THAI SENTENCE SEGMENTATION USING LARGE LANGUAGE MODELS

Mr. Narongkorn Panitsrisit

An Independent Study Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Arts in Linguistics
Department of Linguistics
FACULTY OF ARTS
Chulalongkorn University
Academic Year 2022

การตัดประโยคภาษาไทยโดยใช้แบบจำลองทางภาษาขนาดใหญ่

นายณรงค์กร พนิตศรีสิทธิ์

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

Independent Study Title	THAI SENTENCE SEGMENTATION USING LARGE LANGUAGE MODELS
By	Mr. Narongkorn Panitsrisit
Field of Study	Linguistics
Thesis Advisor	Associate Professor ATTAPOL THAMRONGRATTANARIT, Ph.D.

Accepted by the FACULTY OF ARTS, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Arts

INDEPENDENT STUDY COMMITTEE

..... Chairman
(Associate Professor WIROTE AROONMANAKUN, Ph.D.)

..... Advisor
(Associate Professor ATTAPOL THAMRONGRATTANARIT, Ph.D.)

..... Examiner
(Assistant Professor NATTANUN CHANCHAOCHAI, Ph.D.)

ณรงค์กร พนิตศรีสิทธิ์ : การตัดประโยคภาษาไทยโดยใช้แบบจำลองทางภาษาขนาดใหญ่. (THAI SENTENCE SEGMENTATION USING LARGE LANGUAGE MODELS) อ.ที่ปรึกษาหลัก : รศ. ดร.อรรถพล ธำรงรัตนฤทธิ์

การตัดประโยคภาษาไทยเป็นเรื่องที่มีผู้สนใจอยู่มาก แต่การตัดประโยคโดยใช้แบบจำลองทางภาษาขนาดใหญ่ซึ่งใช้สถาปัตยกรรมทรานส์ฟอร์เมอร์ยังมีผู้ศึกษาไม่มากนัก ผู้วิจัยใช้คลังข้อมูล LST20 เพื่อทำการทดลองจำนวนสามการทดลองโดยประกอบไปด้วย (1) การปรับจูนการจำแนกคำในสถานการณ์ต่าง ๆ ด้วย WangchanBERTa ซึ่งเป็นแบบจำลองทางภาษาขนาดใหญ่ที่ฝึกฝนด้วยข้อมูลภาษาไทย (2) การใช้ Joint Learning สำหรับการตัดประโยคและอนุพากย์ และ (3) การถ่ายโอนข้ามภาษาโดยใช้ XLM-RoBERTa ซึ่งเป็นแบบจำลองหลากภาษา ผลการทดสอบพบว่า WangchanBERTa มีประสิทธิภาพดีกว่าแบบจำลองอื่นในการตัดประโยคภาษาไทย และเมื่อปรับจูนเพิ่มเติมด้วยข้อมูลคำและบริบทจะทำให้แบบจำลองดังกล่าวมีประสิทธิภาพดีขึ้น อย่างไรก็ตาม การถ่ายโอนข้ามภาษาจากภาษาอังกฤษและภาษาจีนไปยังภาษาไทยเป็นวิธีที่ไม่ได้ผลดีนักสำหรับการตัดประโยคภาษาไทย

สาขาวิชา ภาษาศาสตร์ ลายมือชื่อนิสิต

.....

ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก

.....

6480012522 : MAJOR LINGUISTICS
KEYWORD: Natural language processing, Thai sentence segmentation, large language models

Narongkorn Panitsrisit : THAI SENTENCE SEGMENTATION USING LARGE LANGUAGE MODELS. Advisor: Assoc. Prof. ATTAPOL THAMRONGRATTANARIT, Ph.D.

Thai sentence segmentation has been on the topic of interest among Thai NLP communities. However, not much literature has explored the use of transformer-based large language models to tackle the issue. We conduct three experiments on the LST20 corpus, including (1) fine-tuning WangchanBERTa, a large language model pre-trained on Thai, across different classification tasks, (2) joint learning for clause and sentence segmentation, and (3) cross-lingual transfer using the multilingual model XLM-RoBERTa. Our findings show that WangchanBERTa outperforms other models in Thai sentence segmentation, and fine-tuning it with token and contextual information further improves its performance. However, cross-lingual transfer from English and Chinese to Thai is not effective for this task.

Field of Study:	Linguistics	Student's Signature .....
Academic Year:	2022	Advisor's Signature .....

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Associate Professor Attapol Thamrongrattanarit, Ph.D., for his constant support throughout this project. His guidance and patience have been invaluable to me throughout the progress of this study, in both model training and writing. His commitment to supporting me in my academic endeavors has been crucial, and I am sincerely grateful for his mentorship.

I also wish to extend my gratitude to all the professors and faculty members in the Linguistics Department. Their support and teaching have been essential to my academic pursuits, and I am profoundly grateful for the enriching experiences I have had in the department.

To my family, I am eternally grateful for their unconditional love and unwavering support. Their presence has been a beacon of strength, providing me with comfort and inspiration during challenging times.

Lastly, I would like to express my appreciation to all my friends who have shared my struggles and provided me with the support and encouragement to keep moving forward.

Narongkorn Panitsrisit

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1
INTRODUCTION

One of the fundamental tasks in natural language processing (NLP) is sentence segmentation, which involves identifying the boundaries of sentences in a given text. Sentence segmentation is essential for several downstream tasks, including machine translation, text summarization, part-of-speech (POS) tagging, and syntactic and semantic parsing (Read et al., 2012). While sentence segmentation may be straightforward in many languages due to the use of explicit punctuation marks to indicate the end of a sentence, it poses a significant challenge for languages like Thai that lack explicit sentence markers.

In Thai, spaces are used to signal sentence boundaries, but this approach is complicated by the fact that spaces can also appear between clauses, phrases, and words (Office of the Royal Society, 2008). This characteristic of the Thai language makes it challenging to distinguish between spaces that indicate the end of a sentence and those that do not, which poses a significant challenge for accurately identifying sentence boundaries. This difficulty is exemplified in the following text:

> น้ำพริก มีมาตั้งแต่สมัยกรุงศรีอยุธยา โดยคำว่า “น้ำพริก” มีความหมายมาจากการปรุงด้วยการนำสมุนไพร พริก กระเทียม หัวหอม เครื่องเทศกลิ่นแรง มาโขลก บด รวมกัน เพื่อใช้สำหรับจิ้ม โดยมี ดอกแคมะเขือยาว แตงกวา ถั่วฝักยาว มะเขือม่วง ถั่วพู สัตว์น้ำต่าง ๆ เช่น ปลา กุ้ง เป็นต้น น้ำพริก เป็นวิธีปรุงอาหารหรือเครื่องปรุงอาหาร โดยการนำเครื่องปรุงชนิดต่าง ๆ ลงโขลกรวมกันในครก [*Nam Phrik* has been around since the Ayutthaya period, with the term “Nam Phrik” deriving from the preparation method of pounding and grinding a mixture of herbs,

chilies, garlic, shallots, and strong-smelling spices, to be used for dipping with agati flowers, long green eggplants, cucumbers, cowpeas, eggplants, winged beans, and aquatic animals such as fish and shrimp. Nam Phrik is a method of cooking or seasoning food by pounding together various types of seasonings in a mortar.] (Namprikprakdee, 2020)

The absence of punctuation to mark the end of a sentence in Thai, combined with the multiple functions of spaces in the language, makes it difficult to distinguish between sentence boundary (sb) and non-sentence boundary (nsb) spaces. Even in the aforementioned text, which appears to be a single paragraph, different interpretations of sentence boundaries could result in more than one sb space.

Furthermore, it is crucial to consider situations where spaces are completely absent in the text, particularly when the input originates from speech recognition output. This scenario adds another layer of intricacy to the process of sentence segmentation. In the context of the example below, the text was generated using the speech-to-text feature in Google Docs to transcribe an interview clip of a Thai local political candidate.

ก็พยายามคิดว่าเขาต้องได้อะไรสักเรื่องนอกจากตลกก็คือเราก็เอากฎหมายเนี่ยมาย่อยจากภาษากฎหมายที่ทุกคนอ่านยากแล้วก็มาแปลให้เป็นภาษาชาวบ้านที่ทุกคนอ้อแค่นั่งฟังขำๆแต่ว่าความรู้กฎหมายซึมเข้าไปในหัวด้วยในตัวอะไรแบบเนี้ยแล้วก็มีฐานพวกเนี้ยมากพอสมควรแล้วเค้าก็จะรู้ว่าจูลี่จะไม่ใช่คนแค่ตลกอย่างเดียวคือเบื้องลึกเบื้องหลังก็คือเป็นนักกฎหมายด้วยอะไรแบบเนี้ย [I think that they should get more than just entertainment. So, I take complicated legal jargons and break them down into everyday language that everyone can understand and have a good laugh while gaining knowledge of the law. There are plenty of supporters who know

that Juree is not just being funny, but actually a lawyer.]
(WorkpointOfficial, 2022)

The absence of spaces poses a significant difficulty, as the absence of explicit boundaries makes it even more challenging to accurately identify sentence boundaries. Dealing with such instances becomes crucial in developing robust models for sentence segmentation in Thai text.

Over the past three decades, researchers have been working to tackle this issue using computational techniques, despite the scarcity of annotated data and Thai language corpora. Initially, rule-based methods that relied on linguistic knowledge were introduced. Subsequently, statistical models such as Winnow, N-gram, and Maximum Entropy gained prominence in the field. Recently, with the success of deep learning, models such as CRF, CNN, LSTM, and BERT have surged in popularity due to their ability to achieve state-of-the-art performance.

Expanding on the approach proposed by Yuenyong and Sornlertlamvanich (2022) of utilizing transformer-based models, our research aims to develop a model that exceeds the limitations of utilizing transformers in a simplistic and conventional manner. We aim to create a novel approach that outperforms standard transformer-based methods in terms of effectiveness. This will be accomplished through three experiments conducted on the LST20 corpus (Boonkwan et al., 2020), a recently released dataset of annotated sentence boundaries in Thai, using the WangchanBERTa model (Lowphansirikul et al., 2021), a transformer-based large language model that has been pre-trained specifically on Thai data. Additionally, we investigate the joint learning approach for clause and sentence segmentation, along with exploring the potential of cross-lingual transfer using the XLM-RoBERTa model (Conneau et al., 2020).

Our experiments showed that the WangchanBERTa model, pre-trained on a diverse range of Thai data sources, outperformed the XLM-RoBERTa model, which was pre-trained on data from multiple languages in the task of Thai sentence segmentation, and that cross-lingual transfer may not be as effective a method for Thai sentence segmentation. We also find that fine-tuning WangchanBERTa by incorporating information about the structure of the text can significantly improve the model's ability to identify sentence boundaries.

CHAPTER 2
BACKGROUND AND LITERATURE REVIEW

Chapter 2 offers a background and literature review on sentence segmentation. It covers fundamental concepts such as sentences, sentence boundaries, and elementary discourse units (EDUs), as well as providing an overview of existing research and approaches in NLP for sentence segmentation in Thai and other languages.

2.1 Sentences

Traditional grammar provides the definition of a sentence as a group of words that expresses a complete thought or idea (Crystal, 2008). Such sentences consist of syntactic units or constituents, including words that are assigned to grammatical categories or parts of speech (POS), as well as phrases, such as noun phrases and verb phrases.

Predicates and arguments are essential components in determining the structure of a sentence. The predicate is the part of a clause that expresses the action or state of being and takes one or more arguments, also known as complements. Complements generally follow the verb and provide additional information about the action or state. For example, the sentence “Aegon chased the bird” contains the subject *Aegon*, the predicate *chased*, and the complement *the bird*. More precisely, *Aegon* and *the bird* are arguments of the predicate *chased*. Sentences can also contain adjuncts, which are optional elements that provide additional information about the time, place, manner, or purpose of the action or state expressed by the predicate. In short, a clause is a predication structure that consists of a subject and a predicate and may include complements and adjuncts (Radford, 2009).

