2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC) | 978-1-6654-8559-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITC-CSCC55581.2022.9894948

A Comparative Empirical Evaluation of Neural Language Models for Thai Question-Answering

1st Fangyi Zhu
Institute of Data Science
National University of Singapore
fyzhu@nus.edu.sg

2nd Nasith Laosen
Faculty of Science and Technology
Phuket Rajabhat University
nasith.l@pkru.ac.th

3rd Kanjana Laosen
College of Computing
Prince of Songkla University, Phuket Campus
kanjana.l@phuket.psu.ac.th

4th Kannikar Paripremkul
College of Computing
Prince of Songkla University
kannikar.par@gmail.com

5th Aziz Nanthaamornphong
College of Computing
Prince of Songkla University, Phuket Campus
aziz.n@phuket.psu.ac.th

6th See-Kiong Ng
Institute of Data Science
National University of Singapore
seekiong@nus.edu.sg

7th Stéphane Bressan
School of Computing
National University of Singapore
steph@nus.edu.sg

***Abstract*—Despite engineers and researchers' significant and continuing efforts in developing natural language processing tools for the Thai language, the Thai language is, alongside many others, a de facto low-resource language. Can unsupervisedly trained neural language models come to the rescue? The remarkable success of transformer-based language models in most natural language processing tasks promises the advent of a much needed polyglot panacea. It seems, unfortunately, that powerful enough models are not yet available for most other-than-English languages. To assess the situation, we propose to empirically and comparatively evaluate the performance of existing neural language models for the task of extractive question-answering for the Thai language.**

***Index Terms*—Thai NLP, Question-Answering, Transformer**

I. Introduction

There are approximately fifty million Thai speakers worldwide. The Thai language, like many other languages and despite the significant and continuing efforts of engineers and researchers in developing natural language processing tools, is still a de facto low-resource language.

Transformer-based language models [1] have been remarkably successful in handling various downstream natural language processing tasks. While most results concern the English language, several works try and exploit transformer-based language models for other languages. For the Thai language, BERT-th [2] and WangchanBERTa [3] are currently the main models. Meanwhile, several multilingual language models are also available that include the Thai language.

Are neural language models realising the universal natural language processing solution that their successes suggest?

To assess the situation for the Thai language, we propose to empirically and comparatively evaluate the performance of existing neural language models for extractive question-answering. We review the main transformer-based neural monolingual and multilingual language models, and the available question-answering corpora. We empirically evaluate and compare the state-of-the-art neural languages models on the available corpora to establish a baseline performance for question answering for the Thai language. We analyse the results and discuss the remaining challenges in light of a comprehensive empirical evaluation results.

II. State-of-the-Art

In order to design and implement a state-of-the-art solution for extractive Thai question-answering, one needs, beside the elementary natural language processing tools for Thai, a neural language model for Thai or a neural multi-lingual language model that was trained with or is adapted to Thai, and an annotated corpus in Thai, sufficient in size and quality, for fine-tuning the language model to the task.

The first building blocks required for the processing Thai texts are word and morphological segmentation algorithms and tools. The written Thai language displays no explicit boundary between words. Processing Thai texts requires a language-aware word segmentation such as PyThaiNLP [4] and LexTo [5].

A language model is essential for solving downstream natural language processing tasks such as machine translation, natural language inference, and question answering. Neural transformer-based neural language models, monolingual language models and multilingual language models [6], have been particularly successful. Most language models, such as BERT and ALBERT [7], are monolingual language models. Several works attempted to train models for or adapt models to other languages following the spectacular success of neural language models for English. BERT-th is a BERT-based model trained on Thai Wikipedia. It underperforms traditional RNN-based models due to its limited training data. WangchanBERTa, a RoBERTa-based language model for the Thai language is trained with large amounts of texts from social media and news articles. Multilingual language models, such as multilingual BERT (mBERT) are pre-trained on raw texts from multiple languages and fine-tuned for downstream tasks [8]. These multilingual models can be generalised to other languages than those used for the initial training, even though they have never seen labelled data in those languages before. Several multilingual language models have been trained for Thai: multilingual BERT (mBERT), cross-lingual language models XLM [9]and XLM-R [10] Table I presents the architecture (encoder-only or encoder-decoder) of the most popular multilingual language models, indicates the number of their pre-trained languages, and whether they include the Thai language. Many of these models are trained using Wikipedia articles and consider the one hundred mostly used languages according to the online encyclopedia. Detailed lists of the pre-trained languages can be found in the reference for each work.

