ 1 ในทวีปเอเชีย) ที่ได้จัดการแข่งขันกีฬาโอลิมปิกฤดูร้อนมากกว่า 1
ครั้ง รวมถึงกรุงโตเกียวก็ได้รับเกียรติเป็นเจ้าภาพกีฬาพาราลิมปิกฤดูร้อน 2020 สำหรับนักกีฬาคนพิการเช่นกัน

Segmentation Result:

['กรุงโตเกียวได้รับเกียรติเป็นเจ้าภาพกีฬาโอลิมปิก เมื่อวันที่ 7 กันยายน พ.ศ. 2556
ในประชุมคณะกรรมการโอลิมปิกสากล สมัยที่ 123 ณ กรุงบัวโนสไอเรส ประเทศอาร์เจนตินา', 'นับเป็นครั้งที่ 3
ที่กรุงโตเกียวได้รับสิทธิ์เป็นเจ้าภาพโอลิมปิก ครั้งแรกเมื่อ ค.ศ. 1940',
'ได้รับสิทธิ์เป็นเจ้าภาพโอลิมปิกฤดูร้อนครั้งแรกของทวีปเอเชีย และเมืองซัปโปโรสำหรับโอลิมปิกฤดูหนาว',
'แต่ได้ถอนตัวจากการแข่งขันเนื่องจากสงครามระหว่างจีนและญี่ปุ่น',
'และกลับมาเป็นเจ้าภาพอีกครั้งในกีฬาโอลิมปิกฤดูร้อน 1964 (พ.ศ. 2507)', 'ซึ่งครั้งนี้
กรุงโตเกียวเป็นเมืองที่ 5 (และเมืองที่ 1 ในทวีปเอเชีย) ที่ได้จัดการแข่งขันกีฬาโอลิมปิกฤดูร้อนมากกว่า 1
ครั้ง', 'รวมถึงกรุงโตเกียวก็ได้รับเกียรติเป็นเจ้าภาพกีฬาพาราลิมปิกฤดูร้อน 2020
สำหรับนักกีฬาคนพิการเช่นกัน']
```

Figure A1. Segmentation example one

```
Input Paragraph:

ในวันที่ 24 มีนาคม พ.ศ. 2563 จากสถานการณ์การระบาดทั่วของโคโรนาไวรัสในทั่วโลก
ทำให้คณะกรรมการโอลิมปิกสากล (IOC) โดยโทมัส บัค ประธานคณะกรรมการโอลิมปิกสากล ได้ปรึกษาหารือกับชินโซ
อาเบะ นายกรัฐมนตรีของประเทศญี่ปุ่น
ก่อนจะตัดสินใจร่วมกันในการเลื่อนการแข่งขันโอลิมปิกและพาราลิมปิกฤดูร้อนออกไปในปี พ.ศ. 2564
และออกแถลงการณ์ยืนยันเลื่อนจัดการแข่งขันโอลิมปิก 2020 และพาราลิมปิก 2020 ที่กรุงโตเกียว ประเทศญี่ปุ่น
ออกไปเป็นเวลา 1 ปี อย่างเป็นทางการแต่ไม่ช้ากว่าปี พ.ศ. 2564
เพื่อความปลอดภัยของนักกีฬาและทุกฝ่ายที่เกี่ยวข้องกับโอลิมปิกและประชาคมโลก แต่ยังคงเป็นชื่อเดิม คือ
โตเกียว 2020 ต่อไป

Segmentation Result:

['ในวันที่ 24 มีนาคม พ.ศ. 2563 จากสถานการณ์การระบาดทั่วของโคโรนาไวรัสในทั่วโลก
ทำให้คณะกรรมการโอลิมปิกสากล (IOC) โดยโทมัส บัค ประธานคณะกรรมการโอลิมปิกสากล ได้ปรึกษาหารือกับชินโซ
อาเบะ นายกรัฐมนตรีของประเทศญี่ปุ่น',
'ก่อนจะตัดสินใจร่วมกันในการเลื่อนการแข่งขันโอลิมปิกและพาราลิมปิกฤดูร้อนออกไปในปี พ.ศ. 2564',
'และออกแถลงการณ์ยืนยันเลื่อนจัดการแข่งขันโอลิมปิก 2020 และพาราลิมปิก 2020 ที่กรุงโตเกียว ประเทศญี่ปุ่น
ออกไปเป็นเวลา 1 ปี อย่างเป็นทางการแต่ไม่ช้ากว่าปี พ.ศ. 2564
เพื่อความปลอดภัยของนักกีฬาและทุกฝ่ายที่เกี่ยวข้องกับโอลิมปิกและประชาคมโลก', 'แต่ยังคงเป็นชื่อเดิม คือ
โตเกียว 2020 ต่อไป']
```

Figure A2. Segmentation example two

