Scientific Question Answering with AAN

Keaton Mueller, Alexander R. Fabbri, and Dragmoir Radev

Department of Computer Science, Yale University, New Haven, CT

LILY Lab

Introduction

This project aims to create a Question Answering system for queries related to AI and NLP using the AAN data set. QA is a task within IR and NLP which seeks to systematically derive an answer to a question asked in a natural language based on a knowledge source. Open domain QA systems take in natural language questions (e.g. "What is an LSTM?") and produce short text responses. Researchers at Facebook have made impressive strides in long-form question answering using their ELI5 data set. Their abstractive, multitask Seq2Seq model outperformed traditional Seq2Seq models, language modeling, and a strong extractive baseline. In a later paper, they refined their approach by using a knowledge graph representation of the supporting documents, allowing them to reduce redundancy while not sacrificing answer token coverage. This approach to long-form question answering is one we will potentially model our own after, but using our own data set of NLP related questions.

Materials and Methods

We gathered queries from Quora, Reddit, and primarily StackOverflow. We filtered StackOverflow posts based on relevancy and were left with 10,874 of the initial 79,688 posts. To compile support documents for each query, we used two approaches: a Solr-based statistical retrieval as well as a k-nearest neighbors approach based on sentence embeddings. We utilized Google's Universal Sentence Encoder to convert the queries into sentence embeddings and the Faiss library to perform a similarity search. For the Solr-based retrieval, we leveraged an existing tool for the AAN data set to interface with Solr. From the documents returned by the Solr retrieval, we performed TF-IDF extraction to gather the most relevant sentences. We then analyzed token overlap with the answers to compare the performance of each method. As another potential source of queries and answers, we looked to Wikipedia. The thought process was that using an article's title and section headings, we could create templates to generate questions. From the initial dump of 5,350,767 Wikipedia articles, we gathered 8,277 relevant articles by searching for members of relevant categories like "Artificial Intelligence" and their subcategories. We then performed an analysis on the most frequent section titles and the most common words in the section titles.

Method	Both (Micro)	Both (Macro)	All (Micro)	All (Macro)
Solr	53.49%	57.38%	39.45%	41.42%
Embedding	47.24%	51.07%	47.22%	51.03%

Table 1: Comparison of retrieval methods

Occurrences	Section Title
8366	introduction
7452	references
4939	external links
4320	see also
1378	history
803	further reading
552	notes
497	applications
361	reception
343	overview
277	plot
212	bibliography
208	examples
187	background
171	features
169	cast
160	production
155	definition
146	example
143	description
142	career
142	biography
131	development
124	research
120	awards
113	education
112	technology
112	software
96	publications
91	release
86	algorithm

Table 2: Most Common Section Titles

Occurrences	Section Word
8390	introduction
7567	references
4972	external
4971	links
4326	see
4325	also
2132	and
1645	history
1416	of
830	further
820	reading
813	the
667	applications
650	in
625	notes
414	overview
395	reception
373	research
350	other
333	plot
301	examples
295	to
292	awards
289	career
269	-
267	features
261	education
247	life
247	definition
245	software
245	development

Table 3: Most Common Section Words

Results

Table 1 shows the results of performing token overlap given the Solr-based and embedding-based retrieval methods. Because the Solr retrieval method sometimes returns zero documents while the embedding-based method always returns documents, the table shows statistics over queries for which both methods returned results and statistics over all queries. Given the 10,874 initial queries, the Solr retrieval did not return results for 3,027 (27.8%) of them.

Table 2 and Table 3 show the most common section titles and most common section words (words appearing in a section title) respectively from the compiled Wikipedia data.

Conclusion

Our results show that when Solr works, it generally outperforms the embedding-based approach. However, the embedding-based method has the added advantage of being guaranteed to produce output. Given the fact that the Solr retrieval failed to return results for 27.8% of the given queries, we conclude that the embedding-based approach is far better suited for more general use.

There is a lot of exciting work left to be done on this project. Now that we have compiled an early form of our data set, further work can be done to annotate it to judge the quality of the questions. We also have concerns about only using the post title as our query. Questions gathered from Quora would likely perform better in this case since on Quora questions are limited to the title.

And of course, we still need to create our Seq2Seq model for generating answers to queries given the query itself and the supporting documents. For this step we would likely mirror much of what Facebook did in their ELI5 paper, as well as incorporate methods such as graph-based knowledge representation in order to reduce the dimensionality and redundancy of the supporting documents.

Acknowledgement

Huge thank you to Alex Fabrri for all of his help and to Professor Radev for advising!