End-to-end text processing: CoQA - A Conversational Question Answering Challenge

Dik Medveščcek Murovec
Faculty of computer science and informatics
University of Ljubljana
dm3825@student.uni-lj.si

Matija Kljun
Faculty of computer science and informatics
University of Ljubljana
mk5282@student.uni-lj.si

***Abstract*—For humans exchanging information is based on conversations involving a sequence of related questions and answers. For machines to be able to do the same it is essential for them to be able to answer conversational questions. The CoQA dataset [1] for building Conversational Question Answering systems contains 127k questions with answers, obtained from 8k conversations about text passages from seven diverse domains. With this dataset we are able to train a model that will be able to answer conversational questions. Human performance on the task gives a F1 score of 88.8%, while the best score from the article that is based on the combination of the DrQA and PGNet model [1] gives a score of 65.1%. Since the lunch of the CoQA challenge many improved models have been built, currently the best score on the leaderboard is 80.2% (D-AoA + BERT (single model)). In our project we will research the current solutions and try build our own model based on the knowledge gathered. We will present our solution and results.**

***Index Terms*—NLP, end-to-end, text processing, CoQA**

I. Introduction

A series of questions and answers is normal in any human conversation. It is a way to learn build up knowledge. Based on questions and their answers we gain a deeper insight into the subject at hand.

This is simple for humans, but hard for computers.

CoQA is a large-scale dataset for building Conversational Question Answering systems. The goal of the CoQA challenge is to measure the ability of machines to understand a text passage and answer a series of interconnected questions that appear in a conversation.

Characteristics that CoQA systems should embody:

- Understanding questions and providing answers based on inherent knowledge of a system
- Upgrading knowledge based on questions and their respective answers
- Ensuring the naturalness of answers provided
- Ensuring robustness across different knowledge domains.

The formal definition of the task at hand is as follows:
Given a passage p, the conversation history $q_1, a_1, ...q_{i-1}, a_{i-1}$ and a question q_i, the task is to predict the answer a_i [1]

II. Related work

Several approaches have been attempted to solve the CoQA challenge. We will introduce three.

The current best solution to the CoQA Challenge uses a combined model as described below [2].

Conversational model

Sequence-to-sequence (seq2seq) models have shown promising results for generating conversational responses. We append the passage, the conversation history and the current question as p <q> q_{in} <a> a_{in}... <q> q_{i1} <a> a_{i1} <q> q_i, and feed it into a bidirectional LSTM encoder, where n is the size of the history to be used. We generate the answer using a LSTM decoder which attends to the encoder states [1].

Reading comprehension model

The state-of-the-art reading comprehension models for extractive question answering focus on finding a passage span that matches the question best [1].

Combined model

In the combined model, we use the comprehention model to first point to the answer eidence in the text and seq2seq to normalize the text in the form of an answer [1].

III. Methods

We will try to model the task from two possible perspectives, as a conversational response generation problem and as a reading comprehension problem. The first approach is based on sequence-to-sequence models with LSTM autoencoder. The second approach based on a reading comprehension model. We will then try to combine both models and try to achieve the same results as in the paper [1].

We will then try to further improve our results by researching similiar solutions [3] in this field and try to implementing these tecniques in our model. A possible improvement to the combined model will be the use of inter-attention and self-attention to comprehend conversation context and extract relevant information from passage [2].

Baselines

CoQA baselines represent several different models posed by the challenge creators we try to beat.

To better understand the subject of CoQA we downloaded them from GitHub and followed the steps covered in their instructions. They provided us with insight into preprocessing of the relevant data and training three different models. Conversational models, reading comprehension models and the pipeline models.

We ran the models and trained them using $n_history = 0$. The results ended up a bit worse, but were comparable to the baselines. Therefore we are taking the numbers presented in the original paper as baselines.

A. BiDAF++

To improve on the baselines we first looked at BiDAF++. A system for granulating text and returning query answers. Due to it's modularity it works for basically any text input.

It's composed of 6 layers:

- Character level: Mapping text to characters.
- Word level: Mapping text to words.
- Contextual embeddings: Mapping text to context.
- Bi-directional attention flow: Obtains a query-aware context representation
- Modelling: Takes query and returns relevant information
- Output: Application-specific layer for modelling answers according to the application requirements

BiDAF is presented with a different dataset and different outputs. We tried to modify the model to take CoQA text as input and the output layer to return answers in the CoQA format.

We failed to do so. We introduce FlowQA.

FlowQA

In order to improve the coqa baseline, we used the FlowQA model that uses the flow mechanism that can incorporate intermediate representations generated during the process of answering previous questions, through an alternating parallel processing structure. In the research paper the FlowQA models gives an overall score of 75% F1. That is a 7.2% improvment over the combined CoQA baseline model. We trained the model for just one epoch since the training time was extremely long (more than one day) on our machine. The trained model therefore give the F1 score of just 62%.

The FlowQA model is a model designed for conversational machine comprehension. It consists of two main components: a base neural model for single-turn machine comprehension and a FLOW mechanism that encodes the conversation history. The model is fed with the entire hidden representations generated during the process of answering previous questions, these hidden representations captures additional related information for answering the previous questions and help to detect what the current conversation is revolving around.

The FLOW mechanism can be viewed as stacking single-turn QA models along the dialog progression (i.e., the question turns) and building information flow along the dialog. This information transfer happens for each context word, allowing rich information in the reasoning process to flow. This design is analogous to recurrent neural networks, where each single update unit is now an entire question answering process. Because there are two recurrent structures in our modeling, one in the context for each question and the other in the conversation progression, a naive implementation leads to a highly unparallelizable structure. To handle this issue, we propose an alternating parallel processing structure, which alternates between sequentially processing one dimension in parallel of the other dimension, and thus speeds up training significantly [4].

IV. Results

Using Baselines, BiDAF and FlowQA we came to a few different results. In table I we show the best F1 results.

Model	Best F1
seq2seq	20.9
seq2seq copy	45.2
DrQA	55.6
Pipeline	65.0
BiDAF++	/
FlowQA	68.8

TABLE I
Best F1 across all models.

We trained to FlowQA models, since the training time was very long (one day per epoch) we managed to train the first model for just 1 epoch and the second model for two epochs. The results are displayed in the table 2. We can see that even with just one epoch training the results are better than the separate baseline models and in two epochs the result of the second model is better than the combined baseline model.

	F1 score	
	1 epoch	2 epoch
FlowQA model 1	62.0	/
FlowQA model 2	63.0	68.8

TABLE II
Results of our trained FlowQA models.

Github

https://github.com/matijaklj/
End-to-end-text-processing-Naloga

References

[1] Siva Reddy Danqi Chen Christopher D. Manning, “CoQA: A Conversational Question Answering Challenge,” August 2018.

[2] Chenguang Zhu, Michael Zen, Xuedong Huang, ”SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering,” December 2018.

[3] Mark Yatskar, ”A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC,” September 2018.

[4] Hsin-Yuan Huang and Eunsol Choi and Wen-tau Yih, ”FlowQA: Grasping Flow in History for Conversational Machine Comprehension,” October 2018.

@inproceedings anonymous2019flowqa:, title=FlowQA: Grasping Flow in History for Conversational Machine Comprehension, author=Anonymous, booktitle=Submitted to International Conference on Learning Representations, year=2019, url=https://openreview.net/forum?id=ByftGnR9KX, note=under review