Machine Reading Comprehension for Multiclass Questions on Thai Corpus

Mr. Theerit Lapchaicharoenkit

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2019

การอ่านทำความเข้าใจด้วยเครื่องสำหรับคำถามหลายประเภทบนคลังข้อความภาษาไทย

นายธีรทธิ์ ลาภชัยเจริญกิจ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

Thesis Title	Machine Reading Comprehension for Multiclass Questions on Thai Corpus
By	Mr. Theerit Lapchaicharoenkit
Field of Study	Computer Science
Thesis Advisor	Assistant Professor PEERAPON VATEEKUL, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

..... Dean of the FACULTY OF ENGINEERING

(Professor SUPOT TEACHAVORASINSKUN, Ph.D.)

THESIS COMMITTEE

..... Chairman

(Professor BOONSERM KIJSIRIKUL, Ph.D.)

..... Thesis Advisor

(Assistant Professor PEERAPON VATEEKUL, Ph.D.)

..... Examiner

(DUANGDAO WICHADAKUL, Ph.D.)

..... External Examiner

(Thanapat Kangkachit, Ph.D.)

ธีริทธิ์ ลาภชัยเจริญกิจ : การอ่านทำความเข้าใจด้วยเครื่องสำหรับคำถามหลายประเภทบนคลังข้อความภาษาไทย. (Machine Reading Comprehension for Multiclass Questions on Thai Corpus) อ.ที่ปรึกษาหลัก : ผศ. ดร.พีรพล เวทีกูล

Previous Thai question answering and machine reading comprehension researches focus on small scale dataset and do not utilize the deep learning approach to build the models. In this research, we develop a Thai machine reading comprehension (MRC) model on Thai MRC dataset provided by NECTEC. This dataset consists of 17,000 question-answer pairs and has two classes of questions, which are factoid and yes-no questions. We use BIDAF as the based MRC architecture. We have performed experiments with 3 different multiclass model designs, which includes special tokens, joint, and cascade model. We also utilize contextual embeddings for Thai language to enhance the model's performance. As the results suggest that cascade architecture has the best F1 performance. We then incorporate transfer learning and modify the attention mechanisms to increase the model's accuracy on yes-no questions.

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต .....
ปีการศึกษา 2562 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6170932121 : MAJOR COMPUTER SCIENCE
KEYWORD: deep learning in NLP

Theerit Lapchaicharoenkit : Machine Reading Comprehension for Multiclass Questions on Thai Corpus. Advisor: Asst. Prof. PEERAPON VATEEKUL, Ph.D.

งานวิจัยที่เกี่ยวข้องกับการถามตอบและการอ่านทำความเข้าใจก่อนหน้านี้นั้นถูกทำบนชุดข้อมูลที่มีขนาดค่อนข้างเล็กและไม่ได้มีการใช้การเรียนรู้เชิงลึกเข้ามาช่วยในการสร้างแบบจำลอง ในงานวิจัยครั้งนี้ผู้วิจัยได้ทำการสร้างแบบจำลองการอ่านทำความเข้าใจบนชุดข้อมูลการทำความเข้าในภาษาไทยจากศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ชุดข้อมูลดังกล่าวมีจำนวนคู่คำถาม คำตอบทั้งหมด 17,000 คู่ด้วยกัน โดยที่คู่คำถามคำตอบสามารถแบ่งได้เป็น 2 ประเภทด้วยกันคือคำถามข้อเท็จจริง และ คำถามตอบรับหรือปฏิเสธ ผู้วิจัยได้ใช้แบบจำลอง BIDAF เป็นแบบจำลองหลักในการทำงานวิจัย ผู้วิจัยได้ทำการทดลองกับโครงสร้างแบบจำลองสำหรับการตอบคำถามหลายประเภท 3 รูปแบบโครงสร้างด้วยกันได้แก่ แบบคำพิเศษ (special token) แบบจำลองร่วมกัน (joint) และแบบจำลองแบบแยก (cascade) ผู้วิจัยได้ทำการใช้เวกเตอร์คำที่คำนึงถึงบริบท (contextual embedding) เพื่อเพิ่มประสิทธิภาพของแบบจำลอง หลังจากที่ผู้วิจัยพบว่าแบบจำลองแบบแยก (cascade) มีประสิทธิภาพที่ดีที่สุด ผู้วิจัยได้ทำการใช้การส่งต่อการเรียนรู้ (transfer learning) และทำการดัดแปลงกลไกการสนใจ (attention mechanism) เพื่อเพิ่มความสามารถของแบบจำลองบนคำถามแบบตอบรับหรือปฏิเสธ

Field of Study: Computer Science Student's Signature .....
Academic Year: 2019 Advisor's Signature .....

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Asst. Prof. Peerapon Vateekul, Ph.D., who guided and coached me throughout the 2 years of my graduate study. He accepted me as advisee even though I do not have a computer science or background. He introduced me to the field of deep learning and natural language processing, emphasize the importance of peer reviews and the necessity of having a lab community. Speaking of which, I also would like to take this opportunity to express my gratefulness to all my lab peers for helping and supporting me throughout the years.

One of the most important aspects for researching in deep learning or machine learning field is data availability. This research is made possible with the Thai question answering dataset provision from NECTEC institution. NECTEC also provides funding to this research as we require the dataset through the NSC competition.

I would like to thank you all the committee's member for all the valuable feedback and comments for this research.

Theerit Lapchaicharoenkit

TABLE OF CONTENTS

Page

List of Tables

List of Figures

1. Introduction

Machine reading comprehension (MRC), one of the many natural language processing (NLP) tasks, involves making machines that possess the ability to read, understand, and comprehend the human languages. One possible way to formulate the MRC problem is through question answering (QA). After reading the given articles or documents, the machine must be able to answer the questions (or queries) related to the assigned literature.

Similar to the breakthroughs of deep learning image classification on ImageNet dataset [1], large and representative datasets are required to build or improve neural machine reading comprehension models. In 2016, Stanford University introduced a large question answering dataset, which consists of more than 100,000 questions created by crowdsourcing workers to read and create questions from Wikipedia articles. The dataset is called SQuAD [2], an abbreviation of Stanford Question Answering Dataset. The dataset is publicly available so there is healthy competition to increase the performance of the reading comprehension model. Currently, the best performers already exceed human capability whose accuracy is at 86 %.

After the release of SQuAD [2], researches in English NLP has continued to grow as different types of question answering datasets have been released. Examples of such datasets include conversational question answering [3], [4], multi-hop reasoning dataset [5], and visual question answering [6].

For the Thai language, a moderate scale dataset has been released in question answering from the Thai Wikipedia competition in 2018 with 4,000 factoid questions, we refer to this dataset as NECTEC V1. The competition was held by the National Electronics and Computer Technology Center (NECTEC). In said competition, the competitors employed various deep learning machine reading. In 2019, NECTEC has held another competition of Thai question answering with a larger dataset, increasing number question-answer pairs to 17,000. We refer to this newer version of the dataset, which is also the dataset of focus in our study as NECTEC V2. Another challenge that was added to this new competition is the introduction of yes-no questions. This new challenge provides an excellent opportunity to develop a reading comprehension model that can handle multiple types of questions. An example of a factoid question of the dataset can be seen in Figure 1.

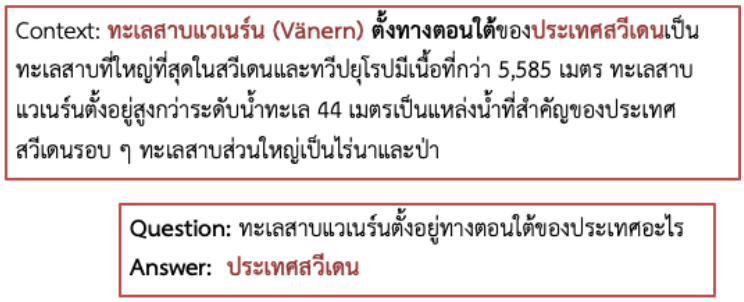


Figure 1: Example of a factoid question. Keywords can be found in bold letters.

In terms of past Thai reading comprehension researches, some researches focus on converting natural language into a structured query language then answering the questions based on structured data [7, 8] or using lexicon rules to answer the queries [9]. In this work, we aim to develop a deep learning reading comprehension model that can answer multiple types of questions based on unstructured text data.

With a larger scale dataset and a multiclass question setting, we propose to develop a novel Thai machine reading comprehension that has can process both factoid and yes-no questions. We propose to use BIDAF [10] as our baseline for this research since it acts as baselines for many newer reading comprehension datasets such as [3] and [4]. Our implementation of BIDAF also performs better compared to the winner of the previous competition on a similar dataset, NECTEC V1. (section 6.1).

In our research, we experiment with various multiclass architecture designs in section 4.2.2. We also utilize contextual embeddings as discussed in section 4.2.1. After we have found that the cascade model has the best performance in section 6.3, we shift our focus to further increase the model's performance on yes-no questions as the model's accuracy on this type of question was still low. We then continue the experiments with the techniques that can improve the performance on yes-no questions as discussed in sections 4.3 and 4.4. Our work has the contributions as follow:

- We have performed experiments with various multiclass architecture to select the most suitable architecture for the Thai NECTEC MRC dataset. (section 4.2),
- We integrate contextual embeddings constructed from a large scale pre-trained language model to the MRC model and compare its performance to static word embedding in the Thai multiclass MRC setting. (section 4.2.1),
- Transfer learning from Thai natural language inference task is used to increase the model's performance on yes-no questions, and
- We modify bidirectional attention mechanisms to further enhance the model's performance on yes-no questions.

1.1 Objectives

We propose to develop a machine reading comprehension based on deep learning model for Thai corpus. Our model will support two types of questions: (1) factoid questions and (2) yes-or-no questions.

1.2 Scope of Works

- The experiment is conducted on the NECTEC Thai question answering dataset from Question answering program from the Thai Wikipedia competition under the National Software Contest (NSC Thailand).
- Our task focuses on questions whose answers appear as spans of text in the document. Each question requires information from a single Wikipedia article only.
- We aim to develop a Thai reading comprehension model that can answer factoid and yes-no questions provided that the documents are given to the model.
- Contextual embeddings will be integrated into our proposed reading comprehension model to provide contextual information of the documents.
- The performances of our proposed technique will be compared to baseline methods, such as BIDAF [10].
- Evaluation of the model's performance will be conducted on token levels.

1.3 Step of Works

1. Literature review
2. Request for Thai MRC dataset
3. Dataset exploratory analysis
4. Define the research problem
5. Implement and establish the baseline
6. Perform preliminary experiments and discuss the results
7. Thesis proposal
8. Continue the experiments
9. Publish the paper and conference preparation
10. Thesis document preparation and submission
11. Final thesis defense

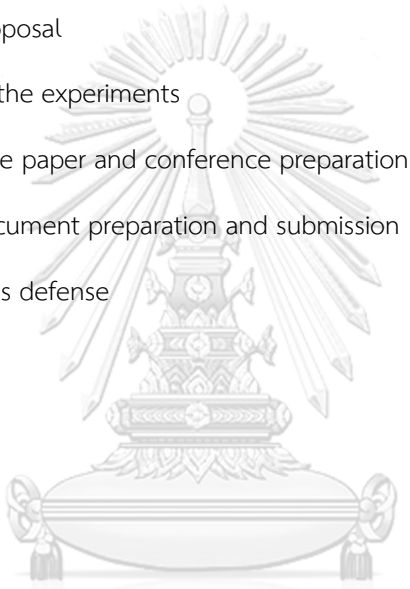
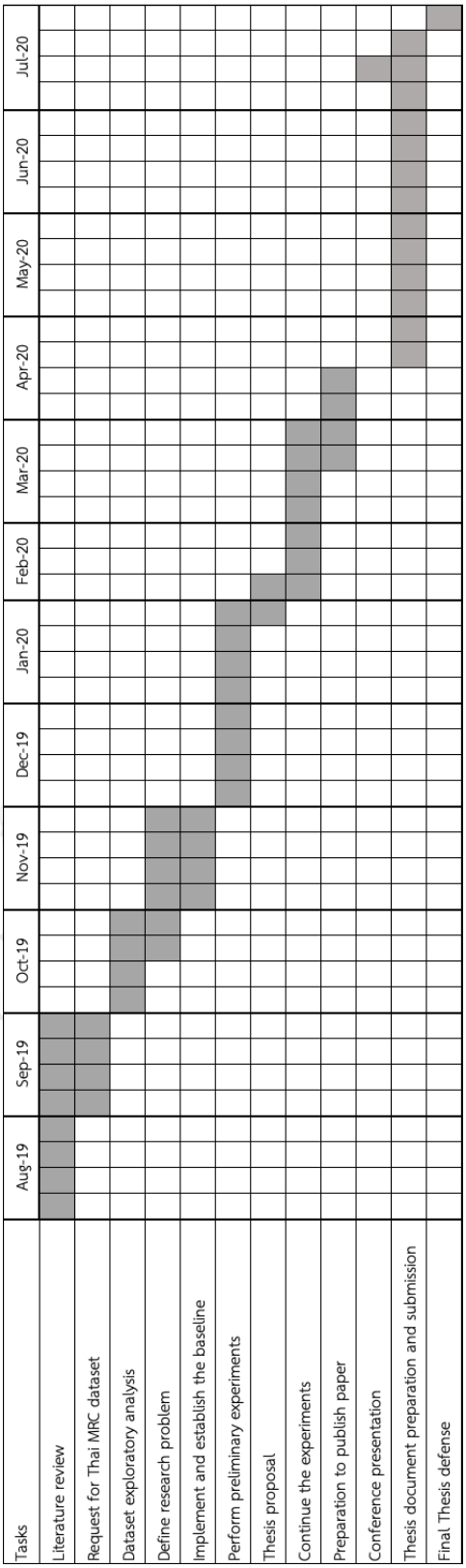


Table 1. Research plan



1.4 Publications

"Machine Reading Comprehension on Multiclass Questions Using Bidirectional Attention Flow Models with Contextual Embeddings and Transfer Learning in Thai Corpus" by Theerit Lapchaicharoenkit, Peerapon Vateekul in the International Conference Proceedings Series by ACM (not published) conference takes place in Singapore, July 17-19, 2020.

2. Background Knowledge

This chapter covers the theory and related knowledge required to conduct this research. We will discuss the MRC task, word vector representations, contextual embedding, recurrent neural network and its components, attention mechanism, architecture of deep learning model used in machine reading comprehension research and concept of transfer learning.

2.1 Machine Reading Comprehension Tasks (MRC)

MRC is the task of teaching machine the ability to read and understand the given questions then answer the questions based on provided documents. Various MRC tasks exist and can be broken down into 4 main categories: (1) cloze task; (2) multiple choices; (3) span extraction; (4) free form generation [11].

Cloze test task was amongst the first MRC task supported by large scale datasets. In this setting, the model is tasked to pick the correct entities or words that appear in the context passage to fill in the missing blanks of the query. Hermann, et al. [12] crafted the large scale cloze test dataset and developed a machine reading comprehension model that can perform such task.

Multiple-choice question setting asks the model to pick the candidate answer based on the provided questions and passages similar to examination for students in real life. An example of a multiple-choice question includes [13] which is a collection of English examination questions in China.

In the span extraction task, the model must extract the correct span of tokens or words which can be found in context passage. Factoid questions in our work can be classified as this type of reading comprehension task as well.

Freeform answer generation requires a more complex model as the answer does not necessarily have to be located in the passage. Text generation technology is commonly used with question answering techniques to successfully deal with this type of task. CoQA [4] is one of the datasets where the answers to questions are not required to appear in the document.

2.2 Static Word Embeddings

Word embedding is one of the methods used to represent words by static dense vectors and are employed by researchers to represent natural language for intelligent agents and various NLP models. Examples of static embedding words vector are word2vec [7] and Glove [8]. One possible way to construct a word vector is to create word vectors with the ability to predict the nearby words. Large corpus such as Wikipedia can be used in the process.

Sleep = [0.05, 0.6, -0.11, 0.05, -0.23, 0.16]

is = [0.8, -0.7, 0.4, 0.95, 0.87, 0.34]

The size of the vector dimension used to represent each word is a design choice and can be adjusted. Embedding word vectors can capture the semantic relationship between each word and many NLP researches utilize this ability. Our work uses Thai2fit word embedding for this type of word embedding.

2.3 Contextual Embeddings

Peters et al. [14] have pointed out that representations of words should also have the ability to vary across different contexts. For example, the word 'left' plays a different role in 'she left me for him.' and 'The book is on the left shelf'. Peters et al. [14] have developed a pretraining method that constructs a language model that can capture and represent contextual relationships among words in sentences. This method of pre-training model is also known as ELMo. The representations of the same word will vary from passage to passage which is different from word embedding where the representational vector of the same word stays the same regardless of the surrounding contexts. These representations can be applied to various downstream NLP tasks such as question answering, named entity extraction, and sentiment analysis.

Besides ELMo, other pre-trained language models also can construct deep contextualized embeddings, some of them include BERT [15] and ULMFiT [16].

2.4 Recurrent Neural Network and Long Short-Term Memory (LSTM)

A recurrent neural network is designed to deal with sequential or time-series data. Information will flow through a series of nodes from one direction to another direction. The processed output from a node will be fed into the node as well as input data itself.

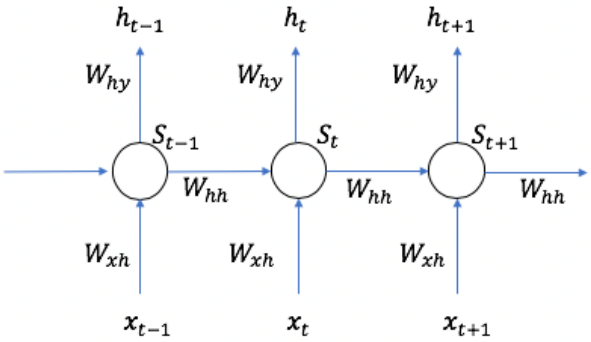


Figure 2. RNN processes data sequentially which can deal with textual data.

Equations governing RNN are listed below.

$$s_t = f(h_{t-1}, x_t) \quad (1)$$

$$s_t = g(W_{hh}s_{t-1} + W_{xh}x_t + b) \quad (2)$$

$$h_t = W_{hy}s_t \quad (3)$$

s_t represents a hidden layer, h_t represents the output of the hidden layer s_t. x_t represents the input of hidden state, s_t at time step t. W_{xh}, W_{hh} andW_{hy} represent parameter in computation from input to hidden state, hidden to hidden state, hidden state to output state respectively. b is an optional bias unit.

One thing worths noting about RNN is the nature of backpropagation in this type of neural network. Similar to ANN, back-propagation is required to compute errors and gradients, which are necessary components in the model's optimization process. Backpropagation in RNN tends to suffer from 'vanishing gradient' and 'exploding gradient' phenomena where gradients of the model get recurrently larger and larger (explode) or get smaller and smaller (vanish). This makes it difficult to optimize RNN as the parameters inside the network cannot be updated properly.

LSTM is a variation of RNN that can address the gradient issues with the cost of being more complex than the traditional RNN. LSTM composes of smaller units referred to as 'cell states'. Each cell state can be viewed to consist of three gates. Forget gate (f_t) decides which information from previous cell states should be carried over. Update gate (i_t) chooses and computes the information for the current cell state. Output gate (o_t) computes the output of the current state as well as information that will be passed to the next cell states.

$$f_t = \sigma(W_{f1}h_{t-1} + W_{f2}x_t + b_f) \quad (4)$$

$$i_t = \sigma(W_{i1}h_{t-1} + W_{i2}x_t + b_i) \quad (5)$$

$$\hat{C}_t = g(W_{C1}h_{t-1} + W_{C2}x_t + b_i) \quad (6)$$

$$o_t = \sigma(W_{o1}h_{t-1} + W_{o2}x_t + b_i) \quad (7)$$

$$h_t = o_t * g\left(f_t \circ C_{t-1} + i_t \circ \hat{C}_t\right) \quad (8)$$

2.5 Attention Mechanism

The attention mechanism is the concept first introduced in [17]. The attention mechanism was used to help RNN components of the neural machine translation model accurately focuses on the words that are crucial to the translation of the next word. However, attention mechanism is proved to be a general and powerful concept that can be readily applied to various deep learning tasks including many NLP related works such as [18], [15], and [19]. In machine reading comprehension, attention mechanisms can be used to align or capture the relationship between the question and the context passage.

The example of passage to query attention in MRC is illustrated in Figure 3. It can be observed that the model comprehends that the span ‘มงคลสมรสระหว่างเผิงคุณกับเกียรติพงษ์ รัชตเกรียงไกร’ (announcement of weddings between volleyball player Feng Kun and Kiattong Radchatagriengkai) is highly correlated to the word span ‘สมรสกับ’ (who does Fengh Kun marry to?). It also can be pointed out that the mentioned passage span is the answer to the question.

