RESOLVING THAI ZERO PRONOUN USING MASKED LANGUAGE MODEL

Miss Sumana Sumanakul

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

ไขสรรพนามไร้รูปภาษาไทยโดยใช้แบบจำลองทางภาษาแบบพรางคำ

น.ส.สุมนา สุมนะกุล

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

Independent Study Title	RESOLVING THAI ZERO PRONOUN USING MASKED LANGUAGE MODEL
By	Miss Sumana Sumanakul
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 THEERAPORN RATITAMKUL, Ph.D.)

สุมนา สุมนะกุล : ไขสรรพนามไร้รูปภาษาไทยโดยใช้แบบจำลองทางภาษาแบบพรางคำ. (RESOLVING THAI ZERO PRONOUN USING MASKED LANGUAGE MODEL) อ.ที่ปรึกษาหลัก : รศ. ดร.อรรถพล ธำรงรัตนฤทธิ์

การไขสรรพนามไร้รูปเป็นหนึ่งในงานที่ท้าทายในการประมวลผลภาษาธรรมชาติในภาษาไทย อย่างไรก็ตามงานศึกษาในหัวข้อดังกล่าวในทางภาษาศาสตร์คอมพิวเตอร์นั้นยังไม่เป็นที่แพร่หลายและยังไม่มีการนำข้อมูลภาษาไทยมาทดลองด้วยวิธีการใหม่ ๆ จากวิทยาการทางด้านนี้ ด้วยเหตุนี้ผู้วิจัยจึงสนใจประยุกต์แบบจำลองทางภาษาที่ผ่านการฝึกฝนมาแล้วจากสถาปัตยกรรมแบบทรานฟอร์เมอร์ ซึ่งเป็นวิธีใหม่ที่มีความแม่นยำสูงที่สุดในการทำงานประมวลผลภาษาธรรมชาติรูปแบบต่าง ๆ และยังสามารถใช้งานกับชุดข้อมูลที่หลากหลาย เพื่อมาใช้ในการไขสรรพนามไร้รูปภาษาไทย ผู้วิจัยทำการทดลองกับชุดข้อมูลขนาดเล็ก โดยออกแบบเป็น 2 การทดลอง คือ (1) ใช้แบบจำลองทางภาษาแบบพรางคำที่ผ่านการฝึกฝนมาแล้วเพื่อทำนายคำสรรพนามไร้รูป และ (2) ปรับแต่งการจำแนกคำในโมเดล Wangchanberta เพื่อให้จำแนกบุรุษของสรรพนามไร้รูป ผลลัพธ์จากการทดลองทั้งสองแสดงให้เห็นถึงประสิทธิภาพของแบบจำลองทางภาษาที่ผ่านการฝึกฝนมาแล้ว ที่ไม่เพียงแค่สามารถจับคุณลักษณะทางไวยากรณ์ของคำสรรพนามไร้รูปในภาษาไทยได้ แต่ยังสามารถเข้าใจระบบการเลือกใช้คำสรรพนามภาษาไทยในระดับปริจเฉทอีกด้วย

สาขาวิชา	ภาษาศาสตร์	ลายมือชื่อนิสิต .....
ปีการศึกษา	2565	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6380049722 : MAJOR LINGUISTICS
KEYWOR
D:

Sumana Sumanakul : RESOLVING THAI ZERO PRONOUN USING MASKED LANGUAGE MODEL. Advisor: Assoc. Prof. Attapol Thamrongrattanarit, Ph.D.

Zero pronoun resolution is an actively challenging NLP task in Thai. However, only a few previous studies have focused on this topic. Therefore, we explore a modern approach that could outperform existing state-of-the-art methods on various datasets and downstream tasks, the transformer-based, pre-trained language model, to apply to the Thai zero pronoun resolution task. We conduct two experiments on a small corpus, which are (1) using a pre-trained masked language model to predict zero pronominal expressions and (2) fine-tuning Wangchanberta on a token classification task to classify persons of pronouns. Based on our experiments, the results demonstrate the effectiveness of the pre-trained language model (1), which successfully encodes not only the grammatical features but also the system of Thai pronoun usage at the discourse level.

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

ACKNOWLEDGEMENTS

I would like to extend my gratitude to my advisor, Associate Professor Attapol Thamrongrattanarit, Ph.D. This project would not have been possible without your exceptional support, and your useful advice and comments were tremendously helpful for the development of this study. In this regard, I am eternally grateful to you.

And I would like to thank all the professors and faculty members in the Linguistics Department for the opportunity to learn and gain knowledge and skills in linguistics, which fulfils both my academic and career aspiration.

This project could not have been accomplished without the support of Mr. Apiwat Jaroonpol and Mr. Nozomi Yamada. Thank you very much for your valuable guidance and suggestions regarding coding. My special thanks go to Miss Manta Klangboonkrong for her guidance and comments in the writing of this project. I also would like to thank all my friends who eagerly lent their help and support during my academic years.

Last but not least, I am extremely grateful to my loving, caring and supportive family: my parents, my aunt and my brother for their understanding and financial support throughout my academic pursuit.

Sumana Sumanakul

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Like Japanese and Chinese, Thai is a pro-drop language. These languages allow pronouns, which are arguments of predicates (or verbs), to be omitted in the finite clause on the surface structure. We call it *'Zero Pronoun (ZP)*, which is analyzed as one of '*Empty Category (EC)*[1], where ZP is considered as a pro (little pro or small pro) in generative grammar. As a result of being [+pronominal], a 'pro' observes principle B in the binding theory[2]. It is free in the governing category (Aroonmanakun, 2002). Sometimes it is represented by *Φ* in the phrase structure tree, for example,

วันนี้ Φ_1 เลิกกี่โมง Φ_2 จะชวน Φ_1 ไปเล่นเกม

What time do you get off work today? I'd like to invite you to play some games afterwards.

The above utterance contains two zero pronouns shown as Φ_1 and Φ_2, of which their non-zero forms are เธอ (you) which refers to the listener and ฉัน (I) which refers to the speaker, respectively.

From a computational point of view, this linguistic feature presents many challenges for the downstream tasks in Natural Language Processing as a lot of NLP

[1] 'Empty Category (EC)' is the term used in generative grammar that refers to an element that does not have a lexical item or any phonological content in a certain syntactic position.

[2] The interpretation and distribution of pronouns and anaphors contains three principles:

Principle A: An anaphor must be bound in its binding domain.

Principle B: A pronoun must be free in its binding domain.

Principle C: An R-expressions must be free (Chomsky, 1981, 1986).

tasks require us to identify all of the arguments for each predicate. For example, relation extraction, information extraction or text categorization (Yeh & Chen, 2003) depends on the patterns of the dependency of the verbs and their arguments. Hence, we would typically find many missing arguments for the verbs due to empty categories. Especially in machine translation (Taira, Sudoh, & Nagata, 2012), one must recover and resolve the missing arguments (or the theta roles) when translating from pro-drop languages to non-pro-drop languages, which more strictly require all of the arguments to be present on the surface. In other words, the source language contains less information of arguments, i.e., subject or object, and could potentially cause mistranslation as illustrated,

The source language (Thai):

พี่สาวเรา$_1$ ทำ ขนมจีน$_2$ อร่อยมาก แต่ช่วงนี้ Φ_1 ไม่ค่อยได้ทำ Φ_2 เลย

Translation (English):

My sister$_1$ made Kanom Jeen$_2$, very delicious, but *I*$_3$ haven't made *it*$_2$ lately.

Translation (German):

Meine Schwester$_1$ hat Kanom Jeen$_2$ gemacht, sehr lecker, aber *ich*$_3$ habe *es*$_2$ in letzter Zeit nicht gemacht.

Translated by Google Translate

In the source language's sentence, the subject or Agent role in the subordinate clause refers to *พี่สาวเรา (my sister)* in the main clause but it is not shown on the surface. However, in both translated versions from Google Translate in English and German, the non-pro-drop languages that do not permit clauses without subjects, the Agent is misinterpreted as *I*, which makes the sentence become semantically wrong compared to the original meaning. Accordingly, the data preprocessing step dealing with zero pronoun is seemingly required in a pro-drop language before being used as an input to some other downstream NLP applications afterwards.

In NLP, the task of recognizing dropped pronouns and their antecedents is called '*Zero Pronoun Resolution*[3].' This task is generally composed of two steps i.e. zero pronoun detection and referent identification (Yoshida & Nagata, 2009). However, from our viewpoint, the process of resolving Thai zero pronoun can be divided into three subtasks: detecting the missing arguments, predicting the right pronouns, and determining their antecedents. In this work, we focus on the last two steps, assuming that we are given a zero element in the sentence. Pronouns in Thai encode more than grammatical features typically coded in pronouns such as person, numbers, and genders. Pronouns in Thai also encode the roles, relationships of the speech participants, their relative social status, the context of discourse, etc. (Hoonchamlong, 1991). These factors must be taken into consideration when finding the right antecedents.

