Ensemble ALBERT and RoBERTa for Span Prediction in Question Answering

Sony Bachina, Spandana Balumuri and Sowmya Kamath S
Healthcare Analytics and Language Engineering (HALE) Lab,
Department of Information Technology,
National Institute of Technology Karnataka, Surathkal, Mangalore 575025, India
{bachina.sony, spandanabalumuri99}@gmail.com
sowmyakamath@nitk.edu.in

Abstract

Retrieving relevant answers from heterogeneous data formats, for given for questions, is a challenging problem. The process of pinpointing relevant information suitable to answer a question is further compounded in large document collections containing documents of substantial length. This paper presents the models designed as part of our submission to the DialDoc21 Shared Task (Document-grounded Dialogue and Conversational Question Answering) for span prediction in question answering. The proposed models leverage the superior predictive power of pretrained transformer models like RoBERTa, ALBERT and ELECTRA, to identify the most relevant information in an associated passage for the next agent turn. To further enhance the performance, the models were fine-tuned on different span selection based question answering datasets like SQuAD2.0 and Natural Questions (NQ) corpus. We also explored ensemble techniques for combining multiple models to achieve enhanced performance for the task. Our team SB_NITK ranked 6th on the leaderboard for the Knowledge Identification task, and our best ensemble model achieved an Exact score of 58.58 and an F1 score of 73.39.

1 Introduction

In recent years, deep learning based transformer models like BERT have accelerated research in the Natural Language Processing (NLP) domain, due to their outstanding performance in various NLP tasks like summarization, machine translation etc, against state-of-the-art models. Question-answering is one such text based Information Retrieval framework, focusing on generating relevant answers to natural language questions presented by humans. Extractive Question Answering models leverage document context to make decisions while identifying the most relevant answer and its location in a given passage or document. The applications of question answering systems include chat bots in medical science, search engines, personal assistants etc.

Several researchers have addressed the problem of answer generation for a given question, especially focusing on the challenge of dealing with descriptive answers. Some works deal with this challenge in a two-phased approach - first, classifying the question into opinion-based or yes/no questions and secondly, dealing with the issue of lengthy questions and generating relevant answers for them using deep neural models like LSTMs for the question answering task Upadhya et al. (2019). Agrawal et al. (2019) proposed a Question Answering model built on BiLSTMs pre-trained on the SquAD dataset, to obtain appropriate ranks for answers corresponding to a given question at hand. Additionally, ensemble techniques have proven well-suited due to better prediction performance, while reducing the variance and bias. Adopting ensemble techniques to combine multiple models can provide better predictions and boost the performance of Question Answering systems. As shown in Fig. 1, pretrained encoders such as BERT (Devlin et al., 2019) with an additional linear layer on top to predict spans have been shown to provide the advantage of transfer learning as they are pretrained on large, open datasets.

The DialDoc21 shared task aims to encourage the development of models that can detect the most relevant details in the grounding document and predict agent responses close to common human responses. DialDoc21 is composed of two different shared tasks –

- *Subtask1 - Knowledge Identification* : The aim of the task is to find where the answer is present (a text span) in the document context for the next agent turn. F1 metrics and Exact Match are the evaluation metrics for Subtask1.

Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 63–68
August 5–6, 2021. ©2021 Association for Computational Linguistics

- *Subtask2 - Text Generation* : The task aims to generate responses close to human spoken language. The assessment metrics for Subtask2 are sacrebleu and individual evaluations.

Most recent Extractive Question Answering systems are predominantly BERT based models. Transformers like BERT, RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2020) are experimented for Extractive Question Answering task on datasets from multiple domains and languages like Stanford SQuAD v1.1 (Rajpurkar et al., 2016) and v2.0 (Rajpurkar et al., 2018), Natural Questions (Kwiatkowski et al., 2019), NewsQA (Trischler et al., 2017) and HotpotQA (Yang et al., 2018). Dua et al. (2019) experimented on the scenario of multiple answer spans.