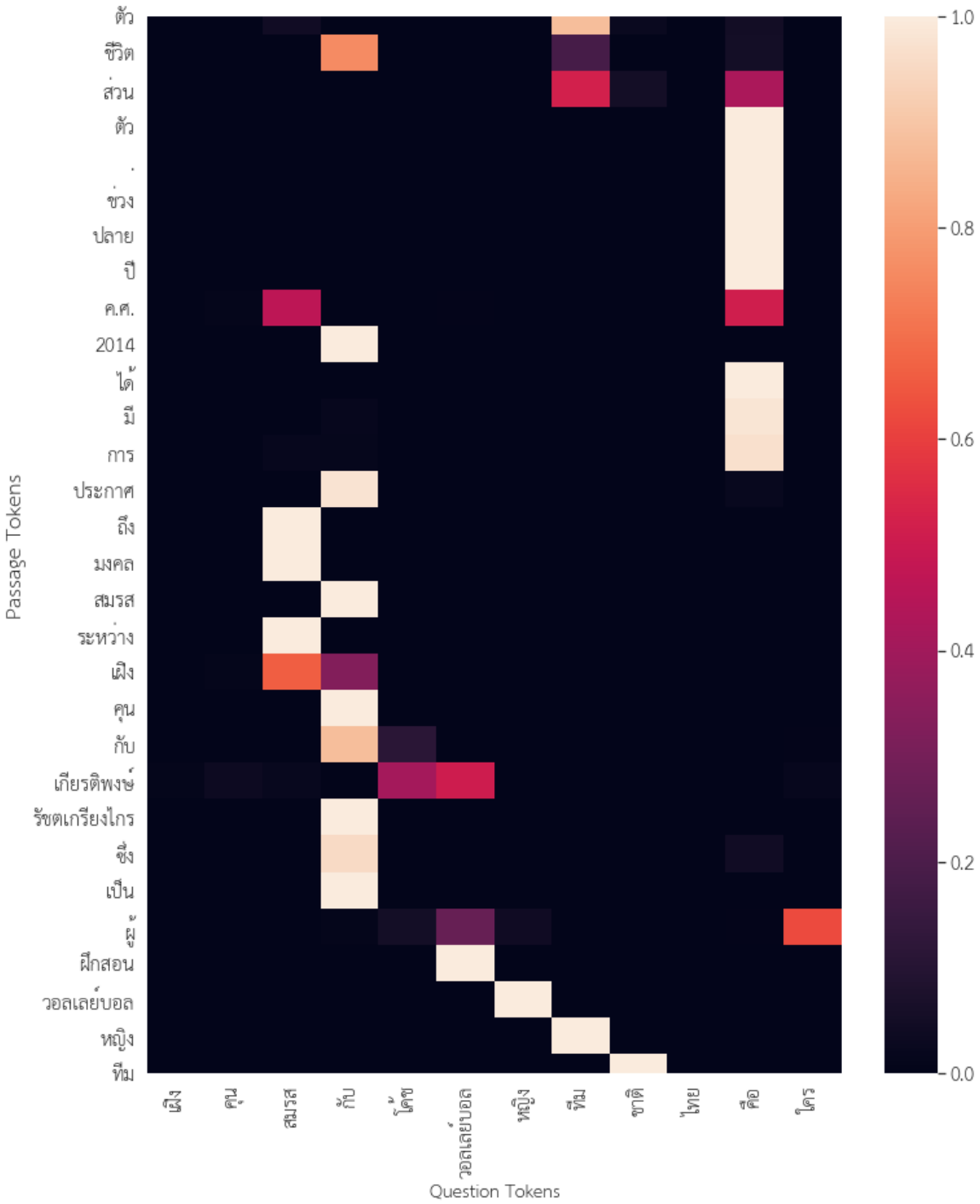


Figure 3. Passage-to-query or context-to-query attention heatmap. The right scale from 0 to 1 indicates the degree of importance of each question tokens to the context tokens.

2.6 General Architecture of Machine Reading Comprehension Model

There are many types of researches in the area of MRC in the English language. Liu et al. [11] have conducted a comprehensive review and survey of methods and trend in reading comprehension and has pointed out the general architecture of the deep learning models that are used in this field of research.

2.6.1 Embedding Layer

This layer serves the purpose of encoding or mapping natural language into meaningful dimensional space. Word vectors, contextual embeddings, and linguistic features like part-of-speech and name entities are normally used as input vectors for this layer.

2.6.2 Feature Extraction Layer

Various deep learning techniques and architecture can be used to extract the information from the input embeddings in both question and context vector. Commonly used methods include the utilization of RNN or CNN while some recent researches employ a purely attention-based architecture in this layer [20].

2.6.3 Context Passage and Question Interaction Layer

In this step, the attention mechanism has been a widely used technique in capturing the interaction between context vectors and question vectors. There are various kinds of attention mechanisms that can be utilized in context and question interaction layer. Lie et al. [11] classifies them into 2 main categories: unidirectional and bidirectional attention. Unidirectional attention mostly employs attention from query to context only while bidirectional uses the attention from both context to query and query to context direction. Bidirectional attention is proven to be better and examples of works that rely on such mechanisms include [10] and [21].

2.6.4 Answer Prediction Layer

This module usually varies from task to task based on the type of answers that the model needs to predict. Examples include prediction of the single word [12], spans of words [2], [3], selecting the correct answer from multiple choice [13], or free form text generation. Our work mainly deals with the span prediction type. For span prediction, the boundary method which is a method of selecting answer spans by predicting the start position and end position of the answers is normally used.

2.7 Pre-training and Transfer Learning

Transfer learning is a learning paradigm that utilizes the ability of the models or agents that were trained for one task to another task. In NLP, a large amount of unlabeled corpus was used to create representations of words or sentences through unsupervised learning setting. This step can be referred to as a pre-training step. Such representations can then be incorporated and help aid the model learning in a supervised learning task.

Word2vec [22] and Glove [23] were one of the first examples of transfer learning in NLP. Examples of more recent approaches of pre-training and transfer learning include ELMo [14], and BERT [15]. Transfer learning also plays a significant role in machine reading comprehension research. Question answering is also normally used as a performance benchmark of many pre-training models. This phenomenon also contributes to the fact that many of the top performers in English machine reading comprehension utilize major pre-training and transfer learning researches.

3. Related Works

In this section, we review the works related to the machine reading comprehension dataset, related Thai machine reading comprehension research, the architectural detail of BIDAF [10], and explore researches that attempt to modify attention mechanisms that are better suited for MRC task.

3.1 Machine Reading Comprehension Dataset

Various English MRC datasets were curated to drive the research. The majority of the datasets were created through the mean of crowdsourcing. SQuAD [2] is one of the first large scale dataset that is suitable for building data-driven deep learning architecture. SQuAD 2.0 [24] introduces an addition of unanswerable questions to the original dataset. QuAC [3] and CoQA [4] aim to introduce the task of conversational machine reading comprehension. In this task, the model must answer a series of questions that mimic the real-world conversations, so the questions are present in the form of multiple turn questions. shARC [25] also tasks the model to answer questions from a series of conversations but focusing more on answering questions related to regulations and rules. HotpotQA [5] introduces a dataset consisting of questions that require reasoning from multiple evidence to support the answer. RecipeQA [6] combines NLP and image processing and task the model to answer questions based on both textual information and pictures related to cooking recipes.

SQuAD [2] and SQuAD2.0 [24] are similar to the dataset in our research in the sense that they both require the model to perform span extraction and the model must answer different types of questions. Another MRC dataset that is related to our research is BoolQ [26]. Clark et al. [26] has curated a dataset that contains only naturally occurring yes-no questions and pointed out the challenges in the task. BoolQ differs from our work as the task does not require the model to support factoid questions. Clark et al. [26] also discovered that for ‘yes-no’ questions, transferring knowledge from inference task yield better result than transferring knowledge from span retrieval question answering task. Conversational reading comprehension is also related to our work because some of the questions found in the dataset are yes-no questions.

3.2 Multiclass Questions Reading Comprehension

In English MRC research, multiclass questions in reading comprehension tasks can be found in conversational reading comprehension datasets such as CoQA [4] and shARC [25]. In these datasets, some questions can be answered by ‘yes’ and ‘no’ which resembles our research area. In QuAC [3], different types of questions exist including yes-no, dialog act, answerable and unanswerable questions. The difference between these datasets and our focus

dataset (NECTEC) is that questions are conversational-based and some questions may relate to other prior questions in the conversation. In our research, we focus on single-turn questions.

Choi et al. [3] have modified BIDAF to answer different types of questions by appending special tokens of 'no answer' to the context passage. The model predicts the special tokens or predict the span of texts depending on the class of questions. Zhong e al. [27] employed a transformer-based model to encode document vectors, question vectors, scenario information, and historical conversation dialog altogether before passing to later layers of the model. Ohsugi et al. [28] concatenated hidden representations used for start and end prediction then pass the concatenated vectors to dense layer for answer type predictions. Ju et al. [29] have also utilized 3 different dense layers for outputting 'yes', 'no', and 'unknown' based on classes of questions. In our work, we propose to integrate the question classification (section 4.2.2.3) module into the machine reading comprehension model to help guide the model during the prediction term.

Examples of multiclass questions in our research can be seen in Figure 4. The factoid question translates to 'Where does the 2010 Women's Futsal World Tournament take place?' and the yes-no question translates to 'Does diamond cutting originate from Germany in 1375?'.

Factoid Question: การแข่งขันฟุตซอลหญิงโลกในปี ค.ศ. 2010 จัดขึ้นที่ประเทศใด

Yes-no Question: การเจียระไนเพชรถูกริเริ่มขึ้นโดยชาวเยอรมันในปีค.ศ.1375ใช่หรือไม่

Figure 4. Example of multiclass questions.

3.3 Boolean Question Answering (BoolQ Dataset)

We dedicate this subsection to discuss about Boolean question answering dataset called, BoolQ [26]. Clark, et al. [26] has curated an English dataset solely comprises of 16,000 naturally occurring Boolean or yes-no questions. The questions are pooled from records of Google search engine's queries hence providing elements of natural occurrence. Other preceding MRC datasets may contain some yes-no questions but are not the majority type of questions. Examples of such Boolean questions are shown below in Figure 5.

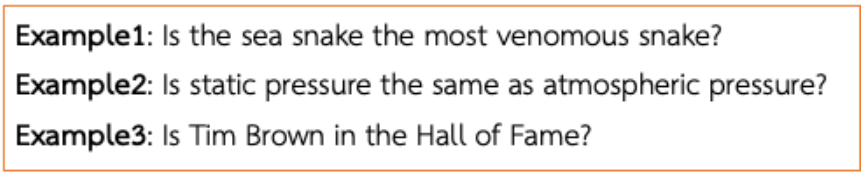


Figure 5. Examples of Boolean questions which are sampled from Table2. in [26]

Apart from the dataset curation, Clark, et al. [26] have also established various baselines including reader models with transfer learning from other MRC datasets or reader models from pre-trained LMs such as and ELMo [14], OpenAI GPT [30], and BERT [15]. In [26], it is found that using transfer learning from natural language inference (NLI) dataset, such as MNLI [31], provides a better result than transfer learning from extractive QA dataset like SQuAD 2.0 [24] or multiple-choice QA like RACE [13] despite having move similar format than NLI dataset. Based on this discovery, we also perform similar experiments to boost the performance in Thai MRC yes-no questions as discussed in section 4.3.

3.4 Thai Question Answering Research

Decha et al. [9] have developed a Thai question answering system focusing on factoid questions from Wikipedia using the pipeline concept. First, the system performs word and sentence segmentation using a machine learning approach, a trained artificial neural network that can predict the sentence boundary. In the next stage of the system, the questions are classified into different categories based on lexical rules. Keywords from question can then be extracted and used to retrieve answer candidates from the passages. The retrieved candidates will be used in word order consistency function to select the best alternative. Word order consistency is a heuristic function used to measure sentence structure similarity between the questions and the source documents.

In our research, we propose to create an end-to-end deep learning model that can support both the question and answer while [9] applied machine learning only in word segmentation step and sentence segmentation.

Another Thai question-answering research was conducted by Kongthon et al. [8]. The research focused on answering tourism-related questions by querying the ontology. Natural language questions are converted into query language format to find the answers. Jitkrittum et al. [7] developed a Thai question answering system, which also relies on structured data query language. Our research will be based upon finding the answers span from unstructured Wikipedia articles and not focus on tourism-related questions.

3.5 Comparison of Deep Learning Thai NLP Researches

As there are many publicly available Thai NLP deep learning models for various NLP tasks. Jettakul et al. [32] have conducted a comprehensive survey of the performance of different models to Thai NLP tasks, including Tokenization, part-of-speech tagging, named-entity recognition. Noisy nature of natural language data was artificially created and injected into the experimental datasets. The study has revealed that Synthai performs the best in terms of tokenization, V-BLSTM-CRF is the best model on named-entity recognition, and BLSTC-CRF is the best in part-of-speech tagging task.

3.6 BIDAF

BIDAF has been used as a baseline in many reading comprehension datasets that were released after SQuAD, as can be seen in [3], [4], and [33].

The key contribution of BIDAF lies in its design of the attention flow layer. In BIDAF, the attention vectors are concatenated with the embeddings from the previous context embedding layer and flow through the downstream layer of the models. This flow of attention vector is a contrast to the design of attention mechanism in other reading comprehension models in which the attention vector is used to summarize the question and context vectors.

3.7 Attention Mechanisms in MRC

In the field of MRC researches, many authors have explored the possibility of modifying or introducing new attention mechanisms to aid the model's performance on the MRC task. Qu, et al. [34] has developed global self-attention and unidirectional attention mechanism for Conversational MRC task while previous works utilize RNN to forward question-turn level information. Xie, et al. [35] incorporated semantic features and metadata features into the calculation of attention mechanism to help the model in answer selection task and solve the attention divergence problem. [36] introduced a new attention mechanism called extAdditive for cloze style question answering. In our work, we focus on testing if the yes-no questions require different attention mechanisms from factoid questions and check this idea by omitting the context-to-question attention mechanism, which is present in the design of BIDAF [10]. We discuss this experiment thoroughly in section 4.4.

4. Methodology

In this work, we aim to develop a novel Thai machine reading comprehension model that supports multiclass questions. Our proposed models have the ability to handle factoid questions and yes-no questions of questions through the integration of question classifier module and yes-no prediction module. Multi-task learning is employed to help train the joint architecture model. We also propose to enhance the models with contextual embeddings constructed from a pre-trained language model, namely BERT [15]. Since our experiments suggest that cascading architecture has the best performance, we further enhance the cascade model with a transfer learning scheme designed specifically to increase the model's accuracy on yes-no questions. Lastly, to increase the yes-no prediction accuracy further we modify the attention mechanism by omitting the context-to-question attention mechanism from the network.

In this section, we discuss the dataset preprocessing steps in section 4.1, our proposed multiclass model architecture in section 4.2, transfer learning scheme in section 4.3, and modification of attention mechanism in section 4.4. The overall workflow of our proposed work is shown in Figure 6.

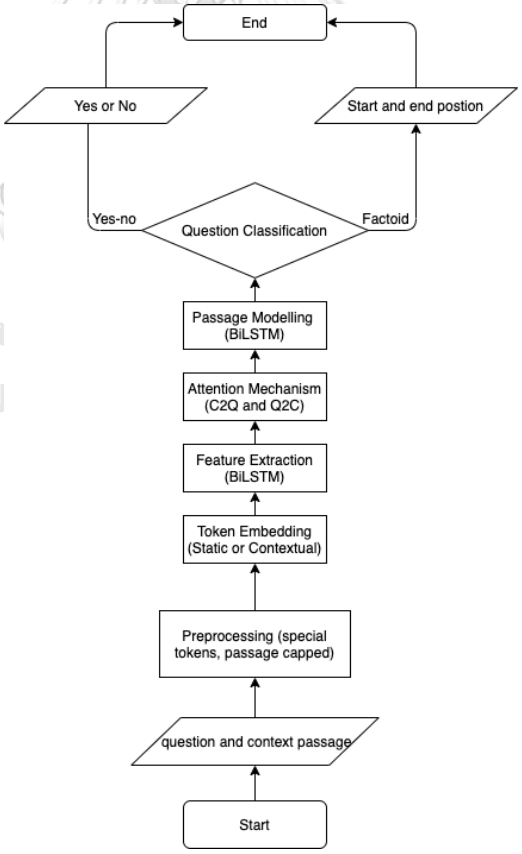


Figure 6. A high-level overview of our proposed model.

4.1 Dataset Preprocessing

We discuss the steps of dataset preprocessing before we pass the textual information to the MRC models in this section.

4.1.1 Match Wikipedia article with questions

The questions provided in the dataset by NECTEC do not have accompanying context passage but can be mapped manually with Wikipedia using provided article ID.

4.1.2 Remove HTML tag

Thai Wikipedia corpus contains HTML tag are needed to be removed before passing the context passage to the model as these HTML tags do not provide any information related to the questions.

4.1.3 Appending YES/NO tokens (Special token model)

After removing the HTML tag from the context passage, we added YES/NO tokens to the passage. The added tokens will serve as answer spans for yes-no questions. The answer start and end position are shifted accordingly to preserve the position of original ground truth tokens for factoid questions if the YES/NO tokens are added to the start of the context passage. This method of adding special token is similar to QuAC [3] where the authors of the dataset append special token [UNS] to the passage. This method of appending special token applies to the model in section 4.2.2.1 only. We have experimented with both adding YES/NO token to both at the start of the context passage and the end of the context passage. We have found that the latter performs significantly better than appending to the start of the passage. This result can be found in appendix A.

4.1.4 Start and end positions of ground truth answer

The start and end position of the answers for each factoid question is based on character positions. As the evaluations of the models are conducted at the token spans level rather than the character's level, we need to properly map those character positions into positions of token spans. For example, if the answer to a certain question is "โรงเรียน" (school) and the context passage is "นักเรียนไปโรงเรียน" (student goes to school). The start position is 11 at vowel "โ", and the end position is 19 at alphabet "น". As F1 evaluation metric is calculated based on token levels. We also need to map character positions into token positions as well. In the previous example, if we tokenize "นักเรียนไปโรงเรียน" (student goes to school) into "นักเรียน" (student), "ไป" (goes to), and "โรงเรียน" (school). The ground truth tokens will be at position 3.

Possibility of wrong word tokenization exist, which could result in some tokenized words do not accurately match with the actual ground truth answers. Another possible scenario is when the character starting positions of the answers do not match with the start positions of the tokenized passage tokens. For example, the actual answers tokens are [นักเรียน, ชอบ, ดื่ม, กาแฟ]

while the tokenized passage tokens are [แต่นัก, เรียน, ชอบ, ดื่ม, กา, แฟมาก]. It can be observed that the first character of the answer ‘น’ does not appear exactly at the start of a token as it should and the final character ‘ฟ’ also does situated at the end of the token as well. In such case, we will mark the first token ‘แต่นัก’ as the start of the answer and the final token ‘แฟมาก’ as the end of the ground truth token spans for the training process.

4.2 Proposed Multiclass Machine Reading Comprehension (MRC) Model

In this section, we explain our proposed multiclass MRC models designed to support factoid and yes-no questions in Thai MRC corpus. We start by describing the method of integrating contextual embeddings to the MRC models in section 4.2.1. Then, the designs of different architecture are explained in section 4.2.2. We present 3 different architectures which are special token, joint architecture, and cascade architectures.

4.2.1 Integration of Contextual Embeddings

To incorporate contextual information for the reading comprehension model, we propose to replace the normal word embeddings with the contextual embedding before inputting the vectors into the model. This should help model capture and understand complex context-dependent information and ultimately perform better in our question-answering task. As the size of the dataset is not large, transfer learning from these models that were pre-trained on large Thai corpus should also be beneficial to the model’s performance. This method of integration is inspired by [14] where the authors incorporate contextualized embeddings from ELMo architecture and they boosted question answering performance of BIDAF by 4%. We also would. Like to note that previous Thai question answering researches and competitions have not incorporated contextual embeddings into the models. We have also shown examples where the model with contextual integration performs better than static word embedding in sections 7.1 and 7.2. The illustration of our proposed implementation can be seen in Figure 7.