In traditional grammar, the structure of sentences can be classified based on the number and type of clauses they contain. A simple sentence comprises a

single clause, whereas a compound sentence is composed of two or more simple sentences joined by coordinating conjunctions. On the other hand, within complex sentences, clauses follow a hierarchical order, with one clause being the main clause, also known as the independent clause. The main clause contains the primary predicate and expresses the primary idea of the sentence, while other clauses are subordinate clauses, or dependent clauses, which serve as complements or adjuncts to the main clause. Complex sentences can be joined by subordinating conjunctions (Prasithrathsint et al., 2011; Radford, 2009).

The phrase structure rules proposed by Chomsky (1957) provides a structural description of sentences. According to this framework, it posits that a sentence is composed of constituents, which are constructed using a set of rules. These rules describe how a complete sentence is formed from phrases and how these phrases are combined from different words. Symbols such as NP for Noun Phrase, VP for Verb Phrase, and N for Noun are used in the rules to define sentence structure. Chomsky illustrated some examples of these rules as follows:

(1) Sentence → NP + VP
(2) NP → T + N
(3) VP → Verb + NP
(4) T → *the*
(5) N → *man*, *ball*, etc.
(6) Verb → *hit*, *cook*, etc.

The rules are hierarchical and specify the order in which constituents are combined to form a sentence. For instance, the rule “Sentence → NP + VP” means that a sentence consists of a Noun Phrase followed by a Verb Phrase, while the rule “NP → T + N” means that a Noun Phrase is composed of a determiner (T) followed by a noun. To demonstrate, we can apply these rules to the sentence “The man hit the ball.” as shown below (Chomsky, 1957):

Sentence	
NP + VP	Rule 1
T + N + VP	Rule 2
T + N + Verb + NP	Rule 3
the + N + Verb + NP	Rule 4
the + man + Verb + NP	Rule 5
the + man + hit + NP	Rule 6
the + man + hit + T + N	Rule 2
the + man + hit + the + N	Rule 6
the + man + hit + the + ball	Rule 5

The application of grammar rules, known as derivation, can be depicted by a tree diagram, which presents a visual illustration of sentence structure, as illustrated in Figure 1. In addition to phrase structure rules, the tree structure offers a constituent structure analysis of sentences, and has influenced the development of various models of grammar and syntax (Prasithrathsint et al., 2011).

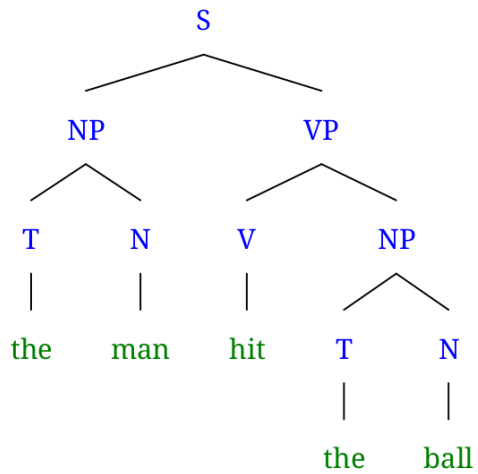

Figure 1: A tree structure of the sentence “The man hit the ball.”

Traditional grammar and phrase structure rules present a brief overview of the idea that a sentence is composed of constituents. Specifically, a sentence consists of a subject, represented by a noun phrase, and a predicate, represented by a verb phrase. These phrases can take multiple arguments and can be further broken down into smaller constituents, allowing for a more detailed analysis of sentence structure.

Moving beyond the theoretical aspects of sentence structure, it is important to consider how sentence boundaries are marked in written language. One way this is achieved is through the use of punctuation. Proper use of punctuation, such as periods, commas, and semicolons, can help to disambiguate sentences and enhance readability.

2.2 Sentence Boundaries and Punctuation

Punctuation marks are essential components of written language systems. They serve to clarify lists and sequences, convey tone and meaning, and signal the end of sentences. In Western European languages such as English, Spanish, French, German, Italian, Portuguese, and Russian, the use of terminal punctuation marks, such as the period (.), question mark (?), and exclamation mark (!) is prevalent in indicating sentence boundaries. Similarly, many modern languages across regions of Asia also employ punctuation marks. For example, the writing systems of Chinese, Japanese, and Korean feature a systematic use of terminal punctuation marks, with Chinese and Japanese using a small circle (。) to represent the period, while Korean uses the Western period variant.

Burmese, on the other hand, has a unique symbol for the period, namely the "။" (ပုဒ်မ [pouʔmá]), which is equivalent to the English period. The traditional Burmese punctuation system only includes two marks, the aforementioned "။", and "၊" (ပုဒ်ဖြတ် [pouʔpʰyaʔ]), which is similar to the

English comma. However, the use of Western punctuation marks such as the exclamation mark and question mark is inconsistent and less common (Jenny & Tun, 2016).

While the consistent application of terminal punctuation marks in some languages may facilitate identifying sentence boundaries, in others, it remains a challenge. Even in cases where punctuation marks are used consistently, they may not always be consistent in their functions. One example of this is the Tibetan punctuation mark "།" (ཤད *shad*), which can be used after a word, phrase, or sentence. As a result, it can be unclear whether a shad is meant to indicate the end of a sentence or not (Li et al., 2022).

Some literature suggests that certain languages, such as Arabic, Lao, and Khmer, lack any clear marker for sentence boundaries (Charoensuk et al., 2005; Saetia et al., 2021). However, this claim is not entirely accurate, as these languages do employ punctuation marks to denote the end of a sentence. In Arabic, the period serves this function much like in English, but a comma can also be used in its place (Alqinai, 2015). Additionally, as Arabic is written from right-to-left, a reversed question mark (؟) is used instead.

Similarly, Lao has a range of punctuation marks, including a period at the end of a sentence or paragraph, a question mark, and an exclamation mark (Simmala & Poomsan Becker, 2003; Srisawang, n.d.). However, one limitation of the available literature is the lack of clear evidence on the consistent use of punctuation marks in Lao, which makes it challenging to draw a definitive conclusion on this matter.

For Khmer, the use of "។" (ខណ្ឌ *khan*) to indicate sentence boundaries remains a significant aspect of Khmer writing, even though it is not as commonly used as the full stop in English. This punctuation mark has its roots in the country's religious heritage, as it is a symbol borrowed from Pali and Sanskrit, the languages used in Buddhist texts (Thong, 1985). However, it typically indicates the conclusion of a paragraph that may encompass a singular

sentence or multiple sentences relating to the same subject matter (Huffman, 1970).

Conversely, while some languages have a standardized system of punctuation marks, many speakers of those languages choose not to use them, which can lead to uncertainty and ambiguity in recognizing sentence boundaries. Thai is one such language, which has a standardized set of guidelines for using punctuation marks to indicate sentence boundaries (Office of the Royal Society, 2008). The Thai full stop serves the same function as its Western counterpart in marking the end of a sentence. Furthermore, the guidelines mandate the inclusion of spaces to indicate boundaries at different levels of linguistic units, including sentences, clauses, phrases, words, and punctuation marks. Specifically, a space is required after each sentence, as well as after independent clauses linked with conjunctions, and after ว่า ([wâː] that) when introducing a new clause or phrase. Spaces are also needed after proper nouns, and around numbers and list items. Additionally, the use of spaces is also mandatory for some punctuation marks such as parentheses and quotation marks.

Despite these guidelines, Thai speakers typically choose not to use full stops. This is true across different domains and registers, from novels to scientific articles and social media as the use of full stops to delimit sentence boundaries is virtually non-existent in contemporary Thai texts (Rojana-Anun, 2019). Instead, Thai speakers rely on spaces as a cue to separate sentences. It should be noted that while the use of spaces to indicate phrases, words, and punctuation marks is mandatory, their use for sentence and clause boundaries is optional (Ngarmwirojki & Luksaneeyanawin, 2013). This optional usage of spaces creates inconsistency and subjectivity in identifying sentence boundaries, which depends on each individual stylistic preference.

The lack of consistency in punctuation usage poses a challenge for NLP tasks that rely on punctuation to identify sentence boundaries in Thai. Without proper punctuation or consistent use of spaces, sentence segmentation or sentence boundary detection becomes more difficult and error-prone. As a result, research in Thai NLP has been focused on developing more robust models that can handle this challenge.

2.3 Elementary Discourse Units

The challenges of identifying sentence boundaries in Thai was investigated in the study of Aroonmanakun (2007). The study involved a small experiment where Thai native speakers were asked to segment one page of Thai translated text and one page of Thai source text. The author proposed using a parallel corpus of English and Thai texts to identify sentence boundaries by aligning Thai segments with English sentences. The preliminary results revealed that there was no consensus among the participants on sentence segmentation. However, the author observed that substantial agreements among sentence boundaries were found at the beginning of a discourse segment when the topic shifted, and when the topic continued with an overt noun phrase or a pronoun. Additionally, the use of conjunctions was not found to be an indicator for sentence boundaries.

The paper suggested that, due to the fuzzy nature of sentence boundaries in Thai, identifying clauses as the basic syntactic unit instead of sentences might be a more useful approach. However, identifying clauses in Thai is not a straightforward process due to the complexity of the language's grammar. The paper suggests a sequence of segmentation steps, including identifying discourse markers and the discourse topic, identifying the head of each segment using a dependency parser, and combining the segments into a discourse structure based on their relations to one another. Essentially, this approach suggests viewing a discourse as a composition of clauses instead of sentences.

The approach of identifying clauses as the basic unit of texts is rooted in the Rhetorical Structure Theory (RST) proposed by Mann and Thompson (1988). RST explains how parts of text are organized and formed into a larger structure of text, represented as a tree structure. RST employs discourse units to represent pieces of information and explains the relations between adjacent non-overlapping units, with one unit being the essential, called the nucleus, and the other functioning as a supporting text, called the satellite. Relations between units of text are, for example, Elaboration, Condition, Interpretation, Evaluation, Summary, and Contrast.

Carlson et al. (2001) utilized the RST framework and introduced the concept of the elementary discourse unit (EDU) as the minimal unit of discourse in their RST corpus annotation work. They considered clauses as EDUs and allowed some phrasal EDUs with strong discourse markers, such as *because, in spite of, and as a result of,* and *according to*. However, clauses that function as subjects, objects, or complements of a main verb are not treated as EDUs.

What constitutes EDUs in English differs from that of Thai discourse structure, and Thai discourse structure presents challenges when identifying EDUs due to several characteristics. Firstly, Thai written texts lack explicit punctuation to indicate word, clause, or sentence boundaries. Secondly, some Thai named entities (NE), noun phrases, and verb phrase sequences share the same structure, making it difficult to distinguish them. Additionally, zero anaphora is common in Thai structures where subjects and/or objects are omitted. Furthermore, Thai text structures often contain many embedded clauses, especially relative clauses indicating noun modifications. Next, some Thai words have multiple functions depending on their position and expression within a unit, resulting in part-of-speech ambiguity. Moreover, some Thai words with the same part-of-speech may have different meanings, causing

word-sense ambiguity. Finally, Thai serial verb constructions, where verbs may form a sequence to express meanings and relationships with respect to the main verb, also pose challenges for segmenting Thai EDUs (Ketui et al., 2013).