Question-answering, also known as machine reading comprehension, finds answers to questions about a corpus. Extractive question-answering proposes to output the text span corresponding to the answer in a document passage in the corpus. Extractive question-answering in English is one of the eleven tasks successfully tackled by BERT [17]. Some authors attempted to use automatic translation to transfer the task from a different language to English and back. Obviously, for question-answering tasks, the boundaries of the passages may not survive the translations. The authors of [18] proposed

Table I
TRANSFORMER-BASED MULTILINGUAL LANGUAGE MODELS

Model	Architecture	Pre-trained Languages	Work on Thai
mBERT [2]	Encoder-only	104 languages	Yes
XLM [9]	Encoder-only	100 languages	Yes
XLM-R [10]	Encoder-only	100 languages	Yes
mBART [11]	Encoder-Decoder	25 languages	No
ProphetNet-Multi [12]	Encoder-Decoder	100 languages	Yes
mT5 [13]	Encoder-Decoder	101 languages	Yes
mT6 [14]	Encoder-Decoder	101 languages	Yes
IndicBERT [15]	Encoder-only	11 Indian languages and Indian English	No
IndT5 [16]	Encoder-only	10 Indigenous languages and Spanish	No

Table II
COMPARISONS OF THAI QUESTION ANSWERING CORPORA

Corpus	Task	Thai	Size	Available
ThaiQA	Extractive QA	Yes	4074	Yes
XQuAD	Extractive QA	Yes	1190	Yes
TYDI QA	Extractive QA	No	204K	Yes
MKQA	Knowledge-based QA	Yes	10K	Yes
Thai WIKI QA	Extractive QA	Yes	15K	No
iApp Thai Wiki QA	Extractive QA	Yes	7242	Yes

the Translate Align Retrieve to translate the Stanford Question Answering corpus (SQuAD) to Spanish, which includes passage translation with a neural machine translator, context-alignment via a statistical unsupervised word alignment model, and answers retrieval with alignment. Subsequently, many works followed translating English SQuAD to other languages. However, such approaches rely on the availability of large parallel corpora to train a well-performed translator and the alignment model. Table II summarises the main Thai natural language processing corpora and compares them in terms of task, language, size, and availability.

III. METHODOLOGY AND EXPERIMENTAL SETUP

In the light of the above state-of-the-art, considering the few annotated corpora and the neural language models, limited in number and quality, available for the Thai language, we propose to design and implement and empirically and comparatively evaluate the currently possible solutions for Thai extractive question-answering.

Given a natural question q, a question-answering system locates a text span in the passage p with the start position a_s and the end position a_e as the answer. A modern question-answering system has two main components: a backbone

neural language model and a neural answer prediction module. First, the concatenated sequence $[q; p]$, consisting of the question and passage, is fed to the backbone neural network. Next, the latent representations of the token, words or stems, in the sequence are output by the neural language model and fed to the answer prediction module to predict the start position and the end position of the answer in the passage. The latent representation obtained by the backbone is fed into two linear layers trained to predict the start position and the end positions of the answer.

We consider the following transformer-based neural language models for the backbone: a multilingual language model mBERT-base, a cross-lingual language model XLM-R, and a monolingual language model for the Thai language WangchanBERTa released by VISTEC-depa AI Research Institute of Thailand on Huggingface. During training, we use AdamW optimizer [19]. We set a learning rate to 3e-5. We train the model with a mini-batch size of 100 for 7 epochs.