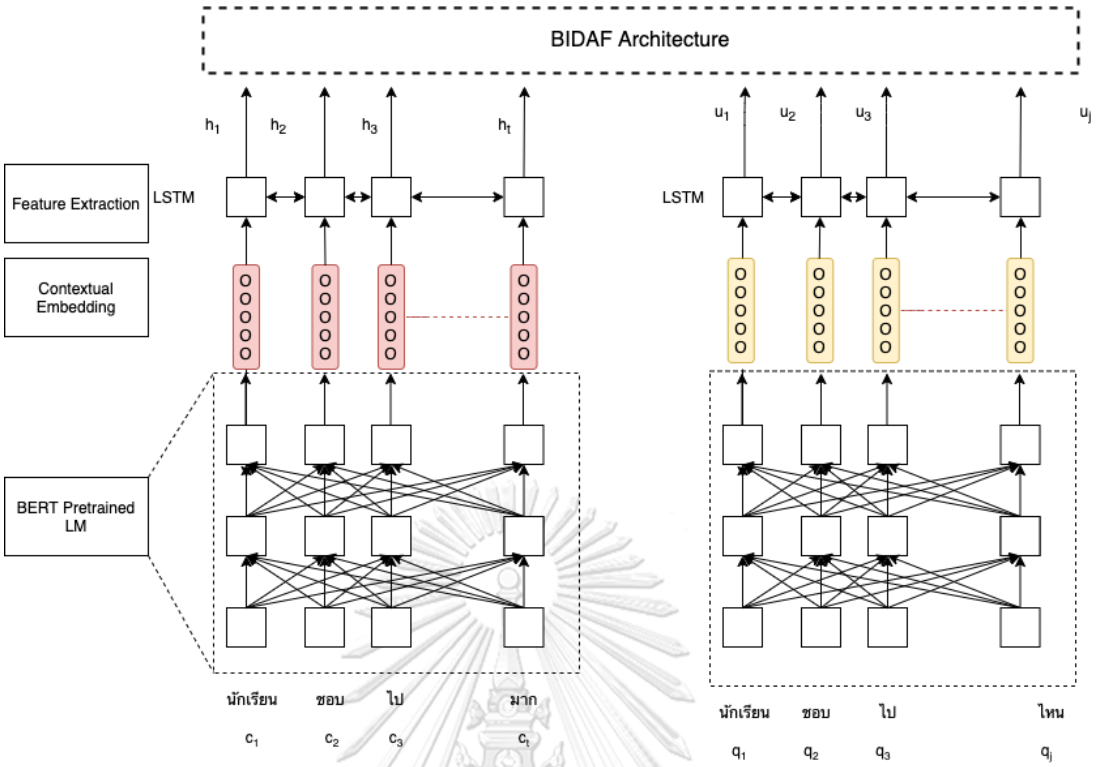


Figure 7. Integration of Contextualized Embeddings into our BIDAF model.

We choose Thai monolingual BERT as the pre-trained language model for contextual embeddings extraction. Our method of utilizing BERT [15] as contextual embeddings extractor is similar to the method proposed in [37], where the authors use incorporate English version of BERT for English conversational machine reading comprehension (CMRC). In our study, we use BERT to support multiclass machine reading comprehension by constructing contextual embeddings for the BIDAF models.

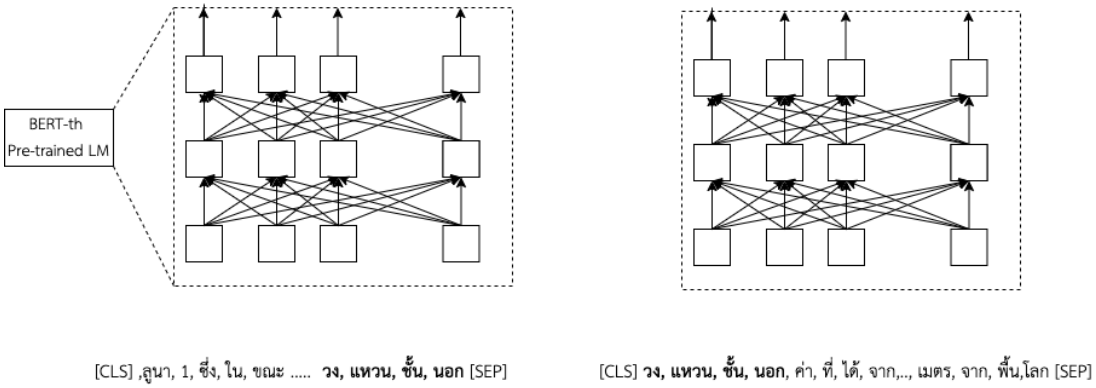


Figure 8. Illustration of input of BERT when using sliding windows, tokens in bold are overlapped between 2 windows.

We apply BERT to context passages and questions separately. We append BERT specific token [CLS] to the start and [SEP] for the sequence of texts that are passed into BERT. If the context passages tokens are longer than 510 tokens which is the size of sequence length in normal BERT architecture, we use the sliding window approach to help extract the context passage. The input into BERT with the usage of the sliding window is shown in Figure 9. The length of each sliding window is 128 BERT token positions. In each window, we use the centermost sequences as the contextual representations as illustrated in Figure 9. As BERT utilizes the self-attention mechanism in transformer architecture [18], the centermost representations are the ones that are the most exposed to surrounding context information.

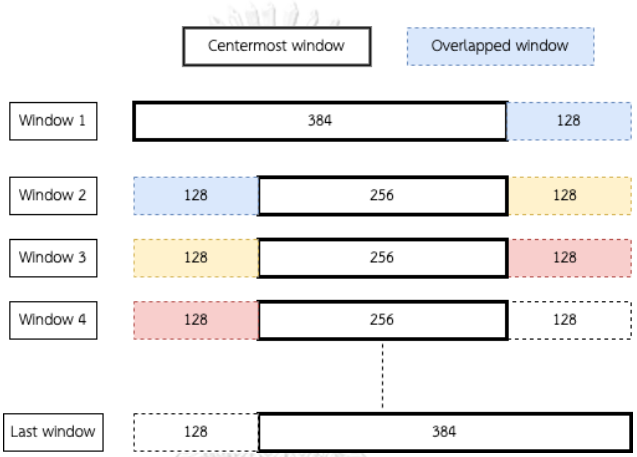


Figure 9. Illustration of the sliding window position in BERT.

We sum all 12 layers of hidden representation from BERT to be used as contextual embeddings for the BIDAF model. As BERT uses different tokenizer from the one we use in our reading comprehension model, we map the representations from BERT tokenized word to Bailarn tokenized words by choosing the last sub-token of the same token to represent the token. For example, if the BERT tokenized tokens are [‘กา’, ‘แฟ’] and Bailarn tokenized tokens is [‘กาแฟ’], we use the representation of BERT tokenized token ‘แฟ’ to represent the word ‘กาแฟ’.

4.2.2 Multiclass Question Architecture

In addition to the existing architecture of BIDAF [10] that deals with span prediction task, we augment the model with classifiers which deal with different types of questions. We propose 3 different architectures which are special token (section 4.2.2.1), joint model (section 4.2.2.2), and cascade models (4.2.2.3).

4.2.2.1 Special Token Architecture

In this variation, we will input both types of questions into our model block indiscriminately. For factoid questions, the model will have to predict the spans that contain the correct answer while for yes-no questions, the model needs to point the start and end position to the special token at either 'YES' or 'NO'. The model diagram can be seen in Figure 10. This design is based on a modification of BIDAF in [3], which is designed to deal with unanswerable questions. We have found that appending special tokens 'YES' and 'NO' to the end of the passage performs better than appending to the beginning of the passage. This model acts as a baseline in terms of multiclass architecture because there is no significant modification done to the model, the modification is done in the input side only.

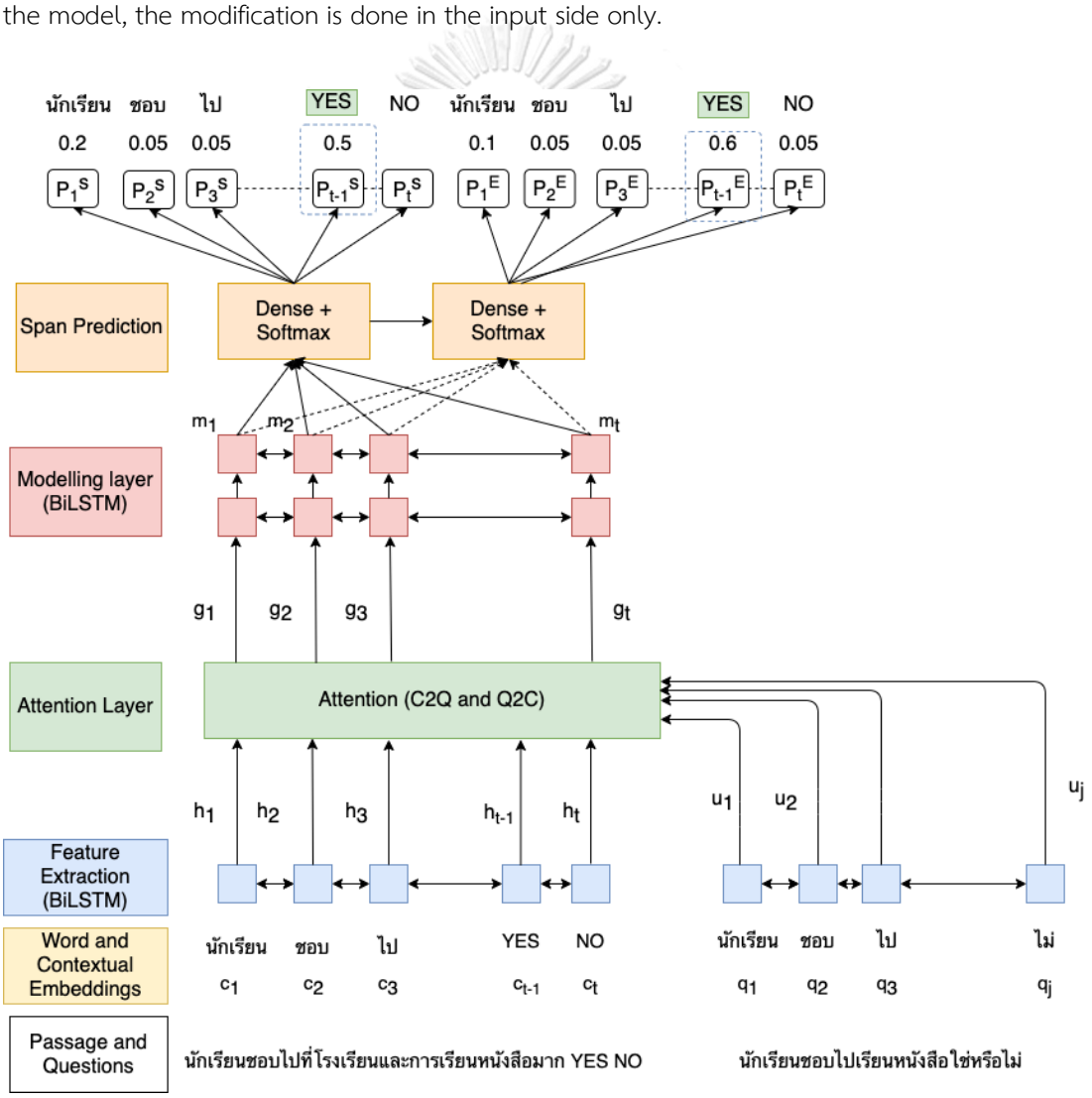


Figure 10. The special token architecture. In our example question, 'do the students like to study?' is a yes-no question, so the model needs to predict special token 'YES' or 'NO'.

Since this model's prediction mechanism is span retrieval. The objective function for this model is the span prediction loss only, similar to the original BIDAF [10].

$$L_{span} = \frac{-1}{N}\sum_{i}^{N} \log\left(P^{S}{}_{y_i^S}\right) + \log\left(P^{E}{}_{y_i^E}\right) \quad (1)$$

L_{span} denotes a loss of the span retrieval module. This loss is essentially cross-entropy losses that are calculated across tokens found in context passages. L_{span} is the sum of the loss of actual start and end indices. N represents the number of learning examples. P^S and, P^E are the predicted start and end indices respectively.

4.2.2.2 Joint architecture

In this architecture, we utilize the question classifier module to classify different types of questions. We hypothesize that using information from the question side alone is enough to classify their types. In the proposal, we proposed to classify types of questions using a fully connected layer. However, we have found that we can simply differentiate types of questions by checking Thai keywords, so we replace the fully connected layer with this method instead. The incorporation of question classifier can be seen in Figure 11. We describe the keywords used for classifying types of questions in section 4.2.2.3.

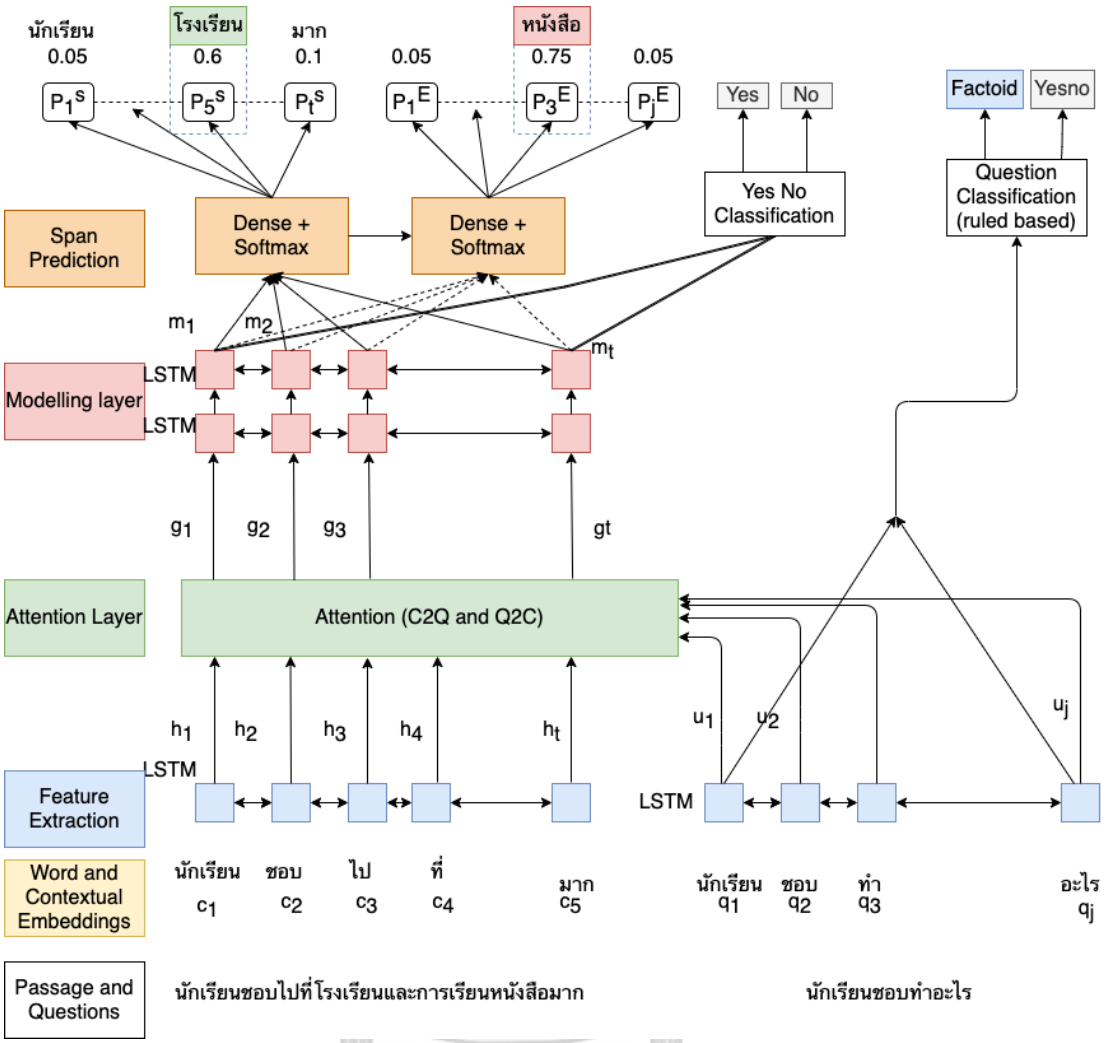


Figure 11. Our proposed joint architecture. In this example query, the question is 'what do the students like to do?' (นักเรียนชอบทำอะไร), which is a factoid question.

In this architecture, we also introduce a 'yes-no classifier' module to predict the answer of yes-no questions. This new module is located at the prediction level, which is the same level as the span retrieval module and utilize the enhanced context representations which have incorporated information from both context and query passages. The equation for the objective function of the yes-no classifier can be seen in equation (2).

$$L_{YN} = \frac{-1}{N}\sum_{i}^{N}((1 - y_i^{YN})\log(1 - q_i^{YN}) + y_i^{YN}\log(q_i^{YN})) \quad (2)$$

$$q^{YN} = softmax(W^{T}{}_{YN} \max_{1 \le i \le d}[M; G]_{:dt}) \quad (3)$$

L_{YN} represents the loss function of the yes-no classifier, which is binary cross-entropy loss. We compute this loss for yes-no questions specifically. y_i^{YN} is a binary indicator of yes-no question's answers and, q_i^{YN} denotes the yes-no class probability. $W_{YN} \in R^{10d}$ denotes trainable weights for the yes-no classifier. $M \in R^{2d\,x\,T}$ denotes the enhanced context representations, d is the dimension of the hidden vector while T is the number of tokens in context passages. $G \in R^{8d\,x\,T}$ is the output of the attention layer. The equation (3) uses element-wise max to construct a single vector that represents context passage. This representation vector can be then used for yes-no prediction.

$$L_{total} = L_{Span} + L_{YN} \qquad (4)$$

During the training process, we compute loss for the span extraction module for factoid questions only. If the question is a yes-no question, the only loss for the yes-no classifier is computed. During the inference time, we use either span retrieval module or yes-no module depending on the output of the question classifier. We then combine the loss of two tasks as shown in equation (4). We stick with the combination of 1:1 as our preliminary experiment suggests that such combination has the best performance in Appendix B. Our proposed question classification method shares some similarities with [38]. In [38], The authors focus on modeling different types of questions to answer factoid questions in SQuAD [2] but in our proposed research, we aim to utilize question classification to help guide the prediction of multiclass questions, namely factoid and yes-no questions.

4.2.2.3 Cascade Architecture

In this architecture, we utilize a question classifier, which is similar to the question classifier module in joint architecture (section 4.2.2.2), to classify types of questions beforehand. After the classification, we will train separate models for different types of questions separately. The overall architecture is illustrated in Figure 12. BIDAF for span prediction has only span prediction module and uses the same objective function as equation (1). On the other hand, BIDAF for yes-no prediction has only the yes-no classifier and uses the yes-no objective function described in equation (2).

For the question classifier, we have found that we are able to differentiate yes-no questions from factoid questions simply by searching for keywords that do normally appear in the yes-no question. The keywords used for distinguishing types of questions are 'ใช่หรือไม่', 'ใช่ไหม', 'ใช่มั้ย', 'ใช่หรือไม่ใช่', and 'หรือไม่'.

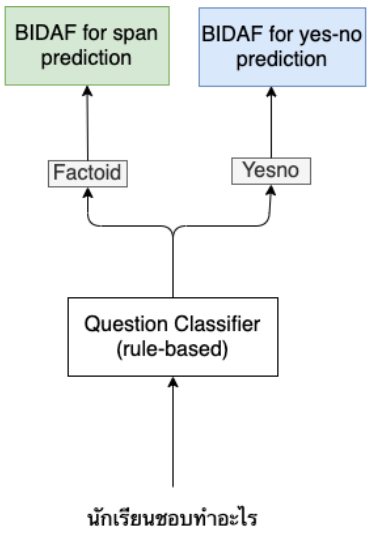


Figure 12. Architecture of the cascade model. Different BIDAF models are trained for different types of questions.

4.3 Transfer Learning from Natural Language Inference (NLI) Dataset

After we have found that cascade architecture achieves the best performance (section 6.3). The model's accuracy on the yes-no question still has room for improvement. We further enhance the cascade architecture (section 4.2.2.3) with transfer learning from natural language inference. Our idea is based on Clark, et al. [26], where the authors had demonstrated the effectiveness of using transfer learning from different MRC datasets to the Boolean dataset. Clark, et al. [26] had observed that transferring from natural language inference yields a significant increase to the model's performance on Boolean questions. In their recurrent models, this accuracy is increased by 5.97%, when comparing to training from scratch. The increase in performance is also greater than transferring from extractive MRC datasets like SQuAD [2], where the performance is increased by 3.18%.

Inspired by these findings, we also compare the effectiveness of utilizing transfer learning from extractive MRC dataset to yes-no questions against transferring from the NLI dataset to yes-no questions in the Thai MRC dataset. For extractive MRC, we use the factoid questions in the NECTEC V2 dataset. We select Thai sentence pairs found in XNLI [39] dataset as the NLI dataset to be transferred to yes-no questions. We refer to Thai sentence pairs from the XNLI corpus as the XNLI-th dataset. More detail of the XNLI dataset is explained in section 5.1.2.

We compare various transfer learning schemes to assess the effectiveness of transfer learning from each dataset. Figure 13. highlights different scenarios of transfer learnings in our experiments.