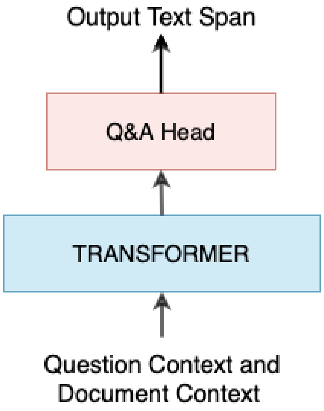

Figure 1: Base architecture for span prediction task using Transformers

In this paper, we describe various models and experiments that were developed and tested for the Knowledge Identification subtask. The models are built on fine-tuned, pretrained transformers like RoBERTa, ALBERT (Lan et al., 2020) and ELECTRA (Clark et al., 2020). We adopt ensembling techniques on multiple models to boost the prediction performance further. As part of our approach, we pretrained the transformer models considered on various question-answering datasets like SQuAD2.0, Natural Questions corpus, and CORD-19 dataset (Wang et al., 2020) prior to fine-tuning them for our dataset, and observe their performance.

The rest of this article is organized as follows. In Section 2, we provide information about the data used such as description of dataset and dataset pre-processing. Section 3 gives an overview of models used and their different versions. In Section 4, we describe the ensemble technique used and the various ensemble models submitted. Section 5 describes the system flow and experimental setup. In Section 6, we list the results and compare the performance of our proposed models in detail, followed by conclusion and directions for future work.

2 Data

2.1 Dataset Description

The organizers of the shared task provided the necessary training and testing data. The training data is taken from Doc2Dial dataset (Feng et al., 2020) which includes dialogues between an agent and an enduser, along with their base information in the associated documents provided. These documents were collected from different social websites such as `ssa.gov`, `va.gov`, `studentaid.gov` and DMV portal. The test dataset includes an unseen COVID-19 related domain's data (cdccov19), in addition to other domains that are available in the training dataset. Therefore, the unseen domain helps in testing the model performance on an unknown domain data.

2.2 Dataset Preprocessing

The average sequence length of the grounding document is 880 which is higher than the maximum sequence length of transformers. As a result the document text has been truncated into sliding windows with a stride value of 128. Each input sample to the encoder includes dialogue context, which is a combination of all previous utterances in reverse order and the corresponding document trunk. An example (a combination of a question and a document) can have multiple features (a pair of a question and a document trunk) in case of a lengthy context. Therefore, we have a map that links each feature to its associated example. Additionally, an offset map is maintained from each token to the position of the character in the actual context.

3 Models

For the DialDoc21 knowledge identification shared task, we experimented with various versions of three different transformer models by fine-tuning them on the Doc2Dial dataset. The details of these implementations are discussed in detail in subsequent sections.

3.1 RoBERTa

Facebook's RoBERTa (Robustly optimized BERT-pretraining approach) transformer model considers previously unexplored architecture options in BERT pre-training. To boost the training process, RoBERTa adopts dynamic masking approach. We experimented with the three different RoBERTa variants as listed below.

1. *roberta-large-squadv2* : RoBERTa large fine-tuned on SQuADv2.0 dataset.
2. *roberta-base-squadv2-nq* : RoBERTa base fine-tuned on NQ and SQuADv2.0 datasets.
3. *roberta-base-squadv2-covid* : RoBERTa base fine-tuned on SQuADv2.0 and CORD-19 datasets.

3.2 ALBERT

The key goal of Google's ALBERT(A Lite BERT) is to minimise the number of parameters in BERT (340M parameters) as training large models like BERT is computationally expensive and time-consuming. ALBERT implements 2 different techniques like factorization of the embedding parameterization and distributing all of its parameters across layers in order to decrease the number of parameters used in training. In our work, three ALBERT models were considered for the analysis.

1. *albert-base-squadv2* : ALBERT base fine-tuned on SQuADv2.0
2. *albert-xlarge-squadv2* : ALBERT xlarge fine-tuned on SQuADv2.0 dataset
3. *albert-xxlarge-squadv2* : ALBERT xxlarge fine-tuned on SQuADv2.0

3.3 ELECTRA

ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) uses replaced token detection (RTD), which trains a bidirectional model such as an MLM while also learning from the input positions similar to an LM (Language Model). We considered two variants of ELECTRA for the benchmarking experiments.