Within the recent years, resolving zero pronoun through computational approaches has been studied broadly in other pro-drop languages such as Japanese, Chinese and Korean, and could be done in multiple ways using rule-based model, machine learning or deep learning, for instance, encoding zero pronouns and their candidate antecedents by LSTM, Attention or BERT (Chen et al., 2021), zero pronoun resolution with attention-based neural network (Yin, Zhang, Zhang, Liu, & Wang, 2018) and using ranking rules combined with machine learning (Isozaki & Hirao, 2003). To the best of our knowledge, Thai zero pronoun resolution with a computational approach has not yet been extensively explored. Despite being a fundamental task in pro-drop languages, no previous studies applied the modern deep learning approaches. Hence, we aim to test a large pretrained language model and contextualized word embeddings for Thai zero pronoun resolution. Considering that contextualized language models could capture the relational knowledge between pronouns and antecedents (or postecedents), we therefore use masked language model (MLM) to resolve zero pronouns. However, in this paper, we focus only on the process of resolving zero pronouns to the right preceding or following entities and

[3] The task of recognizing the antecedents of zero pronouns is generally called Zero Anaphora Resolution (ZAR); however, in our works, we use the term Zero Pronoun Resolution.

assume that the positions of zero pronouns are already determined. We explore how MLM could be applied for unsupervised zero pronoun resolution and further examine how pre-trained language models could possibly become an effective model to be developed as a part of an end-to-end system to resolve zero pronouns.

Our research contribution is summarized as follows:

- Our study is the first to attempt to apply MLM to solve Thai zero pronoun resolution and finds that this unsupervised approach has a potential to be improved further to perform this task as it can resolve inter-sentential, intra-sentential and exophoric cases.
- We find that the pre-trained language model could capture not only the grammatical relation of zero pronoun and its antecedent in Thai but also encode the system of Thai pronoun usage identified by sociolinguistic factors.
- Our analysis reveals that the process of detecting the grammatical person of Thai pronouns could be compiled with the end-to-end model of Thai zero pronoun resolution.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

This chapter reviews the background of zero pronoun in linguistic point of view and computational point of view, as well as previous works on zero pronoun resolution task in Thai and other pro-drop languages.

As mentioned, the phonetically null pronominal argument in a certain syntactic position is called *Zero Pronoun*. Zero pronouns could refer to a previously mentioned entity or an existing real-world entity. It is thus called *zero anaphora* which is one of the anaphora types.

In linguistics, *anaphora* is the coreference of an expression with its backward or forward reference which provides the information necessary for the expression's interpretation. Anaphora is a use of a pronoun referring to another expression which could be words or phrases. In a narrower sense, anaphora is the use of a pronoun that its referent precedes an anaphor. In contrast, *cataphora* is the use of a pronoun that an anaphor precedes its referent, i.e., postcedent. In NLP, the task of identifying the referent of an anaphor[4] is called *anaphora resolution*.

According to the government and binding theory (Chomsky, 1981, 1986), zero pronouns are analyzed as one of the empty categories (ECs), which are categorized in four types: wh-trace (variable), NP-trace, pro, and PRO, where zero pronoun is considered as a *pro* (*little pro* or *small pro*). Unlike *PRO* (*big PRO*), another type of pronominal ECs that represents non-overt DPs (or NPs) in non-finite clauses e.g. clauses headed by subject or object control verbs, the *pro* illustrates the omission of pronoun classes in finite sentences and only occurs in some languages called pro-drop languages. Here are some examples of PROs and pros,

[4] 'Anaphor' is a word or phrase used to refer to a word or phrase mentioned earlier in an utterance. Anaphor must have an antecedent.

Sam$_1$ promised *PRO$_1$* to study.

Risa$_1$ knows how *PRO$_{arb}$* to dance.

pro$_1$ เพิ่งไปกินข้าวมายังอิ่มอยู่เลย

เธอ$_1$ หลับไปก่อนได้เลยนะ เดี๋ยวตอนบ่ายเรา$_2$จะเดินไปปลุก *pro$_1$*

According to Chomsky's DP type (1982), pro is characterized by the features [-anaphor, +pronoun] that are similar to pronoun. It can be said that it is a null version of pronoun as it also behaves like an overt pronoun. Due to its feature of +pronominal, a pro falls under Principle B: A pronoun must be free in its binding domain.

Regarding Thai zero pronoun, Hoonchamlong (1991) pointed out that "Thai allows, quite freely, the omission of pronouns in both subject and object position" (Hoonchamlong, 1991, p. 71). Thai ECs in both positions do not seem to show a different distribution. In contrast to Chinese, Korean and Japanese, the occurrence of zero pronouns in object position has a more limited distribution than zero subject pronouns (Huang, 1984 as cited in Hoonchamlong, 1991). This argument was supported by the evidence from topicalization and relative clauses that proved that object zero pronouns are not variable but are pronouns. Due to the unrestricted use of zero pronouns, "a Thai sentence can be formed by only a verb phrase" (Kongwan, Kamaruddin, & Ahmad, 2022) which could become a problem in computational linguistics.

In addition, the analysis of Thai R-expression needs to be reviewed in this work. When considering the pronominal expression in Thai, R-expression needs to be included. R-expression or referring expression is an expression that describes noun phrases "e.g. proper names, particular professions (e.g., teacher) and kinship terms" (Larson, 2006). In English, R-expression, containing features [-anaphor, -pronoun], fall under Principle C: an R-expression must be free, but Thai R-expression behaves differently as it can receive bound readings as pronouns. According to Larson (2007), Thai bound R-expressions arise frequently but not in all domains because R-expression cannot be locally bound, that is, it can receive a bound reading when it is not a coargument with its antecedent, for example, **ศักดิ์$_1$ ตี ศักดิ์$_1$*. Apart from this,

exact copies of the antecedents are neither necessary nor sufficient for binding R-expression. Instead of full copies, the bound R-expression in Thai can be only a copy of their antecedent's head, for example *อาจารย์สิทธิ์1 บอกว่า อาจารย์1 ไม่ว่างพรุ่งนี้*. The head *อาจารย์* receives bound reading. This analysis of bound R-expression will become useful when evaluating the experiment results.

In NLP, zero pronoun resolution is an active challenging task in such pro-drop languages, yet this task is not extensively explored in Thai (Aroonmanakun, 2000). . Therefore we briefly introduce previous works on zero pronoun resolution in Thai, Chinese, Korean, Japanese and Spain through various methods.

Aroonmanakun (2003) has studied Thai zero pronoun resolution with a focus on resolving the missing pronouns on discourse structure. Based on the existing conventional approach for reference resolution in a discourse segment, Centering Theory, he extended centering algorithms as he indicated that antecedents of Thai zero pronouns are not always in the immediately preceding utterance, so Centering Theory alone is not sufficient. Developed with the concept of hierarchical structure, his extension was performed on the same corpus that was previously used for the former version of algorithms. However, it cannot be strongly concluded whether the hierarchical structure of discourse in the extended model is relevant with the long-distance discourse anaphora as most of antecedents in the corpus occur immediately in the preceding sentence. Consequently, further studies about Thai zero pronoun in NLP field are still challenging.

More details from a linguistic perspective are given in Referent resolution for zero pronouns in Thai, where Aroonmanakun (2002) provided the overview of identifying referents of Thai zero pronouns that could be solved in two levels: the sentence level and the discourse level. As stated, Thai zero pronoun is analyzed as Empty Categories (ECs) in Chomsky's government and binding theory framework, which is considered a gap in the s-structure. Zero pronoun is categorized as pro (little pro or small pro). In determining its antecedent, Principle B of binding theory could be therefore applied to pro as it is free in the governing category. However, zero pronouns in the discourse level e.g., pros and arbitrary PROs, that cannot be solved by the government and binding theory, could be resolved with discourse principles, the

Centering Theory. Centering Theory is a computational model with a focusing mechanism that is used for keeping track of salient entities called Centering, which exhibits local coherence in a discourse segment. Based on syntactic properties, centering rule and ranking of centerings lead to the identification of zero pronoun's referent. Understanding the linguistic knowledge of zero pronouns could lead to the development of an effective model by implementing these principles as part of the algorithm.