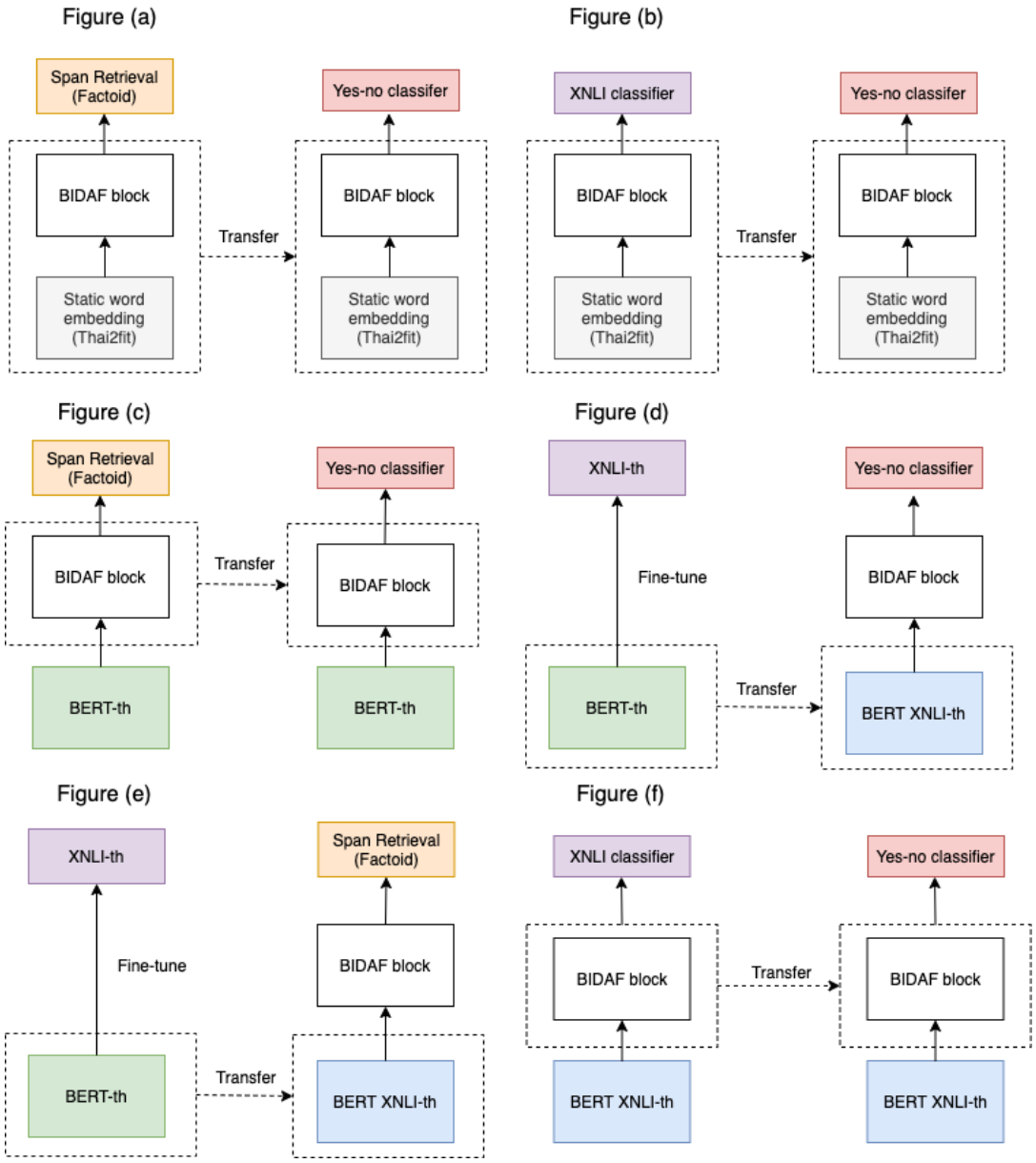


Figure 13. Illustration of different transfer learning setting. (a) top-left: from factoid to yes-no, (b) top-right: from XNLI-th to yes-no, (c) middle-left: from factoid to yes-no using contextual, (d) middle-right: transfer via fine-tuned BERT, (e) bottom-left: use fine-tuned BERT for factoid and (f) bottom-right: pre-training BIDAF with fine-tuned BERT.

4.3.1 Transfer learning from Factoid Questions with Static Word Embeddings

We transfer the weights of BIDAF for span prediction in cascade architecture (section 4.2.2.3) to be used as initial weights for yes-no prediction. The weights that are transferred are from the token embedding layers up to the modeling layer as in Figure 13 (a). This variation serves as a comparison for transfer learning from XNLI-th.

4.3.2 Transfer learning from NLI with Static Word Embeddings

We modify cascade architecture (section 4.2.2.3) to handle the XNLI-th by changing the number of classifier outputs from 2 to 3 as XNLI-th has 3 classes to predict. We discuss the detail of the XNLI-th dataset in section 5.1.2. The objective function of the model is also updated to be cross-entropy loss. The rest of the model architecture is the same as the cascade architecture used for yes-no prediction. Similar to 4.3.1, we transfer the weights up to the modeling layer, and initialize new weights for yes-no classifier task. This setting is shown in Figure 13 (b).

4.3.3 Transfer learning from Factoid Questions with Contextual Embedding

This is essentially similar to the method discussed in section 4.3.1, except we change the static word embeddings to BERT-th, as displayed in Figure 13 (c).

4.3.4 Transfer learning from NLI with BERT fine-tuning

Another approach to use BERT, besides contextual embeddings extractor, is to fine-tune BERT to specific NLP tasks such as text classifications and natural language inference task. We fine-tune BERT-th on the classification task in XNLI-th. The method of fine-tuning in this setting is similar to the way Devlin, et al. [15] fine-tune BERT for MNLI [31], which is also a corpus for NLI task.

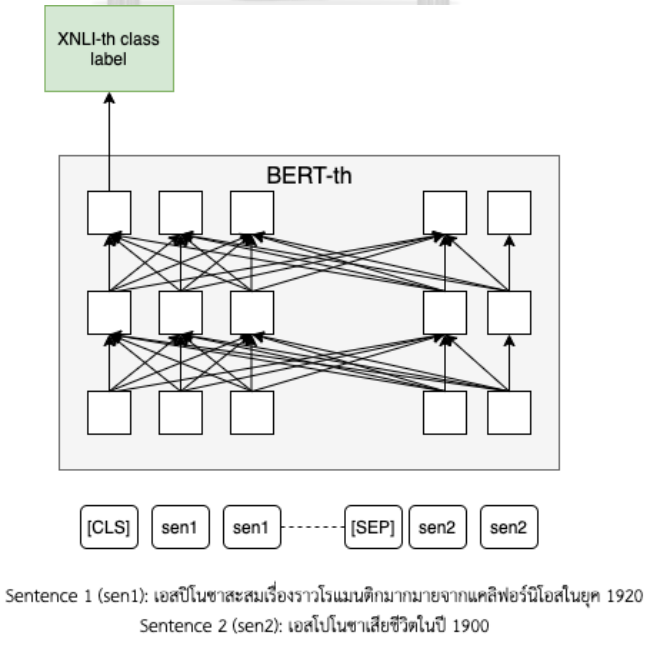


Figure 14. Illustration of fine-tuning BERT on XNLI-th task.

In NLI, the sentence pairs are divided by BERT's special token '[SEP]' which is illustrated in Figure 14. The main idea of fine-tuning BERT for the NLI task is to add a fully connected layer or dense layer for class prediction. The representation vector of the special token '[CLS]', which can be viewed as a single representation vector of the sentence pair, is passed into the fully connectedly. Representation from the topmost layer of BERT is fed to the classifier layer.

$$q^{XNLI} = softmax({W_{XNLI}}^{T}\mathrm{C}) \quad (5)$$

$C \in R^{H}$ represents the last layer representation of token [CLS], $W \in R^{K\,x\,H}$ represents the trainable weights of the XNLI classifier and K is the number of class which is 3 for our case. q^{XNLI} is the output classes of the XNLI-th task, which can be one of the following: entailment, neutral, and contradiction.

After the fine-tuning, we can then use the fine-tuned BERT-th to create contextual embeddings for our MRC model similar to the approach discussed in section 4.2.1. We replace the BERT-th in Figure 8 with BERT-th that was fine-tuned on XNLI-th. We hypothesize that BERT-th will have the ability to construct contextual embeddings that are more suitable to the yes-no prediction task and will increase the MRC model's performance on the task. Transfer learning for this setting is shown in Figure 13 (d).

4.3.5 Transfer learning from NLI with BERT fine-tuning and BIDAF pre-training

This transfer learning setting uses BERT that was fine-tuned on XNLI-th, which is similar to setting in 4.3.4. Before we train the model on the target yes-no questions in NECTEC V2, we first pre-train BIDAF on yes-no questions like in 4.3.2 but we use BERT XNLI-th to create the contextual embeddings. Figure 13 (f) demonstrates this transfer learning scenario.

4.4 Dropping Attention Mechanism for yes-no questions

BIDAF was designed to have 2 attention mechanisms which are context-to-query and query-to context. BIDAF was built to deal with the extractive task which is equivalent to the task of answering factoid questions in our study. We suspect that, for answer yes-no questions, the model may not need both of the attention mechanisms to perform well. To answer yes-no questions, the reader must evaluate if the questions, which can be viewed as some forms of statements, is true or not based on the provided context passage. Based on this intuition, the context-to-query mechanism should not be as important as query-to-context. We design an experiment to omit the context-to-query attention mechanism and observe if the model performs better on yes-no questions.

The original implementation combines embeddings from contextual layer and attention vectors to yield G, which is defined as

$$G_{:t} = \beta(H_{:t}, \widetilde{U}_{:t}, \widetilde{H}_{:t}) \quad (6)$$

$$\beta(h, \tilde{u}, \tilde{h}) = [h; \tilde{u}; h \circ \tilde{u}; h \circ \tilde{h}\,] \quad (7)$$

G represents intermediate context vectors (t is the length of the context passage tokens). $\widetilde{U}_{:t}$ denotes context-to-query attention vector and $\widetilde{H}_{:t}$ denotes query-to-context attention vector. β is a trainable function used to fuse 3 vectors: $\beta \in R^{8d\ x\ T}$.

For our proposed attention mechanism modifications, we redefine G and β as per below.

$$G_{:t} = \beta(H_{:t}, \widetilde{H}_{:t}) \quad (8)$$

$$\beta(h, \tilde{h}) = [h; h \circ \tilde{h}\,] \quad (9)$$

As we reduce the attention vector dimensions, the size of fuse function β must also change: $\in R^{4d\ x\ T}$. The trainable weight vectors used in the span prediction layer and the yes-no prediction layer are also needed to be resized accordingly, from $10d\ x\ T$ to $6d\ x\ T$.

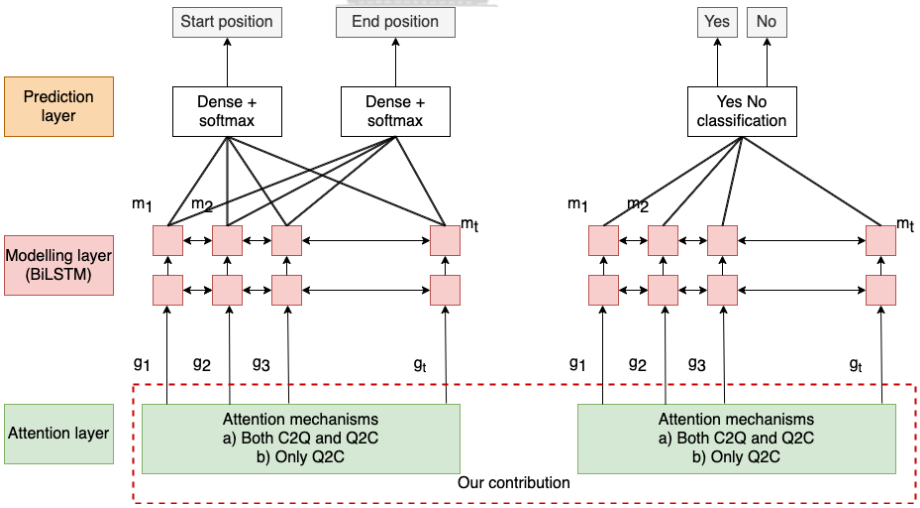


Figure 15. Modification of attention mechanisms for yes-no questions. We propose to use a) both attention mechanisms and b) only the question-to-context mechanism.

Similar to the method proposed in applying transfer learning (section 4.3), we perform this experiment on cascade architecture by modifying the attention mechanism in BIDAF for yes-no prediction. To confirm that dropping context-to-query mechanism does indeed benefit only yes-no questions, we also experiment with a variation of cascade architecture that drops context-

to-query attention in BIDAF for span predictions and BIDAF for yes-no predictions alike. The illustration of our experiments is shown in Figure 15.

5. Experiments

In this chapter, we describe the detail and implementation of our experiments. First, we present the statistics of the datasets used in this research, which are NECTEC and XNLI [39] datasets in section 5.1. Next, we present the detail of our implementations in section 5.2. The detail of the statistic test is in section 5.3 and the evaluation metrics are described in section 5.4.

5.1 Dataset Statistics

5.1.1 Question Answering Program from Thai Wikipedia

We use the dataset from the "Question Answering Program from Thai Wikipedia" competition under the Twenty-Second National Software Contest (NSC 2019). In this competition, the model must also have the ability to query for the Wikipedia article as well as answering the questions. In our work, we will focus on the machine reading comprehension aspect only. The statistics of the dataset is shown below in Table 2. We refer to the Thai datasets as NECTEC V1 and NECTEC V2.

Table 2. Comparison of Thai dataset statistics and it's English counterpart.

Dataset	NECTEC V1.	NECTEC V2.	SQuAD V1 + V2
Number of questions	4,000	17,000	161,560
- Factoid questions	4,000	15,000	107,785
- Yes-no questions	0	2,000	0
- Unanswerables	0	0	53,775
Average context passage length (tokens)	936	736	140
Average question length (tokens)	12.2	15.4	11.2
Average answer length (tokens)	2.4	1.11	1.61

Table 3. Yes-no Class Distribution.

NECTEC V2	Number of questions
Yes-no Questions	2,000
- 'Yes' as answer	994
- 'No' as answer	1,006

Bailarn tokenizer [32] was used for Thai tokenization. It can be observed that numbers of learning instance available in Thai dataset is still significantly smaller than ones in the English counterpart. Another observation worth noting is that the average tokens found Thai passage is significantly larger than its English peers. For yes-no questions, t\he answer class is balance with

questions with ‘yes’ as answers constitute 49.7% of total yes-no questions and questions with ‘no’ as answers constitute 50.3%.

From Figure 16, we can observe that majority of the context passage length is under 5000 tokens. Figure 17 demonstrates that the distributions of context passage length are highly skewness with the skew value of 5.75 for factoid questions and 8.72 for yes-no questions.

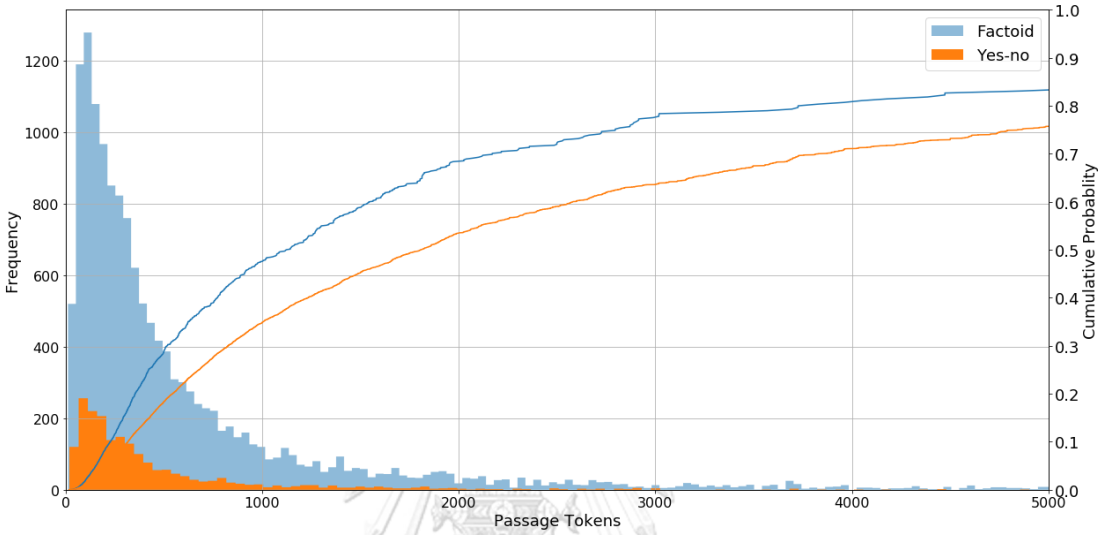


Figure 16. Context passage length (in number of tokens) distribution. We only show the context passage with token lengths of less than 5,000 tokens in this visualization.

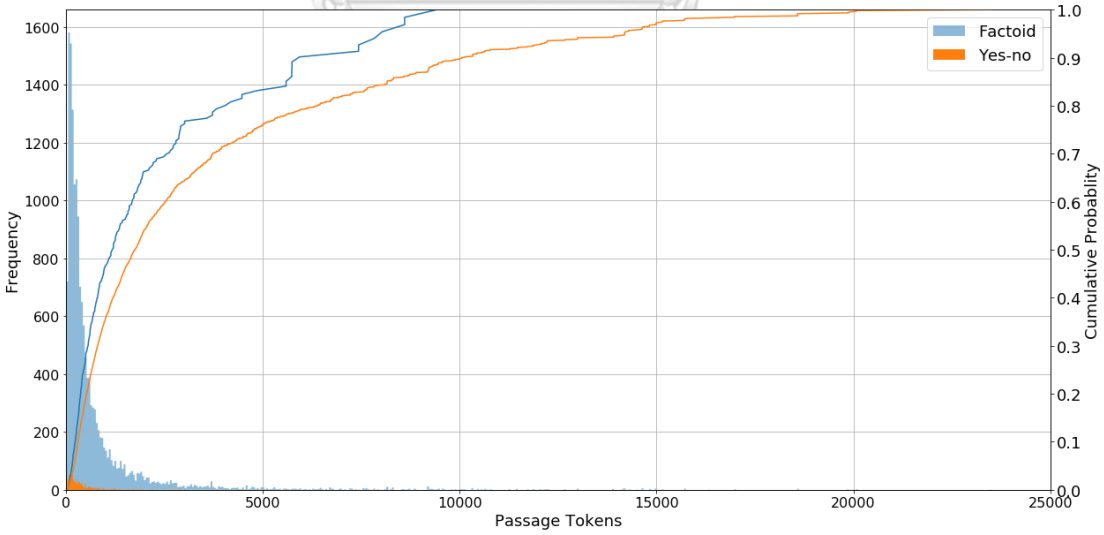


Figure 17. Context passage length (in number of token) distribution

Questions tokens length distributions are less skew than context length. Skewness of factoid questions is 0.815 while skewness of yes-no question is 0.129. The distribution is shown in Figure 18.

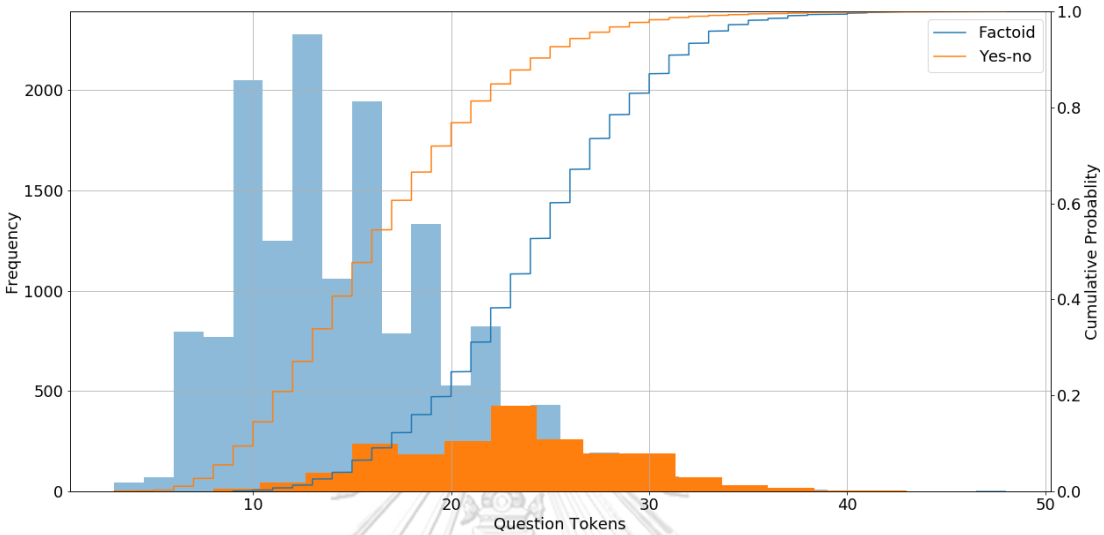


Figure 18. Question length distribution

The starting and ending positions of answers are shown in Figure 19. Similar to context passage length distribution, starting and ending positions are also highly positively skew. As Table 2 points out, the average value of answer spans is just 1.12 tokens so the cumulative probability distribution for starting position almost coincides with the cumulative probability distribution of ending position.

