Study the XLM-RoBERTa, ByT5, mT5 question and answer methods for Lao language in specific areas of internet network equipment

[1]Lathsamy CHIDTAVONG, [2]Sommith THOUMMALY, [3]Outhigone LABOUNTHANH, [4]Soulith SENGMANOTHUM, [5]Amone CHANTHAPHAVONG

Computer Science Department
Faculty of Natural Sciences
National University of Laos.

Abstract- **The purpose of this research is to study Question Answering method using XLM-RoBERTa, ByT5 and mT5 Models for closed domain network device in Lao language, the aim is to develop Question answering system to help answer question more quickly and close to human as much as possible in Lao language. Therefore, the dataset is retrieved and created from www.huawei.com by using web scraping technique, and then translated to Lao language. The dataset is in SQuAD format that consists of 921 articles (samples), where 330 articles is translated to Lao language remaining 591 articles keep in english, dataset has question 989 samples in Lao language, answer has 989 samples (393 Lao samples).**

The result of the research shows that training time of the XLM-RoBERTa model takes 58.3 minutes, the evaluation result by exact match is 51.51% and F1 Score is 78.38%. For the ByT5 model, training time is 120.76 minutes, evaluation result by exact match is 29.29% and F1 Score is 62.42%. The final model, the mT5 takes 82.28 minutes for training time, the exact match is 3.03% and F1 Score is 38.01%

Key words: **XLM-RoBERTa, ByT5, mT5, SQuAD, Machine Reading comphehension.**

1. INTRODUCTION

Now a day, information technology is important and has an impact on our human being in the perception of information, communication and trade, such as online social media, online trade (e-commerce). Data and information on the Internet world which is huge data (Big data), occurring every second and communication is available 24 hours a day. Therefore, government, agencies, companies, individuals, legal entities must find a way to quickly manage, deal with and interact with this problem. The use of artificial intelligence to help in questioning and answering is one of the tools that can help solve this problem, among which there is a popular technique such as Machine Reading comprehension using models XLM-RoBERTa, ByT5 and mT5 models are able to support more than 100 languages, including Lao language, which is ranked as having less vocabulary resources, and there are still few research experiments with Lao language. Therefore, we have the idea for doing a research on question and answer techniques using XLM-RoBERTa, ByT5 and mT5 techniques to study question-and-answer in Lao language.

From the research of Pranav Rajpurkar (2016) Stanford Question Answering Dataset (SQuAD) is a dataset for building, testing and evaluating a question-and-answer system that comes from more than 100,000 questions that bring information from Wikipedia articles. The answers will be in the article. The evaluation is using the F1 score method, the evaluation value of the dataset is to use the method of asking the person to read the answer and find the answer. The model of the question-and-answer system requires an evaluation value of more than 86.8%.

From Ashish Vaswani's (2017) research paper, the Transformer model is a model that uses the Attention-mechanism method to change from the sequence of word groups in a sentence represented by numbers or Sequence to a sequence of other word groups using an Encoder or Decoder and a Decoder as in Figure 1. The Encoder is on the left and the Decoder is on the right, which both parts can be connected to each other in multiple layers, which is defined by the Nx representation as in Figure 1. Attention function is a pair of query vector values, keys and values to calculate the output value. query is a vector input that replaces the value of all sentences, Value is a vector value that replaces the value of all words as a result of the answer, keys is a vector value that replaces all possible words.