Apart from Centering Theory, there is a recent work exploring zero pronoun resolution that applies a different approach. Kongwan et al. (2022) has experimented with resolving Thai anaphora in *Anaphora Resolution in Thai EDU Segmentation*, where the zero anaphora (ZA) is defined in one of the four types of Thai anaphora that the researchers aimed to resolve: zero anaphora, pronominal anaphora, nominal anaphora, and ellipsis of the owner. Their methodology is divided into two main parts, which are Thai morphological analysis for data preparation and the anaphora resolution. Thai morphological analysis is formerly processed from Thai word segmentation to Thai EDU segmentation. The latter part is related to the algorithm of the entire process of anaphora resolution, which are (1) the rule-based anaphora algorithm used for indicating the anaphora types, and (2) the resolution for non-referential anaphora or referential anaphora used for tagging referential or non-referential anaphora and finding the reference. In this process, the semantic concepts such as the ontology of meronymy and hyponymy were utilized as features in the ranking model. Their overall result achieved a precision of 0.77, a recall of 0.84 and a good F1 score of 0.81, where the lowest scores belong to the zero anaphora type with a precision of 0.75 and a recall of 0.82. Although resolving all types of Thai anaphora is beyond the scope of our work, this work presents difficulties in the task of resolving Thai zero anaphora over other types of Thai anaphora.

Compared to Thai language, previous work in zero pronoun resolution abounds in Chinese and has leveraged more modern approaches. For example, Yin et al., 2018 utilized an attention mechanism to develop their self-attentive neural network architecture for anaphoric zero pronoun resolution modeling. Employing attention technique and using embeddings from the context around them, their model in recurrent neural network architectures was trained to attend important informative

parts of the mentions (noun phrases) in order to generate the attention-weight vectors of zero pronouns and encode their candidate mentions in the preceding and following sentences. To select the best scored candidate, they calculated the probability score of each antecedent candidate by using a two-layers feed-forward neural network. Their model yielded the best performance and surpassed the existing Chinese zero pronoun resolution baseline systems. The compelling point from this study is the attention mechanism that enhances the performance of the model by helping it focus on the informative parts of the contexts that represent zero pronouns while the previous works using deep neural network methods tried to encode zero pronouns into the semantic vector-space by additional elements and underutilized context of zero pronouns.

Another study that focuses on Chinese zero pronoun resolution was done by Chen et al. (2021). The researchers investigated the possibility of solving zero pronoun resolution and coreference resolution jointly via a neural model, while most existing works at that time tried to solve these two problems separately. To solve zero pronoun resolution, they introduced a gap-masked self-attention model to get the representation embeddings for tokens and gaps, which are believed to contain the contextual information from the surrounding tokens. They are also treated as candidate zero pronouns. After the model computes unit representations using transformer and gap-masked self-attention, unit score for each unit and a pairwise score for each pair of units and antecedents will be calculated. Considering the exclusive relationship between zero pronouns and mentions, the researchers proposed a two-stage interaction mechanism that adds relevant units to the unit scoring function to capture the interaction among unit representations and the interaction among unit scores. Their end-to-end neural model achieved state-of-the-art performance on both tasks.

Zero pronoun resolution is also an active research topic in Japanese. While the previous state-of-the-art works were done using statistical machine learning on the written corpus, Yoshida and Nagata (2009) conducted the experiment of statistical zero pronoun resolution on the corpus of spoken texts, where first-person and second-person subjects and objects are frequently omitted as in Thai. The researchers therefore particularly focused on distinguishing first-person subjects from others. The

corpus of spoken text was first annotated in each type of tags including predicate tag, event noun, coreference tag, bridging reference tag and case alternation tag. For anaphora classification, the labels of person are designed to attach to all first-person endophoras[5] and exophoras[6]. With the verbal features such as verbal semantic attribute (VSA), auxiliary verb, semantic category of the predicate and adnominal predicate, Support Vector Machine (SVM) with a linear kernel was used to identify persons of zero pronouns for binary labels. For zero pronoun resolution, statistical machine learning and ranking SVM are adopted to deal only with the process of referent identification. This process also utilizes the verbal features provided in the former experiment. Their results show that the accuracy of zero pronoun resolution can be improved by identifying the person of pronouns and adding the verbal features.

As for neural network approach, Iida, Torisawa, Oh, Kruengkrai, and Kloetzer (2016) used a Multi-column Convolutional Neural Network (MCNN), a variant of a Convolutional Neural Network (CNN), for predicting zero anaphoric relations for Japanese intra-sentential subject zero anaphora. The researchers only focused on intra-sentential subject zero anaphoric relation, in which a zero anaphor and its antecedent appear in the same sentence, because the inter-sentential and exophoric cases were extremely difficult to resolve at that time. Regarding the fact that the relative occurrence positions of a predicate and antecedent could be either anaphoric or cataphoric, the training instances were divided into two sets and trained independently. After extracting every pair of predicate and potential antecedent in the target sentence, MCNN predicts probability that indicates the likelihood of a zero anaphoric relation and rank the probabilities of all pairs. Compared to previous works that used word sequences as linguistic features, this work used word sequences as one of the independent columns in MCNN. Motivated by Centering Theory, the recency and saliency properties of a candidate antecedent are used as clues. In choosing the best candidate, following the greedy-style methods, the candidate from the top pair will be filled in the zero anaphora position. Despite having lower F-score than the state-of-the-art method, their proposed method achieved much higher precision in

[5] Pronouns that refer to something in the same text

[6] Pronouns that refer to something outside the text

strong baselines. Although MCNN can handle only intra-sentential cases, the zero anaphora resolution for inter-sentential and exophoric cases were performed later with a new approach.

Masked language model (MLM), which is a modern approach which has contributed to impressive performances of resolving zero pronouns, was further researched by Konno, Kiyono, Matsubayashi, Ouchi, and Inui (2022). In their study, the researchers designed a new pre-training task that specialized MLMs for Japanese zero anaphora resolution (ZAR) called the pseudo zero pronoun resolution (PZERO). Assuming that the same noun phrases that appear multiple times in a text were coreferent, the researchers masked one of them as a pseudo zero pronoun and the others as pseudo antecedents. Afterwards, the MLM was then trained to select the pseudo antecedent of the masked pseudo zero pronouns. With this task, MLM could specifically learn anaphoric relational knowledge more effectively. Moreover, they also proposed a new ZAR model with minimal architectural modifications and pre-trained parameters from MLM, for the model to utilize anaphoric relational knowledge for argument selection during fine-tuning. Their AS-PZERO model outperformed the previous state-of-the-art models in ZAR.

For Korean, a machine-learning-based model was also studied to resolve zero pronoun like other pro-drop languages. For example, Park, Lim, and Hong (2015) has examined a machine learning method to resolve Korean zero object in spoken texts. They proposed eight features based on this linguistic phenomenon to apply in ML method, such as parallelism of syntactic function, semantic relation between predicates, sentence distance, headedness of candidate antecedent and the most salient candidate antecedent. With these features, their accuracy was higher than the baseline.

To enhance the performance of Korean zero anaphora resolution technique, Kim, Ra, and Lim (2021) explored the transformer-based model like bidirectional encoder representations from transformers (BERT) to classify whether each of the input word can be an antecedent of zero pronoun subject. While zero pronoun detection relies on the syntactic analysis of sentences or binary classification of predicates, antecedent search seems to be more challenging and requires more intelligence. The researchers therefore focused on only the process of antecedent search. They developed an antecedent-search module by fine-tuning a pre-trained

BERT on a Korean ZAR tagged corpus, in which ZP predicates[7] and their antecedents were annotated manually. Based on a deep-learning model with a neural-network architecture on top of BERT, their search module performed binary classification for each input token into A (antecedent) or N (not antecedent). The input was a sequence of words and the output was a label sequence assigned to individual tokens. To summarize, their model performed better than other models in Korean ZAR and significantly improved the performance of ZAR, which could demonstrate that BERT is a powerful model that can capture the predicate-argument semantics of a language and deal with long-distance dependencies.

For a European language with pro-drop parameter like Spanish, Ferrández and Peral (2000) proposed the first algorithm for the resolution of Spanish zero-pronouns in texts which achieved better results over several earlier baselines on pronominal anaphora resolution. As zero pronouns are limited to only subject position, the partial parsing was applied on the text to detect the zero pronoun from the syntactic structure. To illustrate, when one cannot get the full syntactic information e.g., subject noun phrase of a clause, that position is detected as zero pronoun. After detection, the missing position was inserted by the pronoun through the computational system with indicative information such as person and number obtained from the clause verb and gender from the copulative verb. In terms of zero pronoun resolution process, the researchers adopted and improved the restrictions and preferences concept[8] to their algorithm of recovering the best candidates. First, the set of restrictions e.g., person and number agreement, c-command constraints and semantic consistency, were applied to the candidates. And the improved set of preferences e.g., preference for candidates in the same sentence, preference for proper nouns and preference for noun phrases that were not included in a prepositional phrase, were applied when the candidate remained more than one. Their approach was evaluated through detecting

[7] The predicate at which a ZP occurs is called the ZP predicate Kim et al. (2021).