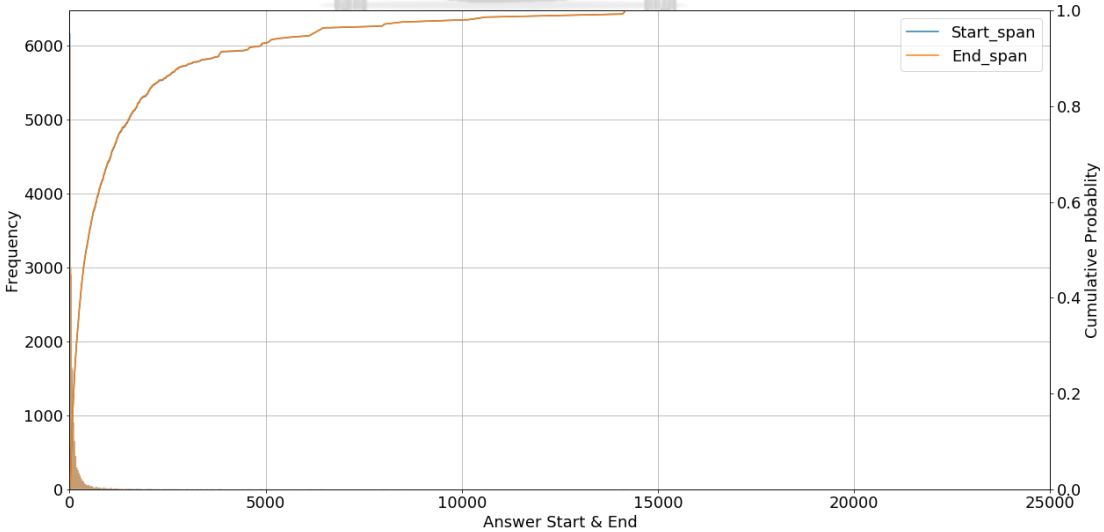


Figure 19. Starting and Ending Positions of Answers

5.1.2 XNLI-th

XNLI [39] is a corpus designed to help aid the research in the area of cross-lingual understanding (XLU) and is an extension of the MNLI [31] dataset, an abbreviation of multi-genre natural language understanding. MNLI is the dataset of the natural language inference task and is also the part of GLUE [40], general language understanding evaluation, benchmark. While MNLI focuses on the English language only, XNLI extends the scope to include 14 other languages including Thai. Conneau, et al. [39] achieved this by translating 7,500 sentence pairs from dev and test set of MNLI into other languages. The translations of validation and test sets were done by human experts. In addition to the translation of development and test sets, [39] also used a machine translation system to translate the sentence pairs in the training set into different languages to be used in one of their baseline methods.

Since we focus on the monolingual setting in our research, we incorporate only translated Thai sentence pairs in our transfer learning experiment (section 4.3). Table 4 describes the dataset statistics of XNLI-th.

Table 4. Statistics of XNLI-th. N stands for Neutral, E stands for Entailment and C denotes Contradictory classes in XNLI-th class distribution

XNLI-th	Number of sentence pairs	Average premise tokens	Average hypothesis tokens	Answer Distribution		
				N	E	C
Training set	386,442	25.2	12.07	128,842	128,828	128,772
Validation set	2,490	24.0	11.5	830	830	830
Test set	5,010	24.1	11.6	1,670	1,670	1,670
Total	393,942	25.2	12.1	131,342	131,328	131,272

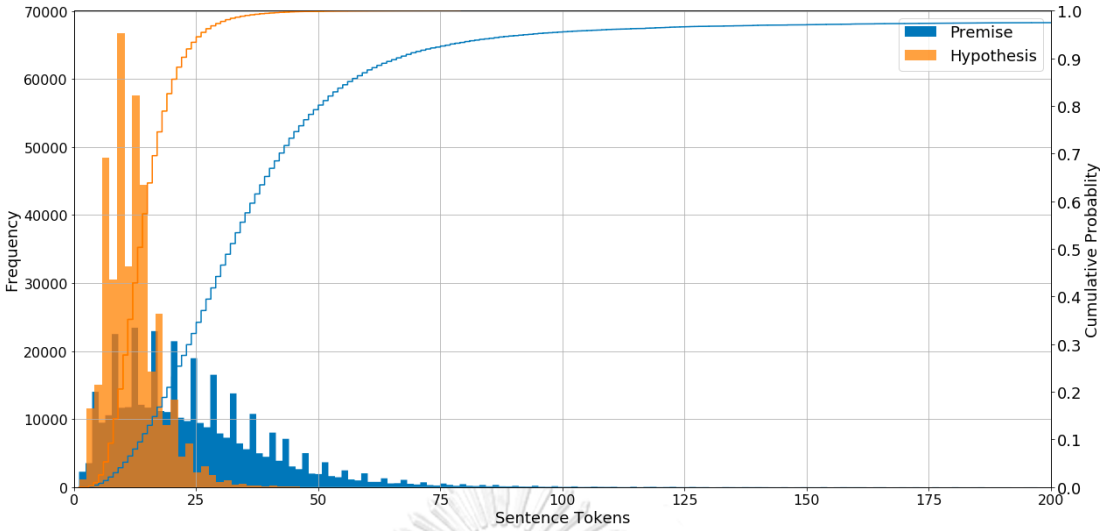


Figure 20. Distribution of number tokens in premise and hypothesis of all datasets in XNLI-th

Table 4 shows us that the class distribution is balanced across training, validation, and test dataset of XNLI-th. Another interesting aspect of this distribution is the number of tokens found in the sentence pairs, which is lower than the number of tokens found in context passage of the NECTEC dataset as shown in Table 2. Figure 20. shows us that the tokens in the premise sentences are more skew than tokens in hypothesis with the skewness of 73 vs 1.3 respectively.

5.2 Implementation Detail

We use the PyTorch library [41] for general deep learning implementation. The AllenNLP [42] framework is also used to aid the implementation of BIDAF and our modified versions of BIDAF in section 4.2.2. For BERT related implementation, we use HuggingFace's Transformer [43] framework. For the weights of pre-trained Thai monolingual BERT, we refer to this work[1].

For all experiments involving the NECTEC V2 question answering dataset, we use 3-fold cross-validation with stratified sampling. We then elaborate on the hyperparameter of our proposed models in the following section. For the experiment involving attention mechanism modification, we use the same hyperparameters as sections 5.2.1 and 5.2.2, the difference is the reduction in attention mechanism as explained in section 4.4.

5.2.1 Multiclass Architecture Hyperparameters

Table 5 describes the hyperparameters detail of each multiclass architecture. We did not include a highway layer in our implementation. Similar to original BIDAF implementation, Dropout is used for LSTM layers and prediction layer. We also use dropout at yes-no prediction layers. Epochs with the best performance of overall F1 on the validation set are used to evaluate the

[1] https://github.com/ThAIKeras/bert

performance on the test set. We freeze weights of BERT-th for experiments with contextual embedding.

Table 5. Hyperparameters for multiclass architecture. Asterisk indicates the implementation in the contextual embedding setting.

Hyperparameters	Special Token	Joint/Cascade
Word Embedding	Thai2fit v0.1	Thai2fit v0.1
Word Embedding (dimension)	300 / 768*	300 / 768*
Contextual Embedding	BERT-th	BERT-th
Tokenizer	Bailarn (Synthai)	Bailarn (Synthai)
BiLSTM Hidden dimension	100	100
Batch Size	10 / 5*	10 / 5*
Optimizer	Adadelta	Adadelta
Learning Rate	0.1	0.1
Training epochs	40	40 (for joint and BIDAF for factoid) 100 (for BIDAF for yes-no)
Passage length (tokens)	5000 / 2500*	5000 / 2500*
RNN components	BiLSTM	BiLSTM
Vocabulary	Only include pre-trained word + minimum of 3 occurrence	Only include pre-trained word + minimum of 3 occurrence
Similarity Function	DotProduct Similarity	DotProduct Similarity
Dropout	0.2	0.2
Yes-no Dense layer (s)	-	2 layers with ReLU
Yes-no Dense layer hidden dimension	-	200

5.2.2 Transfer Learning Hyperparameters

For transfer learning from XNLI-th, the overall model's hyperparameters are the same as section 5.2.1 except for the learning rate. For transfer learning settings that BIDAF parameters are transferred, discussed in section 4.3.1, 4.3.2, 4.3.3, and 4.3.5, we change the optimizer from Adadelta to SGD and learning rate from 0.1 to 0.00001 with momentum of 0.9, and learning rate of 0.001 for parameters in yes-no classifiers. The rest of the hyperparameters are the same as Table 5.

For transfer learning schemes in section 4.3.4 and 4.3.5, we only change BERT-th into BERT-th that was fine-tuned on XNLI-th. Table 6 shows the detail of the hyperparameters of fine-tuning BERT-th on XNLI-th.

Table 6. XNLI-th BERT fine-tuning hyperparameters.

Hyperparameters	XNLI-th fine-tuning
Batch Size	32
Tokenizer	Pre-trained Thai SentencePiece [2]
Optimizer	AdamW
Learning Rate	$3 * 10^{-5}$
Training epochs	2
Passage length (tokens)	128
Dropout (Classification layer)	0.1

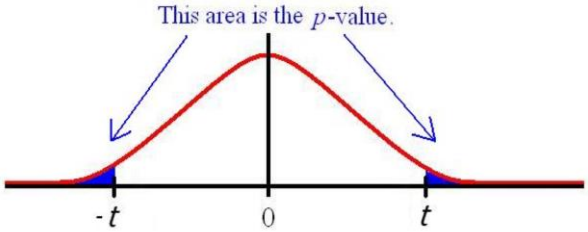
5.3 Statistical Hypothesis Test

We use stratified 3-fold cross-validation to evaluate the model's performance. We use the paired student's t-test for statistics test. We employ the modified version of the student t-test to address the violation of the data sampling assumption. This modified version of the paired student's t-test is proposed by [44]. In the research, they propose the proper method to correct the variance estimate which is shown in Figure 21.

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad \bar{\sigma}^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}$$

$$\bar{\sigma}^2_{Mod} = \left(\frac{1}{n} + \frac{n_1}{n_2}\right)\bar{\sigma}^2$$

$$t = \frac{\bar{d}}{\bar{\sigma}_{Mod}}$$



n_1: Number of sample datapoints used for training
n_2: Number of sample datapoints used for testing

Figure 21. Illustration of variance estimate [44]. Figure is retrieved from [45]

[2] https://github.com/ThAIKeras/bert

The paired student's t-test, similar to other statistical tests, is conducted to compare the difference between two population means. Unlike other statistical tests, paired student's t-test is done when the sets of observations can be paired with another group of observations, for instance, the pre-course examination scores and post-course examination of the same group of students. In this research, we use paired student's t-test to compare the performance of the 2 models on the same test fold.

Table 7. Examples of differences of F1 score calculation across different folds. We use the calculation of paired student's t-test between special token (model 1) and cascade model (model 2) performance as discussed in section 6.2 as an example.

Testing Fold	**Model 1 Overall F1 %**	**Model 2 Overall F1 %**	**F1 score Difference**
Fold 1	63.53	63.74	0.210
Fold 2	62.94	64.13	1.199
Fold 3	62.69	63.37	0.680

Table 8. Step-by-step calculation of each statistical parameter shown in Figure 21.

Statistical Parameter	**Value**
Mean of difference $(\bar{d})$	0.0069
Variance (σ^2)	$1.601 * 10^{-5}$
Modification of Variance (σ^2_{mod})	$8.009 * 10^{-6}$
t-test	2.449
P-value	0.0142

Table 7 and Table 8 demonstrates the steps of paired student's t-test calculation. We first calculate the difference of overall F1 score, which is our main evaluation metrics as discussed in section 5.4. This is shown in Table 7. Table 8 shows the calculation of each statistical parameters that lead to the p-value of the comparison. In comparison of each model pair, the difference in the model's performance has statistical significance when the p-value is below 0.05.

5.4 MRC Evaluation Metrics

In this section, we present the evaluation metrics for the MRC tasks. F1 score and EM are factoid question's metrics while yes-no accuracy is the yes-no question's metric. Since the F1 score is calculated on token levels, the choice of tokenizers affects on the F1 score calculation. In Thai NLP, tokenization aspect is more difficult that English as there are no clear word boundaries. In this study, we select Bailarn tokenizer [32] for both preprocessing and model evaluation. This choice of tokenizer may be different from the ones used in the competition, which use the maximum matching algorithm. We have found that Bailarn tokenizer tokenize passage more correctly, especially when the answers are name-entities, comparing to the maximum matching algorithm which is not a deep learning-based tokenizer. The examples of tokenization are shown in Appendix C.

5.4.1 MRC Evaluation Metrics

F1 score is the harmonic average between precision and recall of the model. In our context, precision and recall are measured based upon numbers of correctly predicted tokens compared to the ground truth token span. For yes-no questions, the F1 score will be either 0 or 1, depending on whether the model predicts the answer correctly. F1 score is an evaluation metric for factoid questions.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{11}$$

TP denotes true positive and represents numbers of predicted tokens that are also appeared in the ground truth. FP denotes false positive, which is an indicator of numbers of predicted tokens that are not in the ground truth. FN represents the false negative, which is the number of actual ground truth tokens not retrieved by the model.

5.4.2 Exact Match (EM)

The exact match score is a binary indicator whether the model correctly retrieves the same span as ground truth span. Like the F1 score, the EM is an evaluation metric for factoid questions.

5.4.3 Yes-no Accuracy and Question Accuracy

Yes-no accuracy is the percentage of correctly predicted yes-no questions, while question accuracy measures the percentage of correctly classified question types (factoid or yes-no).

5.4.4 Overall F1 (%)

This metric is a weighted average of F1 score from factoid questions and yes-no accuracy from yes-no questions. This serves as a general measurement of how well the model performs on overall questions in the dataset.

6. Experiments Results and Discussion

In this chapter, we discuss the results of each experiment in our research. We first explain the baseline establishment which was conducted on NECTEC V1 in section 6.2 After section 6.2, we conduct our experiment on NECTEC V2. Next, we describe the performance of each multiclass architecture along with the effect of integrating contextualized embeddings to the model in sections 6.2, 6.3 and 6.4. We then present the results of using transfer learning to increase yes-no questions accuracy in cascade architecture in sections 6.5 and 6.6**.** Section 6.7 describes the result of attention mechanism modification and section 6.8 concludes the experiment results with the ablation study.

6.1 Baseline Establishment on Factoid Questions in NECTEC V1

First, we replicate the winner's results of NECTEC V1. which consists solely of factoid questions. From the result of the competition on NECTEC V1. This is to ensure that our based model, BIDAF, can achieve similar results to the winner of the analogous dataset. The winner of employs an MRC model called WabiQA [46]. WabiQA is based upon DrQA [47] whose work focuses on both information retrieval and reading comprehension aspects.

As we do not have access to test set used in the first competition (NECTEC V1), we replicated the results on the validation set of which is a 10% split from the available development dataset. This is to align the training method and allow for a fair comparison. The hyperparameter detail of the model used to compare with the result in NECTEC V1 is similar to the ones listed in Table 5. but we use batch size of 6.

Table 9. Result on NECTEC V1. Our implementation exceeds WabiQA both in terms of F1 and EM.

Model	Validation Score	
	EM (%)	**F1 (%)**
WabiQA	45.50	58.25
Our implementation of BIDAF	49.00	63.37

We hypothesize that the reason BIDAF performs better than WabiQA is that BIDAF employs bidirectional attention, both from query-to-context and vice versa while WabiQA utilizes only query-to-context attention. This bidirectional flow of information should help capture the necessary information better. It also can be observed that WabiQA employs attention at a lower level just after the token embedding layer. Passing the token embeddings through 1 layer of

LSTM before applying attention could be more beneficial as the output representation vectors of LSTM contains more contextual information than token embedding vectors.

6.2 Multiclass Architecture Performance with Static Word Embeddings

We now turn our attention to the performance comparison for each proposed multiclass architecture which is located in Table 10.

Table 10. Performance of each multiclass architecture in the static word embedding setting. Column with asterisk indicates the main evaluation metric. All p-values are reported against the special token model. Bold and italic row highlights the model with the best average performance.

Multiclass Architecture	Overall F1 (%)*	Factoid		Yes-no	Question Accuracy (%)	P-value
		EM (%)	F1 (%)	Accuracy (%)		
Special token	63.05	49.35	64.62	51.25	99.15	-
Joint	***63.92***	***50.20***	***65.32***	***53.41***	***99.81***	***<0.001***
Cascade	63.75	49.69	65.16	53.15	99.81	0.014

From Table 10., we can observe that both joint architecture and cascade architecture have better performance than the special token model across all evaluation metrics. This shows that having a dedicated module to handle yes-no questions is better than modify and integrate the yes-no task into the span prediction task. Even though both joint architecture (section 4.2.2.2) and cascade architecture (section 4.2.2.3) performance are greater than the baseline special token (section 4.2.2.1) with statistical significance. When we assess the performance of joint architecture and cascade architecture, joint architecture's overall F1 is not greater than cascade's overall F1 with statistical significance with a p-value of 0.62. According to this statistics test, we cannot conclude that joint architecture performs better than cascade architecture. For the explanations on why joint architecture performs better than cascade architecture on average, we suspect that the model gains benefit from utilizing shared weights between span prediction task and yes-no classification task.

6.3 Multiclass Architecture Performance with Contextual Embeddings

We describe the performance of different multiclass architecture when enhanced with contextual embedding in Table 11.

Table 11. Effect of integrating contextual embeddings to multiclass architecture. A column with an asterisk is the main evaluation criteria and row with bold and italic indicates the best performance. All p-values are reported against the special token model.

Multiclass Architecture	Overall F1 (%)*	Factoid		Yes-no	Question	P-value
		EM (%)	F1 (%)	Accuracy (%)	Accuracy (%)	
Special token	66.27	54.38	64.62	51.06	98.81	-
Joint	67.05	54.40	68.64	55.06	99.81	0.233
Cascade	***67.51***	***54.72***	***69.11***	***55.56***	***99.81***	***0.016***

According to Table 11, joint architecture (section 4.2.2.2) does not statistically perform greater than the baseline special token (section 4.2.2.1) while cascade architecture with contextual embedding performs better than special tokens method with statistical significance. From experiment with static word embeddings in section 6.2, we cannot conclude if joint is better than the cascade model, in this setting, however, we have found that cascade architecture performs better than joint architecture with the p-value of 0.007. With this information, we conclude that cascade architecture performs better than both joint and special token architectures.

For cascade architecture, we hypothesize that having a separate model for each task may perform better since the objective functions are separated. However, to achieve this effect, the model must be provided for sufficient training data for each task. The population of yes-no questions in our study is only 2,000, which is a relatively small number of training data for a deep learning model. This effect is mitigated by utilization contextual embeddings, which provides adequate transfer learning that's why we cascade architecture has the best performance in this setting. This contrasts with results from section 6.2 where, on average, joint architecture with static word embeddings performs better than cascade architecture. We suspect that static word embeddings do not provide transfer learning as adequate as contextual embedding so the model with joint architecture performs better as the model has the benefit of utilizing shared weights between two tasks.

6.4 Effects of Contextual Embedding Integration

Now we evaluate the performance's improvement when applying contextual embedding to the model. This is shown in Table 12.

Table 12. Comparison of contextual embeddings and static word embeddings, Column with an asterisk is the main evaluation criteria and row with bold and italic indicates the best performance. All p-values are reported against their respective static counterpart.

Multiclass Architecture	Word Embedding Setting	Overall F1 (%)*	Factoid		Yes-no	Question Accuracy (%)	P-value
			EM (%)	F1 (%)	Accuracy (%)		
Special token	Static	63.05	49.35	64.62	51.25	99.15	-
	Contextual	***66.27***	***54.38***	***64.62***	***51.06***	***98.81***	***<0.001***
Joint	Static	63.92	50.20	65.32	53.41	99.81	-
	Contextual	***67.05***	***54.40***	***68.64***	***55.06***	***99.81***	***<0.001***
Cascade	Static	63.75	49.69	65.16	53.15	99.81	-
	Contextual	***67.51***	***54.72***	***69.11***	***55.56***	***99.81***	***<0.001***

We can observe from Table 12 that all architectures gain a substantial increase in almost all performance metrics with statistical significance. However, we also note that yes-no question accuracy drops in the special token model when we apply contextual embeddings. A possible explanation is that adding artificial token, such as 'YES' and 'NO' into the original context passage, does not make contextual sense and introduce unoriginal words into the passages, so the contextual embedding does not work as well as expected. This finding is also similar to the results from [14] and [37], in which the model also observe a gain in performance when enhanced with contextual embeddings.

6.5 Results of Transfer Learning from XNLI-th to *Yes-no Questions*

We now discuss or findings on the effect of using transfer learning from the NLI dataset such as XNLI. We apply transfer learning to cascade architecture only. The results of the experiment are described in Table 13. We present only performance in yes-no questions only as the factoid question component remains similar to components discussed in previous experiments (section 6.2 and section 6.3). In this experiment we use fine-tune BERT-th on the XNLI-th dataset then use it as contextual embeddings extractor. The accuracy of fine-tuning on the XNLI-th dataset is 68%. More details of the result of pre-trained models can be found in Appendix D.

