PDF ChatBot

Sanketh Venkataswamy
Purdue University Fort Wayne
venks06@pfw.edu

Abstract

Identify a reliable and efficient model capable of extracting entities from diverse PDF documents.Address challenges specific to PDFs, such as text layout variations, embedded tables, and lack of consistent formatting. And also see how the NER, Zero-Shot Classification, BERT based transformer models and LLM's perform on the question answering capabilities. And to come up with a best solution possible.

1 Introduction

As reviewing a recent textbook ,faced it really helpful. But sometimes we wished for something extra – like a chatbot. It could summarize each paragraph in simple terms and answer any questions I had about the book. Just imagine how much easier studying would be with that! It would make learning more interactive and fun. So, integrating a chatbot with this book could make studying a lot more interesting and easier to understand. With this tool, grasping complex concepts and staying engaged with the material would become effortless, ultimately enhancing the learning experience for students like me.

2 Related work

There are some Bert based chatbot pre-trained on specific dataset which can do question clssification and answering [2]. And in [4] developed a QA for the medical uses wwith CovidQA and CovidGQA datasets. But they had pre-trianed the data only the covid 19 data that are mostly a medical related onces.

3 Methodology

Explore existing benchmark datasets like SQuAD or CORD-19 for general NER training. Utilize domain-specific PDF datasets focused on finance, legal, or medical documents, depending on the target application.Augment existing datasets with synthetic PDF generation tools to increase data diversity and address specific document formats. For the bart based NER will be using CoNLL-2003 Named Entity Recognition[7] dataset. And for the zero-shot classification will be using [8] and for DeBERTa-v3-base-mnli-fever-anli model will be using [3] NER and Zero-shot classification models were used during the experimental process but later used the transformer model.

As the task in this paper is to answer the questions within the pdf the transformer models were tried making use of SQuAD (Stanford Question Answering Dataset): The SQuAD is a popular dataset for question answering tasks, where the model is tasked with answering questions based on a given context passage. The SQuAD is a collection of question-answer pairs derived from Wikipedia articles. In SQuAD, the correct answers of questions can be any sequence of tokens in the given text.[5] Because the questions and answers are produced by humans through crowdsourcing, it is more diverse than some other question-answering datasets. SQuAD 1.1 contains 107,785 question-answer pairs on 536 articles. SQuAD2.0 (open-domain SQuAD, SQuAD-Open), the latest version, combines the 100,000 questions in SQuAD1.1 with over 50,000 un-answerable questions written adversarially by crowdworkers in forms that are similar to the answerable ones. The distilbert/distilbert-base-cased-distilled-squad (Db cased), distilbert/distilbert-base-uncased-distilled-squad (Db uncased) and Intel/dynamic_tinybert (Dy tiny) were used. The distilbert-base-cased-distilled-squad model is a specific variant available in the Transformers library. DistilBERT aims to retain much of BERT's performance while being smaller and faster, making it more suitable for deployment in resource-constrained environments.[6] Like BERT, DistilBERT is pre-trained on large corpora of text using unsupervised learning techniques such as masked language modeling and next sentence prediction. The distilbert-base-cased-distilled-squad variant is fine-tuned specifically for the SQuAD task. Fine-tuning involves training the model on SQuAD data, which consists of context passages along with corresponding questions and answers. Compared to the original BERT model, DistilBERT is significantly smaller due to knowledge distillation techniques used during training. This makes it faster to train and deploy while still maintaining good performance on downstream tasks like question answering. The "cased" variant of DistilBERT retains the case information of the input text, meaning it distinguishes between uppercase and lowercase characters. This can be important for languages like English where case can change the meaning of words. The "uncased" variant of DistilBERT converts all input text to lowercase during tokenization. This can be useful for tasks where case information is not crucial, as it reduces the size of the vocabulary and can improve generalization.

The model was pre-trained with these three objective: Distillation loss: the model was trained to return the same probabilities as the BERT base model. Masked language modeling (MLM): this is part of the original training loss of the BERT base model. When taking a sentence, the model randomly masks 15% of the words in the input then run the entire masked sentence through the model and has to predict the masked words. This is different from traditional recurrent neural networks (RNNs) that usually see the words one after the other, or from autoregressive models like GPT which internally mask the future tokens. It allows the model to learn a bidirectional representation of the sentence. Cosine embedding loss: the model was also trained to generate hidden states as close as possible as the BERT base model.

Dynamic TinyBERT aims to provide a compact and efficient version of BERT while maintaining competitive performance on various natural language processing (NLP) tasks. Dynamic TinyBERT significantly reduces the size of the original BERT model by employing techniques such as knowledge distillation, network pruning, and quantization.

4 Experiments

After extracting the text from the PDF, first I examined different models required for NER and well as Zero-Shot Classification. The bert-based-NER model have f1 score as mentioned in Table 1.

metric	dev	test
f1	95.10	91.3
precision	95.00	90.7
recall	95.30	91.9

Table 1: Eval Results of the bert-base-NER.