```
Input Paragraph:

ฟาบินโญ กองกลางคนสำคัญขยายสัญญากับ ลิเวอร์พูล ออกไปจนถึงปี 2026 เป็นที่เรียบร้อย สำหรับ ฟาบินโญ ย้ายจาก โมนาโก
มาอยู่กับ ลิเวอร์พูล เมื่อปี 2018 ซึ่งเจ้าตัวก็ถือเป็นกำลังสำคัญของทีมมาตลอดหลังลงสนามไปทั้งหมด 122 เกมทำได้ 3
ประตูกับ 7 แอสซิสต์ นอกจากนั้นดาวเตะทีมชาติบราซิล ยังช่วย “หงส์แดง” กวาดแชมป์มาครองถึง 4 รายการ ประกอบด้วย
ยูฟ่า แชมเปียนส์ ลีก, ยูฟ่า ซูเปอร์ คัพ, พรีเมียร์ลีก อังกฤษ และฟีฟ่า คลับ เวิลด์ คัพ ก่อนหน้านี้มีข่าวว่า
ลิเวอร์พูล เตรียมต่อสัญญาระยะยาวกับ ฟาบินโญ กระทั่งล่าสุด “หงส์แดง” ก็ได้แถลงว่าแข้งวัย 27
ปียืดสัญญากับทีมออกไปเป็นที่เรียบร้อย หลังก่อนหน้านี้เพิ่งขยายสัญญากับ เทรนต์ อเล็กซานเดอร์-อาร์โนลด์ ถึงปี
2025 โดย ฟาบริซิโอ โรมาโน ผู้สื่อข่าวคนดัง ระบุว่า ฟาบินโญ ต่อสัญญากับทีมดังในถิ่น แอนฟิลด์ ไปถึงเดือน มิ.ย. ปี
2026 ทั้งนี้มีรายงานเพิ่มเติมว่าผู้เล่นคนต่อไปที่จะขยายสัญญาระยะยาวกับ ลิเวอร์พูล คือ อลิสซง เบ็คเกอร์
ผู้รักษาประตูทีมชาติบราซิล

Segmentation Result:

['ฟาบินโญ กองกลางคนสำคัญขยายสัญญากับ ลิเวอร์พูล ออกไปจนถึงปี 2026 เป็นที่เรียบร้อย', 'สำหรับ ฟาบินโญ ย้ายจาก
โมนาโก มาอยู่กับ ลิเวอร์พูล เมื่อปี 2018', 'ซึ่งเจ้าตัวก็ถือเป็นกำลังสำคัญของทีมมาตลอดหลังลงสนามไปทั้งหมด 122
เกมทำได้ 3 ประตูกับ 7 แอสซิสต์', 'นอกจากนั้นดาวเตะทีมชาติบราซิล ยังช่วย “หงส์แดง” กวาดแชมป์มาครองถึง 4 รายการ
ประกอบด้วย ยูฟ่า แชมเปียนส์ ลีก, ยูฟ่า ซูเปอร์ คัพ, พรีเมียร์ลีก อังกฤษ และฟีฟ่า คลับ เวิลด์ คัพ',
'ก่อนหน้านี้มีข่าวว่า ลิเวอร์พูล เตรียมต่อสัญญาระยะยาวกับ ฟาบินโญ', 'กระทั่งล่าสุด “หงส์แดง”
ก็ได้แถลงว่าแข้งวัย 27 ปียืดสัญญากับทีมออกไปเป็นที่เรียบร้อย', 'หลังก่อนหน้านี้เพิ่งขยายสัญญากับ เทรนต์
อเล็กซานเดอร์-อาร์โนลด์ ถึงปี 2025', 'โดย ฟาบริซิโอ โรมาโน', 'ผู้สื่อข่าวคนดัง ระบุว่า ฟาบินโญ
ต่อสัญญากับทีมดังในถิ่น แอนฟิลด์ ไปถึงเดือน มิ.ย. ปี 2026',
'ทั้งนี้มีรายงานเพิ่มเติมว่าผู้เล่นคนต่อไปที่จะขยายสัญญาระยะยาวกับ ลิเวอร์พูล คือ อลิสซง เบ็คเกอร์
ผู้รักษาประตูทีมชาติบราซิล']
```

Figure A3. Segmentation example three

```
Input Paragraph:

สธ.ทยอยจัดส่งวัคซีนไฟเซอร์แล้ว ไป รพ.พื้นที่ กทม.และปริมลฑล ฉีดบูสเตอร์โดสบุคลากรด่านหน้า
และจะส่งให้ครบทุกจังหวัด พื้นที่ไหนพร้อมฉีดก่อนได้เลย เมื่อวันที่ 4 ส.ค. นพ.โอภาส การย์กวินพงศ์
อธิบดีกรมควบคุมโรค กล่าวถึงการจัดส่งวัคซีนไฟเซอร์ล็อตบริจาค 1.5 ล้านโดส ว่า วันนี้ได้เริ่มทยอยจัดส่งไปยัง รพ.ใน
กทม.และปริมณฑลแล้ว เพื่อฉีดเป็นบูสเตอร์โดสหรือเข็มที่ 3 ให้กับบุคลากรทางการแพทย์และสาธารณสุขด่านหน้า
โดยประสานกับ รพ.รามาธิบดี รพ.ศิริราช และ รพ.จุฬาลงกรณ์ ถ้ามีความพร้อมก็จัดส่งได้เลย ส่วนในต่างจังหวัดประสานกับ
รพ.ในปริมณฑล ด้าน นพ.สุระ วิเศษศักดิ์ รองปลัดกระทรวงสาธารณสุข (สธ.)
ในฐานะประธานคณะทำงานด้านบริหารจัดการการให้บริการวัคซีนไฟเซอร์ กล่าวว่า ในวันนี้เริ่มมีการกระจายวัคซีนลงไป
คาดว่าพื้นที่ กทม.และปริมณฑลจะได้รับวัคซีนก่อน ทั้งนี้ ทั่วประเทศจะได้รับวัคซีนครบทุกจังหวัด
เพื่อเร่งฉีดกระตุ้นเข็ม 3 ให้กับบุคลากรทางการแพทย์และสาธารณสุขด่านหน้า ซึ่งจะทยอยส่งไปให้ เบื้องต้นจะส่งไปจำนวน
50% ของบุคลากรที่ได้ลงชื่อตามความสมัครใจไว้ในแต่ละ รพ. คาดว่าจะมีการดีเลย์ฉีดพร้อมกันทั่วประเทศใน 2-3 วันนี้
แต่พื้นที่ไหนที่ได้รับวัคซีนแล้วก็ฉีดก่อนได้เลย “ความโปร่งใสของข้อมูลการฉีดวัคซีนไฟเซอร์ในแต่ละรพ.
สามารถเช็กรายชื่อบุคลากรที่มีความประสงค์รับวัคซีนกระตุ้นเข็มที่ 3 ได้
ส่วนประชาชนทั่วไปก็สามารถตรวจสอบข้อมูลนี้ได้เช่นกัน เนื่องจากเมื่อมีการฉีดวัคซีนทุกโดส
จะต้องมีการรายงานผลเข้ามาที่ MOPH IC ของ สธ.” นพ.สุระกล่าว

Segmentation Result:

['สธ.ทยอยจัดส่งวัคซีนไฟเซอร์แล้ว ไป รพ.พื้นที่ กทม.และปริมลฑล', 'ฉีดบูสเตอร์โดสบุคลากรด่านหน้า',
'และจะส่งให้ครบทุกจังหวัด', 'พื้นที่ไหนพร้อมฉีดก่อนได้เลย', 'เมื่อวันที่ 4 ส.ค. นพ.โอภาส การย์กวินพงศ์',
'อธิบดีกรมควบคุมโรค กล่าวถึงการจัดส่งวัคซีนไฟเซอร์ล็อตบริจาค 1.5 ล้านโดส', 'ว่า วันนี้ได้เริ่มทยอยจัดส่งไปยัง
รพ.ใน กทม.และปริมณฑลแล้ว', 'เพื่อฉีดเป็นบูสเตอร์โดสหรือเข็มที่ 3 ให้กับบุคลากรทางการแพทย์และสาธารณสุขด่านหน้า',
'โดยประสานกับ รพ.รามาธิบดี รพ.ศิริราช และ รพ.จุฬาลงกรณ์', 'ถ้ามีความพร้อมก็จัดส่งได้เลย',
'ส่วนในต่างจังหวัดประสานกับ รพ.ในปริมณฑล', 'ด้าน นพ.สุระ วิเศษศักดิ์ รองปลัดกระทรวงสาธารณสุข (สธ.)',
'ในฐานะประธานคณะทำงานด้านบริหารจัดการการให้บริการวัคซีนไฟเซอร์ กล่าวว่า ในวันนี้เริ่มมีการกระจายวัคซีนลงไป',
'คาดว่าพื้นที่ กทม.และปริมณฑลจะได้รับวัคซีนก่อน', 'ทั้งนี้ ทั่วประเทศจะได้รับวัคซีนครบทุกจังหวัด',
'เพื่อเร่งฉีดกระตุ้นเข็ม 3 ให้กับบุคลากรทางการแพทย์และสาธารณสุขด่านหน้า ซึ่งจะทยอยส่งไปให้',
'เบื้องต้นจะส่งไปจำนวน 50% ของบุคลากรที่ได้ลงชื่อตามความสมัครใจไว้ในแต่ละ รพ.',
'คาดว่าจะมีการดีเลย์ฉีดพร้อมกันทั่วประเทศใน 2-3', 'วันนี้', 'แต่พื้นที่ไหนที่ได้รับวัคซีนแล้วก็ฉีดก่อนได้เลย
“ความโปร่งใสของข้อมูลการฉีดวัคซีนไฟเซอร์ในแต่ละรพ.',
'สามารถเช็กรายชื่อบุคลากรที่มีความประสงค์รับวัคซีนกระตุ้นเข็มที่ 3 ได้',
'ส่วนประชาชนทั่วไปก็สามารถตรวจสอบข้อมูลนี้ได้เช่นกัน', 'เนื่องจากเมื่อมีการฉีดวัคซีนทุกโดส',
'จะต้องมีการรายงานผลเข้ามาที่ MOPH IC ของ สธ.”', 'นพ.สุระกล่าว']
```

Figure A4. Segmentation example four

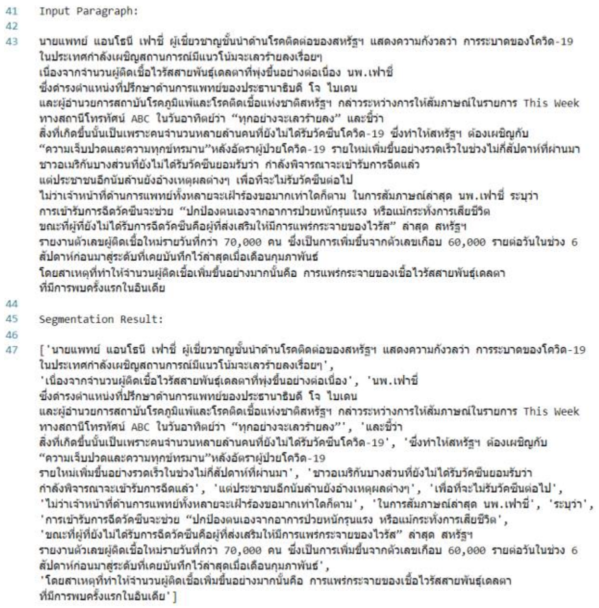

Figure A5. Segmentation example five

Appendix B

Here we provide the English translation for the examples/captions we used in the paper. The translations were chosen to be as literal as possible to preserve the structure of the Thai sentence(s).

B.1 Translation for Figures 1 and 2

"The Hubble Space Telescope is a space telescope that was launched into low Earth orbit in 1990 by the Discovery Space Shuttle. The Hubble telescope is named after astronomer Edwin Hubble. It was not the first space telescope, but is one of the most important scientific instruments in the history of Astronomy that had led to many discoveries. The Hubble Space Telescope is a cooperation between NASA and the European Space Agency. It is one of NASA's Great Observatories, along with the Compton Gamma Ray Observatory, the Chandra X-ray Observatory, and the Spitzer Space Telescope."

In Figure 1, the yellow part is "The Hubble Space Telescope" in the beginning of the paragraph. In Figure 2, the yellow part is "The Hubble Space Telescope is a space telescope that was launched into low Earth orbit in 1990 by the Discovery Space Shuttle. The Hubble telescope is named after astronomer Edwin Hubble."

Note that this translation is different from the English Wikipedia of the same article.

B.2 Translation for Figure 3

On the left panel, sequence A is "On" and sequence B is "this pass January 1st".

On the right panel, sequence A is "...causing the price to have gone up." and sequence B is "Investors should study the information....".

B.3 Translation for Figure 5

"Tokyo was honored to host the Olympic Games on September 7, 2013 at the 123rd session of the International Olympic Committee in Buenos Aires. Argentina This is the third time Tokyo has been granted the right to host the Olympics. For the first time in 1940 it was granted the right to host the first Asian Summer Olympics. and Sapporo for the Winter Olympics. But has withdrawn from the competition due to the war between China and Japan. This time, Tokyo is the fifth city (and the first city in Asia) to host more than one Summer Olympics. Tokyo has also been honored to host the 2020 Summer Paralympic Games for athletes with disabilities."