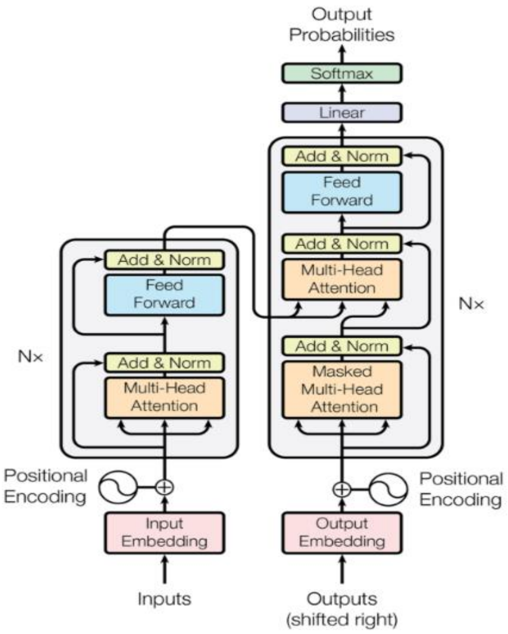

Figure 1: Transformer model from Mr. Ashish Vaswani's paper (2017) Attention is all you need.

In the Attention function, the equation will be used to find the Attention weight, which is:

(1)

Scale Dot-product Atter ıtent from the formula:

$$Attention(Q,K,V) = softmax_k\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

(2)

Multi-head attention inc ntion, Concat and Final linear layer. Each Multi-head attention will input Q (query), K (key), V (value) through the linear layer as shown in Figure 2.

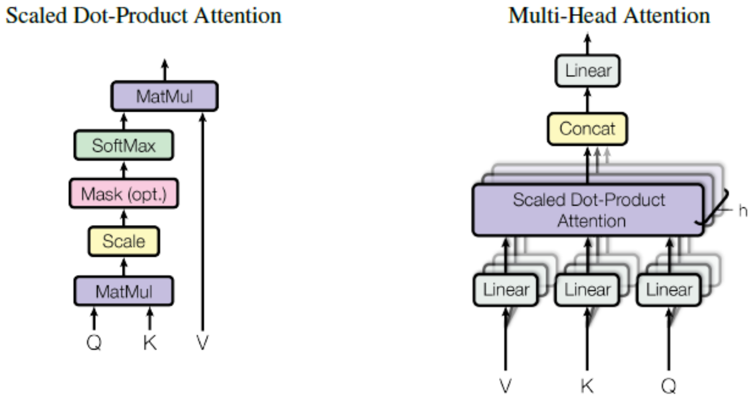

Figure 2: Scale Dot-production and Multi-head Attention Mr. Ashish Vaswani (2017) Attention is all you need.

Multi-head attention enables the model to input multiple values simultaneously using the following formula below:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \qquad (3)$$

$$\text{where head}_i = \text{Attention}(QW_i^Q ;KW_i^K ; VW_i^V)$$

Feed-Forward Networks will have Encoder and decoder layers connected to the Feed-Forward Network to calculate according to the formula below:

$$\text{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2 \qquad (4)$$

The words will be converted into a vector and find the position of the word in the sentence using the following formula below:

$$PE(pos,2i) = \sin(pos/10000^{2i/dmodel)} \qquad (5)$$

$$PE(POS,2i+1) = \cos(pos/10000^{2i/dmodel}) \qquad (6)$$

Optimizer in model training will use Adam optimizer with learning rate by equation:

$$lrate = d_{model}^{-0.5} * \min(step\_num^{-0.5},step\_num.warmup\_steps^{-1.5}) \quad (7)$$

From the research of Jacob Devlin (2018), the BERT model is a learning method that has been learned before (Pre-Train) from a large dataset that does not have an answer to (unlabeled), the time to use the model that is Pre-train must learn another layer (Output layer) with a specific dataset to be used in a specific task or called Fine-tune, for example: a question and answer system. The results of testing and evaluating the model with the SQuAD v1.1 dataset show that the F1 value is 93.2% while the SQuAD V2.0 F1 value is 83.1%