1. *electra-base-squadv2* : electra-base fine-tuned on SQuAD 2.0 dataset.
2. *electra-large-squadv2* : electra-large language model fine-tuned on SQuAD2.0 dataset.

4 Ensemble Models

To further enhance the performance of the proposed models for the Knowledge Identification task, we employed ensembling techniques for leverage the predictive power of all the transformer models considered for the experiments. Various combinations of the models were designed and experimented with, in order to improve the predictions. The confidence score from different models is used as a measure to combine models. The predictions of the model with maximum confidence score is treated as best prediction and the same is considered for evaluation. Initially, various models from same groups of RoBERTa, ALBERT and ELECTRA are ensembled together, and based on the predictions on the validation set, few other models are added. The following are the different combinations of models submitted for testset evaluation.

1. *roberta-ensemble* : roberta-large-squadv2 + roberta-base-squadv2-nq + roberta-base-squadv2-covid
2. *albert-ensemble* : albert-base-squadv2 + albert-xlarge-squadv2 + albert-xxlarge-squadv2
3. *electra-ensemble* : electra-base-squadv2 + electra-large-squadv2
4. *Ensemble1* : alberta-ensemble + roberta-base-squadv2-covid
5. *Ensemble2* : alberta-ensemble + roberta-base-squadv2-nq + roberta-base-squadv2-covid

The performance of all the proposed models was measured using standard metrics, for both the validation and test set, the details of which are presented in Section 6. We also employed metrics like Exact Match and F1 score for individual pretrained models, and also *alberta-ensemble* with *roberta-base* pretrained on the *nq* and *covid* datasets.

5 Model Fine-tuning

In order to improve Exact Match and F1 scores, fine-tuning and ensemble techniques were considered. The dataset provided for the test-dev phase of the shared task is considered as validation set, which shares the same grounding document as training dataset. The test dataset is provided during the test phase of the task and contains 787 end-user questions.

During the training phase, the document trunk along with the corresponding dialogue context was input to the encoder model. The output obtained from the linear layer is a tuple representing the probabilities of the position being the start and end of the corresponding span. In case the ground truth span is not inside the considered trunk, the *begin* and *end* positions are taken as the start of the sequence. In the decoding phase, the probability tuples of all the trunks are considered by the model to obtain the optimum span.

We also conducted experiments by varying hyperparameters such as maximum sequence length (384, 512), batch size (4, 8, 16) and learning rate(3e-5, 1e-4) to select best performance. Each model has been trained for 5 epochs and during training, checkpoints were generated for every 2000 steps. Loss reduction was examined at every checkpoint to pick the best optimal checkpoint that could potentially generate optimal predictions for each model. All experiments were performed on NVIDIA V100 GPUs with 32GB RAM. In Section 6, the details of experiments conducted and the observed performance are described.

6 Experimental Results and Discussion

In this section, we discuss various versions of models submitted for consideration in the final phase on the leaderboard. Table 1 presents the observed results for the proposed transformer models for the metrics Exact Match and F1-score. It can be observed that, among the various RoBERTa models, fine-tuning *roberta-base-squadv2-nq* on Doc2Dial achieved the best performance, which proves that increasing the scope of data gives better results. From Table 1, it can seen that amongst the ALBERT models, fine-tuning *albert-xlarge-squadv2* resulted in improvements in performance when compared to those of *albert-base-squadv2*.

Table 2 tabulated the results of benchmarking of proposed ensemble models on test dataset. It is evindent that, combining different models through maximum confidence score ensembling technique helped in achieving increased performance when compared to the performance of individual models (Refer Table 1). The *albert-ensemble* model was the best-performing model among ensemble models belonging to same group such as *roberta-ensemble*, *albert-ensemble* and *electra-ensemble*. Therefore, we decided to combine cross-group models with *albert-ensemble* to improve predictions.

