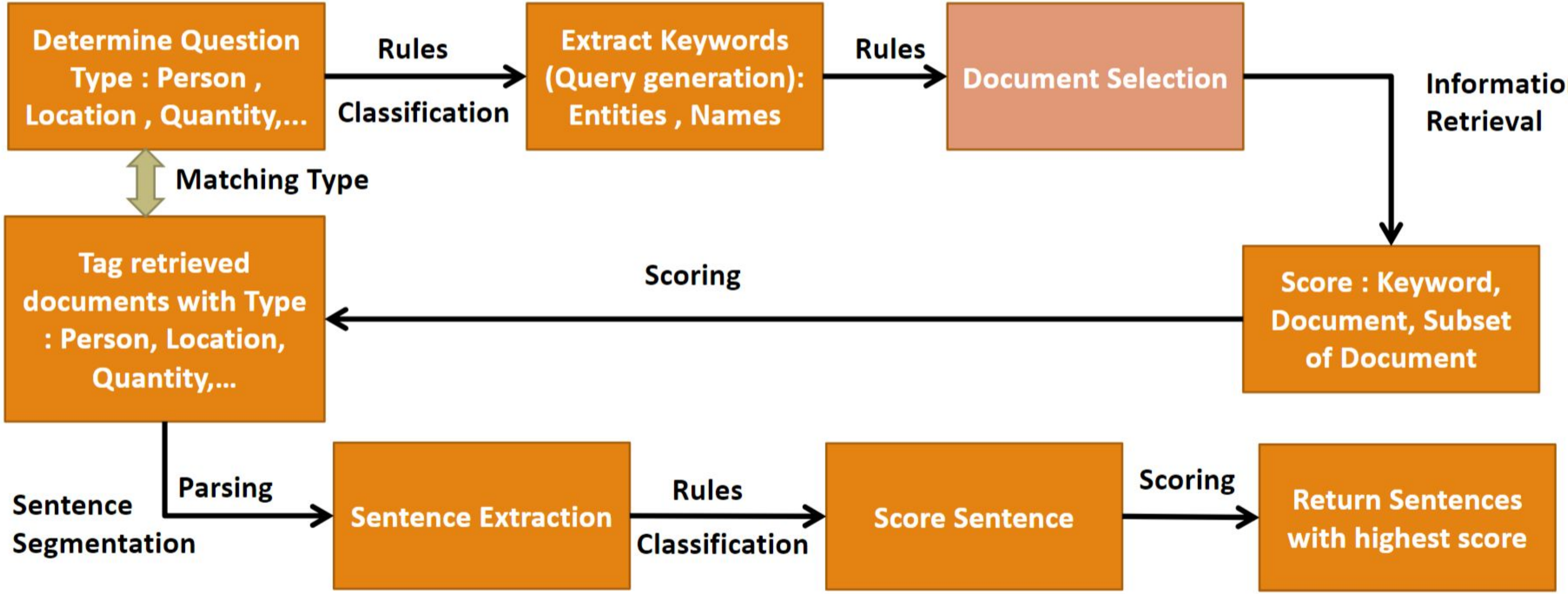
Question and Answer System

Amy Wang

Methodology



Determine Question Type : Person , Location , Quantity,...

Rules

Classification

Extract Keywords (Query generation): Entities , Names

Rules

Document Selection

Information Retrieval

Score : Keyword, Document, Subset of Document

Scoring

Tag retrieved documents with Type : Person, Location, Quantity,...

Matching Type

Sentence Segmentation

Parsing

Sentence Extraction

Rules

Classification

Score Sentence

Scoring

Return Sentences with highest score

Pre-processing step

- Transform the corpus into a document-term matrix
  - Each document is a news article
  - 35898 entries in total
- Store the matrix into Elasticsearch database
  - Index by date + article number of the day
  - Localhost:9200 to access it

Question Classification

- Automatic classifier approach
  - How similar the input question is to all four sample questions
    - Use the maximum cosine-similarity score to determine the type
  - Sample questions are:
    - Question 1 = "which companies went bankrupt in month x of year y?"
    - Question 2 = "what affects gdp?"
    - Question 3 = "what percentage % of drop or increase change in gdp is associated with results from this property consumption consumer spending government investment imports exports foreign trade?"
    - Question 4 = "who is the ceo of company x?"

Query Formation and Document Retrieval

- Use “Query string” type for Elasticsearch
- Queries are usually keywords
    - For example, for company’s ceo name, the query will be “CEO and XXX(company)”
- Use elasticsearch to retrieve top candidates of the document

Answer generation

- Select candidate sentences based on keyword-filtering
- Use NER to tag nouns
  - For CEO, looking for Person
  - For Bankruptcy, looking for Organization and GPE (geopolitical entity)
- For GDP reasons, tf-idf for terms are calculated as NER will not work
- For the impact of the specific reason for GDP, regex is used
  - "\d+(?:\.\d+)?(?:%| percent(?:(?:age|ile)? points)?)”
  - Min value in the list of percentages is used because some extreme large percentages are not for the specific reason
    - For example, “40% of China’s GDP came from Shanghai” will be extracted
    - Taking average will be inaccurate as these extreme values will shift the true value to a bigger one

Interface

- Use the function “answer_the_question”
- If downloaded as “.py” file, then can use command line to enter the function

Business Insights

- Rule-based approaches are easier to execute
- Factual questions are easier to answer, and yield accurate results
  - Even if we don’t have prior knowledge about the topic
- Closed-domain allows the search to be quicker
- If datasets are downloaded locally, no web connection is necessary for answering the questions

Further steps

- Based on this corpus
    - Other factual questions can be developed for analysis
    - Market sentiment analysis can also be analyzed from the text
- If possible, enlarge the corpus
    - To get a broader range of questions answered
- Get a training dataset of questions
    - To formally train the dataset in order to classify the input string more accurately
- A better interface
    - Use Python GUI packages to make it more user-friendly to input the question