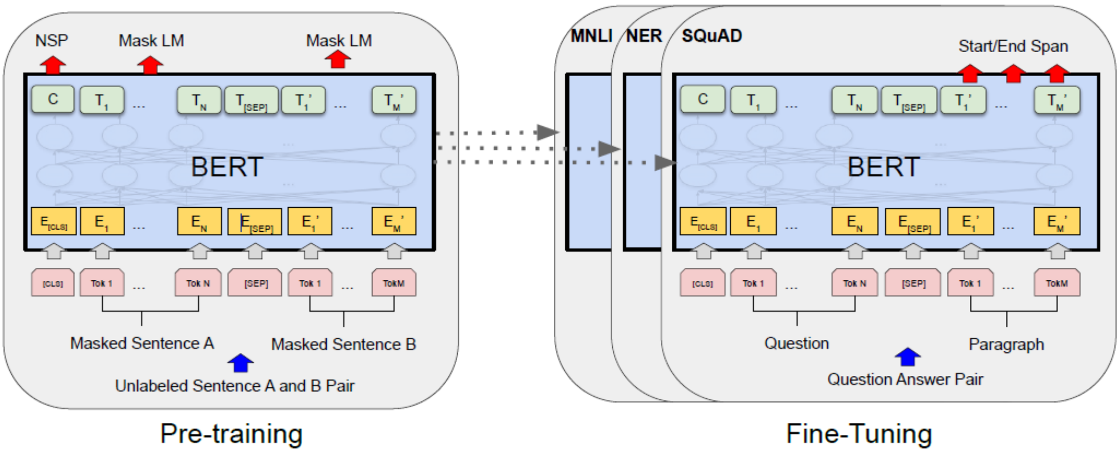

Figure 3: The overall process of BERT from Pre-training to Fine-tuning

From Figure 3, BERT Pre-training will be trained with each language dataset that requires a large amount of data, then it can be trained to use in other specific tasks such as: NER, SQuAD, MNLI. Which will be called Fine-Tuning.

From the research article of Alexis Conneau (2019) Cross-Lingual Language Model Robustly Optimized BERT Pre-training Approach is a pretraining model that has learned multiple languages at least 100 languages from the information of various articles on public websites such as Wikipedia. The model supports the Lao language whose F1 value from the XNLI data set is 14.6% higher than mBERT. However, the model has not yet published a document on the evaluation and test results of the Lao language.

From Linting Xue's (2022) research paper, pre-trained models will mostly use tokens to match words or sub-words. Researchers have provided a modeling technique that uses Byte Unicode UTF-8 of the teaching text instead of using words like other models to reduce the problem of interference and errors caused by the words of each language. The model is basically developed from the T5 model and uses the GLUE, SuperGLUE, XNLI and TweetQA datasets in the experiments. The research results show that the small model has better results in evaluating the XQuAD dataset (F1/EM) 74.0%/59.9% compared to mT5 71.3%/55.7%.

The model researched by Mr. Linting Xue (2021) is based on the T5 model by taking the information of articles in each language from the website or the C4 data set of 101 languages to train the said model which includes the Lao language. The results of the experiment show that the question and answer of the model can make F1 evaluation value of 82.5%, EM value is 66.8% compared to XLM-R with F1 value of 76.6% and EM value is 60.8%.

2. METHODOLOGY

In this research, we studied how the system answers the questions of the network equipment by using the model to read the article to find the answer, namely XLM Roberta, ByT5 and mT5. The dataset of questions and answers are retrieved from the website www.huawei.com and then selected 8 topics: CCTV, Fusion Cloud, Router, Switch, Server, Firewall, IPCC and Storage. The dataset contains 921 articles, including 330 articles in Lao, 591 English, 989 questions in Lao, and 989 answers.

Table 1: sample question-answer in Lao

Lao language article	Lao language questions
ເຈົ້າບໍ່ສາມາດເບິ່ງ alarms report ຂອງກ້ອງ IPC6231-WD-VRZ ໃນ web page ໄດ້, ເພາະວ່າກ້ອງບໍ່ມີ log records ເນື່ອງຈາກບໍ່ມີ SD card. ເຈົ້າສາມາດເບິ່ງ alarms report ຂອງກ້ອງ IPC6621-Z30-I ຜ່ານ alarm logs ທີ່ຖືກ ດຶງຂໍ້ມູນໃນ log query ຂອງໜ້າ web page.	ຂ້ອຍສາມາດຊອກ alarm report ຂອງກ້ອງ ipc6231-wd-vrz ແລະ ipc6621-z30 -i ໃນ web page ໄດ້ແນວໃດ?