[8] “Restrictions tend to be absolute and, therefore, discard any possible antecedents, whereas preferences tend to be relative and require the use of additional criteria” Ferrández and Peral (2000).

verbs and categorizing them into verbs with omitted subjects and verbs with overt subjects. The results showed that the detection of the first category of verbs performed better than the second for all first-person, second-person and third-person pronouns. To evaluate the anaphora resolution, the third-person zero pronouns were classified into three classes: cataphoric, exophoric and anaphoric. The accuracy result, compared to previous baselines on anaphora resolution, accomplished the improvement by their approach. Moreover, their proposed application can be adapted and applied to different genres of Spanish text.

Inspired by the recent and remarkable achievement of MLM applied to resolve zero anaphora (Konno et al., 2022), we focus on employing pre-trained MLM to resolve Thai zero pronouns due to the composition of contextualized representation as well as attention mechanisms which are believed to capture semantic properties and relations between null subjects or objects and their references as reviewed previously. Furthermore, since Yoshida and Nagata (2009) has confirmed that identifying persons of zero pronouns is a crucial step to increase the performance of ZAR, we also decided to perform an experiment on classifying persons of Thai zero pronouns. We fine-tune a pretrained RoBERTa-based models, one of the most popular variants of the BERT, for token classification task, as Kim et al. (2021) has described that BERT can capture the predicate-argument semantics of a language and the deep-learning attention mechanism can model long-distance dependencies in the text. Besides, Ferrández and Peral (2000) provided us with the significant idea of detecting the grammatical position of zero pronouns which we adapt to apply to our data.

CHAPTER 3

APPROACHES

In our scope, we confine the process of zero pronoun resolution to identify its referential entity and insert the right pronominal expression i.e., pronouns or R-expressions. Generally, detecting zero pronouns and inserting pronouns in their positions were examined with the same process, and resolving anaphora, where the antecedent was revealed e.g., selecting the best references from the candidates, was done afterwards (Ferrández & Peral, 2000). As the zero pronoun's position can be detected by an algorithm based on syntactic analysis or binary classification of predicates (Kim et al., 2021), in this work, assuming that the zero pronoun detection was done earlier, the process of recovering the overt pronoun from the detected position could help indicate its reference by the clues of agreement and other characteristic factors of Thai pronouns. Furthermore, we aim to explore whether the relational knowledge between zero pronouns and their references e.g. persons, could be captured by the contextualized language models. We then define our task as zero pronoun resolution, which is investigated in two experiments: (1) predicting the masked pronouns and (2) predicting the person of the masked pronouns.

In our first experiment, we use pre-trained MLM from transformer-based model to predict the masked token inserted among the zero pronoun positions on the preprocessed dataset and evaluate whether the pre-trained MLM can potentially replace the tokens with pronominal expressions i.e., pronouns or R-expressions. Concisely, mask language modeling is a pre-trained task of language modeling[9] in the transformers-based language model (LM) such as BERT. To make predictions, the modeling procedure in MLM allows the model to learn to attend to the surrounding contexts in bidirectional. We can directly utilize MLM in the pipeline module of the transformer model without fine-tuning data, as we assume that the corpus used for

[9] The task of fitting a model to a corpus

pretraining covers all domains in our dataset. With MLM, we expect the pronominal expression from the prediction, since all the masked tokens in our dataset represent zero pronouns. MLM takes sentences with masked tokens as input and generates the most probable substitution as candidate tokens with input ids ordered by probability scores as illustrated in Figure 1.

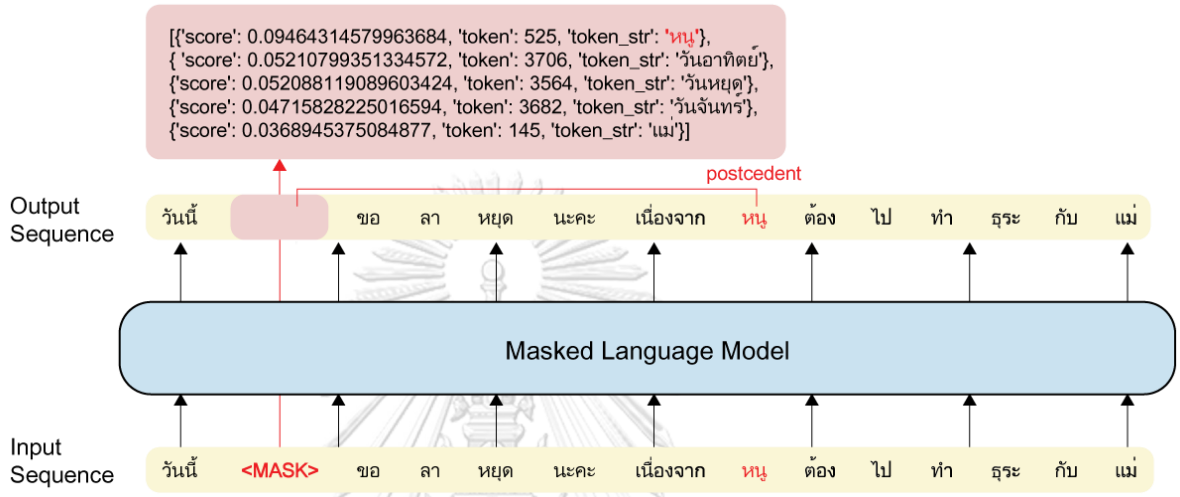

Figure 1 Configuration of MLM

In our second experiment, we fine-tune a pre-trained transformer-based language model for token classification to distinguish persons of pronouns. The masked tokens in our dataset are further annotated to represent the first-person, second-person, and third-person pronouns. In this task, the model assigns a label to individual tokens in the sequences. However, we focus only on our target labels for personal zero pronouns and ignore the outside labels when evaluating. We aim to test whether the model could particularly understand this grammatical feature of pronouns when they are dropped. Figure 2 depicts the overall configuration of the token classification task, from which the preprocessing step in detail will be described in the experiment chapter.

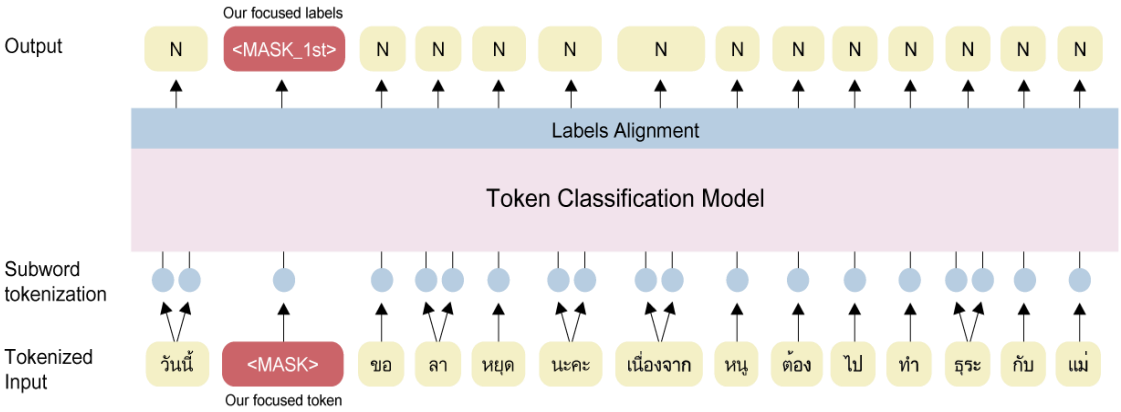

Figure 2 Overall configuration of the token classification task

CHAPTER 4

EXPERIMENTS

We conduct our two experiments separately using the same dataset. Our data consists of a total of 300 utterances with zero pronouns in Thai language, whose antecedents or poscedents are found both inside and outside the text. With no personally identifiable information or sensitive context, 200 sequences were collected from the researchers' archived chat logs from March to August 2022 as well as from online articles and interviews posted in 2022 from the Facebook pages 'Poetry of Bitch', 'The Standard' and 'Popcornfor2', and the others were manually authored. Our data contains diverse topics and length from 7 tokens to 37 tokens. Some of them consist of more than one sentence. Besides, these utterances have various levels of formality and social situations which could reflect the choice of Thai pronoun usage.