Charoensuk et al. (2005) proposed a classification system for Thai EDUs, dividing them into Basic and Embedded types. Basic EDUs consist of clause structures such as simple sentences or phrases that begin with a strong discourse marker, while Embedded EDUs contain clause or phrase structures and appear in the middle of a Basic EDU. Each EDU is marked in square brackets, as demonstrated in Table 1.

Types of EDUs		**Example**
Basic	Clause/ Simple sentence	[กะหล่ำปลีมีสีเขียว] [The cabbage has green color.]
	Noun phrase	โรคระบาดพบในภาคกลาง [เช่น ปทุมธานี, นครปฐม] [Epidemics was found in the middle region] [such as Pathumtani, Nakornpathom.]
Embedded	Clause	กะหล่ำปลี [ที่ถูกทำลาย] จะมีสีเหลือง The cabbage [that was destroyed] will have yellow color.
	Noun phrase	เกษตรกรควรใส่ปุ๋ยไนโตรเจน [เช่น ปุ๋ยแอมโมเนียมซัลเฟต หรือยูเรีย] ลงในแปลงด้วย Agriculturist should put Nitrogen fertilizer [such as Ammonia fertilizer, Urea] into the plot.

Table 1: Thai EDU types proposed by Charoensuk et al. (2005)

The study suggests a three-step process for EDU segmentation. Firstly, the input undergoes word segmentation, POS tagging, NE extraction, and compound noun extraction. The next step involves using machine learning rules to identify the starting and ending of Basic and Embedded EDUs. The C4.5 decision-tree learning system is utilized, along with five categories of features: Discourse Segmentation Cues, Correlative Discourse Markers, Blank, WORD POS, and Phrase Boundary Features. Finally, some heuristic rules are

employed to resolve redundant cues and improve the accuracy of the segmentation. The proposed segmentation method attained an F1 score of 80.49%.

Alternative criteria for determining EDUs in Thai was proposed by Ketui et al. (2013). According to their definition, each EDU should contain one verbal unit or verb phrase as the core verb. Consequently, serial verbs are broken down into multiple units. The authors used syntactic and semantic patterns to define six syntactic units for detecting Thai EDUs (T-EDUs), and classified two types of Thai Non-EDUs (T-Non-EDUs). Simple clauses, subject zero-anaphora clauses, clauses with attribution verbs, comparative clauses, question clauses, and embedded conjunction clauses are considered as T-EDUs, while T-Non-EDUs are used to distinguish between clausal subjects/objects and synthetic nominal compounds. These rules are then used to create a set of context-free grammar (CFG) rules and a chart parser is used to detect T-EDUs in a text, including their structures.

The study evaluated the proposed method in four environments, including close tests with pre-chunked and running text, and open tests with pre-chunked and running text. The highest F-scores were obtained in the close test with running text, where the longest-matching (LM) and maximum-matching (MM) constraints achieved 92.76% and 92.39%, respectively. In the open test with pre-chunked text, the LM and MM constraints achieved F-scores of 58.01% and 60.75%, respectively.

Given the inconsistent guidelines for identifying Thai EDUs in previous studies, there is a lack of clarity in the definition on this topic. Charoensuk et al. (2005) included clauses and phrases marked by strong discourse markers as EDUs, while Ketui et al. (2013) only considered clauses as EDUs. To address this issue, Intasaw (2013) and Intasaw and Aroonmanakun (2013) aimed to establish clear principles for determining Thai EDUs in a consistent manner.

Their study proposes guidelines for determining minimal units in Thai discourse structures, with clauses containing finite verbs and noun phrases with strong markers classified as EDUs. The classification of EDUs and non-EDUs is shown in Table 2.

Structures	**EDU**	**Non-EDU**
Finite clauses	Independent clauses	Dependent clauses • subject/object clauses
	Dependent clauses • Finite relative clauses • Adverbial clauses • Coordinate clauses	
Non-finite clauses		Non-finite relative clause
Clausal complements	Finite clausal complements of attributive verbs	Non-finite clausal complements
		Clausal complements of noun
Serial verb construction (SVC)	SVC with attributive verbs	SVC
Cleft	Identificational cleft	Contrastive cleft
Phrases with strong markers	Noun phrases with strong markers	
	Noun phrases in the form of parentheticals	
	Names of titles and authors	
Same unit construction	Construction with relative clauses, appositives, and parentheticals.	
Punctuation		Punctuation marks

Table 2: The classification of Thai EDUs from Intasaw (2013)

The study classifies finite clauses as independent or dependent, and treats independent clauses as EDUs. Dependent finite clauses, including subject/object clauses, finite relative clauses, adverbial clauses, and coordinate

clauses, are subcategorized. Among them, subject/object clauses are not considered as EDUs because they lack the function of modifying any part of the text. In contrast, finite relative clauses are treated as EDUs since they function as noun modifiers, while adverbial clauses provide additional information and coordinate clauses hold elaboration relations between independent clauses and are treated as EDUs. For non-finite clauses, including non-finite relative clauses, they are not considered EDUs due to the non-finite status of their verbs.

Finite clausal complements of attributive verbs, such as ยอมรับ ([jɔːm ráp] accept), คิด ([kʰít] think), เชื่อ ([tɕʰɯ̂ə] believe), เสนอ ([sanə̌ː] propose), ถาม ([tʰăːm] ask), สงสัย ([sŏŋ săj] doubt), are treated as EDUs and may be introduced by a complementizer, such as ว่า ([wâː] that) or ที่ ([tʰîː] that). However, non-finite clausal complements are not treated as EDUs.

Serial verb constructions (SVC) are treated as a single EDU since they express a single event and represent one piece of information. However, if there is an attributive verb within the SVC, it should be segmented into separate EDUs to ensure proper identification of EDU boundaries.

In Thai, cleft constructions are classified as either contrastive cleft or identificational cleft. The former, which is made up of the definite marker ที่ ([tʰîː] that) and the copula เป็น ([pen] be), is not considered as a separate EDU. In contrast, the latter is comprised of the copula คือ ([kʰɯː] be) followed by a cleft clause featuring the definite marker ที่ ([tʰîː] that). It is treated as two EDUs as the cleft clause has an elaboration relation with the noun it describes. To clarify the distinctions between the two constructions, examples from Intasaw (2013) are illustrated as follows:

Contrastive cleft: [นิกที่เป็นคนทำจานแตก]1

[ník tʰîː pen kʰon tʰam tɕaːn tɛ̀ːk]

[Nick, who was the one who broke the plate]1

Identificational cleft:[นั่นแหละคือสิ่ง]1[ที่ทำให้เกิดภาพเหมือน]2

[nân lɛ̀ʔ kʰɯː sìŋ tʰîː tʰam hâj kɛ̀ːt pʰâːp mɯ̆ən]

[That's the thing]1 [that creates the portrait.]2

Phrases with strong discourse markers serve as strong connectors between discourse units and are considered EDUs. Examples of these markers include เช่น...ฯลฯ ([tɕʰên ... lá] for example ... etc.), ได้แก่...เป็นต้น ([dâj kɛ̀ː ... pen tôn] for example ... etc.), ยกตัวอย่างเช่น ([jók tua jàːŋ tɕʰêːn] for example), อย่างเช่น ([jàːŋ tɕʰên] for example), and เพื่อ ([pʰɯ̂ə] for). Additionally, noun phrases in the form of parentheticals and the name of the title and author may also be considered EDUs.

The same unit construction refers to when a clause is split by the insertion of another clause, but both parts are treated as a single EDU. This construction can be found in relative clauses, appositives, and parentheticals.

Finally, punctuation marks are not considered as separate EDUs. While some punctuation marks can be used to identify EDU boundaries, such as the question mark, parenthesis, and quotation marks, others such as the dash, comma, period, and colon usually appear within the EDU and do not play a role in identifying EDU boundaries.

Intasaw (2013) applied the support vector machine (SVM) model to a corpus of Thai academic written language containing 8,102 clauses. The SVM model employed various features, such as POS tags, discourse markers, spaces, and punctuation marks, to identify clause boundaries. The best feature pattern was a combination of all linguistic features, achieving an F-measure of 81.17%. The system's performance further improved to 84.74% when the exponent D of the kernel parameter value was adjusted to 4.

2.4 Sentence Segmentation in Thai

Thai sentence segmentation, which is also referred to as sentence boundary detection (SBD), has been a subject of research for almost thirty years in the field of NLP, with various methods being developed and studied. The existing literature can be divided into three main categories: (1) rules-based approaches that rely on linguistic knowledge of Thai, including its grammatical rules and the characteristics of spaces, (2) statistical approaches such as Winnow, N-gram, and Maximum Entropy models; and (3) deep learning approaches that include models such as Convolutional Neural Network (CNN), Conditional Random Fields (CRF), Bidirectional Long Short-Term Memory (BiLSTM), and Transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) which have achieved high accuracy in Thai sentence segmentation. Many of these approaches treat sentence segmentation as binary classification or sequence labeling tasks to predict whether a token is the end of a sentence.

The computational approach to Thai sentence segmentation was first introduced by Longchupole (1995), who proposed a three-level algorithm based on Thai grammatical rules. The algorithm begins with tokenization of words from the paragraph, followed by analysis of sentence structure to identify the head verb, and then sentence segmentation from the paragraph input.

In the first level, the input sequence undergoes Possible Word Matching, a word segmentation process that forms all combinatorial possibilities. For example, the input "เรือโคลงเนื่องจากโคลงเรือ" ([rɯa kʰloːŋ nɯ̂ɔŋ tɕàːk kʰoː loŋ rɯa] The boat swayed because a cow boarded it.) can be tokenized to form four possible combinations (Longchupole, 1995):

(1) เรือ โคลง เนื่องจาก โคลง เรือ
(2) เรือ โคลง เนื่องจาก โค ลง เรือ
(3) เรือ โค ลง เนื่องจาก โคลง เรือ
(4) เรือ โค ลง เนื่องจาก โค ลง เรือ

Next, each token is labeled with a POS tag using a set of rules from Dependency Grammar to identify the head verb of the input sequence. Phrase Structure Rules are then used to identify the specifier and complements of the head verb, followed by Case Grammar to establish possible relations between the head and other words to determine the correct sentence structure. Finally, each token undergoes an inspection process to determine its inclusion in the sentence. If a token fails to meet the criteria for sentence membership, the algorithm verifies if it is labeled as a conjunction and subsequently merges it with the preceding sentence to create a single sentence.

The algorithm's performance was evaluated using 11 short paragraphs containing 34 sentences, and the study reported that 9 out of 11 paragraphs and 32 out of 34 sentences were correctly segmented, resulting in an accuracy of 81.18% and 94.2%, respectively. However, this method faces limitations when dealing with longer paragraphs.

Moving on to statistical machine learning models, Mittrapiyanuruk and Sornlertlamvanich (2000) and Wang et al. (2019) explored the use of n-gram models for Thai sentence segmentation, albeit with different models and techniques.

Mittrapiyanuruk and Sornlertlamvanich (2000) proposed a method for extracting sentences from Thai paragraphs using a probabilistic POS trigram model. The authors approached the problem as a binary classification task, classifying spaces as either sb or nsb. To train and test their algorithm, they used the ORCHID corpus, a Thai corpus with POS annotation (Sornlertlamvanich et al., 1997).

The algorithm involves reconstructing adjacent tokens to form a word sequence with spaces in between, which are then classified using a statistical POS tagging approach. The most probable POS sequence is determined using a trigram model and the Viterbi algorithm. Then, the algorithm constructs a sentence by taking the first word of the sequence and continuing until the word before the sb space. The remaining words after the sb space are then used as the previous token in the next iteration. By using this approach, the algorithm is able to extract sentences by scanning tokens instead of the whole paragraph. Using the *space-correct* metric[1], the accuracy of classifying sb was tested on the corpus and resulted in an 85.26% rate.

