Dialectal Extractive Question Answering (DialQA)

Adel Alkhamisy
Department of Computer Science
George Mason University
aalkhami@gmu.edu

Aniket Pandey
Department of Computer Science
George Mason University
apandey7@gmu.edu

Pankaj Jatav
Department of Computer Science
George Mason University
pjatav@gmu.edu

Yash Bagal
Department of Computer Science
George Mason University
ybagal@gmu.edu

1 Introduction

1.1 Task / Research Question Description

A natural language Question Answering system takes questions in natural language and outputs answer. In this system, there is interaction between the system and the user. This assists in establishing a connection for finding accurate results more quickly. The extractive approach-based QA system comprises of a reader and retriever. Instead of storing the answers to the questions in database, the system loads the document related to the user's question and then the reader extracts the answer. The current QA systems do not consider the faults that might be introduced by speech recognition models and do not take the dialect in to consideration. Through this work, our aim is to make a QA system which understands various languages and then provide a result. In order to solve this issue we will implement a extractive QA system. The Dialectal Extractive Question Answering systems will be built on the existing QA benchmarks TyDi QA (Clark et al., 2020) and SD QA (Faisal et al., 2021). We will make use of parts of the SD QA dataset, with recorded dialectal variations of TyDi QA.

1.2 Motivation and Limitations of existing work

In recent times QA systems are a rapidly growing field of study and the main motivation behind them is to answer the human question prompts which not only support the users to find answers but also aid users with visual and motor impairments to interact with devices.

The majority of evaluation benchmarks of existing QA systems rely on text-based which are noise-free. However, the real word usage of such systems involves input with noise. Faisal et al. proposes a QA system based on the SD-QA dataset that aims to mitigate the effects of multi-dialect of users on the QA system. SD-QA covers five languages and twenty-four dialects. However, the user accent is different even among users who speak the same dialect and this challenge is one of the limitations of this study. Another gap in this study is the difference between reading speech and spontaneous speech (Faisal et al., 2021). Also, in (Batliner et al., 1995) has recorded readings of text questions that have different characteristics of spontaneous speech. In order to enable researchers to create multilingual models that are effective across several languages Clark et al. proposed large-scale multilingual corpora. However, as the content is not written in various languages, cross-language answer retrieval and translation are required. Another limitation is the error introduced by automatic speech recognition (ASR) which is experimented in (Sidiropoulos et al., 2022).

1.3 Proposed Approach

Research work performed in DialQA (Faisal et al., 2021) implements a minimal BERT model which is also used by the TyDi QA (Clark et al., 2020). We will start with implementing the minimal BERT model Devlin et al. (2019) and match our result with the baseline result.

In addition to minimal BERT, to improve the baseline we will either try BERT with Language-Clustered vocabulary as proposed in (Chung et al., 2020) or Poolingformer (Zhang et al., 2021), Long Document Modeling with Pooling Attention. According to the experiments performed in (Chung et al., 2020) show improvements across languages on key multilingual benchmark tasks TyDi QA (+2.9 F1) and (Zhang et al., 2021) outperforms previous state-of-the-art model by 1.9 points (79.5 vs. 77.6) on TyDi QA passage answer, and 1.6 points (67.6 vs. 66.0) on TyDi QA minimal answer. The choice of method will depend upon the timeline and result obtained.

1.4 Likely challenges and mitigations

The main challenge while working on this problem statement is taking into consideration the users with different dialects and grouping them together. The user accent is different even among users who speak the same dialect. So, designing a system which is robust and can handle different variations of dialects is in itself challenging. Our goal is to first reach the baseline and then work on it to improvise. If things go side way we may still have a lot of learning and understanding of the work that we could later work on. We believe every experiment is in itself a success. That being said if the experiment doesn't go as planned we will try to reach as close as possible to the baseline model implementation.

2 Related Work

In recent times many related work have been done to take us from text based QA to voice based QA machines. The latest work is done by Faisal et al. in which they discuss about errors introduced by speech recognition models and language variations (dialects) of the users. Through their experiment they try to improve the existing TyDi QA (Clark et al., 2020) dataset to form a multi-dialect, spoken QA on five different languages. TyDi QA is a question answering dataset covering 11 typologically diverse languages with 204K question-answer pairs. With regard to typology it is diverse and contains language phenomena that would not be found in English-only corpus.

