Vietnamese Legal Information Extraction

Hung Long Tran
Pho Thong Nang Khieu
Ho Chi Minh city, Vietnam
hunglongtran2004@gmail.com

Abstract

Legal question answering is an important task in Vietnamese Language Processing as advancements in this task help many people to find instructions for their legal problems. This is important for a developing country like Vietnam. In 2021, Zalo AI challenge, an annual Deep Learning competition for Vietnamese tasks, focus on this task. In this report, we will experiment with two famous information extraction systems. We acknowledge our limitations in this research due to the shortage of time and knowledge. We will discuss next works we will carry out in the future to further improve our contribution to the research community.

1 Introduction

The settings of this Zalo AI challenge is similar to open-domain question answering, a task of building computer systems that automatically answer questions given by human users in natural language (English, German, Spanish, etc), usually rely on general ontologis and world knowledge **cite here**. However, the systems are not required to output the exact answer, but only the passages that contain answers to the given question. Passages must be extracted from a large corpus, which has, in this challenge, $130,000$ different passages. There is a severe problem that every researcher has to deal with when encounter this task is the time limit. We cannot run a super Deep Learning model throughout the whole corpus with 130,000 passages every time the system is given a new question. This will cost us a large amount of time and computation to perform the task. Therefore, in this report, our approach to this task is similar to approach that prominent researchers in Natural Language Processing approach Open-Domain Question Answering.

2 Related Work

Open-Domain Question Answering has long developed before the onset of Deep Learning (Simmons, 1965; Green et al., 1961; Kirsch, 1964). Early categories of Question Answring systems include list-stuctured database systems (Organizing knowledge in list database), Graphic database systems (map text and graphic data to the same logical representations), Text-based systems (matching questions and text in a corpus to find answers), Logical inference systems (textual entailment, answering science text book questions). However, these early researches in Question Answering have limited success due to the fact that Question Answering is too complicated to be dealt with a mostly rule-based system. In recent years, the rapid development of Deep Learning has empowered the development in Question Answering with the pioneer Watson system from IBM (Ferrucci, 2012).

The development of Open-Domain Question Answering in current years takes place with various approaches. The very first approach that set new state of the art on Open Domain Question Answering benchmark is two-stage retriever-reader with the pioneer is Chen et al. (2017). This approach use two independent systems, a document retriever and a document reader, to perform two independent tasks. The main task for first phase system (document retriever) is to find passages, from a large corpus, that contain answers to the given question. The document retriever is not trainable, which is beneficial for researchers in saving computation. On the other hand, the document reader is trainable and be trained on predicting the answer from the extracted passages. The document reader also need to reject answering question if there is no answer to the given question. Following the success of Chen et al. (2017), many other researchers have approached Open Domain Question Answering using

two-stage system such as Raison et al. (2018); Yang et al. (2019); Clark and Gardner (2017); Wang et al. (2019).
Other approaches to Open Domain Question Answering are dense retriever end-to-end training (Yih et al., 2011), which, instead of using non-trainable retriever, use a deep learning based retriever, and retriver free, which has no retriever, with the famous representatives are members in the family of GPT (Radford et al., 2019; Brown et al., 2020).
The development of Open-Domain Question Answering take place with the existence of high quality question answering datasets (Rajpurkar et al., 2016, 2018; Dua et al., 2019; Saha et al., 2018; Lai et al., 2017; Dasigi et al., 2019; Yang et al., 2015).
Works have been done in Vietnamese Language Processing to contribute to the development of Vietnamese question answering including datasets (Nguyen et al., 2020b,a,c; Luu et al., 2021) and systems development (Van Nguyen et al., 2021, b,a).

3 Experimented Systems

Our experiments are inspired by work of DrQA by (Chen et al., 2017). We design our systems as two phase systems:

- In the first phase, our systems will try to limit the number of passages that is related to the questions. In other words, in this phase, our system will try to reduce the number of passages that will be processed by Deep Learning models by using different statistical methods. In this work, we will try BM25 and TF-IDF.

- In the second phase, our systems will try to validate whether each passage extracted by the first phase part contains answer for the given question. In this part, we need a super model as the number of passages that model has to process given a question is constant. This means the running time of the phase 2 part at that time will be $O(n)$ with n is the number of questions. Due to the limit of research fund, we cannot carry out our works on different super-models to compare the performances between models. Our chosen model for this phase is XLM-Roberta. This model is state-of-the-art on Vietnamese Question Answering reported by (Nguyen et al., 2020a).