Wang et al. (2019) revisited the n-gram language model's ability to classify sentence boundaries in Thai. To train the model, the authors employed maximum likelihood estimation on the frequency and probability statistics of Thai words in the corpus. The experimental findings indicated that the model's sentence segmentation performance improved with the expansion of the context window. In order to mitigate the effects of sparse data, the authors increased the back-off model for smoother data and adjusted relevant algorithm parameters. Ultimately, the authors found that the optimal performance for Thai sentence segmentation was achieved with a 13-gram model, utilizing a trigram back-off model, and resulted in a classification accuracy of 85.43% for sb space in the ORCHID corpus.

Another method proposed for Thai sentence segmentation task is the Winnow algorithm, as introduced by Charoenpornsawat and Sornlertlamvanich (2001). This algorithm is a neuron-like network which involves specialist nodes examining specific attributes of the target concept, which is either a

[1] The overall classification accuracy, calculated by (#correct sb + #correct nsb)/(total # of space tokens), is used as the evaluation metric instead of F1 score in this paper. For consistency, we will report only the accuracy of subsequent models that are tested on the ORCHID corpus, if available.

sentence break or non-break space, and voting for an outcome based on their specific expertise. The global algorithm then utilizes these weighted-majority votes to predict the value of the target concept.

To train and evaluate the Winnow algorithm, the authors used the ORCHID corpus and formed features from the context around the sb or nsb space, including collocations and the number of words on the left and right of the target space. By utilizing these features, the Winnow algorithm can predict whether a given space is classified as sb or nsb. The experimental results showed an improvement over the previously reported trigram approach, achieving an accuracy of 89.13% for sb space classification.

While basic statistical models such as trigram and n-gram have been used for natural language processing tasks, some researchers have employed Maximum Entropy models in improving accuracy and performance.

Slayden et al. (2010) proposed a Maximum Entropy model which utilizes both linguistic and contextual features to predict whether a space token in the input text should be classified as a sb or nsb. The model considers a context window of four space tokens, taking into account the dependencies between adjacent space tokens. The features used in the model include information such as the Thai tokens, numeric digits, the number of tokens since the last sb, and the paired characters that exhibit directional variation, such as brackets, braces, and parentheses. The model applies a label with the highest probability as the prediction for that space token. On the ORCHID corpus, the model achieved an accuracy of 91.19% for sb space classification, surpassing the performance of previous models.

Expanding on this approach, Wang et al. (2020) incorporated a set of Thai grammar rules with the Maximum Entropy model to enhance their sentence segmentation. The authors identified the rules governing the occurrence of sb and nsb spaces. For example, sb spaces can occur after final particles such as จ้ะ [tɕâʔ], ค่ะ [kʰâʔ], ครับ [kʰráp], นะ [náʔ] in affirmative

sentences or after terminal punctuations, while nsb spaces can occur before or after the iteration “ๆ” [máːj já mók] or abbreviation “ฯ” [paj jaːn nóːj] mark. To classify space characters in Thai sentences, the proposed method utilized a maximum entropy classifier based on context features. The experimental results show that the proposed method has achieved better experimental results than the traditional n-gram model and the Maximum Entropy method, achieving 94.16% for sb space classification accuracy on the ORCHID corpus.

The increasing popularity of deep learning models in natural language processing, owing to their improved performance across a range of tasks, has prompted researchers to shift their focus towards the development of deep learning models that can achieve state-of-the-art results.

Zhou et al. (2016) proposed a word labeling approach for sentence segmentation that considers it as a sequence labeling task, and investigates the contribution of POS information on the task. They proposed three different models: Isolated, Cascade, and Joint models, using linear-chain CRF (LCRF) and factorial CRF (FCRF) for sentence boundary detection and POS tagging. Isolated and Cascade models use LCRF, while Joint models use FCRF.

The Isolated models label words with sb if they begin a sentence; otherwise, they are labeled as nsb, with each word labeled with one of 35 POS tags. Features include the current word and surrounding words within a specified window, and the word type (English, Thai, punctuation, digits, or spaces). Cascade models incorporate additional features such as the POS tag of the current and surrounding words, and detect sentence boundaries before POS tagging. Joint models use FCRF to perform sentence boundary detection and POS tagging simultaneously, and can either use all 35 POS tags (1-step Joint approach) or first predict 12 top categories of the POS tags and then restore them back to the original POS tags (2-step Joint approach).

The paper compares the performance of these three models on Thai sentence segmentation and POS tagging tasks and shows that FCRF models, specifically the Joint model, outperform LCRF models in terms of accuracy. In the ORCHID corpus, the Isolated models achieved 95.91% accuracy for sb space classification, which is higher than previous work.

Building on previous studies that utilized CRF, recent research has explored the effectiveness of integrating BiLSTM to improve performance on sentence segmentation. BiLSTM models can capture contextual information from both left-to-right and right-to-left, making them particularly effective for this task. In addition, BiLSTM models can also be combined with CNN and CRF for further performance improvements.

Sirirattanajakarin et al. (2020) developed a BiLSTM-based model called BoydCut, which integrates CNN and BiLSTM models for sentence segmentation. In this framework, BiLSTM is utilized for word and POS features in the first layer to learn the sequential data, while CNN is applied to extract character-level features. The output vectors from word, character, and POS are concatenated into one vector and fed to the next BiLSTM layer. Finally, each time step is fed into a dense layer for binary classification. The model was trained on the ORCHID dataset and the English-Thai parallel corpus scb-mt-en-th-2020 (Lowphansirikul et al., 2020).

The study conducted four experiments to compare the performance of different combinations of features and architectures for sentence boundary prediction. The results indicated that the BiLSTM-CNN model with word, character, and POS features achieved the best performance with an F1-score of 81.34%. In contrast, the lowest F1-score of 2.16% was obtained using BiLSTM-CNN with only word and character features.

However, in the study of Thiengburanathum (2021), the CRF model performed better than the BiLSTM-CRF model. In the study, CRF and BiLSTM-CRF were compared for Thai sentence segmentation on textual data

related to beauty products. Each word in a sentence was labeled using the Beginning, Inside, End (BIE) tagging scheme to indicate the start and the end of the sentence. The CRF model outperformed the BiLSTM-CRF model in terms of F1 score on the test set, with higher precision observed for the beginning and end of the sentence. While no clear reasons were provided for the superior performance of the CRF model, feature engineering was identified as a contributing factor, along with the small sample size.

Another BiLSTM approach was proposed by Saetia et al. (2021) in which the model consists of three main modules: a low-level module, a high-level module, and a prediction module. The low-level module contains two structures: local and distant structures. The model takes a sequence of text as input to extract different features. In the local structure, input tokens are used to create n-gram embedding vectors from Word, POS, and Type embeddings at different time steps. These vectors are concatenated and fed to the BiLSTM-CRF model for local presentation to capture word groups near sentence boundaries. In the distant structure, the input sequence of text is incorporated with a self-attention mechanism to obtain distant representation. The output vectors from these two structures are then used in the next module.

The high-level module combines the two low-level representation vectors and uses them as input. They are fed into a stacked BiLSTM (StackBiLSTM) and a self-attention module, which help the model to capture the context from the whole word sequence. This second module creates high-level representation vectors. Finally, the prediction module consists of two layers: a fully connected layer and a CRF layer. The fully connected layer takes the output vectors from the high-level module to create virtual logit vectors, which are then fed to the CRF layer for predicting the token.

The experiments were conducted using two datasets, namely the ORCHID corpus and UGWC (Lertpiya et al., 2018). While all the data in the former were labeled, the latter did not. Therefore, two techniques were

employed to enable the model to leverage the unlabeled data in UGWC for training: Cross-View Training (CVT), utilized as a semi-supervised learning method, and ELMo, a language model that learns contextualized word representations. The proposed model has achieved state-of-the-art performance in Thai sentence segmentation on both the ORCHID and UGWC datasets, with the F1 score of 92.5% and 89.9% respectively. The authors also recommended using the LST20 corpus (Boonkwan et al., 2020) as an additional source of labeled data to address the limited availability of labeled data in Thai.

The latest research on Thai sentence segmentation is presented in the work of Yuenyong and Sornlertlamvanich (2022), which provides a novel approach that eliminates the need for time-consuming annotation of POS tags in previous literature. Instead, the authors propose the TranSentCut model, which relies on a pre-trained transformer model called WangchanBERTa (Lowphansirikul et al., 2021). The WangchanBERTa model is pre-trained on a large collection of Thai texts and is based on the RoBERTa architecture, which utilizes masked language modeling for pre-training.

TranSentCut is designed to predict whether a space in a given input sequence should be segmented or not. To train the model, the authors parsed the ORCHID corpus into a text file where each line represents a complete sentence. The input training example for the model is denoted by special tokens, <s> and </s>, where <s> marks the beginning of the input, and </s> is used both as a separator between two sequences and to indicate the end of input, which looks like the following: <s>sequenceA</s>sequenceB</s>. The model receives a pair of input sequences, one representing the text to the left of a space to be segmented and the other representing the text to the right.

The model achieves competitive performance on in-domain texts, receiving 96% accuracy for sb space classification in the ORCHID corpus. Furthermore, the model exhibits significant improvements on out-of-domain texts when compared to existing approaches.

2.5 Sentence Segmentation in Other Languages

In this section, we present examples of recent studies on sentence segmentation in other languages. Although there exists a body of literature on sentence segmentation in languages such as Chinese and Tibetan, recent works in English have been largely centered on specific text domains, including unpunctuated speech, social media posts, and legal texts, as well as the development of multilingual models.

In their paper, Xue and Yang (2011) addressed the problem of comma disambiguation in Chinese text. While sentence boundary detection is relatively straightforward in Chinese based on orthography, determining the appropriate use of commas can be more challenging, particularly when they are used to mark the end of a sentence. To investigate this issue, the authors utilized a subset of the Chinese Treebank (CBT) 6.0 dataset, extracting features such as lexical, POS, and syntactic features. They were then trained on a Maximum Entropy classifier, which achieved an overall F1 score of 89.2%. The study also found that lexical features were more effective than syntactic features in this task.

Another recent work in Chinese was from Srinivasan and Dyer (2021). They proposed a solution to the challenge of long Chinese sentences in machine translation tasks caused by the ambiguity of English-like sentence boundaries in Chinese. The authors utilized Reinforcement Learning (RL) to learn an optimal segmentation policy, RLSEGMENT, to maximize the BLEU scores and improve the translation quality. The proposed policy was evaluated on the WMT20 Chinese-English dataset and compared with strategies, such as the baseline NOSPLIT (no segmentation), ALLSPLIT (segmentation at all punctuations), and HEURISTIC (hand-engineered length constraints). The experimental results showed that RLSEGMENT outperformed the baseline strategy and other supervised strategies trained on syntactic data. RLSEGMENT improved the BLEU scores on long sentences by more than

three points and the brevity penalty on those sentences by about nine points compared to the baseline.

Li et al. (2022) explored the task of sentence segmentation in Tibetan and developed a deep learning model for sentence boundary detection using a recurrent neural network (RNN) with an attention mechanism called the RNN_Att model. The model is composed of five components: input layer, embedding layer, forward layer, backward layer, and attention layer. They trained the model on 465,670 Tibetan sentences, where each syllable is converted into a real-valued vector using an embedding matrix. The embedding layer uses the continuous bag of words (CBOW) model to map each syllable to a 100-dimensional vector, while the LSTM layer uses BiLSTM to capture the sentence's high-level features. The attention layer produces a weight vector to merge the syllable-level features from each step into a sentence-level feature vector.