Table 13. Yes-no accuracy improvement from transfer learning. P-value for the static word embedding setting is compared against training from scratch while contextual models are compared against normal BERT.

Word Embedding Setting	Transfer Learning scheme	Yes-no Accuracy (%)	P-value
Static	Train from scratch (4.2.2.3)	53.15	-
	Transfer from factoid (4.3.1)	53.84	0.597
	Transfer from XNLI-th (4.3.2)	***54.50***	***0.08***
Contextual	BERT-th (4.2.1)	55.56	**-**
	BERT-th + transfer from factoid (4.3.3)	54.31	0.107
	BERT-th + XNLI-th (4.3.4)	56.90	0.111
	BERT-th + BIDAF + XNLI-th (4.3.5)	***60.18***	***0.063***

It can be seen that applying transfer learning from XNLI-th both word in the static word embedding setting, in which both BIDAF and word vector architectures are pre-trained on the XNLI-th dataset and in the fine-tuning setting where we BERT is fine-tuned on XNLI specific task and later used as a feature extractor. Another observation from Table 13 is that the performance of transfer learning from factoid questions are not as competitive as XNLI-th. Transferring from factoid question in contextual embedding setting (section 4.3.3) even has worse performance than from training from scratch. This could be due to the fact that factoid and yes-no questions have different characteristics, so the transfer learning is not that effective. The setting where we both fine-tune BERT on XNLI-th and BIDAF before transferring both components of the model (section 4.3.5) has the best accuracy.

6.6 Results of Transfer Learning from XNLI-th to *Factoid Questions*

Next, we would like to confirm if the BIDAF for span prediction will also receive performance improvement if transfer learning from XNLI-th is applied. We present the results of the experiment of transferring from XNLI-th to factoid questions in Table 14.

Table 14. Results of applying transfer learning to both factoid questions and yes-no questions.

Transfer Learning scheme	Target Questions	Overall F1(%)	Factoid		Yes-no	P-value
			EM (%)	F1 (%)	Accuracy (%)	
BERT-th + XNLI-th	Yes-no	**67.67**	**54.73**	**69.10**	**56.90**	**-**
BERT-th + XNLI-th	both	67.35	53.93	68.75	56.90	0.3246

We can observe from Table 14 that applying XNLI-th transfer learning to both factoid questions worsens the model's EM and F1 performance. A possible explanation for this finding is that when we fine-tune BERT to specific tasks like XNLI-th, the language model loses the ability to generalize and may perform worse on other tasks that the model is not fine-tuned on. This is also in line with previous observation in [26], in which the author has pointed out that using transfer learning from an extractive dataset (like factoid questions) is not as effective using transfer learning from the NLI dataset to Boolean questions. We suspect that the reverse, applying NLI to extractive questions, also holds true.

6.7 Effects of Modifying Attention Mechanism

Now we move to the last experiment in our study, attention mechanism modification. The results of such modification is shown below in Table 15.

Table 15. Comparison between having 2 attention mechanisms and dropping C2Q. P-Value of Drop C2Q is report against the model with both attention mechanisms for both word embedding settings.

Word Embedding Setting	Attention Mechanisms	Yes-no Accuracy (%)	P-value
Static	Both C2Q and Q2C	53.15	-
	Drop C2Q	***58.31***	***<0.001***
Contextual	Both C2Q and Q2C	55.56	-
	Drop C2Q	***59.21***	***0.003***

Table 15. provides the evidence that dropping the C2Q mechanism does indeed leads to an increase in performance on yes-no questions, supporting our claim that C2Q does not play a critical role in yes-no questions and keeping only Q2C attention mechanism results in superior performance.

Table 16. Effects of dropping C2Q in factoid questions. F1 is the main metric for this table.

Word Embedding Setting	Attention Mechanisms	Factoid		P-value
		F1 (%) *	EM (%)	
Static	***Both C2Q and Q2C***	***49.69***	***65.03***	-
	Drop C2Q	25.06	35.21	<0.001
Contextual	***Both C2Q and Q2C***	***54.72***	***69.11***	-
	Drop C2Q	27.03	37.33	<0.001

Table 16 proves that context-to-query (or C2Q) serves has its purpose in the span prediction task and dropping it causes a serious drop in factoid's performance metric. This supports the claim that, for yes-no questions, only the Q2C mechanism is crucial while both attention mechanisms play an important role in factoid questions.

6.8 Ablation Study

In this section, we now study the contribution of different techniques discussed in this research. We have summarized the effects of each module to the model's overall performance in Table 17.

Table 17. Contribution of each proposed techniques. Row (2) and (3) are compared against the preceding rows. Row (4), (5), and (6) are compared against row (3).

Model Setting	Overall F1(%)*	Factoid		Yes-no	Question Accuracy (%)	P-value
		EM (%)	F1 (%)	Accuracy (%)		
(1) Special token	63.05	49.35	64.62	51.25	99.15	-
(2) Cascade	63.75	49.69	65.16	53.15	99.81	0.014
(3) Cascade + BERT-th	67.51	54.72	69.11	55.56	99.81	<0.001
(4) Cascade + BERT-th + XNLI-th	***68.06***	***54.72***	***69.11***	***60.18***	***99.81***	***0.058***
(5) Cascade + BERT-th + Drop C2Q	***67.94***	***54.72***	***69.11***	***59.21***	***99.81***	***0.002***
(6) Cascade + BERT-th + Drop C2Q + XNLI-th	68.00	54.72	69.11	59.67	99.81	0.004

We can observe that adding contextualize embeddings yields the largest boost in the model's overall performance. In terms of yes-no accuracy, using transfer learning from the XNLI-th gives the best result. Unfortunately, combining both transfer learning and dropping the C2Q attention does not yield us the best result. We suspect we do not have enough data to appropriately re-train the part of the model that has mismatched dimensions as a result of attention mechanism modifications. The performance in row (6) uses the transfer learning setting from Figure 13 (e). as we have found that such setting has better performance than setting in Figure 13 (f). We would like to point out that, for experiments from row (4) to row (6), we focus on introducing techniques that improve the model's performance in terms of yes-no accuracy. We do not realize a significant improvement in overall F1 % as the yes-no questions make up only 11.7% of the whole dataset.

7. Qualitative Analysis of the Proposed Methods

In this section of our thesis, we discuss the results of our proposed methods in qualitative aspects. We start with the comparison of factoid question predictions from a model with static word embedding and contextual embedding in factoid questions (section 7.1) and yes-no questions (section 7.2). We conclude section 7 by with visualization of query-to-context attention heatmap in section 7.3).

7.1 Static Word and Contextual Embeddings Predictions in Factoid Questions

We now aim to discuss some examples question-answer pair that models with contextual embedding correctly predict while the models with static word embeddings do not. We define correctly predicted factoid questions if the predicted EM is 1 and 0 otherwise. Table 18 discusses the examples where some or both versions of the models predict the questions correctly while Table 19 focuses on the questions that both models predict incorrectly.

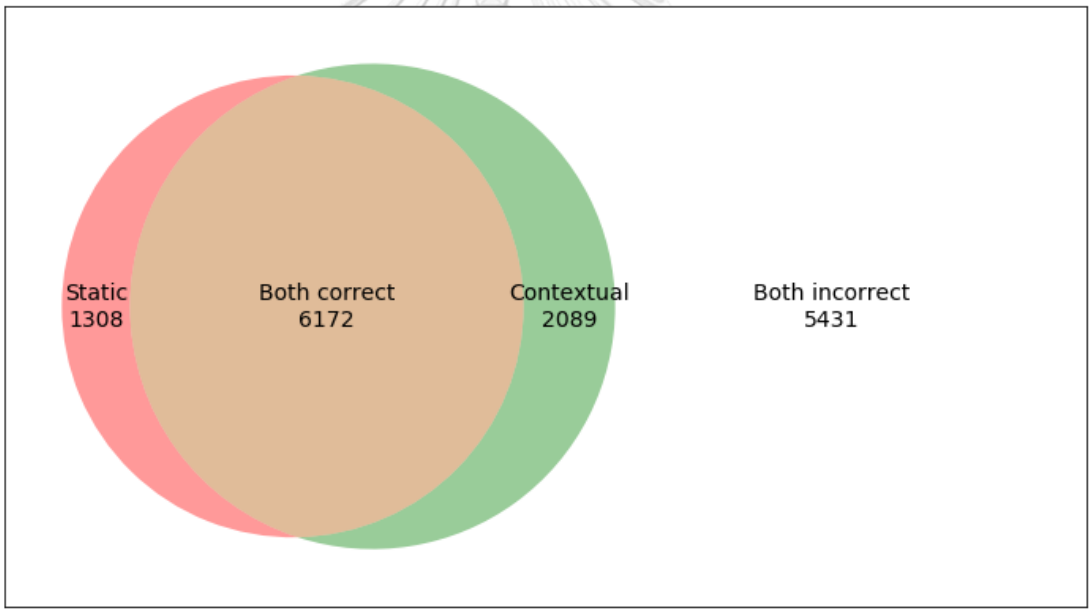


Figure 22. Vein diagram of factoid question predictions from static and contextual embeddings models, we define correct as having EM=1.0 and incorrect otherwise.

Table 18. Comparison of predictions from models with static word and contextual embeddings, we have highlighted a segment of context passage that relates to factoid questions. Green highlights indicate the region of passage that is related to contextual embedding model prediction while the yellow highlights belong to the static word embeddings

Question-answer pair	Embeddings	Prediction
1. Context: ชะพลู ชะพลู หรือ ช้าพลู (ชื่อวิทยาศาสตร์: PipersarmentosumRoxb.) เป็นพืชในวงศ์ Piperaceae มักสับสนกับพลู แต่ใบรสไม่จัดเท่าพลูและมีขนาดเล็กกว่า ชะพลูเป็นพันธุ์ไม้ที่ชอบพื้นที่ลุ่ม มีความชื้น ขยายพันธุ์ด้วยวิธีการปักชำ โดยการเลือกกิ่งที่มีใบอ่อนและใบแก่เด็ดใบแก่ออกและนำไปปักชำได้ ชะพลูมีชื่อพื้นเมืองอื่น ๆ อีกคือทางภาคเหนือเรียกว่า "ผักปูนา" "ผักพลูนก" "พลูลิง" "ปูลิง" "ปูลิงนก" ทางภาคกลางเรียกว่า "ช้าพลู" ทางภาคอีสานเรียกว่า "ผักแค" "ผักปูลิง" "ผักนางเลิด" "ผักอีเลิด" และทางภาคใต้เรียกว่า "นมวา" **Question:** ทางภาคใต้ของไทยเรียกต้นชะพลูว่าอะไร **Reasoning:** Keywords are located quite far apart	Static word	ช้าพลู
	Contextual	นมวา
	Ground Truth	นมวา
2. Context: มิตซูบิชิเอฟ-1 มิตซูบิชิเอฟ-1 (MitsubishiF-1) มิตซูบิชิเอฟ-1 เป็นเครื่องบินรบความเร็วใต้เสียงแบบแรกที่สร้างโดยบริษัท มิตซูบิชิ เฮวี่ อินดัสทรีย์ของประเทศญี่ปุ่น รายละเอียด มิตซูบิชิเอฟ-1 -ผู้สร้าง :บริษัทมิตซูบิชิ เฮวี อินดัสตรี (ประเทศญี่ปุ่น) -ประเภท:เครื่องบินขับไล่สนับสนุนกองกำลังภาคพื้นดินที่นั่งเดียว -เครื่องยนต์:2×-กางปีก: 7.88 เมตร -ยาว: 17.86 เมตร -สูง: 4.39 เมตร -พื้นที่ปีก: 21.18ตารางเมตร -น้ำหนักเปล่า: 6,288 กิโลกรัม -น้ำหนักวิ่งขึ้นสูงสุด: 13,614 กิโลกรัม -อัตราเร็วสูงสุด 1,700 กิโลเมตร/ชั่วโมงที่ระดับความสูง 11,000 เมตร -อัตราไต่/วินาที -รัศมีทำการรบ 556 กิโลเมตรเมื่อบิน สูง-ต่ำ-สูง พร้อมติดขีปนาวุธ ASM-1 จำนวนสองลูกและถังน้ำมันเสริมขนาด 830 ลิตรจำนวนหนึ่งถัง -พิสัยบินไกลสุด: 2,870 กิโลเมตร -อาวุธ: ปืนกลอากาศลำกล้องหมุนเจเอ็ม-61 ขนาด 20 มม. 1 กระบอก -อาวุธปล่อยอากาศสู่อากาศ 4 นัด หรือ อาวุธปล่อยปราบเรือรบ 2 นัด -ลูกระเบิดขนาด 500 ปอนด์8-12 ลูก -สามารถติดตั้งอาวุธได้ 3,629 กิโลกรัม **Question:** เครื่องบินมิตซูบิชิเอฟ-1 มีอัตราเร็วสูงสุดได้กี่กิโลเมตร/ชั่วโมง **Reasoning:** Both pick number as answers but contextual model picks number with correct meaning.	Static word	11,000
	Contextual	1,700
	Ground Truth	1,700
3. Context: ภาษาอัสกุนุเป็นภาษาในอัฟกานิสถานพูดโดยชาวอัสกุนุ ซานุ และกรัมซานา ในหุบเขาเปช รอบ ๆ วามาทางตะวันตกเฉียงเหนือของอซา	Static word	อัฟกานิสถาน

ดาบัดในจังหวัดกุนาร์ ชื่อรวมของทั้งสามเผ่านี้คืออัสกุน ใช้เป็นครั้งแรกโดย George Scott Robertson เมื่อ พ.ศ. 2439 จัดอยู่ในภาษากลุ่มนูริสถาน **Question:** ภาษาอัสกุนุจัดอยู่ในภาษากลุ่มใด **Reasoning:** Keywords are located quite far apart	Contextual	นูริสถาน
	Ground Truth	นูริสถาน
4. Context: ไฮโดรเจนคลอไรด์ (อังกฤษ: Hydrogen chloride) สูตรโมเลกุลว่า HCl เป็นก๊าซมีพิษ ไม่มีสี มีฤทธิ์กัดกร่อน เมื่อสัมผัสความชื้นจะเกิดควันสีขาว ควันนี้จะประกอบด้วย กรดไฮโดรคลอริกซึ่งจะเกิดขึ้นเมื่อไฮโดรเจนคลอไรด์ละลายในน้ำ ก๊าซไฮโดรเจนคลอไรด์และกรดไฮโดรคลอริกเป็นสารเคมีที่มีความสำคัญ ในทาง เคมี วิทยาศาสตร์ เทคโนโลยี และ อุตสาหกรรมมาก **Question:** ไฮโดรเจนคลอไรด์ มีสูตรโมเลกุลว่า HCl เป็นก๊าซมีพิษไม่มีสีมีฤทธิ์กัดกร่อนเมื่อสัมผัสความชื้นจะเกิดอะไรขึ้น **Reasoning:** Contextual model does not answer the whole span	Static word	ควันสีขาว
	Contextual	ควัน
	Ground Truth	ควันสีขาว

We can analyze from examples in Table 18. that the models with contextual embeddings can understand the meaning or the context behind the answer candidates better than the static word embeddings. For question 1 in Table 18, the contextual-enhanced model also can answer the factoid question in which the keywords are located far apart while the model with static word embeddings fails to predict this question correctly. The reasoning behind question 3 in Table 18 is similar to question 1. Another observation from question 2 is that we may see that both versions of the model are able to predict the tokens with correct types of words/ pos tags, both models realize that the number should be the answers to the given question, but the model with contextual embedding is able to correctly select the 11,000 km/hr as an answer since the candidate matches contextual sense with the provided question. Question 4 shows the examples where static word embedding predicts correctly while the contextual embedding model does not. In question 4 example, the full correct answer is “ควันสีขาว” (white smoke), we suspect that the contextual enhanced model thinks that the token “ควัน” (smoke) might appropriate enough as an answer.

Table 19. Examples of factoid questions, which both models fail to predict correctly. We have highlighted the regions of text that are keywords to the questions.

Question-answer pair	Embeddings	Prediction
1. Context: วัดดาวดึงษารามเป็นพระอารามหลวงชั้นตรีชนิดสามัญตั้งอยู่เลขที่ 872 แขวงบางยี่ขัน เขตบางพลัด กรุงเทพมหานคร สร้างขึ้นมาในสมัยพระบาทสมเด็จพระพุทธยอดฟ้าจุฬาโลกมหาราชโดยเจ้าจอมแว่นพระสนมเอกในรัชกาลที่ ๑ สร้างขึ้นทำด้วยเสาไม้แก่นพระอุโบสถก่ออิฐสูงพ้นพื้นดินประมาณ 2 ศอกชาวบ้านเรียกว่า “วัดขรัวอิน” ต่อมาในสมัยรัชกาลที่2ข้าราชการฝ่ายในชื่ออินซึ่งเป็นญาติของเจ้าจอมแว่นได้ปฏิสังขรณ์วัดนี้เหตุด้วยผู้ครองวัดและผู้ปฏิสังขรณ์วัดมีนามเดียวกันว่า “อิน” พระบาทสมเด็จพระพุทธเลิศหล้านภาลัยจึงพระราชทานนามวัดนี้ว่า “วัดดาวดึงษาสวรรค์” **Question:** วัดดาวดึงษารามเขตบางพลัดกรุงเทพมหานครสร้างขึ้นในรัชสมัยใด **Reasoning:** The model may lack sufficient world knowledge	Static word	พระบาทสมเด็จพระพุทธยอดฟ้าจุฬาโลกมหาราช
	Contextual	พระบาทสมเด็จพระพุทธยอดฟ้าจุฬาโลกมหาราช
	Ground Truth	รัชกาลที่ 1
2. Context: ราเดลฟูเอโก หรือชื่อทางการคือ รัฐติเอร์ราเดลฟูเอโก แอนตาร์กติกา และหมู่เกาะในมหาสมุทรแอตแลนติกใต้ เป็นรัฐในประเทศอาร์เจนตินา ที่แยกออกจากแผ่นดินใหญ่ของอาร์เจนตินา ข้ามช่องแคบมาเจลลัน เมืองหลวงชื่อ อูชัวยา เดิมทีรัฐนี้มีชนพื้นเมืองอาศัยมาก่อนมากกว่า 12,000 ปีก่อน ถูกค้นพบโดยชาวยุโรปในปี ค.ศ. 1520 โดย เฟอร์ดินานด์ มาเจลลัน อย่างไรก็ตามชนพื้นเมืองยังคงปกครองดินแดนดังกล่าวจนถูกพิชิตในคริสต์ทศวรรษ 1870 **Question:** บุคคลใดเป็นผู้ค้นพบรัฐติเอร์ราเดลฟูเอโก **Reasoning:** Question contains some ambiguity, the word that models pick as an answer is partially correct	Static word	ติเอร์ราเดลฟูเอโก
	Contextual	ชาวยุโรป
	Ground Truth	เฟอร์ดินานด์ มาเจลลัน
3. Context: พราหมณนั้นเป็นวรรณะหนึ่งในสี่วรรณะของสังคมอินเดีย เป็นผู้สืบทอดวิชาความรู้ ในคัมภีร์ ไตรเวทพิธีกรรม จารีต ประเพณี ศิลปะวัฒนธรรม และคติความเชื่อต่าง ๆ ให้สืบทอดต่อไป ..... แล้วพราหมณ์ผู้ใหญ่จะมอบสายสิญจน์รับพราหมณ์ใหม่ หรือทวิชาติ ซึ่งหมายถึงการเกิดครั้งที่ 2 ซึ่งการบวชพราหมณ์ไม่ได้มีกฎปฏิบัติจำนวนมากเหมือนกับการบวชพระ	Static word	2
	Contextual	2

Question: สังคมอินเดียแบ่งวรรณะออกเป็นกี่วรรณะ **Reasoning:** The context states information that is vital to the question very subtly.	Ground Truth	สี่

Table 19 shows some examples of the questions that both models with static and contextual embeddings fail to predict correctly. Example 1 is the case where both models pick the correct entity, which is the full name of the king (“พระบาทสมเด็จพระพุทธยอดฟ้าจุฬาโลกมหาราช”) but the correct answer is another entity that can generally use to refer to the king (“รัชกาลที่ 1”). We suspect that the models lack the general world knowledge, so the models were unable to pick the latter entity as an answer. For example 2, the answer that contextualized model picks can be considered as partially correct. The question is “who discovered Tierra del Fuego”, in which the contextualized model picks “European” as an answer while, in reality, the actual answer is “Ferdinand Magellan”. The model may think that the word European is highly correlated or refer to the word “Ferdinand Magellan” thus picking the word “European” as an answer. For the final example, the information required to answer the question is not explicitly stated in the context passage. Since the question asks for some number and the actual answers are not explicitly stated, the models retrieve the next available number in the context passage as an answer instead.

From the examples shown in Table 19, we can see that the questions, which both models predict incorrectly are more difficult than the examples. Some questions require the models to have world knowledge (example 1) while other questions may demand the model to have a more complex reasoning skill (example 3).