For availability reasons, we can only compare the following corpora for training and testing the different language models: iApp Thai WIKI QA, ThaiQA, XQuAD. We use the iApp Thai WIKI QA corpus released on https://huggingface.co/datasets/iapp_wiki_qa_squad that consists of 5761/742/739 question-answer pairs from 1529/191/192 articles for training/validation/testing, separately. We use the ThaiQA corpus released on https://huggingface.co/datasets/thaiqa_squad. This corpus only released the training set and the validation set. Thus, we train the model with the original training set consisting of 4k question-answer pairs, and evaluate the model on the validation set consisting of 74 question-answer pairs. We use the XQuAD corpus, we leverage the data released on https://huggingface.co/datasets/xquad. This corpus contains 1190 question-answer pairs in Thai language. We split these question-answer pairs into three sets: the training set, the validation set, and the testing set. There were 876/161/153 question-answer pairs from 34/7/7 articles in the training/validating/testing set respectively. We have uploaded the split corpus to https://huggingface.co/datasets/zhufy/xquad_split.

IV. Experimental Results and Analysis

We compare the effectiveness of the three models on the different training and testing corpora. We use the standard metrics for question answering, namely, F1 measure and exact match (EM). F1 measures the average overlap between the prediction and the ground truth. EM measures whether the prediction exactly matches the ground truth.

Table III
F1 Measure / Exact Match

Model	iApp Thai WIKI QA	ThaiQA	XQuAD	Overall
WangchanBERTa	74.58/49.66	69.24/50.00	41.21/31.37	68.89/46.79
mBERT	75.28/54.13	**78.30/60.81**	**58.79/50.33**	72.90/**54.04**
XLM-R	**79.69/56.16**	75.88/56.76	39.98/33.33	**73.11**/52.59

Table IV
Combined training, F1 Measure / Exact Match

Model	iApp Thai WIKI QA	ThaiQA	XQuAD	Overall
WangchanBERTa	76.46/50.20	72.11/52.70	68.51/54.25	74.87/51.04
mBERT	77.72/**54.26**	77.73/**64.86**	65.10/52.94	75.72/54.86
XLM-R	**79.90**/53.72	**81.85**/63.51	**74.72/62.09**	**79.23/55.80**

Table III reports the F1 and exact match of WangchanBERTa, mBERT, and XLM-R on the three corpora. It can be observed that the multilingual language models always performed better than WangchanBERTa. On the iApp Thai WIKI QA corpus, XLM-R achieved 79.69/56.16 in terms of F1/EM, while mBERT got 75.28/54.13 and WangchanBERTa got 74.58/49.66. On the other two corpora, ThaiQA and XQuAD, mBERT achieved 78.30/60.81 F1/EM and 58.79/50.33 F1/EM, respectively, which significantly outperformed XLM-R and WangchanBERTa. We further reported the overall performance of each approach in the last rows.

To test the effect of more training data, we combined the training data of the three corpora. Totally, there are 10637 question-answer pairs for training, and 4903 question-answer pairs for validation. Table IV reports the models performance under the joint training data in terms of F1 and EM. It can be observed that most models could perform better after adding more training data, except XLM-R dropped a bit on the iAPP Thai WIKI QA corpus. In particular, the performance on the XQuAD corpus achieved significant improvement. For instance, WangchanBERTa obtained a 54.25 EM score after adding more training data, while it only got 31.37 previously.

A micro-analysis found recurring types of errors and this regardless of the size of the training. This errors can be due to word segmentation, ambiguous, imprecise question and synonyms, incomplete composition, date, time, and symbols in the question, or multiplicity of answers. For instance, in Example 1, the word สำนักพิมพ์ (publisher) should not be segmented into the two words สำนัก (office) and พิมพ์ (print). Also, in the question, it is referred to as สำนัก in short.

Example 1:

Context: ...โดย|สำนักพิมพ์|ฮะ|คุ|เซน|ชะ| |ส่วน|ในประเทศ| ไทย|มี|การ|ตี|พิมพ์|ฉบับ|เลิข|สิทธิ์|จัดจำหน่าย|ครั้งแรก|โดย|สำนักพิมพ์ |วิบูลย์|กิจ...

Question: ตีพิมพ์|ครั้งแรก|สำนัก|ไทย|ชื่อว่า|อะไร

Prediction: ของ|ญี่ปุ่น

GroundTruth: สำนักพิมพ์|วิบูลย์|กิจ

Example 2 shows an error occuring when the context and the question use words or constructions with similar but different meaning. The context is "...It has an approximate length of about 60 - 80 centimeters. The biggest was 112 centimeter...". The question asks "How long is the longest blue-gray snapper?". The prediction by the model is "has 60-80 centimetres long", yet the correct answer should be "112 centimetre". There are two possible reasons behind this error. First, the model confuses ยาว (long) and ใหญ่ (big). Second, the model confuses the comparative modifier กว่า (-er) and the superlative modifier ที่สุด (-est).

