Advanced NLP: Major Project
WikiQA-based Open Domain Question Answering
Scope Document*

Sagar Joshi (2020701007) Abhishek Tyagi (20173067)
Gadela Kesav (2018101079)
Team 33: ChatBot Mentor: Priyanka Ravva

November 16, 2021

1 About Open Domain Question Answering

Open-domain question answering is the task of answering questions in natural language without constraints of any specific domain or context. The larger task can involve, depending on the pipeline used, interactions with open-domain knowledge bases such as WikiData[1] and answering questions on general topics such as those found on Wikipedia, and has special implications in advancing conversational agents (such as Alexa[2]). In this project, making use of the WikiQA dataset[6], we will focus on the very specific tasks of answer selection and answer triggering. In the process, we will implement some of the systems giving a solid performance on the dataset.

2 Project Scope

Based on our exploration so far, and consultation with the mentor, the scope of our project will consist of the following deliverables:

- Explore the WikiQA dataset, the baselines, and some of the current benchmarks implemented on the same. Also, in getting familiarity with the literature in this area, know the common metrics used for evaluating the systems using this dataset. Our study for this deliverable has been presented in the remainder of this document.
- Implement one formidable baseline on the WikiQA dataset. For this, we plan to implement the attentive pooling networks described in section 5.1, an improvement over the best performing models used in the WikiQA paper which used just convolutional features.
- The TANDA pretraining approach described in section 5.2 will be our final model, which is the current best performing model on the WikiQA benchmark[3]. We will initially target finetuning base versions of models for ease of computing, if time permits and if it seems computationally feasible, will also try with the large model versions.
- *Optional* additional task: Following the methodology adopted on the WikiQA dataset, the models will also be trained on the TREC-QA dataset[5], which is another popular, but older dataset in answer selection. This will be subject to the availability of time after completion of the necessary components in the above deliverables.

3 WikiQA Dataset[6]

3.1 Dataset Summary

The WikiQA corpus is a publicly available dataset of questions and sentence sets, collected and annotated for performing research on open-domain question answering. To mimic real-life settings, Bing search query logs were used as the source of questions. Each question is then linked to a Wikipedia page that potentially contains the answer. Because the summary section of a Wikipedia page provides the basic and usually most important information about the topic, the authors used all the sentences in this section as the candidate answers. Using crowdsourcing, 3,047 questions and 29,258 sentences were included in the dataset, where only a small subset of 1,473 sentences was labeled as the correct answer sentences to their corresponding questions. So, approximately only 1/3rd of the

*Presentation link
[1] `https://www.wikidata.org/wiki/Wikidata:Main_Page`
[2] `https://developer.amazon.com/en-GB/alexa`
[3] `https://paperswithcode.com/sota/question-answering-on-wikiqa`

total questions have a correct answer in the candidate answer set. In addition, 20.3% of the answers in the dataset share no content words with questions. Table 1 shows some of the statistics of the dataset.

3.2 Tasks

1. Answer triggering: Detect whether there exists at least one correct answer in the set of candidate sentences for the question.

2. Answer selection: If yes, select one or more sentences as the correct answer from the candidate set.

4 Metrics for Evaluation

The MAP and MRR metrics have been traditionally used in literature for evaluating the performance of the QA systems on answer selection. Since the WikiQA dataset introduces a new task of answer triggering, which is a classification task, classification metrics are necessary for evaluating this part.

4.1 Mean Average Precision (MAP)

This metric summarizes the precision-recall curve of the system, taking the mean of the precision of the results achieved at each threshold for answer selection. It is a popular metric used in information retrieval for search result ranking evaluation.

4.2 Mean Reciprocal Rank (MRR)

It is the average of the reciprocals of the rank given to the correct answer by the system. If the sentence containing the answer is ranked first, a full score will be rewarded for the performance at that instance, else the score awarded by the metric will decrease inversely with the assignment of lower rank to the correct answer. For a set of Q questions, with each instance i having its candidate sentence ranked at position $rank_i$, the metric can be formulated as:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{1}$$

4.3 Precision, recall & F1

In particular, these metrics are computed by aggregating all the candidate sentences of a question. A question is treated as a positive case only if it contains one or more correct answer sentences in its candidate sentence pool.

5 Approaches Studied for Implementation

5.1 Attentive Pooling Networks[1]

- Attentive Pooling (AP): A two-way attention mechanism for discriminative model training. In the context of pair-wise ranking or classification with neural networks, AP enables the pooling layer to be aware of the current input pair, in a way that information from the two input items can directly influence the computation of each other's representations.

- Along with the above representations of the paired inputs, AP jointly learns a similarity measure over projected segments (e.g. trigrams) of the pair, and subsequently, derives the corresponding attention vector for each input to guide the pooling.

- The two-way attention mechanism is independent of the underlying representation learning and is applicable to both CNNs and RNNs.

5.2 TANDA[2]

- TANDA is a finetuning strategy for answer selection task, consisting of two stages: first 'transfer' the model on the target domain of the task and then 'adapt' on the specific dataset at hand on the downstream task.

- This was designed especially considering the unavailability of large task-specific datasets in question answering, which would otherwise result in regular finetuning of transformers producing an on-off behavior.

- For the transfer component of the task, a new dataset - Answer-Sentence Natural Questions (ASNQ) - was created for the answer selection task based on the Google Natural Questions (NQ) dataset[3]. It consists of 57,242 and 2,672 distinct questions in the train and dev set respectively.

- Among other analyses, results showing the remarkable stability of TANDA models over different training epochs and robustness in performance on injection of up to 10% and 20% noise in the training data were shown. The metrics of MAP and MRR were used for the same.

- Overall, models finetuned with the TANDA strategy achieve a new state-of-the-art in the answer selection task, with RoBERTa-Large[4] achieving values of 92.0% and 93.3% in MAP and MRR respectively on the WikiQA dataset.

	Train	Dev	Test	Total
# questions	2,118	296	633	3,047
# sentences	20,360	2,733	6,165	29,258
# answers	1,040	140	293	1,473
Average question length	7.16	7.23	7.26	7.18
Average sentence length	25.29	24.59	24.95	25.15
# questions w/o answer	1,245	170	390	1,805

Table 1: Statistics of the WikiQA dataset

References

[1] Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. Attentive pooling networks, 2016.

[2] Siddhant Garg, Thuy Vu, and Alessandro Moschitti. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. *ArXiv*, abs/1911.04118, 2020.

[3] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, March 2019.

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

[5] Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. What is the Jeopardy model? a quasi-synchronous grammar for QA. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 22–32, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[6] Yi Yang, Wen-tau Yih, and Christopher Meek. 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, September 2015. Association for Computational Linguistics.