To assess the effectiveness and generalizability of the model, the authors also evaluated it on three other languages: English, German, and Thai, using 2,000 English and German sentences and 80,000 Thai sentences. The RNN_Att model outperformed several baseline models in all four languages, achieving F1-scores of 95.37%, 93.92%, 97.78%, and 99.15% for Tibetan, English, German, and Thai, respectively. The attention mechanism was particularly effective in improving the model's performance.

In contrast to the Chinese and Tibetan literature, recent studies on sentence segmentation in English have taken a slightly different approach, with a focus on specific text domains and the development of multilingual models.

In the study conducted by Sheik et al. (2022), the aim was to explore the application of deep learning frameworks for detecting sentence boundaries in legal text. The authors utilized a dataset consisting of 80 court decisions from four domains, totaling 26,052 annotated sentences, namely Cyber Crime,

Intellectual Properties, Board of Veterans, and the United States Supreme Court. Although the authors focused on identifying periods as potential end-of-sentence (EOS) markers due to their frequency in legal text, they found that only 40% of period occurrences in the dataset were true boundary delimiters.

The study employed several models for training, including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), Bidirectional GRU (BiGRU), and Convolutional Neural Network (CNN) architectures. Additionally, transformer-based models LEGAL-BERT and XLNet were also used. While the deep learning models used a fixed-size context window, the transformer-based models read input at the subword level. The authors observed that the CNN model outperformed other deep learning models and transformer models, achieving an F1 score of 97.7%. They reported that the CNN model had a decent performance without the need for exhaustive feature engineering.

For sentence segmentation in social media text, Rudrapal et al. (2015) presented an automated method for detecting sentence boundaries that is specifically designed for this type of text. Two different approaches were developed and evaluated, a rule-based system and a machine learning-based system. The study employed three text collections for the evaluation, including a mix of 3,000 tweets and Facebook posts in English, English-Hindi code-mixed Twitter data, and formal English text from the Brown corpus.

The rule-based system is based on handcrafted rules designed to detect sentence boundaries in social media text. The system takes into account the specific characteristics of social media text, such as the use of emoticons, hashtags, and ellipses, and uses regular expressions to identify sequences of characters that indicate the end of a sentence, including multiple question or exclamation marks. The rule-based system achieved an F-measure of 78.7% on social media text.

The machine learning-based system utilized three different classification algorithms, including Conditional Random Fields (CRF), Naïve Bayes, and Sequential Minimal Optimization (SMO). The study found that SMO outperformed the other models, achieving an F-measure of 87.0% on social media text.

For unpunctuated text, the paper by Donabauer et al. (2021) explores the application of transformer-based architectures to address two natural language processing tasks: sentence boundary detection and speaker change detection. These tasks pose significant challenges since they involve both spoken and written text. The authors adopted a binary IO tagging strategy and fine-tuned the BERT model to tackle the problems. To evaluate their approach for sentence boundary detection, the researchers leveraged the Stanford Lectures Dataset and the DailyDialog Dataset. They reported that their approach outperformed or performed comparably to existing state-of-the-art methods for both tasks. The researchers attributed the success of their approach to the powerful contextual encoding capabilities of BERT, which enables it to capture essential contextual information for the two tasks.

Finally, in a study by Wicks and Post (2021), the authors proposed a binary classification approach to sentence segmentation by predicting sentence-internal or sentence-ending positions. They presented a simple context-based model, ERSATZ, which used a two-layer Transformer architecture with 6 tokens of left context and 4 tokens of right context with a 128 embedding size. Additionally, they introduced a multilingual version of ERSATZ that could segment text irrespective of input language. The authors trained the model on three language settings, namely monolingual English, a multilingual setting, and a much larger multilingual setting that includes all languages with at least 10,000 lines in the WikiMatrix dataset.

The authors evaluated the performance of monolingual and multilingual ERSATZ on various languages, including English, French, Chinese, and

Japanese, and compared it with several state-of-the-art sentence segmentation methods, such as the Punkt algorithm and SpaCy. They reported that ERSATZ outperformed all existing methods in terms of F1 score. However, they also observed that some monolingual models performed worse than the multilingual model, possibly due to a lack of data.

Building upon prior research on Thai sentence segmentation, this study intends to leverage transformer-based models for their powerful contextual encoding capabilities, allowing for the capture of vital contextual information necessary for accurate segmentation (Donabauer et al., 2021). We fine-tune pre-trained RoBERTa-based models for our task, which allows us to leverage the use of token inputs without the need for additional features (Yuenyong & Sornlertlamvanich, 2022). Additionally, the effect of cross-lingual transfer will be investigated in Thai sentence segmentation, inspired by Wicks and Post (2021) in multilingual settings. In addition to sentence segmentation, this study will also explore clause segmentation using the same methodology, as suggested by Aroonmanakun (2007).

CHAPTER 3
APPROACHES

This chapter presents our research approaches, which aims to develop a new model that outperforms other methods in terms of effectiveness. Our research focuses on two key tasks in Thai sentence segmentation: space disambiguation and character token classification. Space disambiguation involves distinguishing between sentence boundary (sb) and non-sentence boundary (nsb) spaces, while character token classification aims to identify specific tokens that serve as markers of sentence boundaries.

Through our research, we aim to develop a model that effectively addresses these tasks. To achieve this, we will conduct three experiments: (1) WangchanBERTa fine-tuning, (2) Joint learning for clause and sentence segmentation, and (3) XLM-RoBERTa cross-lingual transfer. Each experiment will explore different approaches to improve the effectiveness of sentence segmentation in Thai.

In the first experiment, we fine-tune the pre-trained WangchanBERTa model (Lowphansirikul et al., 2021) by training it on five different space and token classification tasks. These tasks include: (1) space disambiguation, (2) token classification with spaces, (3) token classification without spaces, (4) overlapping sentences with spaces, and (5) overlapping sentences without spaces.

As mentioned earlier, space disambiguation involves classifying spaces as either sb or nsb, while token classification identifies the tokens which function as the end-of-sentence marker. We explore two approaches for token classification: one with spaces, where all tokens including spaces are considered, and one without spaces, excluding space tokens entirely. Furthermore, we also investigate the task of token classification in overlapping sentences, where each input contains three sentences. The configuration for the

token classification tasks is depicted in Figure 2, illustrating the approach with spaces, while Figure 3 illustrates the approaches without spaces. Tokens which are sentence separators are labeled as ‘SEP’ (separator), while those that do not are labeled as ‘N’ (non-separator). Detailed information regarding the preprocessing steps can be found in the next chapter.

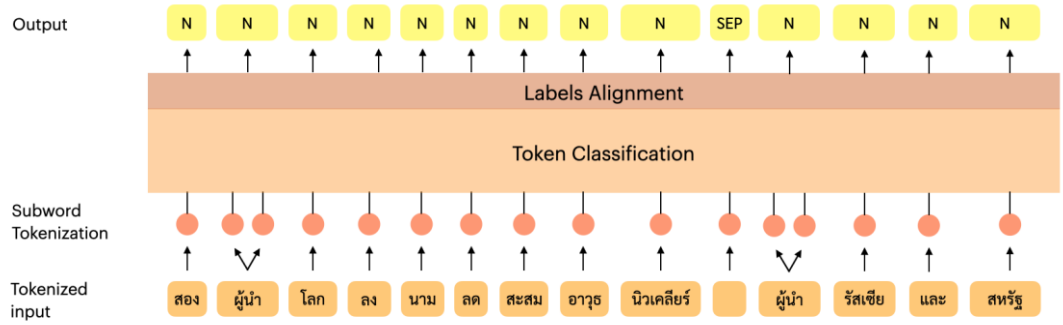

Figure 2: Configuration for the token classification tasks with spaces

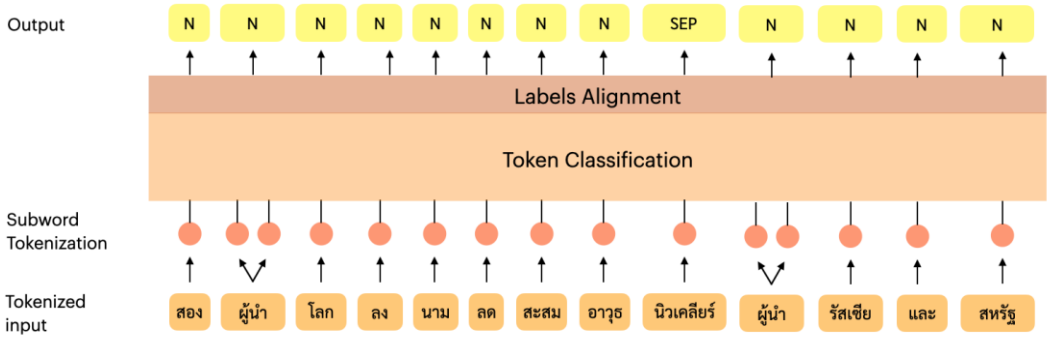

Figure 3: Configuration for the token classification tasks without spaces

For the second experiment, we will attempt joint learning with the objective of segmenting sentences and clauses. This approach will explore clause segmentation, the approach proposed by Aroonmanakun (2007), using similar methodology as the first experiment. Specifically, we will perform the task of token classification without spaces. By removing spaces, both sentence and clause beginnings are represented by the same tokens, posing a greater challenge for the model to distinguish between them.

In the final experiment of our study, we will investigate cross-lingual transfer learning by leveraging the XLM-RoBERTa model (Conneau et al., 2020), a multilingual variant of RoBERTa that has undergone pre-training on a diverse dataset from 100 languages. This approach allows us to capitalize on the knowledge gained from languages with clear sentence boundary markers. For the fine-tuning process, we will specifically focus on token classification tasks involving the distinction between sentence boundaries and non-sentence boundaries.

To ensure consistency of data across languages, we will perform a pre-processing step that involves eliminting final punctuation marks. Additionally, we will remove spaces from the data since they are not considered as tokens in English. For the fine-tuning process, we will incorporate English data from the Brown Corpus[2] (Bird et al., 2009) and GerericsKB_Best (Bhakthavatsalam et al., 2020), as well as Chinese data from a subset of the WMT19 dataset (Wikimedia Foundation, 2019). This approach draws inspiration from the work of Wicks and Post (2021) on multilingual settings, which has demonstrated superior performance. The goal is to investigate whether leveraging information from other languages and increased training data can improve sentence boundary detection in Thai. In terms of model configuration, the setup for this experiment is similar to the first experiment, as depicted in Figure 4.

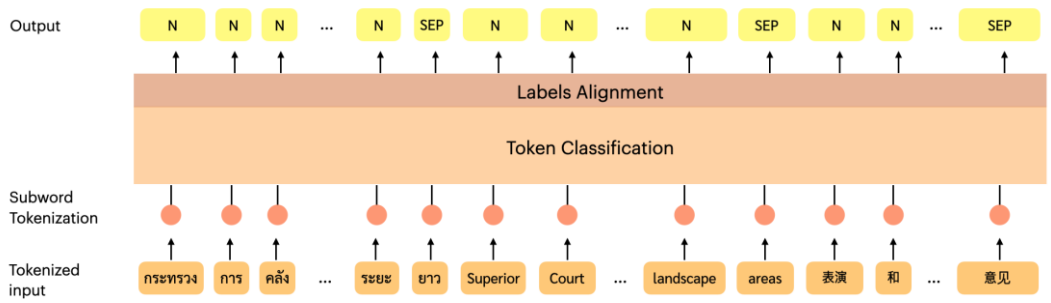

Figure 4: Model configuration for cross-lingual transfer experiment

[2] https://www.nltk.org/book/ch02.html

CHAPTER 4
EXPERIMENTS

We conducted three experiments utilizing the LST20 corpus for training and evaluation. The corpus, recommended by Saetia et al. (2021), provides annotated data of clause and sentence boundaries. It includes five layers of linguistic annotation, including word boundaries, POS tagging, named entities, clause boundaries, and sentence boundaries. The corpus follows the CoNLL-2003 style, featuring four columns for word, POS tag, named entity, and clause boundary separated by a tab. The sentence boundaries are marked by an empty line, and spaces are replaced by underscores.