In case of *Ensemble1* and *Ensemble2*, applying ensembling techniques on best performing RoBERTa models like *roberta-base-squadv2-nq* and *roberta-base-squadv2-covid* with *albert-ensemble* resulted in further improvements as is

Table 1: Exact Match and F1 Scores of different pretrained models on validation set

Model	Exact Match	F1 Score
roberta-large-squadv2	52.02	67.57
roberta-base-squadv2-nq	**55.56**	**69.36**
roberta-base-squadv2-covid	54.54	68.09
albert-base-squadv2	44.44	59.01
albert-xlarge-squadv2	50.00	63.69
electra-base-squadv2	46.46	62.37

Table 2: Exact Match and F1 Scores of proposed ensemble models

Ensemble Model	Validation Set		Test Set	
	Exact Match	F1 Score	Exact Match	F1 Score
roberta-ensemble	56.56	69.91	54.76	70.17
electra-ensemble	53.53	69.47	47.65	65.14
Ensemble1	59.60	73.27	57.94	73.11
Ensemble2	**61.62**	**74.48**	**58.58**	**73.39**

Table 3: Sample question context generated by series of user and agent turns: *"user: what should I do if i go out in public? agent: Call 911 right away user: What if symptoms worsen? agent: you are at higher risk for more serious COVID-19 illness It is very important for you to take steps to stay healthy .s user: what if you are If you are an older adult or someone who has severe chronic medical conditions such as heart or lung disease , or diabetes agent: If you don t have soap and water , use an alcohol - based hand sanitizer with at least 60 % alcohol user: What if i do not have access to soap and water?"*

Model	**Predicted Text Span**
roberta-base-squadv2-nq	keep away from others who are sick, limit close contact, and wash your hands often. Consider steps you can take to stay away from other people. This is especially important for people who are at higher risk of getting very sick.
roberta-ensemble	keep away from others who are sick, limit close contact, and wash your hands often.
electra-ensemble	Avoid crowds as much as possible When you go out in public, keep away from others who are sick, limit close contact, and wash your hands often.
Ensemble1	keep away from others who are sick, limit close contact, and wash your hands often.
Ensemble2	keep away from others who are sick, limit close contact, and wash your hands often.

evident from the tabulated values shown in Table 2. The ensemble models performed well on both validation and test datasets, thus, underscoring the consistent performance of proposed models on different datasets.

Prediction metrics on the test set also emphasize the effectiveness of the proposed models on even completely new and unknown domain data. Table 3 shows predicted text span for a sample question context for different ensemble models. Amongst them, *Ensemble2* gave the optimal span followed by *Ensemble1*. The *Ensemble2* model gives the best scores on both validation and test datasets with an Exact Match score of 58.58 and F1 Score of 73.39 on test dataset which show significant increase over the baseline (Feng et al., 2020) 55.4 Exact Match score and an F1 score of 65.0 respectively.

7 Conclusion and Future Work

In this paper, the application of transfer learning by utilizing transformer models like RoBERTa, ALBERT and ELECTRA for Question Answering Span Prediction task was explored. We also experimented with pretrained models on several other datasets prior to fine-tuning it on the DialDoc21 dataset, provided as part of the DialDoc21 Shared task. Maximum confidence score based ensemble techniques were employed to combine various base transformer models to further boost the performance. We plan to extend our approach and experiment with other ensembling techniques for further enhancing the performance and also explore avenues for improved scalability when applied to larger datasets.

References

Anumeha Agrawal, Rosa Anil George, Selvan Suntiha Ravi, Sowmya Kamath, and Anand Kumar. 2019. Ars_nitk at mediqa 2019: Analysing various methods for natural language inference, recognising question entailment and medical question answering system. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 533–540.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A. Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

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.

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.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Akshay Upadhya, Swastik Udupa, and S Sowmya Kamath. 2019. Deep neural network models for question classification in community question-answering forums. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William Merrill, P. Mooney, D. Murdick, Devvret Rishi, J. Sheehan, Zhihong Shen, Brandon Brandon Stilson Stilson, Alex D Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas A. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.