Study of Question Answering on Legal Software Document using BERT based models

Anonymous ACL submission

Abstract

The transformer-based architectures have achieved remarkable success in several Natural Language Processing tasks, such as the Question Answering domain. Our research focuses on different transformer-based language models' performance in software development legal domain specialized datasets for the Question Answering task. It compares the performance with the general-purpose Question Answering task. We have experimented with the PolicyQA dataset and conformed to documents regarding users' data handling policies, which fall into the software legal domain. We used as base encoders BERT, ALBERT, RoBERTa, DistilBERT and LEGAL-BER and compare their performance on the Question answering benchmark dataset SQuAD V2.0 and PolicyQA. Our results indicate that the performance of these models as contextual embeddings encoders in the PolicyQA dataset is significantly lower than in the SQuAD V2.0. Furthermore, we showed that surprisingly general domain BERT-based models like ALBERT and BERT obtain better performance than a more domain-specific trained model like LEGAL-BERT.

1 Introduction

Question Answering (QA) Systems are an automated method for retrieving correct answers to questions posed by humans in natural language (Dwivedi and Singh, 2013). A subclass of Question Answering systems is machine reading comprehension, whose primary goal is to retrieve answers to a given question in a single paragraph of text. The task has achieved remarkable success in the general domain using transformer-based architecture. However, previous research has demonstrated that utilizing in-domain text for training can benefit more from general-domain language models in specialized disciplines such as biology (Lee et al., 2020). Thus, one can infer that using legal-domain text can leverage the Question Answering performance regarding legal questionnaires.

Legal documents are challenging to understand appropriately without a legal background. The challenges also lie with software companies and software privacy policies and regulations. The exponential growth of applications worldwide and monitoring of our environments, decisions, tastes, and others make it more and more important to be aware of how the data is being managed, shared, and used. Companies must include the stated information in the privacy policies of every application. The challenges lie in the characteristics of legal software documents, which are longevity, ambiguity, and complexity. A high-performance Question Answering system on legal documents of software systems (privacy, policy rules) can have various applications. One example of possible use is that every person can check fast if their queries have an answer in such a long document before signing or agreeing to the terms and conditions.

The big transformer-based BERT model has had a great result on general-purpose dataset SQuAD version 1 (V1), and 2 (V2) (Rajpurkar et al., 2016, 2018). There is limited research on the performance of BERT-model variants on legal software datasets such as PolicyQA (Ahmad et al., 2020; Martinez-Gil, 2021) which reports the results only of the original BERT model.

In this paper, we provide a study of the performance of models like BERT (Devlin et al., 2018), LEGAL-BERT (Chalkidis et al., 2020), ALBERT (Lan et al., 2019), DistilBERT (Sanh et al., 2019), and RoBERTa (Liu et al., 2019) in the SQuAD and PolicyQA datasets. Furthermore, we compare and analyze such results, allowing us to pick the best model to obtain the best results in the PolicyQA dataset using only pre-trained models. Our results indicate that the performance of these models as contextual embeddings encoders in the PolicyQA dataset is significantly lower than

in the SQuAD V2.0. Furthermore, we showed that surprisingly general domain BERT-based models like ALBERT and BERT obtain better performance than a more domain-specific trained model like LEGAL-BERT.

The paper is organized as follows. First, we present a section of related work. The next section elaborates on the methodology followed. Then, in section 3 presents performed experiments. The consequent section 4 details a discussion about the results. The final section concludes the work.

2 Related Work

Researchers have dedicated significantly to the Question Answering systems in the legal domain in recent times. According to the study by (Martinez-Gil, 2021) Deep Learning models achieved the best results. The most recent research success in Question Answering and Legal Question Answering (LQA) has come from Neural Attentive Text Representation. Few Shot Learning in the legal domain and diverse applications of the successful BERT model (Devlin et al., 2018; Martinez-Gil, 2021).

There are various datasets for working with Question Answering tasks in the general-purpose domain, and the SQuAD is the most recognized because of achieving benchmark results (Rajpurkar et al., 2018). In the legal domain, JEC-QA (Zhong et al., 2020), ResPubliQA (Peñas et al., 2009), JRC-ACQUIS Multilingual Parallel Corpus (Steinberger et al., 2006) are well recognized. In the legal domain of Software Development of Question Answering datasets, PolicyQA (Ahmad et al., 2020) is one of the most well-known and compatible with the SQuAD dataset format. However, the study on the performance of some significant benchmarks in general Question Answering in these domain-specific datasets is limited (Martinez-Gil, 2021).

To the best of our knowledge, this is the first work studying the performance on the PolicyQA dataset using as architecture only variants of BERT-based model encoder besides the original BERT. Also, the best results reported have been on the work (Ahmad et al., 2020) using the original BERT model with **29.5** of Exact Match (EM) and **56.11** of F1. This work showed that ALBERT is a better encoder and obtains the best results in our study on the PolicyQA dataset.

3 Methodology

We studied and compared the performance of several BERT-related models in two Question Answering datasets. One dataset is from general domains SQuAD V2.0 and a specific domain in legal text related to software development called PolicyQA.

3.1 Datasets

The SQuAD V2 dataset is a reading comprehension dataset consisting of more than 100,000 questions posed by crowdworkers on a set of Wikipedia articles. The answer to each question is a segment of text from the corresponding reading passage (Rajpurkar et al., 2016, 2018).