To prepare the input for our experiments, we basically perform the preprocessing tasks on our data involving removing unnecessary characters e.g., emojis, duplicate characters, multiple spaces and punctuations, as well as word normalizing using the python toolkit PyThaiNLP 3.1.0. Furthermore, we normalize some non-standard words into standard words e.g., *เลา* to *เรา (I/we)*, *เทอ* to *เธอ (you)* and *เค้า* to *เขา (he/she)*. After the preprocessing steps, we manually detect the zero pronoun positions in each input sequence. In order to detect zero-pronouns, the sentences should be divided into clauses (Ferrández & Peral, 2000). Concerning the theta criterion (θ-criterion), we can identify the missing arguments of the predicates, which could be verbs or prepositions. Through each clause or constituent, we can easily observe noun phrases assigned to theta-roles (θ-roles) and determine the one that violates θ-criterion which actually contains the omitted argument of the predicate completed by zero pronouns. Those predicates are called ZP predicates (Kim et al., 2021).

เจ้าของร้านเป็นคนเกาหลี ฉันรู้จักตั้งนานแล้ว

The shop owner is Korean. I have known Φ for a long time.

From this utterance, we can divide the entire sentence into two smaller sentences: “เจ้าของร้านเป็นคนเกาหลี” and “ฉันรู้จักตั้งนานแล้ว”. In the second sentence, the predicate *รู้จัก (know)* needs the Theme role but there are no surface elements in the clause to be assigned. Hence, it should be filled with a zero pronoun.

Table 1 Theta grid for *รู้จัก* in example (1)

Experiencer	**Theme**
ฉัน (I)	Φ

In addition, seeing that zero pronoun occupies a governed position and receives the θ-role, all of them can be replaced by the overt pronouns. In other words, the sentences are grammatically with or without the overt pronouns in that argument position as shown in the following examples:

(a) นางเอก$_1$โดน คนในบ้านของพระเอก$_2$ กลั่นแกล้งสารพัด แต่ Φ_1 ก็ยังอดทนและผ่านมันไปได้

(b) นางเอก$_1$โดน คนในบ้านของพระเอก$_2$ กลั่นแกล้งสารพัด แต่ เธอ/นางเอก$_1$ ก็ยังอดทนและผ่านมันไปได้

In this case, the subject pronoun in the subordinate clause should be the pronoun ‘*เธอ*’ or R-expresion *‘นางเอก’*, which corefers with the antecedent *‘นางเอก’*, the subject of the main clause, as in (b). However, since Thai language allow null subjects, we can drop the subject pronominal expression ‘*เธอ/นางเอก*’ whereas the semantic associations between the dropped pronoun and its antecedent *‘นางเอก’* remain and the sentence is still naturally acceptable as in (a).

After all these steps, we separately process the data given into each model as discussed further.

4.1 Experiment 1: Predicting the Masked Pronouns

Our study benefits from the pre-trained MLM of WangchanBERTa, a Thai language-specific language model trained on RoBERTa[10]-based architecture. From the python package Transformers library version 3.5.1 and Thai2transformers library version 0.1.2, we use CamembertTokenizer as a tokenizer, Process Transformers as a preprocessor of the input text, and AutoModelForMaskedLM as a model class. We use pre-trained 'wangchanberta-base-att-spm-uncased', which was trained with Assorted Thai Texts containing 382,000,000 training examples or 78.5GB dataset with a large vocabulary size of 25,000 on the RoBERTa-based architecture with SentencePiece tokenizer (Lowphansirikul, Polpanumas, Jantrakulchai, & Nutanong, 2021).

As required, we put the special <*mask*> tokens substituted for zero pronouns that were manually detected earlier. One utterance contains only one masked token. Since this approach is recognized as an unsupervised zero pronoun resolution, all 300 utterances with 300 masked tokens become an input for MLM. From the preprocessing step by Process Transformer, all spaces in the text are replaced with the special token <_>. After that, we put the preprocessed text to the MLM. Table 2 shows the input text to the model and the output tokens with the probability scores.

[10] To exceed the performance, RoBERTa was trained longer with bigger batches, removed the Next Sentence Prediction (NSP) task from BERT's pre-training and applied dynamically changing the masking pattern to the training data so that the masked token changes during the training epochs (Liu et al., 2019).

Table 2 Examples of input and output of MLM

Input	Process Transformer	Output 1	Output 2	Output 3	Output 4	Output 5
ปีเตอร์นะ ฉันได้ยินข่าวว่าจอห์นเพิ่งพา<mask>ไปโรงพยาบาลเมื่อเช้านี้	ปีเตอร์นะ<_>ฉันได้ยินข่าวว่าจอห์นเพิ่งพา**<mask>**ไปโรงพยาบาลเมื่อเช้านี้	**ปีเตอร์ (0.3155 905604 362488)**	จอห์น (0.08104 0777266 02554)	คุณ (0.04978 6604940 891266)	ภรรยา (0.04965 3030931 949615)	ครอบครัว (0.01633 0463811 75518)
เมื่อก่อนตอนเธออ้วน <mask> ดูแก่กว่าปัจจุบันอีก	เมื่อก่อนตอนเธออ้วน<_><mask>ดูแก่กว่าปัจจุบันอีก	**เธอ (0.25907 8830480 57556)**	คุณ (0.0847 429484 128952)	ฉัน (0.03722 3245948 553085)	เธอยัง (0.03490 2937710 28519)	ปัจจุบัน (0.02335 9894752 50244)
ทำไมส่วนใหญ่ญาติต้องอิจฉากันเอง ทำไม**<mask>** ถึงไม่รักกัน	ทำไมส่วนใหญ่ญาติต้องอิจฉากันเอง<_>ทำไม<mask>ถึงไม่รักกัน	**ญาติ (0.17374 3724822 99805)**	พี่น้อง (0.1270 301192 998886)	คนไทย (0.08553 7478327 75116)	ลูกหลาน (0.06577 3956477 64206)	มนุษย์ (0.05521 1346596 479416)
ในหัวตอนแรกก็คิดว่าเขาเป็นคนบ้าพลังคนหนึ่งแต่พอได้มารู้จัก**<mask>**ก็ได้รู้ว่ามันไม่ได้อย่างที่เห็น	ในหัวตอนแรกก็คิดว่าเขาเป็นคนบ้าพลังคนหนึ่ง<_>แต่พอได้มารู้จัก**<mask>**ก็ได้รู้ว่ามันไม่ได้อย่างที่เห็น	จริงๆ (0.4468 560516 834259)	_จริงๆ (0.07134 5143020 15305)	**เขา (0.03578 7761211 395264)**	พฤติกรรม (0.02977 8003692 626953)	ตัวเอง (0.02009 6071064 4722)
ก็หนูเห็นพ่อชอบกินข้าวหมูแดงเจ้านี้ หนูก็เลยซื้อมาฝาก**<mask>**	ก็หนูเห็นพ่อชอบกินข้าวหมูแดงเจ้านี้<_>หนูก็เลยซื้อมาฝาก<mask>	**พ่อ (0.22674 8600602 14996)**	ค่ะ (0.1464 214026 927948)	แม่ (0.06181 5466731 78673)	หนู (0.05317 2364830 970764)	5 (0.03562 6035183 66814)
วันนั้นกูไม่ได้กินเยอะเพราะ<mask>กิน ไรไม่ลงแต่กูหารค่าอาหารได้ แต่กูไม่หารค่าเหล้านะ	วันนั้นกูไม่ได้กินเยอะเพราะ<mask>กินไรไม่ลง<_>แต่กูหารค่าอาหารได้<_>แต่กูไม่หารค่าเหล้านะ	หิว (0.1027 032658 457756)	**กู (0.08759 0351700 78278)**	เครียด (0.06316 7348504 06647)	ปวดท้อง (0.06217 3381447 79205)	ขี้เกียจ (0.05204 1627466 67862)
พอ<mask>ไปถึงรีสอร์ท นุ่นก็สำรวจพูลวิลล่าของนายดีคร่าว ๆ	พอ<mask>ไปถึงรีสอร์ท<_>นุ่นก็สำรวจพูลวิลล่าของนายดีคร่าว<_>ๆ	**นุ่น (0.2089 84375)**	เดิน (0.07221 4387357 23495)	ขับรถ (0.02076 0368555 784225)	พวกเรา (0.01920 9673628 21102)	ทราย (0.01732 5684428 215027)

We evaluate the performance of MLM whether it could be an applicable approach for Thai zero pronoun resolution. Therefore, we detect only the predicted pronominal expressions that precisely correspond to their antecedents or postcedents for our results, which could be further handled with the rule-based algorithm for selecting the best candidates, and observe tokens returned with the high scores that could occur in that position while keeping the utterance understandable and native sounding in spite of other part of speech.

