Question-Answer System for Business Insider

Rosanne Harrison

Executive Summary

The goal of this analysis was to develop a question and answer system capable of answering three distinct categories of questions. The questions relate to identifying CEOs, companies that have gone bankrupt, and how various properties impact GDP. The corpus for this QA system is a collection of Business Insider articles from the years 2013 and 2014.

The analytical tools applied in the following analysis all relate to text mining and the development of a QA system. The exact methodology will be discussed below; however, the main steps in the process were identifying the question, selecting and scoring documents, and then selecting and scoring sentences in those documents to find an answer. This project also utilized the matrices of named entities developed in project 3, in order to easily search for potential answers in the corpus.

After creating a document term matrix, I used rules I developed for each question to select documents and sentences, and then searched through the relevant sentences for answers found in the named entity matrices. In the end, my system is able to accurately answer questions in each 'category' of questions. However, the systems' responses are limited both by the depth and breadth of the corpus, and the capabilities that were able to be built into the QA system.

In analyzing the final results, I found that some questions are able to be answered more accurately, more often; for example, the QA system is extremely good at identifying CEOs of companies. The answers to other questions tend to be accurate less often. The company identification question is limited by the amount of information in the corpus; companies that have gone bankrupt are often only mentioned a couple times (or not at all), and the year and month are not always included. On the other hand, the GDP question is limited by the amount of manual inspection required for an accurate answer to be produced. This question is also not as simple of a 'factoid' question as the others, as GDP is a time-varying value that can be positively or negatively impacted by the same property many times over time. Thus, the results from this question are limited in their business relevance; the answers produced by this part of the QA system would likely not be able to provide real business value due to their lack of context.

Methodology

In order to develop the QA system, I used a process based on the following diagram (taken from IEMS 308 notes):

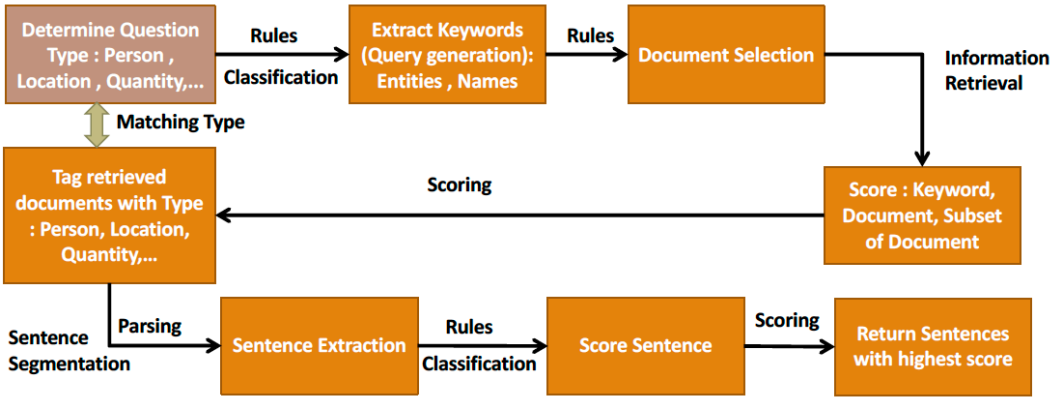

Figure 1: Question and Answer system diagram

Question Analysis

To begin, I determined the question type (of the three potential questions) by looking at the first word of the question. Based on the question type, I next extracted keywords. In order to determine keywords, I used all named entities as well as any other nouns that I believed would help in searching for answers. For some questions I also included another version of a keyword from the question with a different suffix (e.g. bankruptcy) or common synonyms for the word (e.g. growth). The following table shows the keywords used for each question:

Question	Keywords
Who is the CEO of company X?	"CEO" name of company X
Which companies went bankrupt in month X of year Y?	"bankrupt" or "bankruptcy" month (X) year (Y)
What affects GDP?	"GDP"
What percentage of decrease or increase is associated with Z? **increase OR decrease synonyms were used depending on which was in the quesiton	"GDP" "percentage" or "percent" or "%" **"drop" or "decrease" or "drag" **"increase" or "growth" property Z (all words)

Table 1: Keywords used in query generation

I created a single document-term matrix for use in analysis of all questions. The DTM contains every word in the corpus, as well as a count of how many times each word appears in each document. This matrix was utilized in the next step, which involves selecting and scoring documents for use in answer analysis.