In another work by Ravichander et al. in 2021 talks about the introduction of noise and error in QA system by keyboard input, machine translation and speech input. The paper discusses about tests performed which shows that missing punctuation degrade the result by 5.1%. Other factor such as accent is also taken into consideration. There is another similar work in (Peskov et al., 2019) which studies mitigating ASR errors in QA, assuming white-box access to the ASR systems.

Furthermore, in another research Spoken SQuAD (Lee et al., 2018b), the QA is done in text form and comprehensive reading in speech form. The speech data is first converted to text and the output could be either text or audio. This was further modified and augmented in ODSQA (Lee et al., 2018a) where the QA was also given in audio form. In these experiments the ASR errors significantly degraded the performance of reading comprehensive models and then they proposed the use of different kinds of subword units to mitigate the impact of ASR errors.

3 Experiments

3.1 Datasets

The Dialectal Extractive Question Answering Dataset is available on GitHub. See https://github.com/ffaisal93/SD-QA for instructions. The SD-QA (Faisal et al., 2021) dataset extends the TyDi QA (Clark et al., 2020) dataset by incorporating questions, contexts, and responses from five typologically distinct languages.

The dataset has following languages and varieties.

Language	Locations (Variety Code)
Arabic	Algeria (DZA), Bahrain (BHR), Egypt (EGY), Jordan (JOR), Morocco (MAR), Saudi Arabia (SAU), Tunisia (TUN)
Bengali	Bangladesh-Dhaka (BGD), India-Kolkata (IND)
English	Australia (AUS), India-South (IND-S), India-North (IND-N), Ireland (IRL), Kenya (KEN), New Zealand (NZL), Nigeria (NGA), Philippines (PHI), Scotland (SCO), South Africa (ZAF), US-Southeast (USA-SE)
Korean	South Korea-Seoul (KOR-C), South Korea-south (KOR-SE)
Kiswahili	Kenya (KEN), Tanzania (TZA)

Table 1: Languages and sample collection locations in SD-QA dataset.

3.2 Baselines

As a baseline we will compare our model to DialQA (Faisal et al., 2021) and will also use their open-source code available at https://github.com/ffaisal93/DialQA.

3.3 Timeline

We will implement minimal BERT as state of the art baseline model. We will utilize the productivity of each of our team member and all work will be done in collaboration. Additionally, a GitHub

Week	Goal
Week 1	Understanding the dataset and segregating the Data based on language and it's variations.
Week 2	Start with the minimal BERT implementation.
Week 3	Continuing with BERT implementation.
Week 4	Testing our BERT implementation on Hopper and tuning the hyperparameters.

Table 2: Timeline.

repository for the same will be created and step by step progress can be tracked using the git commits. To complete the work successfully we will also employ the practice of pair-programming and every week distribute work among ourselves and collaborate among us if a member needs help.

References

A. Batliner, R. Kompe, A. Kießling, H. Niemann, E. Nöth, A.J.R. Ayuso, and J.M.L. Soler. 1995. *Can You Tell Apart Spontaneous and Read Speech If You Just Look at Prosody?* NATO ASI Series. Universität Augsburg.

Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. 2020. Improving multilingual models with language-clustered vocabularies.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018a. Odsqa: Open-domain spoken question answering dataset.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018b. Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension. In *Proc. Interspeech 2018*, pages 3459–3463.

Denis Peskov, Joe Barrow, Pedro Rodriguez, Graham Neubig, and Jordan Boyd-Graber. 2019. Mitigating noisy inputs for question answering.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. Noiseqa: Challenge set evaluation for user-centric question answering.

Georgios Sidiropoulos, Svitlana Vakulenko, and Evangelos Kanoulas. 2022. On the impact of speech recognition errors in passage retrieval for spoken question answering.

Hang Zhang, Yeyun Gong, Yelong Shen, Weisheng Li, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Poolingformer: Long document modeling with pooling attention.