3.1 Experiment Set-up

Firstly, researchers are not provided with the labels of development and test sets. Therefore, we have to split development set from the training set although we acknowledge that this might underestimate models' performances because of the lack of list of all possible answers. We divided the training set with the proportion of 90% and 10% for train and development respectively.
Secondly, for the evaluation while developing systems, we evaluate two different part of the designed systems. For the first part, we use normal accuracy to evaluate the non-trainable model's performances. On the other hand, to evaluate the performances of the overall system, we use F1-score.

$$Precision = \frac{true\_positives}{true\_positives + false\_negatives}$$

$$Recall = \frac{true\_positives}{true\_positives + false\_positives}$$

$$F1_{score} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

3.2 Non-Deep Learning Phase

In this phase, we focus on running the two different non-Deep-Learning information extraction models, tf-idf based and Okapi BM25. The table 1 report the results of our experiments.

For evaluating the performances of next stage model in collaboration with different non-deep-learning system, we extract the training set from the predictions of these first-phase systems and train second-phase models on these training sets. The F1-score is then reported on the development for the whole systems.

3.3 Deep Learning Phase

Due to the limits of hardware, we cannot experiment with different state-of-the-art models. Therefore, we have to choose one super model to experiment with different train and development sets. At the end, after researching different works (Van Nguyen et al., 2021, b,a; Nguyen et al., 2020b,a,c; Luu et al., 2021) in Vietnamese Question Answering, we decide to use XLM-Roberta (Conneau et al., 2020).

Top k	tf-idf 3 gram	tf-idf 2 and 3 gram	bm25
10	76.62	76.19	66.71
5	64.89	68.77	63.21
3	59.42	62.04	59.79

Table 1: Phase 1 results of different non-trainable systems

4 Discussion

There are many points that we must work on to fully assess the performances of the designed systems.

- We have not try to properly deal with the context that is too long. We must acknowledge that the maximum number of tokens that we can pass into XLM-Roberta is 768. However, there are many passages given by the legal corpus is much longer than this length. We will try our best to improve our coding skills to implement different methods researched in English.

- As the main task of XLM-Roberta in our designed systems is to determine whether a given passage contains answer to the given questions, there are many passages in the training and testing sets not containing answer to the corresponding questions. From this point of view, models that previously trained (pre-trained) on modified version of SQuAD 1.1 might be not enough to optimize the performances of our designed systems. In the near future, we might try to pre-train original XLM-Roberta with the task similar to that of SQuAD 2.0 to prepare models with knowledge of unanswerable questions.

Besides, we only see the importance of developing open-domain question answering systems for Vietnamese Language. Therefore, instead of focusing merely on the legal corpus, in the future, we might shift our focus on datasets that is more general in Vietnamese (Nguyen et al., 2020a). This might be important for the development of question answering tasks in Vietnamese.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Christopher Clark and Matt Gardner. 2017. Simple and effective multi-paragraph reading comprehension.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

D. A. Ferrucci. 2012. Introduction to “this is watson”. *IBM Journal of Research and Development*, 56(3.4):1:1–1:15.

Bert F. Green, Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. 1961. Baseball: An automatic question-answerer. In *Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference*, IRE-AIEE-ACM ’61 (Western), page 219–224, New York, NY, USA. Association for Computing Machinery.

Russell A. Kirsch. 1964. Computer interpretation of english text and picture patterns. *IEEE Trans. Electron. Comput.*, 13:363–376.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Son T. Luu, Mao Nguyen Bui, Loi Duc Nguyen, Khiem Vinh Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Conversational machine reading comprehension for vietnamese healthcare texts. In *Advances in Computational Collective Intelligence*, pages 546–558, Cham. Springer International Publishing.

Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020a. A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kiet Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020b. New vietnamese corpus for machine reading comprehension of health news articles. *arXiv preprint arXiv:2006.11138*.

Kiet Van Nguyen, Khiem Vinh Tran, Son T. Luu, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020c. Enhancing lexical-based approach with external knowledge for vietnamese multiple-choice machine reading comprehension. *IEEE Access*, 8:201404–201417.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Martin Raison, Pierre-Emmanuel Mazaré, Rajarshi Das, and Antoine Bordes. 2018. Weaver: Deep co-encoding of questions and documents for machine reading.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. Duorc: Towards complex language understanding with paraphrased reading comprehension.

R. F. Simmons. 1965. Answering english questions by computer: A survey. *Commun. ACM*, 8(1):53–70.

Kiet Van Nguyen, Phong Nguyen-Thuan Do, Nhat Duy Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. a. Multi-stage transfer learning with bertology-based language models for question answering system in vietnamese.

Kiet Van Nguyen, Nhat Duy Nguyen, Phong Nguyen-Thuan Do, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Vireader: A wikipedia-based vietnamese reading comprehension system using transfer learning. *Journal of Intelligent & Fuzzy Systems*, (Preprint):1–19.

Kiet Van Nguyen, Nhat Duy Nguyen, Phong Nguyen-Thuan Do, Tin Van Huynh, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. b. Xlmrserini: Open-domain question answering on vietnamese wikipedia-based textual knowledge source.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 247–256, Portland, Oregon, USA. Association for Computational Linguistics.