The PolicyQA dataset is a reading comprehension dataset that contains 25,017 reading comprehension style examples curated from an existing corpus of 115 website privacy policies. PolicyQA provides 714 human-annotated questions for a wide range of privacy practices (Ahmad et al., 2020).

Both datasets are designed for the extractive Question Answering, where the answer is a span of text in the passage. Also, the passage might not be related to the question and does not contain the answer.

3.2 Models

We selected a set of the most used BERT-related models with outstanding performance in the SQuAD dataset like ALBERT, RoBERTa, and classic BERT. Additionally, we utilized the DistilBERT model because it is a cheaper and smaller model with competitive capabilities compared to other bigger BERT-based models. Thus, it is feasible to choose the DistilBERT model for speed during inference and usability on devices (Sanh et al., 2019). Moreover, we tested the LEGAL-BERT model, a version of the original BERT model trained from scratch with legal documents (Chalkidis et al., 2020). We compared it with the other general-purpose models to see if we could get better results using a legal-based BERT in the PolicyQA dataset than others trained in general text.

4 Experiments

We conducted our experiment using the pretrained versions of the models BERT, ALBERT, RoBERTa, DistilBERT, and LEGAL-BERT. The benchmark in the Question Answering (QA) task is evaluated

Dataset Name	BERT		ALBERT		LEGAL-BERT		RoBERTa		DistillBERT	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
SQUAD V2.0 5 epochs	71.6	77.39	73.9	77.91	73.5	77.01	**76.9**	**80.1**	65.47	69.27
PolicyQA 5 epochs	**29.5**	56.11	28.76	**57.36**	28.08	54.66	27.23	54.88	25.42	52.34
SQUAD V2.0 10 epochs	71.7	75.39	72.71	**77.21**	71.5	75.30	**75.18**	74.30	64.79	68.93
PolicyQA 10 epochs	29.6	57.02	**29.7**	**58.43**	28.45	55.01	27.85	54.91	25.81	52.48

Table 1: Result of the models on SQUAD V2.0 and PolicyQA

on the EM (Exact Match) and F1 metrics (Jurafsky and Martin).

To run our experiments, we used the open-source code from hugginface/transformers [1], we built two notebooks to run experiments using this code, and we provide an open repository with both notebooks at [redacted]. In this work, we didn't make any changes to the original code, we only plugged in the PolicyQA dataset instead of the usual SQuAD and also the particular model we wanted to test.

Our selected models ran for 5 and 10 epochs on PolicyQA and SQUAD V2.0 datasets. Table 1 compares the models on both datasets and variations of epochs, and the metric for measurement is EM (Exact Match) and F1.

5 Discussion

Our experimental results revealed some interesting aspects. Firstly, The BERT and ALBERT model's pre-trained version performed better than the LEGAL-BERT model in SQUAD V2.0 and PolicyQA datasets. Since LEGAL-BERT was trained on legal text, our assumption was that it would perform better than all the models. However, our assumption did not hold. We believe that there are notable separations among subdomains even inside the legal domain. PolicyQA is legal text in the subdomain of Software Development and privacy in applications. Thus, it is feasible that a more generalized pre-trained model would produce a better result than a model trained in a legal subdomain with high separation from the Software Development legal domain. Secondly, in PolicyQA, the increase of epochs improves the performance of the models to some degree, which suggests that training for more epochs might still improve the results. We will continue increasing epochs, however, every experiment consumes a considerable amount of time.

Another essential aspect to note is the big difference between the results in SQuAD V2.0 and PolicyQA of every model. One first reason is the difference in the size of both datasets. Another reason is the difficult language that comes with legal domain text (Martinez-Gil, 2021).

From these results, our recommendations and, at the same time, our future work are the following. First, train the BERT model from scratch on only data related to the software development legal domain and use that model instead as the base encoder. Furthermore, since ALBERT obtained the best performance, we believe the best would be to train from scratch the ALBERT model directly. Finally, we recommend the use of ensemble methods which have proved to give significant improvements in Question Answering Systems [2].

6 Conclusions

Transformer-based language models have significantly advanced the field of NLP tasks, including Question Answering. This work studies the performance of several BERT model variants in the SQuAD V2.0 and PolicyQA datasets. The results showed LEGAL-BERT did not perform better than general pre-trained models like BERT and ALBERT. Finally, the ALBERT model achieved top results, which makes it a proper choice at the moment as a root and contextual embeddings encoder for a complex model design in the future. Furthermore, our work suggests that the best venue to follow is to train the ALBERT model from scratch on only data related to the software development legal domain and use that model instead as the base encoder.

[1] https://github.com/huggingface/transformers

[2] https://rajpurkar.github.io/SQuAD-explorer/

References

Wasi Uddin Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. Policyqa: A reading comprehension dataset for privacy policies. *arXiv preprint arXiv:2010.02557*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sanjay K Dwivedi and Vaishali Singh. 2013. Research and reviews in question answering system. *Procedia Technology*, 10:417–424.

Daniel Jurafsky and James H Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jorge Martinez-Gil. 2021. A survey on legal question answering systems. *arXiv preprint arXiv:2110.07333*.

Anselmo Peñas, Pamela Forner, Richard Sutcliffe, Álvaro Rodrigo, Corina Forăscu, Iñaki Alegria, Danilo Giampiccolo, Nicolas Moreau, and Petya Osenova. 2009. Overview of respubliqa 2009: Question answering evaluation over european legislation. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 174–196. Springer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jecqa: A legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9701–9708.