4.2 Experiment 2: Predicting the person of the masked pronouns

In order to classify *person*, the grammatical attribute of zero pronoun, we further annotate our masked tokens from the first experiment as *<mask_1st>*, *<mask_2nd>* and *<mask_3rd>*, which refer to the first-person, the second-person, and the third-person pronoun, respectively. The number of sequences and the label distribution are shown in Table 3.

Table 3 The number of sequences and the label distribution

Person	**Qty**
1st-person	91
2nd-person	93
3rd-person	116
Total	**300**

For the token classification task, the sequence of tokens and the sequence of labels assigned to individual tokens need to be prepared. So, we first tokenize the utterance using ‘newmm’ engine from PyThaiNLP adding custom words representing our masked tokens in the dictionary, *mask_1st, mask_2nd* and *mask_3rd*. Accordingly, we have to label every other token outside the zero pronouns as N, which are not going to be evaluated. Table 4 demonstrates the examples of prepared data where the length of tokens is equal to the length of labels.

Table 4 Examples of prepared data for token classification task

Equence of Tokens	Sequence of Labels
พี่,เขา,ไป,ไหน,แล้ว,ไม่,รู้, เมื่อ,กี๊,mask,ยังอยู่,ตรงนี้,อยู่เลย	N,N,N,N,N,N,N,N,N,N,mask_3rd,N,N,N
แม้,mask,จะ,โด่งดัง,ใน,ฐานะ, นักแสดง,แต่,สุดท้าย,เขา,ก็, สามารถ,สาน,ฝัน,การ,เป็น, นักร้อง,ให้,ตัวเอง,จนได้	N,mask_3rd,N,N,N,N,N,N,N,N,N,N,N,N,N, N,N,N,N,N,N
ซวย,แล้ว,วันนี้,mask,ลืม,เอา, แว่น,มา,กรี๊ด	N,N,N,N,mask_1st,N,N,N,N,N,N
แหม,มึง,ทำ,มา,เป็น,เด็กดี, ,เมื่อวาน,mask,ก็,โดด,เรียน	N,N,N,N,N,N,N,N,mask_2nd,N,N,N
ผม,ลืม,บอก,ว่า,ผม,ส่ง,ราคา,เข้าไป,ใน,อีเมล,ของ,คุณ, อิง,แล้ว,นะ,ครับ,ถ้า,mask,ได้รับ, แล้ว,รบกวน,ช่วย,แจ้ง,ผม,หน่อย,นะ,ครับ	N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,m ask_2nd,N,N,N,N,N,N,N,N,N

We perform fine-tuning pre-trained RoBERTa-based models using a pre-trained language model for a token classification task from Transformer library version 4.9.2. Firstly, the input needs to be preprocessed, by which the tokenizer adds the word boundary '▁' into the sequence, tokenizes some tokens into subwords, and converts them into input ids and attention mask embeddings. Consequently, the lengths of tokenized texts become longer than the labels', so we need to prepare the alignment function to map the subwords into the former version of word tokens for the evaluation process. The example below shows the original input text tokens and the outputs from BERT tokenizer.

Tokens: ['ชวย', 'แล้ว', ' ', 'วันนี้', 'mask', 'ลืม', 'เอา', 'แว่น', 'มา', ' ', 'กรี๊ด']

Input ids: [5, 10, 9242, 627, 10, 984, 10, 20113, 10, 1124, 1722, 10, 6239, 10, 26, 10, 17910, 6]

Attention mask: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

BERT Tokens: ['<s>', '▁', 'ชวย', '▁แล้ว', '▁', '▁วันนี้', '▁', 'mask', '▁', 'ลืม', '▁เอา', '▁', 'แว่น', '▁', 'มา', '▁', '▁กรี๊ด', '</s>']

In the training process, we employ the pre-trained 'wangchanberta-base-att-spm-uncased' LM to fine-tune a token classification model. We found these adjusted hyperparameters: 8 of batch size, 17 epochs, 2e-5 learning rate (AdamW optimizer) and weight decay of 0.01 achieve the best performance for this specific task.

Due to our limited data, we adopt K-Folds cross-validation from Scikit-learn library's model_selection module to help measure the model's performance more accurately. We split our data into 3 sections using 3-folds. In each iteration, one section becomes a test set and the others become a training set. We use the classification report from Scikit-Learn to demonstrate the classification metrics including precision, recall and the f1-score by class along with the support. Using K-Folds cross-validation, different loops provide different results. Accordingly, we calculate the mean of the performance scores.

CHAPTER 5

RESULTS AND ANALYSIS

5.1 Predicting the Masked Pronouns

The result gained from the first experiment, predicting the masked pronouns using MLM, shows that MLM has a potential to be a feasible application applied for unsupervised zero pronoun resolution. As Thai has a rich pronoun system and the choice of pronoun usage is determined by many social components (Hoonchamlong, 1991), we cannot evaluate the output as a binary judgment, i.e., whether it is true or false. Therefore, we discuss our supporting statements in the qualitative analysis.

First of all, despite not being fine-tuned, MLM could resolve 195 zero pronouns with acceptable pronominal expressions i.e., pronouns and R-expressions, out of 300 input masked tokens, which accounts for 65% when considering all five candidates. Although, only 103 of them are resolved as the first candidate with the highest scores, which accounts for 34%, from our perspective, it is rather an effective performance from a single pre-trained language model. This finding could also become a baseline for further studies in the future.

Moreover, MLM can resolve zero pronouns in various cases regarding syntactic levels or discourse levels. Following Konno et al. (2022)[11], we analyze the result based on the argument type divided by the positions of the arguments of a given predicate[12]: intra-sentential arguments, inter-sentential arguments and exophoric arguments. The intra-sentential argument could be determined by syntactic structures, whereas the arguments in the inter-sentential and exophoric case should be analyzed at discourse level. As the exophoric arguments are not referenced in the text e.g., author, reader, and general noun phrase, it is therefore recognized as a difficult and

[11] This study adopted three kinds of domain-adaptation techniques and their combinations from Sango, Nishikawa, and Tokunaga (2019)

[12] i.e., the antecedent (or postcedents) of a given zero pronoun

challenging task. The results characterized by the position of references are analyzed as follows.

For intra-sentential and inter-sentential cases, MLM could resolve the omitted pronominal expression due to its reference explicitly appearing on the surface structure called endophoric references. Since MLM is known to predict a masked word based on its context in bidirectional ways – left to right and right to left – MLM could undoubtedly find the correlation between anaphoric or cataphoric arguments and the masked token, which is recognized as the zero pronouns. The following exhibits the example of the resolved pronouns filled in the utterance by MLM, grouped by the condition of arguments given:

(1) when the antecedent or postcedent is a pronoun, the zero pronoun could be solved properly.

“จากที่<คุณ>เล่ามา คุณก็ไม่ได้ทำผิดอะไรมากมายขนาดนั้นนะคะ ไม่ต้องกังวล”

“มาสายอีกแล้วนะเรา ถ้าครั้งหน้า<เรา>มาสายอีก พี่จะไม่ใจดีแบบนี้แล้วนะ”

“เราจะขอยืมหมวกแกหน่อย วันก่อนเห็น<แก>ใส่แล้วสวยดีอ่ะ”

“ถึงนางจะขี้แกล้งชาวบ้าน แต่เอาจริงๆ <นาง>ก็ไม่ใช่คนเลวร้ายอะไร”

“แม้<เขา>จะโด่งดังในฐานะนักแสดง แต่สุดท้ายเขาก็สามารถสานฝันการเป็นนักร้องให้ตัวเองจนได้”

(2) when the antecedent or postcedent is an R-expression, the model can also return a pronoun which corefers with its reference.

“อีพลอยไปด้วยป่ะ กูไม่ค่อยอยากเจอ<มัน>เลยอ่ะ”

“นอกจาก<เขา>จะทำธุรกิจเบียร์นำเข้าแล้ว นายดียังทำธุรกิจบันเทิง ปั้นนักแสดง ศิลปินและไอดอลด้วย”

“น้องนิวใกล้จะถึงหรือยังคะ อีกนานไหม พอดีพี่จะฝาก<หนู>ซื้อน้ำให้ผู้กำกับเขาหน่อยได้ไหมคะ”

“คุณแนนครับผมออกจากบริษัทมาแล้ว ให้ผมไปรับ<คุณ>ตรงไหนดีครับ”

“เรื่องแย่ ๆ ในชีวิตอาจจะเปลี่ยนใครบางคนให้กลายเป็นคนเลวร้าย แต่สำหรับคนบางคน<มัน>ก็กลับกลายเป็นแรงบันดาลใจให้ทำสิ่งดี ๆ ได้เหมือนกัน”

(3) when the antecedent is an R-expression, MLM can return the copied token of the reference which is considered as R-expression which can receive bound reading in Thai.