The LST20 corpus contains over 3 million words, 248,181 clauses, and 74,180 sentences, and is annotated with 16 distinct POS tags. The corpus has already been split into three sets: a training set, a development set, and a test set. Table 3 illustrates an example of the data in CoNLL-2003 format.

"	PU	O	B_CLS
วิเอรี	NN	B_PER	I_CLS
"	PU	O	I_CLS
มี	VV	O	I_CLS
เฮซบ	VV	O	I_CLS
กุหลาบ	NN	B_ORG	I_CLS
ไฟ	NN	E_ORG	E_CLS
คริสเตียน	NN	B_PER	B_CLS
_	PU	I_PER	I_CLS
วิเอรี	NN	E_PER	I_CLS
_	PU	O	I_CLS
หัวหอก	NN	O	I_CLS
จอม	NN	O	I_CLS
เก๋า	VV	O	I_CLS

Table 3: Data from the LST20 corpus in CoNLL-2003 format

The dataset was obtained from a variety of Thai newspapers, including Thairath, Dailynews, Manager, Matichon, Nation, and Prachachat Business, and spans the period from January 2003 to December 2009. It is composed of 15 distinct news genres, such as politics, crime and accident, economics, entertainment, environment, sports, culture, and international news.

The annotation guidelines for the corpus define clause boundaries as parts of sentences containing at least one verb and are identified using syntactic clues such as subordinate connectors, cohesive markers, list markers, particles, and question adverbs, which differ from the EDU segmentation guidelines from Intasaw (2013). On the other hand, sentence boundaries are defined as groups of at least one clause or a phrase acting as a topic, and are identified using topic shifts denoted with cohesive markers, subject shifts between two adjacent clauses, direct and indirect speech, and item lists. Particles are used to indicate breaks in sentence boundaries.

To prepare the input data for our experiments, we preprocess text data and convert it into a more workable format. Specifically, we create a function that reads text files from the LST20 corpus and converts them into a pandas DataFrame with five columns: ‘TOKENS’ for words, ‘POS’ for part-of-speech tags, ‘NER’ for named entity tags, ‘CLAUSE’ for clause boundary tags, and ‘SEP’ for binary tags that indicate whether each token is a sentence separator. The ‘SEP’ column has two possible labels: SEP, indicating that the token is a separator, or N, indicating that it is not.

Each row of the resulting DataFrame contains a maximum limit of 150 tokens. If the number of tokens in a sentence exceeds the specified limit, the function creates a new row in the DataFrame.

The preprocessed data is then ready to be used as input for each of our models. We will discuss each model’s process in more detail in the following subsections. Table 4 depicts the preprocessed text in the DataFrame format.

Tokens	POS	NER	CLAUSE	SEP
[เผย, เจ็ต, , ลี, , และ, แจ็กกี้, , ชาน, , …	[VV, NN, PU, NN, PU, CC, NN, PU, NN, PU, …	[O, B_PER, I_PER, E_PER, O, O, B_PER, I_PER, E_PER, …	B_CLS, I_CLS, I_CLS, E_CLS, O, B_CLS, I_CLS, …	N, N, N, N, N, N, N, N, N, N, …
[เป็น, หนัง, แอ็กชั่น, ที่, มี, ฉาก, แสดง, …	[VV, NN, NN, CC, VV, NN, VV, …	[O, O, O, O, O, O, O, …	B_CLS, I_CLS, E_CLS, B_CLS, I_CLS, I_CLS, …	N, N, N, N, N, N, N, …
[นายก, รัฐมนตรี, ญี่ปุ่น, เรียกร้อง, ให้, จีน, …	[NN, NN, NN, VV, AX, NN, …	[O, O, B_LOC, O, O, B_LOC, …	B_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, …	N, N, N, N, N, N, …
[สาว, ม., ปลาย, สาธารณรัฐ, เช็ก, คว้า, มงกุฎ, …	[NN, NN, NN, NN, NN, VV, NN, …	[O, O, O, B_LOC, E_LOC, O, O, …	[B_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, …	N, N, N, N, N, N, N, N, N, …

Table 4: An example of preprocessed text in the DataFrame

We utilize a Maximum Entropy model to establish performance baselines in three distinct classification tasks: space disambiguation, token classification with spaces, and token classification without spaces. In space disambiguation, the model only considers space tokens as inputs, relying solely on the positioning of spaces to predict sentence boundaries. In token classification with spaces, the model takes all tokens as inputs, including space tokens, leveraging additional information provided by the presence or absence of spaces to better predict sentence boundaries. In token classification without spaces, the model excludes space tokens entirely, evaluating the model's ability to predict sentence boundaries based solely on the content and ordering of non-space tokens. The different classification tasks aim to evaluate the effect of including or excluding space tokens on the model's ability to predict sentence boundaries.

To train the Maximum Entropy classifier, we extract bag-of-word features from the preprocessed input. Specifically, the model takes the current

token, current POS tag, left token, left POS tag, right token, and right POS tag as input and generates corresponding feature vectors. Once the feature vectors have been created, the Maximum Entropy classifier is trained using the feature vectors and their corresponding binary labels of ‘SEP’ and ‘N’. The model is trained to maximize the likelihood of the training data given the feature vectors and their labels.

After the Maximum Entropy model has been trained, it can predict labels for the test data. The accuracy of the model can be assessed by utilizing the classification report functionality provided by scikit-learn, which includes precision, recall, and F1-score.

4.1 WangchanBERTa Fine-Tuning

In our first experiment, we employ the three classification tasks for this experiment: space disambiguation, token classification with spaces, and token classification without spaces. Additionally, we include two more tasks: overlapping sentences with spaces, and overlapping sentences without spaces. In overlapping sentences, each row contains a window of three sentences, allowing for each sentence to see the other two surrounding sentences. By including or excluding space tokens and allowing for overlapping sentences, we could comprehend the effect of different types of input information on the performance of the pre-trained model in predicting sentence boundaries.

In this experiment, we conduct fine-tuning of a pre-trained RoBERTa-based model, using the *wangchanberta-base-att-spm-uncased* model. Our first step involves converting the input data from pandas DataFrames to dictionaries and creating a DatasetDict object for training and evaluation.

The BERT tokenizer then utilizes the SentencePiece algorithm to tokenize the input text into subwords and assign each subword a unique ID. Special tokens are added to indicate the start and end of word boundaries, and an attention mask is generated. This process aligns the labels with the subwords

to indicate sentence boundaries and produces input IDs, attention masks, and tokenized subwords, as well as the original tokens and word IDs for each subword. The example below illustrates the original input tokens and the outputs generated by the BERT tokenizer.

tokens	['และ', 'แผ่น', 'เซรามิก', ' ']
sep	[0, 0, 0, 1]
input_ids	[5, 222, 10, 1093, 10, 793, 2649, 3380, 10, 6]
tokens_bert	['<s>', '▁และ', '▁', 'แผ่น', '▁', 'เซ', 'ราม', 'ิก', '▁', '</s>']
word_ids	[None, 0, 1, 1, 2, 2, 2, 2, 3, None]
labels	[-100, 0, 0, 0, 0, 0, 0, 0, 1, -100]
attention_mask	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

However, because the tokenized text is longer than the original labels, an alignment function is necessary to map subwords back to the original word tokens during evaluation.

Subsequently, the model is fine-tuned with specific hyperparameters, including a batch size of 16, 3 epochs, a learning rate of 2e-5, and weight_decay of 0.01. By using the training data, the fine-tuned model can predict labels for the test dataset. We evaluate the model's precision, recall, and F1-score on the positive class using the classification report feature provided by scikit-learn.

4.2 Joint Learning for Clause and Sentence Segmentation

In our second experiment, we treat the task as token classification without spaces. During the data preprocessing stage, we eliminate all space tokens and incorporate sentence and clause boundary labeling by assigning the 'B_Sentence' label to the first token of each sentence and 'B_CLS' labels to the initial token of each clause that is not a standalone sentence (i.e., a dependent clause), which are added to the 'SEP' column of the DataFrame.

Other tokens are assigned ‘N’ (non-boundary) labels and also added to the column.

Table 5 presents a sample of data extracted from the LST20 corpus. The column labeled ‘Clause’ indicate the annotation of Beginning, Inside, End (BIE) tagging scheme of clauses found within the corpus. Labels for sentence and clause boundary are included in the ‘SEP’ column.

Tokens	Clause	SEP
[…, **ทำให้**, คาดการณ์, **ว่า**, ใน, การ, ประชุม, ก, ., ตร., ใน, วันที่, 13, ส.ค., นี้, **จะ**, พิจารณา, ของ, มติ, ก, ., ตร., **เพื่อ**, ขอ, เลื่อน, การ, ประกาศ, พ.ร.ฎ., แบ่ง, ส่วน, ราชการ, สำนักงาน, ลง, ใน, ราชกิจจานุเบกษา, ออก, ไป, ก่อน, **จาก**, กำหนด, เดิม, ที่, **ก**, ., ตร., มี, มติ, ให้, ประกาศ, ใน, วันที่, 15, ส.ค., **และ**, มี, ผล, ใน, วันที่, 16, ส.ค., **จึง**, เป็น, **ที่**, จับตา, กัน, ว่า, **การ**, แต่งตั้ง]	[…, **B_CLS**, E_CLS, **B_CLS**, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, E_CLS, **B_CLS**, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, E_CLS, **B_CLS**, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, E_CLS, **B_CLS**, I_CLS, I_CLS, E_CLS, **B_CLS**, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, E_CLS, **B_CLS**, I_CLS, I_CLS, I_CLS, I_CLS, I_CLS, E_CLS, **B_CLS**, E_CLS, **B_CLS**, I_CLS, I_CLS, E_CLS, **B_CLS**, I_CLS]	[…, **B_CLS**, N, **B_CLS**, N, N, N, N, N, N, N, N, N, N, N, **B_Sentence**, N, N, N, N, N, N, **B_CLS**, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, **B_Sentence**, N, N, N, **B_CLS**, N, N, N, N, N, N, N, N, N, **B_CLS**, N, N, N, N, N, N, N, **B_Sentence**, N, **B_CLS**, N, N, N, **B_CLS**, N]

Table 5: Sentence and clause boundaries labeled as ‘B_Sentence’ and ‘B_CLS’

In text classification tasks, sentence beginnings and dependent clause beginnings would typically be distinguishable by spaces, either sentence-breaking or non-sentence breaking spaces. By removing all spaces during data preprocessing, we force sentence and clause beginnings to share the same token representation, which makes it more challenging for the model to differentiate between them.

For model fine-tuning, we employ the WangchanBERTa model, following the same procedure as the previous experiment. After predicting the test set, we convert all instances of the ‘B_Sentence’ label to the ‘B_CLS’ label, as sentence beginnings are also clause beginnings. We assess the model’s precision, recall, and F1-score on the positive class using the classification report function in scikit-learn.

Additionally, we conduct another experiment where we only train on the ‘B_CLS’ labels and evaluate the model’s performance. By doing so, we could determine if there is a significant difference in performance between training on only ‘B_CLS’ labels versus training on both ‘B_CLS’ and ‘B_Sentence’ labels, and the contribution of labeling sentence beginnings separately from clause beginnings.