Example 2:

Context: ..."|prion|"| |หมายถึง| |"|เลื่อย|")| |มี|ความ|ยาว| ประมาณ| |60|-|80| |เซนติเมตร| |ใหญ่|ที่สุด|พบ|ยาว|ถึง| |112| |เซน-ติเมตร|...

Question: ปลากะพงเขียว มีความยาวมากที่สุดเท่าไร

Prediction: มีความยาวประมาณ 60-80 เซนติเมตร

GroundTruth: 112 เซนติเมตร

V. Conclusions

We empirically and comparatively evaluated the performance of state-of-art neural language models for extractive question-answering in Thai with the available Thai question-answering corpora. A closer looks at the generally mediocre results, compared to those of the same architecture and comparable models for the English language, but for the size and quality of their training corpora, compels further efforts integrating language- and linguistics-aware mechanisms to handle language-specific morphology, syntax, and grammatical knowledge. We are devising graph neural network architectures that shall allow the combination of statistical information with the prescribed knowledge of symbolic linguistic structures alongside semantics and pragmatics.

Acknowledgment

This research is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Prepositioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," NIPS, pp. 5998–6008, 2017.

[2] Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," ACL, pp. 4171–4186, June 2019.

[3] L. Lowphansirikul, C. Polpanumas, N. Jantrakulchai, and S. Nutanong, "WangchanBERTa: Pretraining transformer-based Thai language models," arXiv preprint arXiv:2101.09635, March 2021.

[4] W. Phatthiyaphaibun, K. Chaovavanich, C. Polpanumas, A. Suriyawongkul, L. Lowphansirikul, and P. Chormai, "PyThaiNLP: Thai Natural Language Processing in Python", Zenodo, June 2016.

[5] C. Haruechaiyasak and A. Kongthon, "LexToPlus: A Thai lexeme tokenization and normalization tool," WSSANLP, Federation of Natural Language Processing, pp. 9–16, October 2013.

[6] S. Ralethe, "Adaptation of deep bidirectional transformers for Afrikaans language," LREC, European Language Resources Association, pp. 2475–2478, May 2020.

[7] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," ICLR, OpenReview.net, April 2020.

[8] Z. Chi, L. Dong, F. Wei, W. Wang, X.-L. Mao, and H. Huang, "Cross-lingual natural language generation via pre-training," AAAI, pp. 7570–7577, April 2020.

[9] G. Lample and A. Conneau, "Cross-lingual language model pretraining," NIPS, Curran Associates Inc., 2019.

[10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," ACL, pp. 8440–845, July 2020.

[11] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," TACL, MIT Press, pp. 726–742, 2020.

[12] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, "ProphetNet: Predicting future n-gram for sequence-to-Sequence Pre-training," Findings of EMNLP, pp. 2401–2410, November 2020.

[13] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," NAACL, pp. 483–498, June 2021.

[14] Z. Chi, L. Dong, S. Ma, S. H. X.-L. Mao, H. Huang, and F. Wei, "mT6: Multilingual pretrained text-to-text transformer with translation pairs," EMNLP, pp. 1671-1683, November 2021.

[15] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," Findings of EMNLP, pp. 4948–4961, November 2020.

[16] E. M. B. Nagoudi, W.-R. Chen, M. Abdul-Mageed, and H. Cavusoglu, "IndT5: A text-to-text transformer for 10 indigenous languages," AmericasNLP (NAACL), pp. 265–271, Jun. 2021.

[17] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," Science China Technological Sciences, pp. 1872-1897, October 2020.

[18] C. P. Carrino, M. R. Costa-juss'a, and J. A. R. Fonollosa, "Automatic Spanish translation of SQuAD dataset for multi-lingual question answering," LREC, European Language Resources Association, pp. 5515–5523, May 2020.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," ICLR, OpenReview.net, May 2019.