As it is an NER model the model classifies into PER, LOC, MISC, and ORG which is person, location, miscellaneous, and organisation. Based on this available classification only certain questions can be answered such which target to these classification. Like we can ask for the author of the chapter or the author of the book. But the model can only classify the text into these classes, and doesn't answer the question. Tested the models on a test data which was a PDF file, the results of certain words were recorded in Table 2. The table 2 shows the some results by the NER classification and its f1 score. The words in table 2 are the person classification results.

words	bert-based-NER	bert-large-NER	distilbert-NER
"Daniel Jurafsky"	0.85	0.98	0.95
"##cognition"	0.49	0.705	-

Table 2: Scores (f1) of certain words with prediction class.

Since we need an application where the classifier classifies the text into multiple classes as in the real word requirements the text might not and will not always based on these four classifications. Could have used C-NER but this works well with structured and labeled data. Where as the Zero-Shot Classification, classifies text data into predefined categories without needing specific training data for those categories. Quickly classifying new types of data you might not have anticipated beforehand.This is used when labeled data is scarce or expensive to obtain. So if we want to quickly classify text into new categories without a lot of data, zero-shot classification might be a better option.[9]

As the zero-shot classification is more likely suitable for the PDF chat-bot the questions asked can be pre-processed, classified and given as parameters to the model. The Table 3 shows the scores of the words(parameter) during testing on bart-large-mnli model. Its a generative model and quite different approach for a chat bot. As we see in table 3 the question asked was "chapter name" and the results it generated were been recorded in table 3 with tits bleu score. As the bleu score is very less and sometimes it doesn't generate an accurate answers this generative method might not suit the chatbot. The PDF used for testing this model was taken from Chapter 16 of Speech and Language Processing (3rd ed. draft) by Dan Jurafsky and James H. Martin

Labels(parameter)	Score (BLEU)
Understanding spoken language	0.347
Speech Recognition Recognition Speech voice NLP	0.203
Speech voice NLP	0.362
Understanding Understanding Understanding	0.333

Table 3: Scores (BLEU) of the parameter in bart-large-mnli model.

For the next task transformer models were used and the whole text was given to the model and the model was then asked for the following question in the Table 4 and the following were its f1 score with which extracted the answer from the given text. The Table 4 shows some question and answer with different models with thier socres and also the human evaluation on these results generated. The model seems to extract some results with a high rate of f1 score but when an human evaluation is done the answer is wrong.

To address this issue later chunking of the text into 100 tokens each were made as a group and the question was looped for all the chunk text. This approach might even be called as brutal chunking, this chunking sometimes might help in the extraction but it chunks off the sentences making it vulnerable during extraction if the answer that we are looking might get chunked. Table 5 shows the performance of the model on specific questions.

5 Results

The following tables 4 and 5 show the results obtained from the bert models with their f1 score and human evaluation.

The github link for the paper [1]

Model name	**Question**	**Answer**	**f1 score**	**Human Evaluation**
Db cased	What is the name of the chapter?	16.8 Summary	0.98	Wrong Answer
Db cased	What is the chapter number?	Chapter 13	0.66	Wrong Answer
Dy tiny	What is the chapter name?	Automatic Speech Recognition and Text-to-Speech	0.59	Right Answer
Dy tiny	What is the chapter number?	16.8	0.82	Wrong Answer
Db uncased	What is the chapter name?	Automatic Speech Recognition and Text-to-Speech	0.93	Right Answer
Db uncased	What is the chapter number?	28	0.94	Wrong Answer

Table 4: Question answers with score (f1) with the whole text.

6 Conclusion

The NER model might perform better at classification and the Zero-Shot Classification can classify the text into the required classes but these might be helpful in text extraction based on the Chat Bot tasks, the use and role of an LLM such as bert based models perform slightly better than the NER or the Zero-shot approach. And as BERT-based text extraction continues to evolve, we can expect chat bots to play an increasingly vital role in various applications, from customer service to education and beyond.

References

[1] https://github.com/sankethvenkataswamy/natural-language-processing-project.

[2] Dimitra Anastasiou. ENRICH4ALL: A first Luxembourgish BERT model for a multilingual chatbot.

Model name	Question	Answer	f1 score	Human Evaluation
Db cased	What is the name of the chapter?	Speech and Language Processing	0.93	Might be right but not exactly right
Db uncased	What is the chapter name?	Automatic Speech Recognition and Text-to-Speech	0.93	Right Answer
Db uncased	What is the chapter about?	Speech and Language Processing	0.37	Right Answer
Db uncased	What is LibriSpeech?	-	-	No Answer
Dy tiny	What is LibriSpeech?	Right reserved	0.23	Wrong Answer

Table 5: Question answers with score (f1) with chunked text.

In Maite Melero, Sakriani Sakti, and Claudia Soria, editors, *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 207–212, Marseille, France, June 2022. European Language Resources Association.

[3] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

[4] Hillary Ngai, Yoona Park, John Chen, and Mahboobeh Parsapoor. Transformer-based models for question answering on covid19, 2021.

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

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[7] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, Edmonton, Canada, 2003. Association for Computational Linguistics.

[8] Adina Williams, Nikita Nangia, and Samuel 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)*, 2018.

[9] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, 2019.