“มึงสั่งรองเท้าจากร้านไหน กูจะไปสั่ง<รองเท้า>ตามบ้าง”

“ลูกค้าทำไฟล์งานขนาดที่ต้องการตัดแบ่งมาให้ทางร้านได้เลยนะคะ เดี๋ยว<ทางร้าน>จะนำงานลูกค้ามาวางลง A3
ให้เองนะค้า”

“ตอนโบกลับถึงบ้านก็ดึกมากแล้ว พอ<โบ>อาบน้ำเสร็จก็ง่วงไม่ไหว”

“คุณลูกบ้านมีพัสดุอยู่ที่นิติสองชิ้น เดี๋ยวนิติจะเอา<พัสดุ>ขึ้นไปให้พร้อมกันด้วยเลยนะคะ”

“เพจดังดึงสติชาวเน็ต หลังพบ<ชาวเน็ต>แชร์ภาพล้อเลียนนักแสดงดัง”

(4) when the antecedent is an R-expression, MLM can provide only a copy of their antecedent’s head instead of a full copy of the noun phrase.

“วันก่อนกูเจอกล่องดนตรีจิ๋วๆ ที่เบาะข้างคนขับ คือ นางแอบวาง<กล่อง>ทิ้งไว้ก่อนไปอังกฤษ”

“อาจารย์สมเกียรติบอกว่าวันนี้<อาจารย์>เข้าประชุมไม่ได้นะ”

“แกไปดูคอนเสิร์ตพี่บอยมานี่ เสียดายเราก็อยากไปอ่ะ <คอนเสิร์ต>ดีป่ะ”

“เสื้อรุ่นนี้เหลือสองตัวที่ร้านแล้วนะคะ ลูกค้าจะรับ<เสื้อ>ไปเลยไหมคะ”

For exophoric cases which contain non-referential zero anaphora, MLM could remarkably resolve some of them. This implies that MLM could understand the context at discourse level as it can replace the masked token with the fitting word. Yet, the masked token representing the exophora that MLM predicts might considerably depend on the pre-trained corpus. The examples of successful uncovered non-referential zero pronouns are shown as follows:

(1) when the zero argument is 1st and 2nd pronoun, MLM can recover it with the right pronouns.

“<คุณ>ไปไหนมาอ่ะ เขาประชุมกันจะเสร็จแล้ว”

“พวกมึงไปกันก่อนเลย <กู>ยังทำงานไม่เสร็จอ่ะ”

“ถ้า<พวกคุณ>ไม่ชอบเราก็ไม่ต้องมากดติดตามจ้า เลิกตามไปได้เลยค่า เราไม่ได้แคร์ค่า”

“<ดิฉัน>ได้ทดลองหยุดสวดมนต์และทำสมาธิไปพักใหญ่เพื่อสังเกตผลที่จะเกิดขึ้นกับสภาวะจิตใจของตัวเองในช่วงเวลาที่ไม่ได้ปฏิบัติอย่างต่อเนื่อง รู้สึกได้ว่าจิตใจไหลเข้าสู่อารมณ์ด้านลบง่าย”

(2) when the zero pronoun receives generic reading, MLM could also replace it with the right pronominal expression.

“ทุกงานจะดูง่าย เมื่อ<คุณ>ไม่ต้องทำเอง”

“แต่ในชีวิตจริงถึงแม้<ทุกคน>จะได้รับการเลี้ยงดูจากพ่อแม่เท่าเทียมกันพี่น้องแต่ละคนก็ไม่สามารถประสบความสำเร็จในชีวิตเท่าเทียมกัน”

Last but not least, MLM can resolve to the right pronominal expressions, as MLM seems to encode various underlying aspects of pronoun usage e.g., the speech participants’ roles, relationships of the speech participants, their relative social status, the context of discourse, etc., for example,

“สวัสดีค่ะ <ดิฉัน>จะขอสอบถามเรื่องบัตรคอนเสิร์ตที่ซื้อไป พอดีมันไม่ขึ้นในระบบให้ค่ะ”

In spite of the non-overt reference, the contextual clue such as *‘ค่ะ’* is contained in the utterance, which implies the formal situation and non-personal relation between the speaker and the listener, and that the speaker is female. Interestingly, MLM returns *‘ดิฉัน’*, a formal first-person feminine pronoun in Thai which is the proper one to be filled in this case.

In contrast with the previous utterance, in this conversation, the first-person pronoun used in the preceding sentence, *‘กู’*, suggests the informal situation and the tight friendship of the speech participants. And MLM could pick the correct pronoun as well.

“กูแต่งตัวรอแล้วจ้า <มึง>จะให้กูออกจากบ้านตอนไหนก็บอกนะ”

Plus, we also observe the unsuccessful cases in resolving zero pronouns, for example, the resolved pronouns refer to the wrong antecedent, the pronouns are resolved with the wrong level of formality, and the masked tokens are replaced with the non-NP expressions.

Apparently, MLM could also predict the wrong antecedent. It reveals the fact that even though the references are obviously presented in the text, it is difficult for the unsupervised model to put the right entity in one clause to the argument of the ZP predicate in another clause. In other words, the model attends to the mistaken salient candidate antecedent. Consequently, deeper semantic knowledge is required for the model to learn more.

“แมรี่ไม่เคยรักพอลเลย <แมรี่>เป็นแค่คนคั่นเวลาเติมเต็มที่ว่างในใจของเธอเท่านั้น”

(The zero should refer to ‘พอล’.)

“ตัวเองรอเขาแป๊ปนึงนะ เขากำลังไปหา<เขา>แล้ว”

(The zero should refer to ‘ตัวเอง’.)

“คนอื่นไม่น่าจะเข้าเฟสบุคมึงได้นะ ถ้า<มึง>ไม่มีพาสเวิด”

(The zero should refer to ‘คนอื่น’.)

Similar to the previous case, MLM also leverages the single pre-trained knowledge by returning the irrelevant entities from the pre-trained corpus instead of the entities that are mentioned in the text. This makes the utterances syntactically correct but semantically incorrect, as shown in this example,

“พี่เอจะถึงหรือยัง จะให้บีออกไปรับ<โทรศัพท์>ตรงปากซอยไหม”

“รถเราที่โดนกรีดวันก่อน วันนี้พ่อเอา<ฟิล์ม>ไปทำสีมาให้ใหม่แล้ว”

“ต้องถามว่าจะให้<อาจารย์>ไปประชุมที่ไหน”

Moreover, the zero pronoun resolves but with the wrong level of formality. Long-distance dependency might be a problematic case where the model could lose the association between antecedents and zero pronouns.

“อ๋อ เรามาทำฟันแถวนี้พอดี <ฉัน>ก็เลยแวะมาหาเธอด้วย”

“ถ้าปกติมึงกินกะเพราไก่ไข่ดาว <คุณ>ก็เปลี่ยนเป็นกะเพราไก่ผัดน้ำเปล่าใส่ไข่ต้มแทน ไรงี้”

These findings confirm that the model understand syntax of Thai language and has the ability to analyze the syntactical structure. However, deeper semantic knowledge is required for the model to learn more.

Besides, there are some cases that MLM returns adverbs, aspect markers or particles etc. instead of pronouns, which could occupy that masked token while the utterance maintains natural and understandable, for instance,

“เครื่องซักผ้ายังเสียอยู่ใช่ไหม เดี๋ยวแม่จะได้เรียกช่างมาซ่อม<ให้เสร็จ>ทีเดียว”

“วันนี้ผมขอไม่เข้าออฟฟิศนะครับ เดี๋ยว<จะ>ไปเข้าวันศุกร์แทน”

“ถ้าเกิดว่า<วันนี้>รู้สึกเหนื่อยเกิน พวกเธอก็สามารถพักกินน้ำกินขนมได้เลยนะ”

“อินุ่นไปลองชุดเจ้าสาวมา กูก็แอบอยากเห็นว่ามันใส่<จริงๆ>แล้วเป็นไง”

“ผมเพิ่งเริ่มฝึกนั่งสมาธิ เดินจงกรม ผลคือหลับสบายขึ้น เมื่อก่อนยังไง<ๆ>ก็ต้องตื่นมาเข้าห้องน้ำทีนึง”

“<เพิ่ง>เห็นที่ผู้บริหารโพสล่าสุด เดี๋ยวน่าจะมีฟ้องกันเดือดๆ แน่”

“วันก่อนอาจารย์บอกว่าวันนี้ให้พวกเราเข้าไปหา<ค่ะ>”

As Thai is a pro-drop language, we believe that the corpus used for pre-training contains a lot of instances with dropped pronouns which reflects the natural use of Thai zero pronouns. The masked token in that position thereby allows other parts of speeches to be filled. We conclude that this problem could be resolved by fine-tuning MLM to the specific task.