Document Analysis

The next step was to select and score documents. I essentially utilized a binary scoring system for document selection; that is, based on their score documents were either selected for

sentence extraction in the next step, or they were thrown out. I implemented a slightly different scoring system for different questions, as some query keywords (and information in general) appeared more often in the corpus.

For each question, I first counted how many times each keyword appeared in each document, and stored this information in a data frame. For the questions "Which companies went bankrupt…" and "What affects GDP/What percentage…", I simply extracted all documents that contained **ALL** of the keywords in the query. The exception was that I only required a document to contain one of any of the keywords listed with an 'or' between them (e.g. "bankrupt" or "bankruptcy").

For the question "What companies went bankrupt…", I also implemented a second document extraction rule for questions inquiring about year 2013 or 2014. If the posed question related to one of those two year, I also extracted all documents from the month and year in question into the relevant document list. If a bankruptcy is discussed during the month it occurred, the month and year are likely not included in the article text.

For the question "Who is the CEO of company X?", I took another step to narrow down the selected documents. I ranked all the documents in the corpus based on the count of appearances of both keywords – thus, each document had two rankings. I then extracted only the documents for which the count of BOTH keywords was in the top 10% of all documents.

Sentence Analysis

After extracting documents to examine, I then proceeded with sentence segmentation and extraction. I used regular expressions to perform sentence segmentation on all selected documents, breaking a sentence up whenever a period, question mark or exclamation mark appeared. I then extracted only sentences that were likely to contain the answer to my question. I used the following rules for sentence extraction, depending on the question that was asked:

Question	Sentence Extraction Rules
Who is the CEO of company X?	*Only sentences that contain both keywords are extracted
Which companies went bankrupt in month X of year Y?	*Only sentences that contain 3 keywords (either "bankruptcy" or "bankrupt") are extracted. *If the year Y (in the question) is either 2013 or 2014, extract all sentences that contain either "bankruptcy" or "bankrupt" from documents written in month X of year Y
What affects GDP?	*Only sentences that contain the keyword
What percentage of decrease or increase is associated with Z?	*Only sentences that contain 4+ keywords (either "percentage"/"percent"/"%" and one of the synonyms for 'increase' or 'decrease')

Table 2: Sentence Extraction rules used in answer generation

NOTE: *For the question "What affects GDP?" I pulled out documents that contained the word GDP more than 5 times. I then performed a manual search of all segmented sentences that contained the word GDP, in order to find properties that affect GDP that are contained in the corpus. Once a list was developed, I programmed the responses to this question to automatically appear once that question was asked.*

Answer Analysis

After extracting the relevant sentences, I searched for the answers to the posed question in these sentences. To do this, I utilized the lists of CEOS, companies and percentages from Assignment 3. I searched through all relevant sentences for any of the CEOs, companies or percentages (whichever was relevant for the question) on this list. I then created a data frame containing all extracted potential answers, as well as a count of how many times each potential answer appeared in the relevant sentences. The potential answer that appeared **MOST often** in the relevant sentences is then returned as the answer to the question.

For the question "What percentage of…", I did a final search through the answer matrix for which potential answers included either "percent", "percentage" or "%", as some answers in the potential answer matrix did not include any these indicators. After potential answers NOT containing any of those terms were eliminated, the final answer is returned based on the process described above.

Conclusions & Business Insights

There are several key business insights that we can take away from the development and analysis of the QA system:

Some questions have a better 'correct answer rate' than others. The effectiveness of the QA system in answering the question has to do with both the complexity of the question and the frequency of the answer terms in the corpus. Generally, the QA system does the best job of finding correct answers for the question "Who is the CEO of company X?", and seems to struggle most with the question "What percentage decrease or increase is associated with this property?". The system's performance is due to several factors, including the complexity of the question and the contents of the corpus.

The effectiveness of the QA system is limited by the size of the corpus and the depth/breadth of its content. For example, CEO names are mentioned at *extremely* high rates in the corpus; as such, it is easy for the QA system to identifying many examples of these names. CEO names are also generally written in the same format (e.g. "CEO Jeff Bezos"), making it easy to find common patterns in the corpus. On the other hand, specific bankruptcies are mentioned much less often in BI articles; as such, there is a lower chance of *any* instances existing in the corpus, as well as a lower chance of the instances that exist appearing in the same sentence.