4.3 XLM-RoBERTa Cross-Lingual Transfer

In our final experiment we investigate the effectiveness of cross-lingual transfer using the *xlm-roberta-base* model (Conneau et al., 2020). In this experiment, we incorporate the training data for English and Chinese, and augment the data for Thai. For English, we utilize the Brown Corpus, which consists of 57,340 sentences that are tokenized and annotated with part-of-speech (POS) tags in tuple pairs. We employ the Natural Language Toolkit (NLTK) library (Bird et al., 2009) to download and preprocess the corpus, converting the data into a pandas DataFrame. Each word is stored in the ‘TOKENS’ column, and the corresponding POS tag is stored in the ‘SEP’ column. We then relabel all non-punctuation POS tags as ‘N’ and punctuation tags as ‘SEP’. The data from the Brown Corpus is presented in Table 6.

Tokens	SEP
[Superior, Court, Judge, Durwood, Pye, to, ...	[N, N, N, N, N, N, N, N, N, N, N, N, N, N, ...
[``, are, outmoded, or, inadequate, and, often...	[N, N, N, N, N, N, N, N, SEP, N, N, N, N, N, N...
[proposed, However, ,, the, jury, said, it, be...	[N, N, N, N, N, N, N, N, N, N, N, N, N, N, ...
[jury, It, urged, that, the, next, Legislature...	[SEP, N, N, N, N, N, N, N, N, N, N, N, N, N, N...

Table 6: Some examples of the preprocessed Brown Corpus data

To ensure proper formatting of the data for cross-lingual transfer, we implement an additional step in which the final punctuation token in each sentence is removed, and the ‘SEP’ label is assigned to the previous token. This step is necessary for all datasets in this experiment, as punctuation is not commonly used to indicate sentence boundaries in Thai. Furthermore, spaces are also removed during this step as they are not considered as tokens in English.

We also incorporate the GenericsKB-Best dataset (Bhakthavatsalam et al., 2020), which contains 1,020,868 untokenized sentences, to investigate the effect of input size on the model’s performance. For this dataset, we use the nltk.word_tokenize() function to tokenize each sentence into a list of tokens. We then converted the data into a pandas DataFrame, with each word stored in the ‘TOKENS’ column. We label the last token of the list as ‘SEP’ if it was a punctuation mark and ‘N’ otherwise, and implement the removal of the final punctuation token in each sentence.

For the Chinese language, we used a subset of the WMT19 corpus (Wikimedia Foundation, 2019), consisting of 1,050,000 sentences. We tokenize each sentence using the jieba Chinese tokenizer (fxsjy, 2020) into a list of tokens and store the resulting tokens in a pandas DataFrame. We follow the same labeling approach as with the GenericsKB-Best dataset. The preprocessed Chinese text can be seen in Table 7.

[不准, 的, 东西, 来, 帮助, 某人, 表演, 基本上, 很, 精彩, --, 我, …	[N, N, N, N, N, SEP, N, N, N, N, N, N, N, N, N...
[结束, 后, , , 众人, 期待已久, 的, 园游会, 终于, 正式, 开锣, , , 美味…	[N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, ...
[有, 一张, 床, , , 一面镜子, , , 一张, 椅子, 以及, 一位, 穿着, 内衣,…	[N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, ...
[的, 嫡传, 技艺, , , 150, 多年, 来, 他们, 家族, 一直, 都, 是, 演…	[N, N, N, N, N, N, N, N, N, N, N, N, SEP, N, N...

Table 7: Some examples of the preprocessed WMT19 data

Since the LST20 corpus only contains 74,180 sentences, to increase the amount of training data, we introduce another dataset for Thai, the machine translation parallel corpus scb-mt-en-th-2020 (VISTEC-depa, 2020), which contains an additional 433,530 Thai sentences sourced from generated reviews. We used the PyThaiNLP library to tokenize the sentences and followed the same approach as for the English and Chinese datasets to store the resulting data in a pandas DataFrame.

Each dataset is divided into three subsets: a training set comprising 80% of the data, a development set with 10%, and a test set with 10%. The same percentage split is applied to all six datasets (including LST20).

The XLM-RoBERTa model is fine-tuned and evaluated using the same methods as in the second experiment, and we only evaluate its performance on the LST20 corpus on the positive class. The evauluation will encompass different settings, including a monolingual and multilingual context. The model will be evaluated on English, Thai-English, Thai-Chinese datasets, for example.

CHAPTER 5
RESULTS AND DISCUSSION

5.1 WangchanBERTa vs XLM-RoBERTa

In our experiments, we observed that the WangchanBERTa model, a large language model pre-trained on Thai data, outperformed other models in terms of F1-score on the positive class on the LST20 corpus. In contrast, the XLM-RoBERTa model, which was pre-trained on 100 different languages, had a slightly lower F1-score than WangchanBERTa. Nevertheless, both large language models surpassed the Maximum Entropy baseline model, which had the lowest F1-score. These results suggest that the contextualized word embeddings in the large language models enable them to acquire more comprehensive and diverse representations of each word via token embeddings.

One possible reason for the weaker performance of the multilingual XLM-RoBERTa model relative to WangchanBERTa is that it was pre-trained on only 71.7GB of Thai data from a single source, CommonCrawl, whereas WangchanBERTa utilized 78.5 GB of Thai data from various sources, such as the Thai-language Wikipedia, news articles, social media posts, subtitles from the OpenSubtitles project, and other publicly available datasets. Moreover, pre-training on data from more than 100 languages can limit the model's ability to capture the nuances of Thai language usage and topic diversity in the training data (Conneau et al., 2020; VISTEC-depa, 2021).

Different pretraining parameters and segmentation may also attribute to the superior performance. WangchanBERTa employed four levels of tokenization, including subword-level tokenization using SentencePiece, word segmentation using maximal matching algorithm, syllable segmentation using maximal matching algorithm, and machine learning-based word segmentation using SEFR. The precision, recall, and F1-score on the postive class for each model are summarized in Table 8.

Models	Conditions	Precision	Recall	F1-score
Maximum Entropy (Baseline)	Only spaces	0.7156	0.3742	**0.4914**
	No spaces	0.7084	0.2456	0.3647
	With spaces	0.7087	0.3721	0.4880
WangchanBERTa	Only spaces	0.4695	0.6605	0.5489
	No spaces	0.8971	0.8762	**0.8865**
	With spaces	0.8558	0.6316	0.7268
	Overlapping no spaces	0.8874	0.7379	0.8058
	Overlapping spaces	0.9120	0.7548	0.8260
	B_CLS	0.8679	0.8527	0.8603
	B_CLS & B_Sentence	0.8710	0.8526	**0.8617**
XLM-RoBERTa	Thai (LST20)	0.8294	0.6498	**0.7287**
	Thai (LST20+VISTEC)	0.8242	0.6423	0.7220
	Thai (LST20) + English (Brown)	0.8289	0.6375	0.7207
	Thai (LST20) + English (Brown+Generics)	0.8231	0.6454	0.7235
	English (Brown)	0.5527	0.1527	0.2393
	English (Brown+Generics)	0.5441	0.3220	0.4045
	Chinese	0.3489	0.3434	0.3461
	Chinese + Thai (LST20+VISTEC)	0.8333	0.6190	0.7104
	All datasets	0.8343	0.6362	0.7219

Table 8: The precision, recall, and F1-score of the models in our experiments

5.2 Input Information on WangchanBERTa

The ‘only spaces’ task refers to the classification of spaces as sentence boundaries, without any information about the tokens within the sentence. Unsurprisingly, this task is the most challenging as it requires the model to determine sentence boundaries purely based on the presence of spaces in the input text, and achieved 54.89% on F1-scores. As a result, it is not surprising that this task obtained the lowest F1 score among all the tasks.

In contrast, we can see that token classification without spaces achieved the F1-score of 88.65% which is highest among all the tasks. This indicates that incorporating token information significantly improves the model’s ability to identify sentence boundaries. While the overlapping tasks with and without spaces also demonstrate that providing the model with more contextual information through overlapping sentences leads to improved performance, achieving more than 80% of F1.

Overall, these results suggest that incorporating token information and contextual information to fine-tune WangchanBERTa can significantly improve its performance of sentence segmentation.

5.3 Sentence vs Clause Labeling

The results of the fine-tuned model indicate that when considering only clause boundaries (‘B_CLS’), the model achieved an F1-score of 86.03%. When both clause and sentence boundaries (‘B_CLS’ and ‘B_Sentence’) were considered, the model achieved a slightly higher F1-score of 86.17%. These results suggest that there is no significant difference in performance between training solely on ‘B_CLS’ labels and training on both ‘B_CLS’ and ‘B_Sentence’ labels. However, labeling sentence beginnings separately from clause beginnings may contribute to better performance in joint learning for clause and sentence segmentation, albeit marginally.

In comparison to Intasaw (2013), where the author achieved an F1-score of 84.74% using the SVM model on a corpus of Thai academic written language consisting of 8,102 clauses, our approach achieved competitive performance with minimal feature requirements. While Intasaw's model relied on various features such as POS tags, discourse markers, spaces, and punctuation marks, our approach does not require many features. However, due to the differences in the datasets and definitions of clause or EDU boundaries used, a direct comparison may not be appropriate. Nevertheless, our results suggest that the fine-tuned model can achieve competitive performance in Thai clause and sentence segmentation.

5.4 Multilingual Approaches

The experiment results indicate that cross-lingual transfer for Thai sentence segmentation does not benefit significantly from additional data from English and Chinese. Surprisingly, the best performing model was fine-tuned solely on the LST20 corpus. This result implies that the additional data from English and Chinese may not be of high quality and the heuristic methods used for tokenization may not be effective. One possible explanation for this is that while languages like English and Chinese may share some syntactic features with Thai, the contextualized syntactic features of these languages may not be useful for Thai sentence segmentation. Thus, the heuristic methods and tokenizers used for these languages may not be sufficient to capture the nuanced features of Thai language.

Regarding the size of the input, the results suggest that larger input sizes do not necessarily lead to better performance. For instance, the addition of 433,530 sentences from the scb-mt-en-th-2020 dataset did not improve the F1-score of the existing LST20 corpus which only consists of 74,180 sentences. This finding indicates that the quality of the corpus is more important than the size. In other words, better quality datasets with dedicated human annotation

may be more effective for cross-lingual transfer than larger datasets that rely on heuristic methods and the publicly available tokenizers.

However, it is worth noting that the model was still able to classify tokens using the English or Chinese data alone, possibly by leveraging the contextualized syntactic features of these languages. Nevertheless, the results suggest that cross-lingual transfer from English and Chinese to Thai may not be as effective for sentence segmentation.

CHAPTER 6
CONCLUSION

In our study, we aimed to assess the performance of large language models in Thai sentence segmentation by conducting three experiments: (1) evaluating the effect of different input information on WangchanBERTa's performance through five classification tasks, (2) performing joint learning for clause and sentence segmentation, and (3) investigating the effectiveness of cross-lingual transfer using XLM-RoBERTa.

We found that transformer-based large language models with contextualized word embeddings outperformed the baseline Maximum Entropy classifier, and language-specific WangchanBERTa outperformed the multilingual XLM-RoBERTa model. Fine-tuning WangchanBERTa by incorporating token and contextual information significantly improved its performance in sentence segmentation, while the model can already achieve competitive performance in both Thai clause and sentence segmentation. However, cross-lingual transfer from English and Chinese to Thai was not effective for the task.

Future research could explore incorporating additional linguistic features annotated in the LST20 corpus, including part-of-speech tags and named entities. Additionally, future studies could also expand the evaluation to include other Thai corpora such as the ORCHID and UGWC datasets and evaluate the generalizability of the models.

