Literature Review

CSCE 798: Directed Study and Research

By: Kunal Ajay Dhavale

Instructor: Dr. Biplav Srivastava

Introduction

This document is a literature review for the papers, implemented systems based on Question Answering systems. Question Answering is under the field of Natural Language Processing and Information Retrieval. The purpose of this document is to read and summarize the existing methods which perform the task and in future try to build a system keeping the underlying model in consideration.

This document is an attempt to explain the core functionality about the following papers/projects:

- NeuralQA: A Usable Library for (Extractive) Question Answering on Large Datasets with BERT (https://github.com/victordibia/neuralqa)
- Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning (https://github.com/salesforce/WikiSQL,https://arxiv.org/pdf/1709.00103.pdf)
- TAble PArSing (TAPAS) By Google (https://github.com/google-research/tapas)

NeuralQA

This paper is an attempt to make Question Answering compatible with existing infrastructure. Automatic QA can reduce the task of finding the answer from huge number of documents. They can give exact answers based on the query. Possible problem with sparse representation could be vocabulary mismatch. Also, the solution to sparse representation i.e. dense representation results in additional complexity and latency. To resolve existing problems the authors introduced an end-to-end library. To handle the vocabulary mismatch problem, contextual query expansion was performed using masked language model. Concept called as *RelSnip,* relevant snippets, was introduced to reduce the latency for chunking the document.

Important modules in NeuralQA are: Retriever, Reader, Expander, User Interface. Retrieving is nothing but a process which returns a list of possible candidates(passage) which can contain the answer based on the query(question). When a query is passed, elasticsearch method is used in order to return the passages. Reader takes the passage which might contain the solution and it predicts where the corresponding answer to the query would be. It uses BERT model for predicting the span of answer in the context(passage). To improve recall, Contextual Query Expansion is performed. User Interface is nothing but adds an aesthetical bonus to this project. It allows the user to check the system on their own customized passage and perform Question Answering over the provided passage.

What I liked about the paper:

Usage of Masked Language Model for contextual query expansion.

RelSnip was a new concept introduced by the authors. It was really interesting to see how it managed to chunk a large document and process each chunks separately. RelSnip was just used to extract some relevant snippets which would be fed into the reader.

To minimize fake information, the idea to use minimum threshold value proved to be very useful according to me.

Seq2SQL

This paper proposed two things. First, Seq2SQL, a deep Neural Network to translate Natural Language to Structured Query Language (SQL). Second, WikiSQL, a dataset which consisted of questions and SQL queries. The basic functionality of Seq2QL is to apply policy based Reinforcement Learning. It takes the input as questions, columns of table. The corresponding query is generated which is then executed against the database to fetch the answer to the query. As the output space of softmax in seq2seq is large and does not fit into this project, the authors use Augmented Pointer Network (APN). APN generates SQL query token by token from the input sequence passed in. The input sequence is concatenation of the column names of the data. SQL query has three components: Aggregation operator, SELECT operator, WHERE operator. Using APN, first the network classifies the aggregation operator for the input query. Then it points to the SELECT the table which contain the data. Then finally the conditions are generated for the question.

WikiSQL is nothing but a database, which is a collection of questions, SQL queries and SQL tables. This is the dataset used in order to compute the evaluation of the Seq2SQL introduced by the authors. It was showed that the model which used Reinforcement Learning proved to be better than the model which didn't used Reinforcement Learning. It was found that when the output space is constrained then it leads to more accurate predictions. Usage of Reinforcement Learning resulted in higher quality of WHERE clauses which helped in accurate predictions.

TAPAS

TAPAS is an approach to Question Answering without generating the logical forms. The authors claim that that the TAPAS model can be trained with weak supervision. It is an extension of BERT architecture. When a query is given to the model, it selects the subsets in the table where the answer would be present. It used BERT encoder with additional positional embeddings. Also, two additional classification layers are used for selecting the cells in tables and corresponding aggregation. The separator is added between questions and tables and not between the cells or rows. TAPAS can predict the corresponding cell and it can also predict which aggregation operator would be used for the operation to complete.

When a question is asked to the model, it encodes the question as well ass the table content row by row using BERT which is extended by special embeddings. The model outputs two things: 1) A probability for each cell which indicates the likelihood that the answer will be in this particular cell, 2) prediction of the aggregation operator for the corresponding query. The model is pre-trained to table text pairs which are extracted from Wikipedia. A key concept about this model is that it learns from it answers. At the time of fine tuning, this model itself learns how to answer the question from the given table. Either strong or weak supervision is used. The authors recommend using weak supervision as it provides the correct answer.