7.2 Static Word and Contextual Embeddings Predictions in *Yes-no Questions*

Similar to the analysis in section 7.1 we now assess the predicted answers in yes-no questions. Table 20 shows questions where some models predict correctly and Table 21 shows examples that both types of models fail to predict.

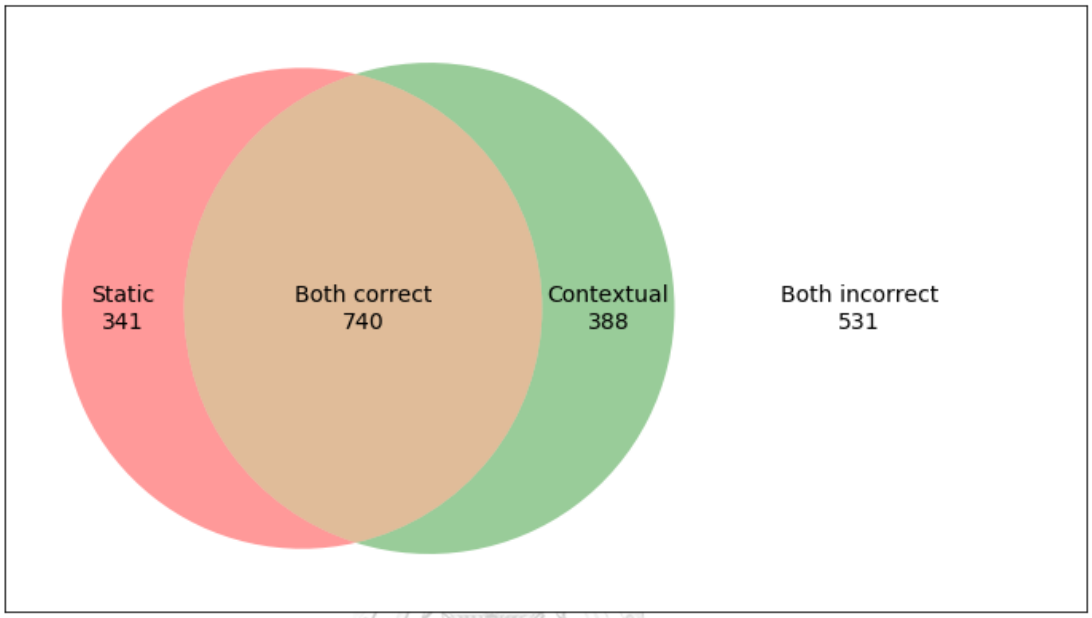


Figure 23. Vein diagram of yes-no predictions from models with static and contextual embeddings

Table 20. Yes-no predictions from static word and contextual embedding models, both of which are shown with the ground truth of the questions. Yellow highlights indicate the area, which is critical to the answers of yes-no questions

Question-answer pair	Embeddings	Prediction
1. Context: เอ็มวีอาร์ดี วีเอ็มวีอาร์ดีวี (MVRDV) เป็นสำนักงานออกแบบสถาปัตยกรรมและผังเมือง ตั้งอยู่ในนครรอตเทอร์ดัม ประเทศเนเธอร์แลนด์ ก่อตั้งมาตั้งแต่ ปีค.ศ.1991 ชื่อของสำนักงานเป็นตัวอักษรย่อของคณะผู้ก่อตั้ง ได้แก่ วินี มาส (M - เกิดเมื่อปีค.ศ.1959) จาคอบ แวน ริจ (VR - เกิดเมื่อปีค.ศ. 1964) และนาตาลี เดอ วรี (DV - เกิดเมื่อปีค.ศ. 1965) มาส กับ แวนริจเคยทำงานที่สำนักงานสถาปัตยกรรมเมโทรโปลิตัน และสำนักงานสถาปนิกของเร็ม คูลฮาส ส่วนเดอ วรี เคยทำงานที่เมคานูก่อนจะมาร่วมกันก่อตั้งเอ็มวีอาร์ดีวี งานชิ้นแรกที่ได้สร้างได้แก่สำนักงานใหม่ของ วีโปร ในเมืองฮิลเวอร์ซัม ประเทศเนเธอร์แลนด์ (ค.ศ. 1993 - ค.ศ. 1997) ผลงานออกแบบอื่นๆที่ได้รับการก่อสร้างได้แก่ อาคารพักอาศัยพักอาศัยนครอัมสเตอร์ดัม (ค.ศ.1994 - ค.ศ.1997) และศาลาดัชท์ที่เอ็กซ์โป 2000 ที่เมืองฮานโนเวอร์ ประเทศเยอรมนี (ค.ศ.1997 - ค.ศ.2000) **Question:** เอ็มวีอาร์ดีวี เป็นสำนักงานออกแบบสถาปัตยกรรมและผังเมืองในนครรอตเทอร์ดัม ประเทศเนเธอร์แลนด์ก่อตั้งเมื่อ ค.ศ.1991 ใช่หรือไม่ **Reasoning:** Yes-no question and context passage do not have exact word to word.	Static word	ไม่ใช่
	Contextual	ใช่
	Ground Truth	ใช่
2. Context: รกลอกตัวก่อนกำหนดเป็นภาวะแทรกซ้อนทางสูติศาสตร์ของการตั้งครรภ์ซึ่งรกได้แยกตัวออกจากผนังมดลูกของมารดาก่อนที่ทารกจะคลอดตามปกติ เป็นสาเหตุที่พบบ่อยสาเหตุหนึ่งของการมีเลือดออกในช่วงท้ายของการตั้งครรภ์ ในมนุษย์ถือว่าการลอกตัวของรกหลังสัปดาห์ที่ 20 ของการตั้งครรภ์และก่อนการเกิดนั้นเป็นการลอกตัวก่อนกำหนด มีอุบัติการณ์ 1% การตั้งครรภ์ทั่วโลกโดยมีอัตราการเสียชีวิตของทารกประมาณ 20-40% ขึ้นอยู่กับความรุนแรงของการลอกตัว **Question:** รกลอกตัวก่อนกำหนดเป็นภาวะที่รกได้แยกตัวออกจากผนังมดลูกของมารดาก่อนที่ทารกจะคลอดตามปกติ ใช่หรือไม่ **Reasoning:** Yes-no question and context do not have exact word to word and keywords are located far apart.	Static word	ไม่ใช่
	Contextual	ใช่
	Ground Truth	ใช่
3. Context: เปปไทด์ เปปไทด์ (มาจากภาษากรีก πεπτίδια) คือสายพอลิเมอร์ของกรดอะมิโนที่มาเชื่อมต่อกันด้วยพันธะเปปไทด์ ปลายด้านที่มีหมู่อะ	Static word	ใช่

มิโนเป็นอิสระเรียกว่าปลายเอ็น (N-terminal) ส่วนปลายที่มีหมู่คาร์บอกซิลเป็นอิสระเรียกว่าปลายซี (C-terminal) การเรียกชื่อเปปไทด์จะเรียกตามลำดับกรดอะมิโนจากปลายเอ็นไปหาปลายซี เปปไทด์ขนาดเล็กหลายชนิดมีความสำคัญในสิ่งมีชีวิต **Question:** เปปไทด์เป็นสายพอลิเมอร์ของกรดอะมิโนที่มาเชื่อมต่อกันด้วยพันธะเปปไทด์ปลายด้านที่มีหมู่อะมิโนเป็นอิสระเรียกว่าปลาย M ใช่หรือไม่ **Reasoning:** Yes-no question has misleading/adversarial word.	Contextual	ไม่ใช่
	Ground Truth	ไม่ใช่
4. Context: มาคะฟุชิกิ แอดเวนเจอร์! เป็นเพลงประกอบการ์ตูนแอนิเมชัน *ดราก้อนบอล* ซึ่งเป็นเพลงแนว เจ-ป็อป ที่ขับร้องโดยฮิโรกิ ทาคาฮาชิ และได้วางจำหน่ายในรูปของแผ่นเสียงในช่วงมีนาคม ค.ศ.1986 และอีกครั้งในรูปของซีดี 8 ซม. เมื่อวันที่ 8 มีนาคม ค.ศ.1998 ที่ประเทศญี่ปุ่น และได้นำเพลง "โรแมนติกอาเงรุ โย" ที่ขับร้องโดย อุชิโอะ ฮาชิโมโตะ มาร่วมเข้าไว้ด้วยกัน **Question:** มาคะฟุชิกิแอดเวนเจอร์เป็นเพลงแนวเคป็อปสำหรับประกอบการ์ตูนแอนิเมชันดราก้อนบอลใช่หรือไม่ **Reasoning:** Yes-no question has misleading/adversarial word.	Static word	ไม่ใช่
	Contextual	ใช่
	Ground Truth	ไม่ใช่

Similar to examples shown in Table 18 of section 7.1, yes-no models, which are enhanced with contextual embedding, can answer yes-no questions where the keywords in context passage are distant which are represented by questions 1 and 2 in Table 20. Contextual embedding also gives the model ability to handle misleading or attacking words as shown in question 3 of Table 20. Question 4 from Table 20. Shows the case where the static word model predicts correctly, and the contextual model does not. The contextual vectors for the token “เคป็อบ” (K-pop) in the question could be similar to the token “เจป็อบ” (J-pop) in the context passage so the contextual model gives ‘yes’ as an answer while static word embedding vector for K-pop and J-pop could be more different and gives ‘no’ as an answer.

Table 21. Examples of yes-no questions that both types of models predict incorrectly. We have highlighted the regions of text that are keywords to the questions.

Question-answer pair	Embeddings	Prediction
1. Context: บทนำทฤษฎีสัมพัทธภาพทั่วไปทฤษฎีสัมพัทธภาพทั่วไปเป็นทฤษฎีความโน้มถ่วงซึ่งอัลเบิร์ตไอน์สไตน์พัฒนาระหว่างค.ศ. 1907 ถึง 1915 ตามทฤษฎีสัมพัทธภาพทั่วไปผลของความโน้มถ่วงที่สังเกตได้ระหว่างมวลเกิดจากการบิดงอ (warp) ของปริภูมิ-เวลาต้นคริสต์ศตวรรษที่ 20 **Question:** ทฤษฎีสัมพัทธภาพทั่วไปเป็นทฤษฎีความโน้มถ่วงซึ่งอัลเบิร์ตไอน์สไตน์พัฒนาระหว่างค.ศ. 190 ถึง 191 ใช่หรือไม่ **Reasoning:** Number were changed.	Static word	ใช่
	Contextual	ใช่
	Ground Truth	ไม่ใช่
2. Context: หมอลำ เป็นรูปแบบของเพลงลาวโบราณในประเทศลาวและภาคอีสานของประเทศไทย สามารถแบ่งออกได้เป็นหลายอย่าง ตามลักษณะทำนองของการลำ เช่น ลำเต้ย ลำพื้น ลำกลอน ลำเรื่อง ลำเรื่องต่อกลอน ลำเพลิน ลำซิ่ง รวมทั้ง ลำตัดในภาคกลางก็จัดได้ว่าเป็นหมอลำประเภทหนึ่ง **Question:** หมอลำเป็นรูปแบบของเพลงลาวโบราณเฉพาะในภาคอีสานของประเทศไทยเท่านั้นใช่หรือไม่ **Reasoning:** Requires complex reasoning.	Static word	ใช่
	Contextual	ใช่
	Ground Truth	ไม่ใช่

From Table 21, question in example 1 changes the number of years that normally appear in the context passage, from 1907 to 190 and 1915 to 191 respectively. In this case, we expect that the models are still not robust against attack on numbers so the models fail to predict this type of question correctly. For example 2, this question requires complex reasoning. The question asks if certain dance style A can only be found in the northeastern region of Thailand and Laos or not. The context passage has a span of text that describes that this style of dance A can be normally be found in the mentioned region. But 2 – 3 sentences later, the context passage mentions that another kind of dance style B can also be classified as style A and is practiced in the central region of Thailand.

7.3 Query-to-Context Attention Heatmap Visualization in *Yes-no Questions*

We illustrate the heatmap of context-to-query and query-to context attention in Figure 24 and Figure 25 respectively.

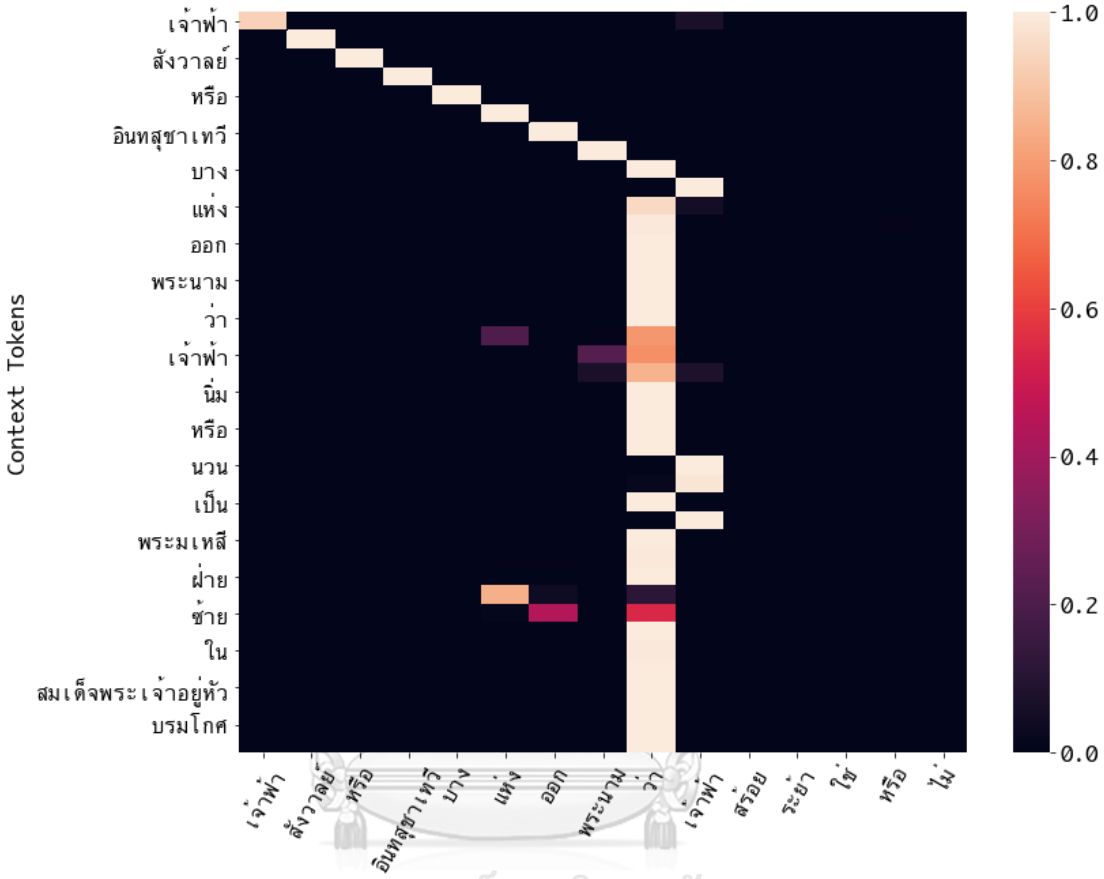


Figure 24. Heatmap from context-to-query attention mechanism in one of the yes-no questions. The lighter shade in the heatmap represents tokens with a higher similarity score. For each row, the summation of the similarity scores equals to one.

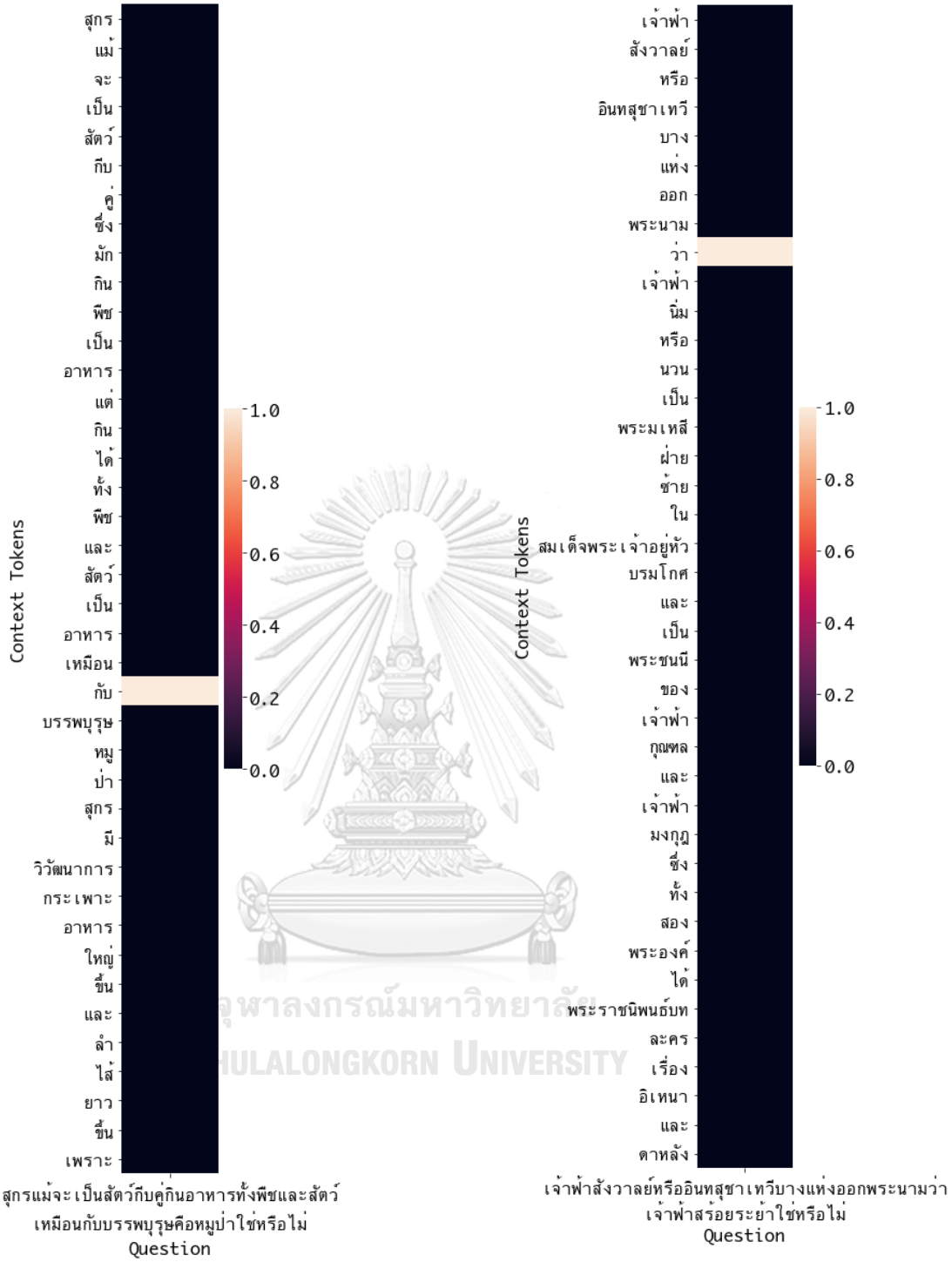


Figure 25. Query-to-context heatmap where the question vector gauges the importance of different context tokens, the lighter shades in the heatmap represent tokens with a higher similarity score. (a): left-side is a yes-no question involving pig's diet. (b): right-side is a yes-no question about Thai poet. Both examples are correctly yes-no questions that are correctly predicted.

From Figure 24, we can see how context-to-query attention work. The mechanism is designed to guide the model to focus gauge the importance of query tokens judging from context tokens. We may observe that some query tokens are deemed important by many context words, e.g. 'ว่า'. Comparing to Figure 25, the query-to-context (Q2C) attention mechanism measures the importance of context tokens from all query tokens as a whole. During the analysis, we have found that most query-to-context heatmap has similar characteristics with Figure 25, in which query tokens assign a high value of similarity score to only a few context tokens. Figure 25 also represents attention heatmap from correctly predicted yes-no questions, we suspect that query-to-context attention allows the model to focus on the context tokens that are located in the area that is most relevant to the meaning of query vectors, making it easier for the model to fact-checking the context passage to answer yes-no questions. From an example shown in Figure 25 (a), attention mechanism guides the model to focus on the context word "กับ" (to). This word acts as preposition between the phrase "กินได้ทั้งพืชและสัตว์เป็นอาหารเหมือน" (is omnivore similar) and the phrase "บรรพบุรุษหมูป่า" (its boar ancestors). These 2 phrases are essential to answer the question "สุกรแม้จะเป็นสัตว์กีบคู่กินอาหารทั้งพืชและสัตว์เหมือนกับบรรพบุรุษคือหมูป่าใช่หรือไม่" (Pigs, even though are hoof animals, are omnivore like its boar ancestors?). The attention heatmap in Figure 25 (b) also works in a similar manner, where the token "ว่า" (is called as) serves as preposition between 2 phrases in the context passage with critical information.