Table 2: sample question-answer in English

Lao language article	Lao language questions
A camera must have an electric lens for performing zoom control on its web page. Camera zoom is to adjust the camera's shooting area. You can click "Zoom in" or "Zoom out" on the web page to adjust the area.	How can I perform zoom control on the web page of a camera?

Software tool to create the XLM roBERTa Large, ByT5 Small and mT5 small models are Nvidia CUDA, Pytorch, Farm Haystack, Huggingface-hub and Haystack Annotation. Detail for creating three models will explain as follow:

2.1 Creating XLM roBERTa Large Model

The dataset for modeling is totally 989 questions, and divided into 3 parts: 80% equal to 791 questions are for model training, 10% equal to 99 questions are for model validation and 10% equal to 99 questions are for model testing. In the process of training the XLM model roBERTa will use the Fine-tuned xlm-roberta-large-squad2 model to train the SQuAD question-answer dataset and as a Tokenizer.

In the Encoder layer, there will be 24 layers, so the total number of parameters will be word embedding 256,002,048 + position embedding 526,336 + token embedding 1,024 + Normalization(bias) 2,048 + encoder (12,596,224x24) + Linear(bias) 2050 1,472 = 558,842,882.

2.2 Creating ByT5 Small Model

The same process to the development of XLM roBERTa Large Model, we use the same dataset with 989 questions for modeling and divided into 3 parts: 80% equal to 791 questions are for model training, 10% equal to 99 questions are for model validation and 10% equal to 99 questions are for model testing. In the process of training the mT5 small model, the Fine-tuned ByT5 small Squad model will be used to train the SQuAD question-answer dataset and as a Tokenizer.

From the number of parameters of the encoder and decoder, each layer has embedding 565,248 + encoder 217,092,224 + decoder 81,415,040 + Linear 1,472 x 384 = 299,637,760 parameters.

2.3 Creating mT5 small Model

We keep use the same dataset (989 questions) and the same division ratio 80:10:10 for mT5 small modelling. In the process of training the mT5 small model, the Fine-tuned mT5 small model will be used to train the SQUAD question-answer data set and as a Tokenizer.

From the number of parameters of encoder and decoder in each layer, it will be embedding 128,057,344 + encoder 18,883,264 + decoder 25,178,816 + Linear 512 x 250.112 = 300,176,768 parameters.

3. Results

3.1 The results of xlm Roberta large Model

After training the XLM RoBERTa large model, then Test Dataset with 99 questions use to test the model. The result shows that the Exact match value is 45.45% and the F1 Score is 72.25%. The training time is 3,502 seconds. The training cycle is 4 cycles, in the table 3 shows numbers of learning cycles with value of training loss and validation loss in each round.

Table 3: training results for 4 training sessions

Epoch	Training Loss	Validation Loss
1	1.190500	1.675504
2	1.153700	1.228011
3	0.660900	1.385429
4	0.471100	2.361260

3.2 The results of ByT5 small Model

The evaluation results of the ByT5 small model is lower than the XLM RoBERTa large model, the Exact match value is 29.29% and the F1 Score is 62.42%. It takes 10 learning cycles (Epoch=10) and takes a total of 7,246 seconds to learn. As illustrated in figure 4 and figure 5 below:

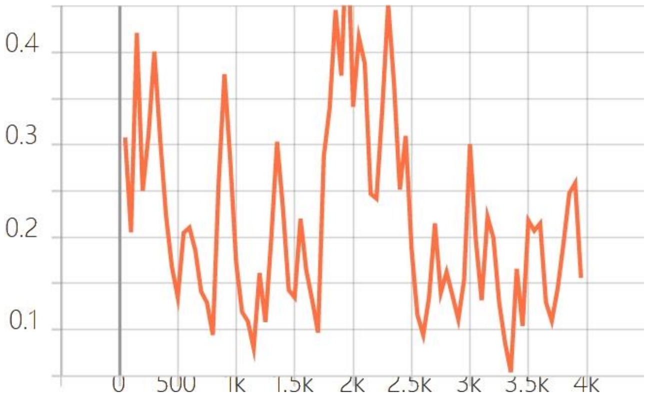

Figure 4: training loss of ByT5 small model

From Figure 4, we can be see that training the ByT5 small model for 10 rounds, the training loss value is decreasing according to the order of rounds.

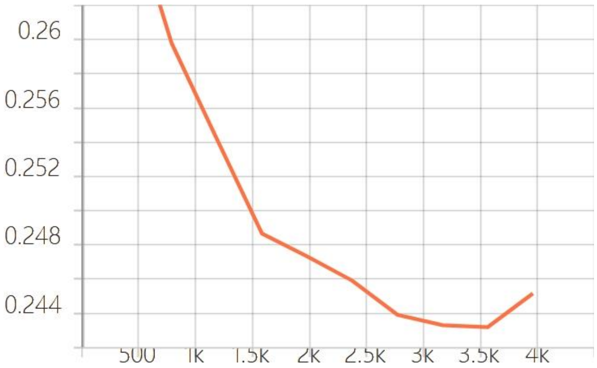

Figure 5: Validation loss of ByT5 small model

From Figure 5, we can be see that training the ByT5 small model for 10 cycles during the steps of 3,500 to 4,000. The value of validation loss is increased, causing overfitting problems. If we continue to increase the training cycle, it is not appropriate.

3.3 The results of mT5 small Model

The model is trained to learn 10 rounds (Epoch=10) takes a total of 4,937 seconds to learn. If we compare to the XLM RoBERTa large and the ByT5 small models, the mT5 small model gets the lowest performance. The Exact match value is only 3.03% and the F1 Score is 38.01%. From Figure 6 and 7, we can be see that the training loss and validation loss are reduced to 10 learning cycles of the mT5 small model.