REFERENCES

Alqinai, J. B. S. (2015). Mediating Punctuation in English Arabic Translation. *Journal of Applied Linguistics and Professional Practice*, *5*(1), 5–29. https://doi.org/10.1558/japl.v5i1.5

Aroonmanakun, W. (2007). Thoughts on Word and Sentence Segmentation in Thai. *Proceedings of the Seventh Symposium on Natural Language Processing, Dec 13-15, 2007, Pattaya, Thailand*, 85–90. http://pioneer.netserv.chula.ac.th/~awirote/ling/snlp2007-wirote.pdf

Bhakthavatsalam, S., Anastasiades, C., & Clark, P. (2020). *GenericsKB: A Knowledge Base of Generic Statements*. https://huggingface.co/datasets/generics_kb

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Boonkwan, P., Luantangsrisuk, V., Phaholphinyo, S., Kriengket, K., Leenoi, D., Phrombut, C., Boriboon, M., Kosawat, K., & Supnithi, T. (2020). The Annotation Guideline of LST20 Corpus. *arXiv*. https://doi.org/10.48550/arXiv.2008.05055

Carlson, L., Marcu, D., & Okurovsky, M. E. (2001). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*. https://aclanthology.org/W01-1605.pdf

Charoenpornsawat, P., & Sornlertlamvanich, V. (2001). Automatic Sentence Break Disambiguation for Thai. *International Conference on Computer Processing of Oriental Languages (ICCPOL)*, *33*, 231–235. http://www.cs.cmu.edu/~paisarn/papers/old/iccpol2001.pdf

Charoensuk, J., Suvakree, T., & Kawtrakul, A. (2005). Thai Elementary Discourse Unit Segmentation by Discourse Segmentation Cues and Syntactic Information. *The Sixth Symposium on Natural Language Processing*. https://citeseerx.ist.psu.edu/pdf/0d21ab73b44872434358d042a4bc8eb4c303c62a

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *arXiv*. https://arxiv.org/pdf/1911.02116.pdf

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics* (6th ed.). Blackwell Publishing.

Donabauer, G., Kruschwitz, U., & Corney, D. (2021). Making Sense of Subtitles: Sentence Boundary Detection and Speaker Change Detection in Unpunctuated Texts. *Companion Proceedings of the Web Conference 2021 (WWW'21 Companion)*, 357–362. https://doi.org/10.1145/3442442.3451894

fxsjy. (2020). *jieba*. https://github.com/fxsjy/jieba

Huffman, F. E. (1970). *Cambodian System of Writing and Beginning Reader with Drills and Glossary*. Yale University Press. http://www.pratyeka.org/csw/hlp-csw.pdf

Intasaw, N. (2013). *Kan yaek anuphak phasathai duai kan chai baepchamlong sapphot wektoe maechin* [Thai Clause Segmentation Using a Support Vector Machine Model] [Master's thesis, Chulalongkorn University]. Chulalongkorn University Intellectual Repository (CUIR).

Intasaw, N., & Aroonmanakun, W. (2013). Basic Principles for Segmenting Thai EDUs. Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27),

Jenny, M., & Tun, S. S. H. (2016). *Burmese: A Comprehensive Grammar*. Routledge.

Ketui, N., Theeramunkong, T., & Onsuwan, C. (2013). Thai Elementary Discourse Unit Analysis and Syntactic-based Segmentation. *Information: An International Interdisciplinary Journal*, *16*(10), 7423–7436.

Lertpiya, A., Chaiwachirasak, T., Maharattanamalai, N., Lapjaturapit, T., Chalothorn, T., Tirasaroj, N., & Chuangsuwanich, E. (2018). A Preliminary Study on Fundamental Thai NLP Tasks for User-generated Web Content. *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (ISAI-NLP)*. https://doi.org/isai-nlp.2018.8692946

Li, F., Lv, H., La, D., Yong, B., & Zhou, Q. (2022). Sentence Boundary Disambiguation for Tibetan Based on Attention Mechanism at the Syllable Level. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(6), 1–18. https://doi.org/10.1145/3527663

Longchupole, S. (1995). *Kan wikhro prayok phasathai chak yona phuea kan plae duai phasa khomphiotoe* [Thai Syntactical Analysis System by Method of Splitting Sentences from Paragraph for Machine Translation] [Master's thesis, King Mongkut's Institute of Technology Ladkrabang]. WebOPAC KMITL Central Library. https://opacimages.lib.kmitl.ac.th/medias/b00148073/ศังกรศรัณย์%20ล่องชูผล.pdf

Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., & Nutanong, S. (2021). WangchanBERTa: Pretraining transformer-based Thai Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2008.05055

Lowphansirikul, L., Polpanumas, C., Rutherford, A. T., & Nutanong, S. (2020). scb-mt-en-th-2020: A Large English-Thai Parallel Corpus. *arXiv*. https://doi.org/10.48550/arXiv.2007.03541

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, *8*(3), 243–281. https://doi.org/10.1515/text.1.1988.8.3.243

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

Namprikprakdee. (2020, May, 8). *Prawat lae khwamsamkhan khong namphrik to sangkhom thai Namphrik mi ma tangtae samai krung si-ayutthaya…* [History and importance of Nam Phrik to Thai society. Nam Phrik has been around since the Ayutthaya period…] [Image attached] [Facebook post]. Facebook. https://www.facebook.com/namprikprakdee/photos/a.108111790897739/108133370895581

Ngarmwirojki, C., & Luksaneeyanawin, S. (2013). Kansueksa khwamsamat thang phasa nai kan wenwak chak kan khian phasathai khong nakrian thai lae yipun [A Study of spacing in Thai writing by Thai and Japanese students]. *Damrong Journal of The Faculty of Archaeology Silpakorn University*, *12*(2), 57–84. https://so01.tci-thaijo.org/index.php/damrong/article/view/21542/18588

Office of the Royal Society. (2008). *Lakken kan chai khrueangmai wakton lae khrueangmai uen uen Lakken kan wenwak Lakken kan khian khamyo Chabap ratchabandittayasathan* [Guidelines for using punctuation and other marks, Guidelines for spacing, Guidelines for abbreviation, Royal Institute Edition] (7th ed.). Aroon Karnpim.

Prasithrathsint, A., Hoonchamlong, Y., & Savetamalya, S. (2011). *Thritsadi waiyakon* [Grammatical Theories] (3rd ed.). Chulalongkorn University Press.

Radford, A. (2009). *Analysing English Sentences*. Cambridge University Press.

Read, J., Dridan, R., Oepen, S., & Solberg, L. J. (2012). Sentence Boundary Detection: A Long Solved Problem? *Proceedings of COLING 2012: Posters*, 985–994. https://aclanthology.org/C12-2096.pdf

Rojana-Anun, S. (2019). Ponctuation en thaï : emplois prescrits et pratique actuelle [Punctuation in Thai: prescribed usages and current practice]. *Bulletin de l'ATPF*, *42*(1), 1–20. https://doi.org/10.14456/bulletin-atpf.2019.1

Saetia, C., Chuangsuwanich, E., Chalothorn, T., & Vateekul, P. (2021). Semi-supervised Thai Sentence Segmentation Using Local and Distant Word Representations. *Engineering Journal*, *25*(6), 15–33. https://doi.org/10.4186/ej.2021.25.6.15

Sheik, R., Adethya, G., & Nirmala, S. J. (2022). Efficient Deep Learning-based Sentence Boundary Detection in Legal Text. *Proceedings of the Natural Legal Language Processing Workshop 2022*, 208–217. https://aclanthology.org/2022.nllp-1.18.pdf

Simmala, B., & Poomsan Becker, B. (2003). *Lao for Beginners*. Paiboon Publishing.

Sirirattanajakarin, S., Jitkongchuen, D., & Intarapaiboon, P. (2020). BoydCut: Bidirectional LSTM-CNN Model for Thai Sentence Segmenter. *2020 1st International Conference on Big Data Analytics and Practices (IBDAP)*, 1–4. https://doi.org/10.1109/ibdap50342.2020.9245454

Slayden, G., Hwang, M., & Schwartz, L. (2010). Thai Sentence-Breaking for Large-Scale SMT. *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing*, 8–16.

Sornlertlamvanich, V., Charoenporn, T., & Isahara, H. (1997). ORCHID: Thai Part-Of-Speech Tagged Corpus. *National Electronics and Computer Technology Center Technical Report*, 5–19.

Srinivasan, S., & Dyer, C. (2021). Better Chinese Sentence Segmentation with Reinforcement Learning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 293–302. https://aclanthology.org/2021.findings-acl.25.pdf

Srisawang, W. (n.d.). *Chut khwamru 5 Akson thainoi kap phasa lao*. http://www.esansawang.in.th/esanweb/es3_text/tn_andlaos.pdf

Thiengburanathum, P. (2021). A Comparison of Thai Sentence Boundary Detection Approaches Using Online Product Review Data. In L. Barolli, K. F. Li, T. Enokido, & M. Takizawa (Eds.), *Advances in Networked-Based Information Systems: The 23rd International Conference on Network-Based Information Systems (NBiS-2020)* (pp. 405–421). Springer. https://doi.org/10.1007/978-3-030-57811-4_40

Thong, T. (1985). Language planning and language policy of Cambodia. In D. Bradley (Ed.), *Papers in South-East Asian linguistics No. 9: Language policy, language planning and sociolinguistics in South-East Asia* (pp. 103–117). Pacific

Linguistics. https://openresearch-repository.anu.edu.au/bitstream/1885/145086/1/PL-A67.pdf

VISTEC-depa. (2020). *English-Thai Machine Translation Dataset*. https://airesearch.in.th/releases/machine-translation-datasets

VISTEC-depa. (2021). *WangchanBERTa: Pre-trained Thai Language Model*. https://airesearch.in.th/releases/wangchanberta-pre-trained-thai-language-model

Wang, H., Wang, J., Shen, Q., Xian, Y., & Zhang, Y. (2020). Maximum entropy Thai sentence segmentation combined with Thai grammar rules correction. *UPB Scientific Bulletin, Series C: Electrical Engineering and Computer Science*, *8*(1). https://www.scientificbulletin.upb.ro/rev_docs_arhiva/fullf88_290347.pdf

Wang, H., Zhang, Z., Shen, Q., Xian, Y., Zhang, Y., & Mao, C. (2019). Thai Language Sentence Segmentation Based on N-Gram Context Model. *IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS)*, 813–817. https://doi.org/10.1109/iucc/dsci/smartcns.2019.00165

Wicks, R., & Post, M. (2021). A unified approach to sentence segmentation of punctuated text in many languages. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, *1*, 3995–4007. https://aclanthology.org/2021.acl-long.309.pdf

Wikimedia Foundation. (2019). *ACL 2019 Fourth Conference on Machine Translation (WMT19), Shared Task: Machine Translation of News*. https://huggingface.co/datasets/wmt19

WorkpointOfficial. (2022). *To nu maem | EP.208 Churi kap prakotkan dang sut chut mai yu | 17 phoyo 65 | Full EP* [ToNuMaem | EP.208 Juree and the Epic Sensation | 17 Nov 22 | Full EP] [Video]. YouTube. https://www.youtube.com/watch?v=PnuEcg9G-AI

Xue, N., & Yang, Y. (2011). Chinese sentence segmentation as comma classification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 631–635. https://aclanthology.org/P11-2111.pdf

Yuenyong, S., & Sornlertlamvanich, V. (2022). TranSentCut – Transformer Based Thai Sentence Segmentation. *Songklanakarin Journal of Science an Technology (SJST)*, *3*(44), 852–860. https://doi.org/10.14456/sjst-psu.2022.114

Zhou, N., Aw, A. T., 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. https://aclanthology.org/C16-1031.pdf

VITA

NAME Narongkorn Panitsrisit

DATE OF BIRTH 19 August 1996

PLACE OF BIRTH Songkhla