All of these statements show that the transformer-based MLM at least understands the relational knowledge between the masked token and its reference. Not only the typical grammatical information of pronouns, but also other dimensions of

Thai pronoun usage can be captured by this unsupervised system. The attention component allows the model to learn contextual relations between words in bidirectional ways, so that MLM could handle both antecedent and postcedent.

5.2 Predicting the Person of the Masked Pronouns

In our extended experiment, predicting the persons of the masked pronouns, the high-performance scores indicate that the pre-trained language model is appropriate to be developed to a full system for resolving zero pronouns.

In this study, we aim to test whether the model that has been fine-tuned with our specific task can capture the grammatical feature, person of zero pronoun alone. Therefore, we annotate the target token to be classified as a masked token with person instead of the explicit resolved zero pronoun, to separate it from other outside tokens including overt pronouns or entities, which are not going to be evaluated. Consequently, the results of predicting the target token are clear-cut with no overlapping among the target classes and the ignored class. From three iterations of a fine-tuning model for a token classification task, the results are consistent with each other. The model achieved high mean scores in precision, recall and F1 for first-person, second-person and third-person masked pronouns in all iterations. The overall results are that the precision is 0.817, the recall is 0.822 and the F1 score is 0.817, as illustrated in Table 5.

Table 5 Precisions, recalls, F1 scores and support for each class

	1st-Person Pronoun				**2nd-Person Pronoun**				**3rd-Person Pronoun**			
	Precision	**Recall**	**F1**	**Support**	**Precision**	**Recall**	**F1**	**Support**	**Precision**	**Recall**	**F1**	**Support**
Iterate 1	0.800	0.774	0.787	31	0.686	0.8889	0.774	27	0.914	0.7620	0.831	42
Iterate 2	0.846	0.880	0.863	25	0.816	0.838	0.827	37	0.833	0.789	0.811	38
Iterate 3	0.848	0.800	0.824	35	0.735	0.862	0.794	29	0.879	0.805	0.840	36
Mean	**0.8315**	**0.8181**	**0.8244**		**0.7456**	**0.8629**	**0.7982**		**0.8755**	**0.7856**	**0.8275**	

The successful results reflect the powerful contextualized word embeddings that let the model learn deep multiple representations of each word through token embeddings, segment embeddings and position embeddings and utilize attention mechanism to pay attention to the useful contexts which help capture the grammatical person of the masked pronouns. We found that the model could basically (1) figure out that the masked token is an argument of ZP predicate (2) leveraging the informative parts from the surrounding tokens and (3) distinguish the difference properties between masked tokens. In addition, the model yields remarkable results, since this task is designed not to be complicated, and our corpus is quite small. We examine only one dimension of the important grammatical features of pronouns. However, this finding could lead to the development of an end-to-end system for resolving zero pronouns. To illustrate, annotating the grammatical person of Thai pronouns must be a useful filtering step to predict Thai abundant pronouns. Especially, in the complicated system of Thai pronoun usage, some pronouns could be used to refer to various persons of pronoun depending on contexts e.g., ‘เรา’ could refer to a first-person or second-person pronoun, ‘เขา’ could refer to a first-person or third person pronoun or ‘เธอ’ could refer to a second-person pronoun or third-person pronoun. The ability to correctly predict the pronouns’ person should potentially help enhance the performance of the overall system.

To further improve the model, one should add negative examples and try exploring results with a larger corpus.

CHAPTER 6

CONCLUSION

In this study, we aim to investigate the effectiveness of pre-trained language model whether it could be an applicable approach for Thai zero pronoun resolution, through (1) applying masked language model to predict zero pronominal expressions and (2) fine-tuning on a token classification task to classify persons of pronouns. Our idea is to exploit the attention mechanism and contextualized embeddings to recover the omitted form of Thai pronominal expressions, as it is believed to capture semantic information and relation between dropped pronouns and their antecedents or postcedents. Our results confirm that MLM is a strong basis to apply modification on for Thai zero pronoun resolution and that the transformer-based model is capable of classifying the grammatical feature of zero pronouns. However, our study is based only on a collected small-sized corpus. To ensure more reliability of the results, we plan to expand our corpus and research on fine-tuning MLM to the specific zero pronoun resolution task that generates only Thai pronominal expressions.

REFERENCES

Aroonmanakun, W. (2000). Zero pronoun resolution in thai: A centering approach. In D. Burnham (Ed.), *Interdisciplinary approaches to language processing: The international conference on human and machine processing of language and speech* (pp. 127-147). Bangkok: NECTEC.

Aroonmanakun, W. (2002). *Referent resolution for zero pronouns in Thai*. Bangkok: n.p.

Aroonmanakun, W. (2003). *Zero pronoun resolution in Thai: A centering approach*. Bangkok: n.p.

Chen, S., Gu, B., Qu, J., Li, Z., Liu, A., Zhao, L., & Chen, Z. (2021). *Tackling Zero Pronoun Resolution and Non-Zero Coreference Resolution Jointly.* Paper presented at the Proceedings of the 25th Conference on Computational Natural Language Learning.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, N. (1986). *Knowledge of langauge: Its nature, origin and use*. New York, NY: Praeger.

Ferrández, A., & Peral, J. (2000). *A computational approach to zero-pronouns in Spanish.* Paper presented at the Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.

Hoonchamlong, Y. (1991). *Some issues in Thai anaphora: A government and binding approach.* (Doctoral dissertation), The University of Wisconsin-Madison, Madison, WI.

Iida, R., Torisawa, K., Oh, J.-H., Kruengkrai, C., & Kloetzer, J. (2016). *Intra-sentential subject zero anaphora resolution using multi-column convolutional neural network.* Paper presented at the Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, Texas.

Isozaki, H., & Hirao, T. (2003). *Japanese zero pronoun resolution based on ranking rules and machine learning.* Paper presented at the Proceedings of the 2003 conference on Empirical methods in natural language processing.

Kim, Y., Ra, D., & Lim, S. (2021). Zero-anaphora resolution in Korean based on deep language representation model: BERT. *ETRI Journal*, *43*(2), 299-312.

doi:10.4218/etrij.2019-0441

Kongwan, A., Kamaruddin, S. S., & Ahmad, F. K. (2022). Anaphora resolution in Thai EDU segmentation. *Journal of Computer Science, 18*(4), 306-315. doi:10.3844/jcssp.2022.306.315

Konno, R., Kiyono, S., Matsubayashi, Y., Ouchi, H., & Inui, K. (2022). Pseudo Zero Pronoun Resolution Improves Zero Anaphora Resolution. *Journal of Natural Language Processing, 29*(1), 243-247. doi:10.5715/jnlp.29.243

Larson, M. (2007). *The Thais that Bind: Principle C and Bound Expressions in Thai*. Bangkok: n.p.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv*, 11692.

Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., & Nutanong, S. (2021). Wangchanberta: Pretraining transformer-based thai language models. *arXiv preprint arXiv*, 09635.

Park, A., Lim, S., & Hong, M. (2015). *Zero object resolution in Korean.* Paper presented at the Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China.

Sango, M., Nishikawa, H., & Tokunaga, T. (2019). Domain adaptation in Japanese predicate-argument structure analysis considering first and second person exophora. *Journal of Natural Language Processing, 26*. doi:10.5715/jnlp.26.483

Taira, H., Sudoh, K., & Nagata, M. (2012). *Zero pronoun resolution can improve the quality of je translation.* Paper presented at the Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, Republic of Korea.

Yeh, C.-L., & Chen, Y. (2003). *Using zero anaphora resolution to improve text categorization.* Paper presented at the Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation.

Yin, Q., Zhang, Y., Zhang, W., Liu, T., & Wang, W. Y. (2018). *Zero pronoun resolution with attention-based neural network.* Paper presented at the Proceedings of the 27th international conference on computational linguistics.

Yoshida, S., & Nagata, M. (2009). *Utilizing Features of Verbs in Statistical Zero Pronoun Resolution for Japanese Speech.* Paper presented at the Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2.

VITA

NAME Miss Sumana Sumanakul

DATE OF BIRTH 13 February 1988

PLACE OF BIRTH Bangkok