Question Answer Modeling using Deep Learning Techniques

Ayush Bhandari
bhandari.ay@husky.neu.edu

Anupam Gupta
gupta.anup@husky.neu.edu

Casey Knox
knox.ca@husky.neu.edu

1 Introduction

Although NLP has made many strides in different fields such as Spam Detection, Parts of Speech tagging, sentiment analysis, information extraction, etc. it is still difficult to build a good question answering model. Question answering (QA) is a well-researched problem in NLP and is applicable to a wide variety of tasks such as information retrieval and entity extraction. A well known use of QA is the development of chatbots and other dialogue systems that simulate human conversation. Recent advances have made good progress in this area due to a shift in paradigm as to how we build NLP models. For a long time, pre-trained word embeddings such as word2vec were used to initialize the first layer of a neural network, followed by a task-specific architecture trained in a supervised way using the original dataset. Recently, several works have demonstrated that we can learn hierarchical contextualized representations on huge (web-scaled) datasets leveraging unsupervised (or self-supervised) signals such as language modeling and transfering this pre-training to downstream tasks.

Our goal for this project is to develop a system which can answer questions not only dependent on a document's corpus but also on the document's semantic meaning.

2 Related Work

There have been many solutions to the SQuAD dataset (1), most of which revolve around different approaches to the attention and embeddings, such using character-level embedding, self-attention, or bi-directional attention (6) (ex. BiDaF, DCN).

Some of the best performing are Hybrid AoA (Attention-over-Attention) Reader (7) which was originally designed for close style reading comprehension task and trained on the daily mail dataset. The main idea of the AoA reader is to create the additional layer of 'attended attention' over standard individual attentions that would involve not only query-to-document attentions but document-to-query attentions. By merging these attentions together the model could introduce a deeper layer of contextual understanding that secures better performance on all metrics

Another high performing model would be R-Net+ (ensemble) (8). R-Net+ was designed to work with SQuAD and Microsoft Machine Reading Comprehension (MS-MARCO) datasets.The R-Net+ was designed as an ensemble model that consists of 18 training runs with the identical architecture and hyper-parameters. At test time, the model selects the answer with the highest sum of confidence scores amongst the 18 runs for each question.

Currently the highest performing model is the ALBERT (ensemble model) which has an EM of 89.731 and F1 score of 92.215 beating human level performance (EM 86.831 and F1 score 89.452) (9).

3 Methodology

Our goal would not be to reach the state of the art accuracy but to explore and test out different methods for this task and see which one performs well and why does it perform better than others. In phase one we would like to build simple models using Infersent (10), which is a sentence embeddings method that provides semantic sentence representations. It is trained on natural language inference data and generalizes well to many different tasks.

And in the second phase we would like to try out more advanced deep learning techniques, specifically sequence modelling. These sequence models typically would consist of an embedding layer,

Preprint. Work in progress.

an encoder layer which would make use of RNN's and then comes the attention layer. This attention layer is the key component in the Question Answering system since it helps us decide, given the question which words in the context should I "attend" to. And then we would finally have our output layer.

4 Datasets and Evaluation

There are typically two types of QA datasets: open and closed. Open QA datasets depend on general knowledge which may not exist in the dataset. Closed datasets on the other hand contain all the information necessary to answer the question within the dataset. In order to avoid scraping and other information retrieval tasks, this project will only use closed datasets.

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

SQuAD tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. SQuAD uses two different metrics to evaluate how well a system does on the benchmark. The Exact Match metric measures the percentage of predictions that match any one of the ground truth answers exactly. The F1 score metric is a looser metric measures the average overlap between the prediction and ground truth answer.

References

[1] SQuAD Dataset

[2] SQuAD Progress

[3] Spracher Jack, Schoenhals M Robert, Fushini Takahiro. *Question Answering System with Deep Learning*
Stanford University, Stanford.

[4] Benjamin Hamel. *Recurrent Neural Networks With Attention For Question Answering.*
Department of Computer Science, Stanford University, Stanford, 2018.

[5] Tang Zhengyang, Li Songze.
Attention-Based Neural Network For Question Answering.
Department of Computer Science, Stanford University, Stanford.

[6] Minjoon Seo1, Aniruddha Kembhavi, Ali Farhadi1, Hananneh Hajishirzi1.
Bi-Directional Attention Flow For Machine Comprehension.
Allen Institute for Artificial Intelligence, University of Washington, 2018.

[7] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu and Guoping Hu. *Attention-over-Attention Neural Networks for Reading Comprehension.* Joint Laboratory of HIT and iFLYTEK, iFLYTEK Research, Beijing, China
Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin, China

[8] *R-Net: Machine Reading Comprehension With Self-Matching Networks.*
Natural Language Computing Group, Microsoft Research Asia.

[9] *Albert: A Lite Bert For Self-Supervised Learning Of Language Representations.*
Google Research Team, Under review as a conference paper at ICLR 2020.

[10] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, Antoine Bordes. *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data.* arXiv:1705.02364