The QA system is highly effective at identifying CEOs. One reason for this is that the question being asked is not very complex, as it only has to use two keywords: 'CEO' and the company name. Additionally, names of CEOs (and their title) appear extremely frequently in Business

Insider articles, which make up the corpus. Not only is the QA system easily able to identify occurrences of CEO names, it is also easily able to determine which of the potential answers are the correct answer, because the correct answer usually appears far more often than other CEO names. The QA system for identifying CEOs certainly has business value, as it is highly accurate and is able to find the names of *many* CEOs.

The QA system is capable of identifying some bankruptcies, however it is limited by the corpus and the structure of the sentences being analyzed. The number of specific bankruptcies that are mentioned in the corpus is much smaller than the number of CEOs mentioned. The contents of the corpus were out of our control for this assignment; however, a corpus containing specifically information about bankruptcies would likely provide much better results. Additionally, this question has more keywords, meaning the likelihood of them all appearing in the same sentence is lower. Some bankruptcies are discussed in the corpus without mention of the month or year in which it occurred, which makes it impossible for our QA system to identify them. This aspect of our QA system has some business worth (for the bankruptcies it is accurately able to identify); however, some improvements to the corpus or search methodology could make it more valuable.

While the QA system is able to identify impacts on GDP, it is highly reliant on manual inspection and the results are limited with regards to their business value. The capability of our system to answer the GDP question is restricted by the amount of manual inspection required for an accurate answer to be produced. Manual inspection is required to identify properties that impact GDP, and to distinguish how these properties are phrased in the corpus. For example, many mentions of percentages refer to the percent *of* GDP a certain property makes up (e.g. consumption is 70% of GDP). This distinction is essential for accurate results. This question is also <u>not</u> a simple 'factoid' question. GDP changes over time, as a result of countless properties shifting in the positive and negative direction. Thus, the answers produced by this part of the QA system would likely not provide real business value due to their lack of context.

Next Steps and Improvement Opportunities

- Develop a more thorough sentence segmentation algorithm, which takes into account bigrams, trigrams, POS tagging and other regular expressions.
- Give the answer analysis portion of the QA system the ability to look for answers in surrounding sentences. For example, the year of a bankruptcy may be mentioned in one sentence, while the month is mentioned in the next.
- Incorporate the use of bigrams, trigrams, POS tagging, etc. for both document selection and sentence selection.
- Determine the feasibility of expanding the corpus to contain more documents, which might contain more easily digestible information regarding bankruptcies and/or GDP.
- Test other sentence scoring algorithms, such as TF-IDF. However, these algorithms may not be as effective due to the small scale of the problem.

Instructions

Using RStudio as the interface:

- Type question into command line as a variable named “question”:
  - question = “Who is…/Which companies…/What affects…?”
- Run all code
- For two-part questions:
  - Type second question into command line in the same manner (question = “…”) once output of first question is produced.
  - Re-run only second chunk of code, labeled “QUESTION PROCESSING”
- To ask another question, type new question into command line in the same manner and re-run the “QUESTION PROCESSING” code

Sample Output

CEOs

- Q: Who is the CEO of Pepsi?
  - A: “indra nooyi”
- Q: Who is the CEO of Tesla?
  - A: "elon musk"
- Q: Who is the CEO of Ford?
  - A: "alan mulally" (CEO until July 2014)

Companies

- Q: Which companies went bankrupt in month September of year 2008?
  - A: “lehman”
- Q: Which companies went bankrupt in month April of year 2014?
  - A: “gox” (referring to Mt. Gox, a Japanese bitcoin exchange)
- Q: Which companies went bankrupt in month November of year 2011?
  - A: “us airways”

GDP

- Q: What affects GDP?
  - A: "tax deal, economic trends, job gains, inflation, employment"
- Q: What percentage of increase is associated with economic trends?
  - A: “2%”
- Q: What percentage of increase is associated with job gains?
  - A: “2%”
- Q: What percentage of decrease is associated with tax deal?
  - A: “1%”