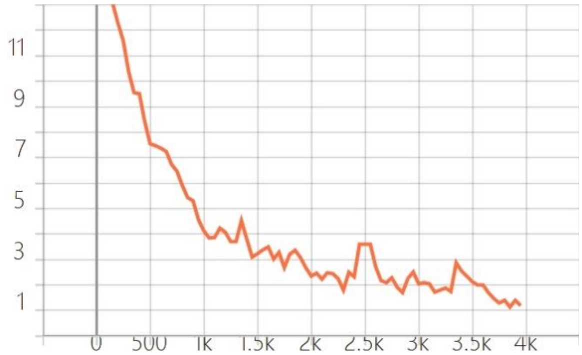

Figure 6: Train loss learning model of mT5 small

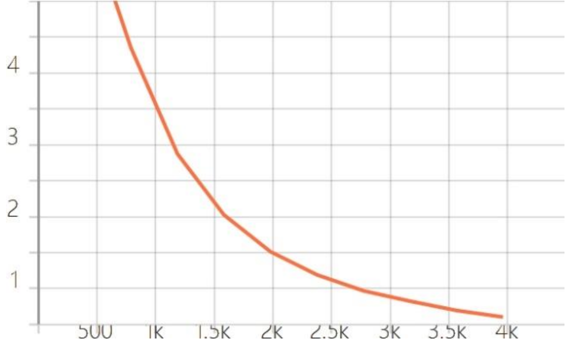

Figure 7: Validation learning loss of mT5 small model

4. DISCUSSION

The research on modelling to questioning and answering in Lao language is not widely investigate among computer scientists. Due to the limitation of dataset. In addition, artificial intelligent and machine learning are new popular research area in Laos. Therefore, it is hardly to find literature review supported to this research. The result of this research will be a good start for Lao computer scientists and others to deeply investigate more. The XLM-RoBERTa, ByT5 and mT5 models work by accepting questions in Lao language (Query) and then read the sentence (key) to find the answer from the set sentence (Value). According to the research result that depicts in figure 8, it is obviously show that the XLM-Roberta model is the most effective in answering questions with an EM value of 51.51% while the F1 Score value is 78.38% followed by ByT5 small with an EM value of 28.28% while the F1 score value is 62.57% and finally mT5 small has a value of EM is 3.03% while F1 score is 38.01%. However, this evaluation result is not high, due to number of dataset in Lao language is limited. The model can be improved and works more efficiency if it has trained and tested with sufficiently dataset.

5. SUMMARY

This research conducts three models for questioning and answering in Lao language, namely the XLM-RoBERTa, ByT5 and mT5 models. The dataset is retrieved from www.huawei.com by using web scraping technique and is transformed in SQuAD format that consists of 921 articles (samples), where 330 articles is translated to Lao language remaining 591 articles keep in english, dataset has question 989 samples in Lao language, answer has 989 samples (393 Lao samples). As the result, the XLM-RoBERTa model is the most efficient model which can be used to develop a system to answer questions in Lao language. For the future work, we can concentrate on some points as follows:

- Add more questions and answers in Lao language to make the model more efficient.
- Add as many Lao words to the model as possible.
- Investigate more models that support the Lao language to compare more efficiently.

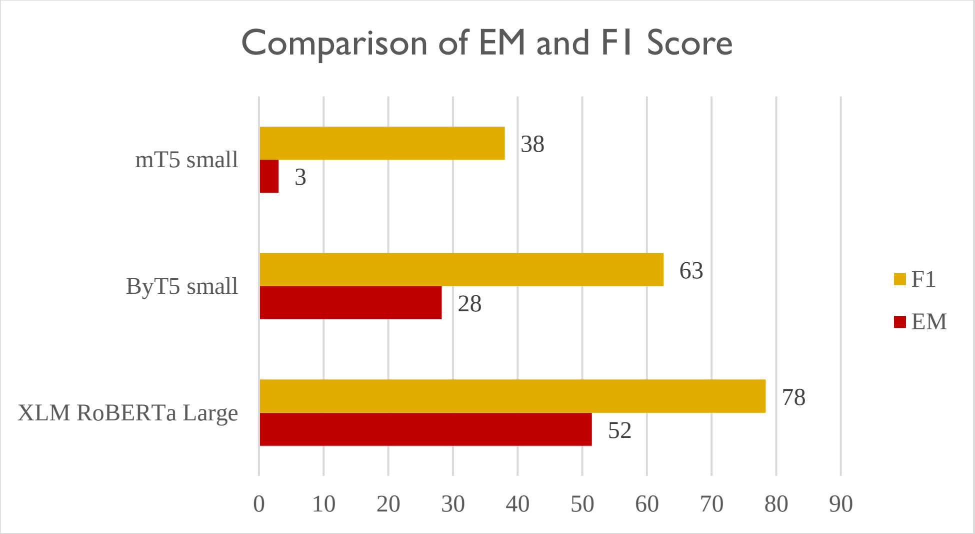

Figure 8: Comparison of EM value and F1 score of XLM-Roberta, ByT5, mT5 models

REFERENCES:

1. Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, Colin Raffel (2022). ByT5: Towards a token-Free Future with Pre-trained Byte-Byte Models. https://doi.org/10.48550/arXiv.2105.13626
2. Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, Colin Raffel (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. https://doi.org/10.48550/arXiv.2010.11934
3. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzm´an, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov (2019). Unsupervised Cross-lingual Representation Learning at Scale. https://doi.org/10.48550/arXiv.1911.02116
4. Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805
5. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin (2017) Attention is all you need. https://doi.org/10.48550/arXiv.1706.03762
6. Pranav Rajpurkar and Jian Zhang and Konstantin Lopyrev and Percy Liang (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. https://doi.org/10.48550/arXiv.1606.05250