In contrast to Figure 25, Figure 26 highlights the heatmap visualization of query-to-context attention in incorrect predicted yes-no questions. In Figure 25, the attention mechanism guides the model to correct region of the context passage, which potentially leads to correct prediction while in Figure 26, the attention mechanism guides the reader to the incorrect regions of the passages that do not necessarily relate to the questions, which ultimately lead to incorrect predictions of yes-no questions.

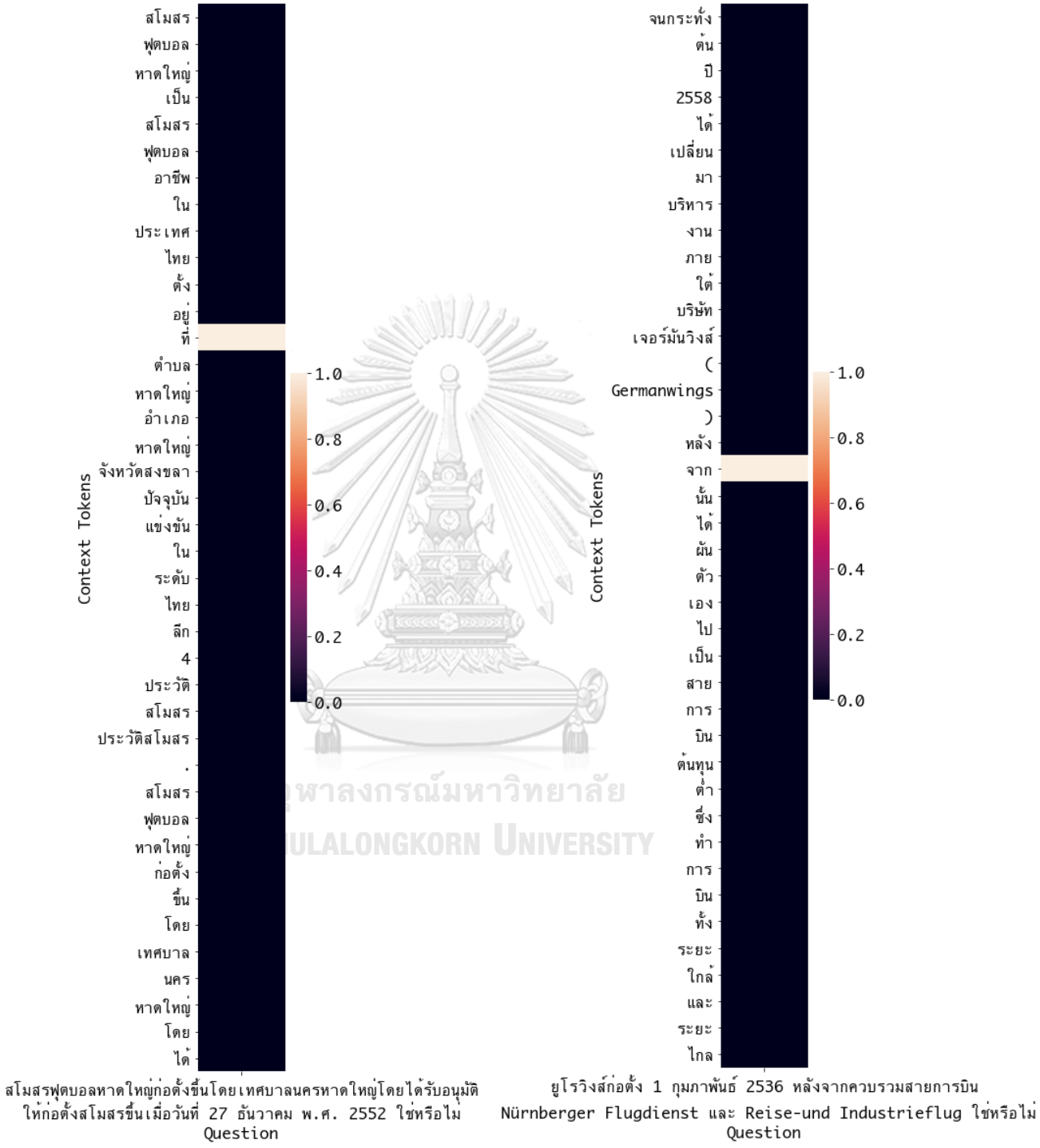


Figure 26. Examples of incorrectly predicted yes-no questions. (a) left-side: a question about football club, (b) right-side: a question about airline company. The lighter the shade of heatmap, the higher the similarity score.

8. Conclusion

In this research, we have built a multiclass machine reading comprehension model for Thai corpus. Our model is built based on BIDAF, which deals with factoid questions only. We enhance BIDAF to support 2 types of questions found in our dataset, which includes factoid and yes-no questions. We compare three types of multiclass architectures, which are special tokens, joint, and cascade model architectures. Both joint and cascade model performs better than the special token model, which serves as a baseline for multiclass MRC in our study.

We conduct our study on the Thai question answering dataset provided by NECTEC in Thailand 2020 National Software Competition (NSC). This dataset consists of 17,000 question-answer pairs and has two types of questions which are factoid and yes-no. Experiments from multiclass architecture with static word and contextual embedding suggest that cascading type has the best performance in terms of overall F1. We then further enhance the cascading architecture by applying transfer learning and attention mechanism modification. We intend to use these 2 techniques to enhance the model's performance on yes-no question specifically and we achieve that objective even though the performance of the model on overall F1 does not increase significantly. Transfer learning from the NLI dataset boosts the model's accuracy on yes-no questions. Pre-training both the MRC reader (BIDAF) and the LM (BERT) is proved to yield the best results. We also have demonstrated that using transfer learning from NLI to factoid questions does not statistically increase the performance. Dropping context-to-query attention mechanisms can also help increase the performance on yes-no questions but greatly hurt the model's ability to answer factoid questions.

For future research direction, there are many areas for Thai MRC to research further. Larger scale and more diverse dataset consisting of different types of questions in one possible area. The number of monolingual Thai MRC datasets is still somewhat limited. Another line of research which focuses on using the benefit of rich resource languages like English by transferring it to other lower resource language, a concept of cross-lingual NLP. Singh, et al. [48] and Conneau, et al. [39] researches involve this aspect of natural language processing tasks. Another future direction that the authors are interested in is the usage of MRC reader models or pre-trained LMs, which are tailored for MRC with long passages since the lengths of our context passages in the NECTEC dataset are very long. Finally, more incorporation of Thai-specific techniques, like usage of Thai dependency parse to help the reader mode, can also be pursued further.

9. Appendix

A. Appending Special Tokens to the Beginning or Ending of the Passages

Table 22 shows that appending the special tokens at the end of the passage works better in terms of overall performance, which is judged from overall F1 (%). A very low performance in yes-no accuracy for the version, which we append the tokens to the beginning of the context passage is noteworthy. The fact that most factoid questions tend to have the answer position located at the earlier portion of the passage could attribute to this phenomenon.

Table 22. Special token model's performance with different special token positions. Column with an asterisk is the main evaluation measurement.

Position of YESNO tokens	Overall F1(%)*	Factoid		Yes-no	Question Accuracy (%)	P-value
		EM (%)	F1 (%)	Accuracy (%)		
At the beginning	61.10	49.42	64.99	20.27	92.65	-
At the ending	***63.05***	***49.35***	***64.62***	***51.25***	***99.15***	***<0.001***

B. Experiments on Loss Combination in Joint Model

We have varied the span retrieval and yes-no classification loss combination in the joint model (section 4.2.2.2) as preliminary experiments. Table 23 reports the results on the validation set. We stick with the combination of 1:1 (no multiplicator factor).

Table 23. Preliminary Experiments on Loss Combination. Column with an asterisk is the main metric.

Span Retrieval Loss	Yes-no Loss	Overall F1 (%)*	Factoid		Yes-no	Question Accuracy (%)
			EM (%)	F1 (%)	Accuracy (%)	
1.0	5.0	63.35	48.17	64.25	56.55	99.74
1.0	0.2	64.43	49.60	65.51	56.34	99.74
1.0	***1.0***	***66.14***	***51.92***	***67.32***	***57.33***	***99.74***

C. Examples of Maximum Matching and Bailarn Answer Tokenization

We now show some results on the tokenized factoid question's answers. Table 24 shows only some of the mismatched tokenized answers from two tokenizers. There are 7,743 miss-matched tokenized answers between Maximum matching and Bailarn tokenizers [32]. For the maximum matching tokenizer, we use PythaiNLP implementation of Newmm.

Table 24. Comparison of Newmm and Bailarn Tokenizers

Newmm	Bailarn
[ฮิ, กกิ้นส์]	[ฮิกกิ้นส์]
[เม, ช, ตา]	[เมชตา]
[ประธานาธิบดี, วิลเลียม, เอช, ., ทัฟต์]	[ประธานาธิบดี, วิลเลียม, เอช., ทัฟต์]
[แคว้น, เอ, มี, เลีย, -, โร, มัญญา]	[แคว้นเอมีเลีย, -, โรมัญญา]
[เมือง, เอ, เม, อ, รี, วิลล์]	[เมือง, เอเมอรีวิลล์]
[บริษัท, เดอะ, วอ, ลต์, ดิสนีย์]	[บริษัท, เดอะวอลต์ดิสนีย์]
[วันที่, 12, เมษายน, พ.ศ., 2539]	[วัน, ที่, 12, เมษายน, พ.ศ., 2539]
[ประเทศอังกฤษ]	[ประเทศ, อังกฤษ]
[จัสติน, ทิม, เบอร์, เลก]	[จัสติน, ทิมเบอร์เลก]
[โป, แลน]	[โปแลน]
[โก, ลดา, เม, อี, ร์]	[โกลดา, เมอีร์]
[สุพรรณ, ษา, เวช, กามา]	[สุพรรณษา, เวชกามา]
[เซอร์, ปีเตอร์, โรเบิร์ต, แจ็กสัน]	[เซอร์ปีเตอร์, โรเบิร์ต, แจ็กสัน]
[ปา, เลม, บัง, บัง, ซุม, เซ, ล, บา, เบล]	[ปาเลมบัง, บัง, ซุมเซล, บาเบล]
[ปา, เลม, บัง, สปอร์ต, ฮอลล์]	[ปาเลมบังสปอร์ต, ฮอลล์]
[สถานีโทรทัศน์, ไทย, ทีวีสี, ช่อง, 3]	[สถานีโทรทัศน์, ไทย, ทีวี, สี, ช่อง, 3]
[แพ, จิน, -, ย็อง]	[แพ, จิน-, ย็อง]
[สะพาน, รุ, สส, กี]	[สะพาน, รุสสกี]
[นายกรัฐมนตรี, รัสเซีย, ด, มี, ตรี, เม, ดเว, เดฟ]	[นายก, รัฐมนตรี, รัสเซียด, มี, ตรี, เมดเวเดฟ]
[ริน, ไค, โฮ]	[ริน, ไคโฮ]
[ปัสกวัล, นี, โก, ลัส, เป, เร, ซ]	[ปัสกวัล, นีโกลัส, เปเรซ]
[โผน, กิ่ง, เพชร]	[โผน, กิ่งเพชร]
[จังหวัด, โล, กรอ, ญโญ]	[จังหวัด, โลก, รอญโญ]
[นะ, งะ, ชะ, กิ]	[นะงะชะกิ]

D. Pre-training results on XNLI-th dataset

Table 25 describes the result of the pre-trained models on XNLI dataset. It can be seen that even though the performances of some pre-trained models are not strong on XNLI, the pre-trained models still increase the performance of the yes-no model as shown in section 6.2.

Table 25. Performance of pre-trained models on XNLI corpus

Model	Test Accuracy (%)
BIDAF with Static embedding (section 4.3.2)	52.83
BIDAF with contextual embedding (section 4.3.5)	67.54
BERT fine-tuning (section 4.3.4)	68.00

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L. Li, L. Kai, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 20-25 June 2009 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[2] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100, 000+ Questions for Machine Comprehension of Text," in *EMNLP*, 2016.

[3] E. Choi *et al.*, "QuAC: Question Answering in Context," in *EMNLP*, 2018.

[4] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A Conversational Question Answering Challenge," *Transactions of the Association for Computational Linguistics,* vol. 7, pp. 249-266, 2019.

[5] Z. Yang *et al.*, "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering," in *EMNLP*, 2018.

[6] S. Yagcioglu, A. Erdem, E. Erdem, and N. Ikizler-Cinbis, "RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes," Brussels, Belgium, 2018: Association for Computational Linguistics, pp. 1358-1368.

[7] W. Jitkrittum, C. Haruechaiyasak, and T. Theeramunkong, "QAST: Question Answering System for Thai Wikipedia," 08/06 2009, doi: 10.3115/1697288.1697291.

[8] A. Kongthon, S. Kongyoung, C. Haruechaiyasak, and P. Palingoon, "A Semantic Based Question Answering System for Thailand Tourism Information," in *Proceedings of the KRAQ11 workshop*, Chiang Mai, nov 2011: Asian Federation of Natural Language Processing, pp. 38-42. [Online]. Available: https://www.aclweb.org/anthology/W11-3106. [Online]. Available: https://www.aclweb.org/anthology/W11-3106

[9] H. Decha and K. Patanukhom, "Development of thai question answering system," presented at the Proceedings of the 3rd International Conference on Communication and Information Processing, Tokyo, Japan, 2017.

[10] M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional Attention Flow

for Machine Comprehension," *ArXiv,* vol. abs/1611.01603, 2016.

[11] S. Liu, X. Zhang, S. Zhang, H. Wang, and W. Zhang, "Neural Machine Reading Comprehension: Methods and Trends," *Applied Sciences,* vol. 9, p. 3698, 09/05 2019, doi: 10.3390/app9183698.

[12] K. M. Hermann *et al.*, "Teaching Machines to Read and Comprehend," pp. 1693--1701, 2015. [Online]. Available: http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf.

[13] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale ReAding Comprehension Dataset From Examinations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, sep 2017: Association for Computational Linguistics, pp. 785-794. [Online]. Available: https://www.aclweb.org/anthology/D17-1082 http://dx.doi.org/10.18653/v1/D17-1082. [Online]. Available: https://www.aclweb.org/anthology/D17-1082 http://dx.doi.org/10.18653/v1/D17-1082

[14] M. Peters *et al.*, "Deep Contextualized Word Representations," New Orleans, Louisiana, 2018: Association for Computational Linguistics, pp. 2227-2237.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Minneapolis, Minnesota, 2019: Association for Computational Linguistics, pp. 4171-4186.

[16] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," Melbourne, Australia, 2018: Association for Computational Linguistics, pp. 328-339.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *CoRR,* vol. abs/1409.0473, 2014.

[18] A. Vaswani *et al.*, "Attention is All you Need," pp. 5998--6008, 2017. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[19] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," pp. 5753--5763, 2019. [Online]. Available: http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf.

[20] A. W. Yu *et al.*, "QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension," *ArXiv,* vol. abs/1804.09541, 2018.

[21] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, "Attention-over-Attention Neural Networks for Reading Comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, jul 2017: Association for Computational Linguistics, pp. 593-602. [Online]. Available: https://www.aclweb.org/anthology/P17-1055 http://dx.doi.org/10.18653/v1/P17-1055. [Online]. Available: https://www.aclweb.org/anthology/P17-1055 http://dx.doi.org/10.18653/v1/P17-1055

[22] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," presented at the Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, Lake Tahoe, Nevada, 2013.

[23] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," Doha, Qatar, oct 2014: Association for Computational Linguistics, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543, doi: 10.3115/v1/D14-1162. [Online]. Available: https://www.aclweb.org/anthology/D14-1162 https://doi.org/10.3115/v1/D14-1162

[24] P. Rajpurkar, R. Jia, and P. Liang, "Know What You Don't Know: Unanswerable Questions for SQuAD," Melbourne, Australia, jul 2018: Association for Computational Linguistics, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 784-789, doi: 10.18653/v1/P18-2124. [Online]. Available: https://www.aclweb.org/anthology/P18-2124 https://doi.org/10.18653/v1/P18-2124

[25] M. Saeidi *et al.*, *Interpretation of Natural Language Rules in Conversational Machine Reading*. 2018, pp. 2087-2097.

[26] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova,

"BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, jun 2019: Association for Computational Linguistics, pp. 2924-2936. [Online]. Available: https://www.aclweb.org/anthology/N19-1300 http://dx.doi.org/10.18653/v1/N19-1300. [Online]. Available: https://www.aclweb.org/anthology/N19-1300 http://dx.doi.org/10.18653/v1/N19-1300

[27] V. Zhong and L. Zettlemoyer, "E3: Entailment-driven Extracting and Editing for Conversational Machine Reading," in *ACL*, 2019.

[28] Y. Ohsugi, I. Saito, K. Nishida, H. Asano, and J. Tomita, "A Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension," in *Proceedings of the First Workshop on NLP for Conversational AI*, Florence, Italy, aug 2019: Association for Computational Linguistics, pp. 11-17. [Online]. Available: https://www.aclweb.org/anthology/W19-4102 http://dx.doi.org/10.18653/v1/W19-4102. [Online]. Available: https://www.aclweb.org/anthology/W19-4102 http://dx.doi.org/10.18653/v1/W19-4102

[29] Y. Ju, F. Zhao, S. Chen, B. Zheng, X. Yang, and Y. Liu, "Technical report on Conversational Question Answering," *arXiv e-prints*. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190910772J

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019.

[31] A. Williams, N. Nangia, and S. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, jun 2018: Association for Computational Linguistics, pp. 1112-1122. [Online]. Available:

https://www.aclweb.org/anthology/N18-1101
http://dx.doi.org/10.18653/v1/N18-1101. [Online]. Available:
https://www.aclweb.org/anthology/N18-1101
http://dx.doi.org/10.18653/v1/N18-1101

[32] A. Jettakul, C. Thamjarat, K. Liaowongphuthorn, C. Udomcharoenchaikit, P. Vateekul, and P. Boonkwan, "A Comparative Study on Various Deep Learning Techniques for Thai NLP Lexical and Syntactic Tasks on Noisy Data," *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE),* pp. 1-6, 2018.

[33] J. Welbl, P. Stenetorp, and S. Riedel, "Constructing Datasets for Multi-hop Reading Comprehension Across Documents," *Transactions of the Association for Computational Linguistics,* vol. 6, pp. 287-302, 2018. [Online]. Available: https://www.aclweb.org/anthology/Q18-1021
http://dx.doi.org/10.1162/tacl_a_00021.

[34] C. Qu *et al.*, "Attentive History Selection for Conversational Question Answering," *Proceedings of the 28th ACM International Conference on Information and Knowledge Management,* 2019.

[35] Y. Xie, S. Liu, T. Yao, Y. Peng, and Z. Lu, "Focusing Attention Network for Answer Ranking," 2019.

[36] C. Fu, Y. Li, and Y. Zhang, "ATNet: Answering Cloze-Style Questions via Intra-attention and Inter-attention," pp. 242-252, 2019, doi: 10.1007/978-3-030-16145-3_19.

[37] C. Zhu, M. Zeng, and X. Huang, "SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering," *ArXiv,* vol. abs/1812.03593, 2018.

[38] J. Zhang, X.-D. Zhu, Q. Chen, L.-R. Dai, S. Wei, and H. Jiang, "Exploring Question Understanding and Adaptation in Neural-Network-Based Question Answering," *ArXiv,* vol. abs/1703.04617, 2017.

[39] A. Conneau *et al.*, "XNLI: Evaluating Cross-lingual Sentence Representations," pp. 2475--2485, 2018, doi: 10.18653/v1/D18-1269.

[40] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-

Task Benchmark and Analysis Platform for Natural Language Understanding," 2018.

[41] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," 2019.

[42] M. Gardner *et al.*, "A Deep Semantic Natural Language Processing Platform," 2017.

[43] T. Wolf *et al.*, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *ArXiv,* vol. abs/1910.03771, 2019.

[44] C. Nadeau and Y. Bengio, "Inference for the Generalization Error," *Machine Learning,* vol. 52, pp. 239-281, 2003, doi: 10.1023/A:1024068626366.

[45] J. Kiani. "Using the Corrected Paired Student's t-test for comparing Machine Learning Models." https://medium.com/analytics-vidhya/using-the-corrected-paired-students-t-test-for-comparing-the-performance-of-machine-learning-dc6529eaa97f

[46] โล่พันธุ์ศิริกุล, "วาบิคิวเอ (WabiQA): โปรแกรมถามตอบจากคลังข้อมูลวิกิพีเดียภาษาไทย Question Answering Program from Thai Wikipedia," NECTEC, 12-Jan-2020., 2018.

[47] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," in *ACL*, 2017.

[48] J. Singh, B. McCann, N. Shirish Keskar, C. Xiong, and R. Socher, "XLDA: Cross-Lingual Data Augmentation for Natural Language Inference and Question Answering," *arXiv e-prints*. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190511471S

VITA

NAME	Theerit Lapchaicharoenkit
DATE OF BIRTH	26 April 1995
PLACE OF BIRTH	Uttaradit, Thailand
INSTITUTIONS ATTENDED	Chulalongkorn University
HOME ADDRESS	1/72-74 Samranrean Rd. Tha-it, Mueng, Uttaradit