Pranav Kaza

Professor Diego Klabjan

IEMS 308: Data Science and Analytics

20 March 2018

Question Answer System Implementation in a Business Context

**Executive Summary**

A question answer system is a program capable of providing answers to factoid questions within a set of given parameters. A typical question answer system will analyze the question being asked by extracting keywords, use search querying methods to extract relevant documents and sentences, and then use a recognition algorithm to extract the answer to the question.

In this case, the documents being analyzed were Business Insider daily news articles from 2013 and 2014, and the questions being asked were of the form “Which companies went bankrupt in month W of year X,” “What affects GDP” with a follow-up question of “What percentage of drop or increase is associated with Y”, or “Who is the CEO of company Z?”

After extracting keywords from the question and preprocessing the keywords as well as the corpus of business articles, the open source querying solution Elasticsearch is used to extract relevant business articles. From there, relevant sentences are extracted using the same process and finally, text recognition techniques are used to extract the answer to the question.

The results were strong when asking for the CEO name of a company, but questionable when searching for answers to the other questions. This is largely because the final text recognition techniques used were not robust, so improvements to the system can be made by developing an improved text recognition classifier.

**Problem Statement**

A question answer system is a program capable of providing answers to factoid questions within a set of given parameters. The assignment was to develop a question answer system capable of answering the following 3 questions:

1. Which companies went bankrupt in month W of year X?
2. What affects GDP? What percentage of drop or increase is associated with Y?
3. Who is the CEO of company Z?

**Assumptions**

1. One big assumption made in developing the system is that the answer is available in the text being searched. Commercially developed solutions such as Apple Siri and Amazon Alexa will typically state if they cannot find the answer to a question, but this system will simply find what appears to be the best answer to the question, regardless of whether it is even close to correct. This is because the search algorithm will simply pick the highest scored documents, then pick the highest scored sentences and extract answers, regardless of the absolute values of the scores.
2. Another assumption is that the answer to the question will be in the same sentence as the keywords being searched for. In many cases, this assumption would likely fail, but the assumption drastically simplifies the scope of the problem computationally by limiting the number of corpus segmentations that must be made.
3. The last assumption being made is specific to the question being asked, based on the named entity recognition method used:
   1. For question 1, the system assumes that each capitalized unigram, bigram and trigram in important sentences is a company name. This is done because the NER classifier developed was found to be very ineffective due to the use of only 1 predictor in the previous text analytics assignment.
   2. For question 2 part 1, the system assumes that all nouns in important sentences are factors that affect GDP. This is done to simplify analysis because relevant answers can be manually picked out from a very small list at this point.
   3. For question 2 part 2, the system assumes that the percentage extracted from the most important sentence is the correct percentage to be identified.
   4. For question 3, the system assumes that the CEO will be mentioned in the form “CEO First Last.” This is done because the NER classifier developed was found to be very ineffective due to the use of only 1 predictor in the previous text analytics assignment.

**Methodology**

The first step in developing the question answer system was to extract keywords from the question asked by the user. In a more complex system, the question would be analyzed using a classification model to determine what exactly is being asked, but the scope of this problem only requires understanding of 3 questions, so a classification model was not necessary in this case. In this scenario, the keywords were manually examined and synonyms and alternate potential keywords were added to enhance the search. For example, the word “declare” was added to the bankruptcy question keywords because “declaring bankruptcy” is a very commonly used phrase.

The keywords are then preprocessed with normalization, stemming, and lemmatization. The same preprocessing is applied to the corpus after importing the documents.

At this stage, the open source querying tool Elasticsearch is loaded into Python to enhance the speed of text searching using powerful indexing. For each word in the keywords, a score is calculated for each document assessing the term frequency in that document, the inverse document frequency of the word in relation to other documents, and the length of the field in which the word is being searched for. The scores are summed over all keywords for each document and the top 20 documents are selected for sentence segmentation.

After segmenting the selected documents by sentence, the sentences are loaded into an Elasticsearch index and the top 10 sentences are chosen based on the same scoring function used for document selection.

Lastly, the selected sentences are analyzed to extract the answers to the questions. For question 1, regular expressions are used to extract all capitalized unigrams, bigrams, and trigrams from sentences containing the word “bankrupt.” For question 2 part 1, all nouns are selected from sentences containing the word “GDP.” For question 2 part 2, regular expressions are used to extract all forms of percentages from sentences containing the word “GDP” and the top ranked percentage (coming from the top-ranked sentence) is used. For question 3, regular expressions are used to extract CEO names in the form of “CEO First Last.”

**Analysis**

Based on the answers given to the sample output questions provided, the only question that can be answered robustly is “Who is the CEO of company Z?” This is because the form in which CEO names are mentioned is more standardized than company names, or percentages in relation to GDP factors. Even this question will not work as well if companies are mentioned to which the CEOs are not as well known, so they are rarely mentioned in the corpus. The key to making this entire system more accurate is to build a more robust named entity recognition classifier for company names and CEOs using more predictors and trying many types of models, as was the task in the third homework assignment.

Another major shortcoming of the process was evident when answering the “What affects GDP” question in that there was no clear way to isolate the answers from this question from unimportant information. Although the methodology used extracted a few important answers, it also selected irrelevant words that were clearly not factors that affect GDP.

A third shortcoming of the process was evident when answering the “Which companies went bankrupt in month W of year X” question in that the month and year were not heavily weighted enough to differentiate the answers between different time periods. For example, the answers outputted from December of 2008 were similar to the answers from March of 2013. This indicates that the date keywords were not heavily weighted enough in the scoring function, or that the bankruptcy dates were likely not included in the same sentence as the company name. This problem could be solved by changing the weighting of the keyword scoring function or by extracting double sentences along with single sentences for search querying (this would work because the field length is considered in the scoring function).

**Conclusion**

Based on the question answer system implemented, more robust changes need to be made before this solution can be implemented in a practical scenario, namely the creation of a more robust NER classifier. However, in several cases, the QA system can accurately identify the CEOs of many well-known companies, including Google, Facebook, and Microsoft.

**Next Steps**

In improving the QA system, there are several major steps to make it more robust:

1. Create a more accurate NER classifier for CEO names, company names, and percentages so that regular expressions do not have to be used.
2. Experiment with segmentation of double sentences or triple sentences instead of single sentences to ensure that the keywords are in the same field as the answer to the question.
3. Experiment with varying the weights of different keywords depending on what results are being fetched.
4. Modify the system to account for absolute values of scores, so that incorrect answers are not given to the user. In cases where the correct answer cannot be found, ensure that the system will output that it cannot find the answer, rather than providing an incorrect answer.