Question Classification in Question-Answering Systems

Pujari Rajkumar

Question-Answering

- Question Answering(QA) is one of the most intuitive applications of Natural Language Processing(NLP)

- QA engines attempt to let you ask your question the way you'd normally ask it .

- More specific than short keyword queries

  - Orange chicken

  - What is orange chicken?

  - How to make orange chicken?

- Inexperienced search users

Types of QA Systems

- Two types of QA Systems:

  - 1. Open domain QA Systems

    Should be able to answer questions written in natural language similar to humans
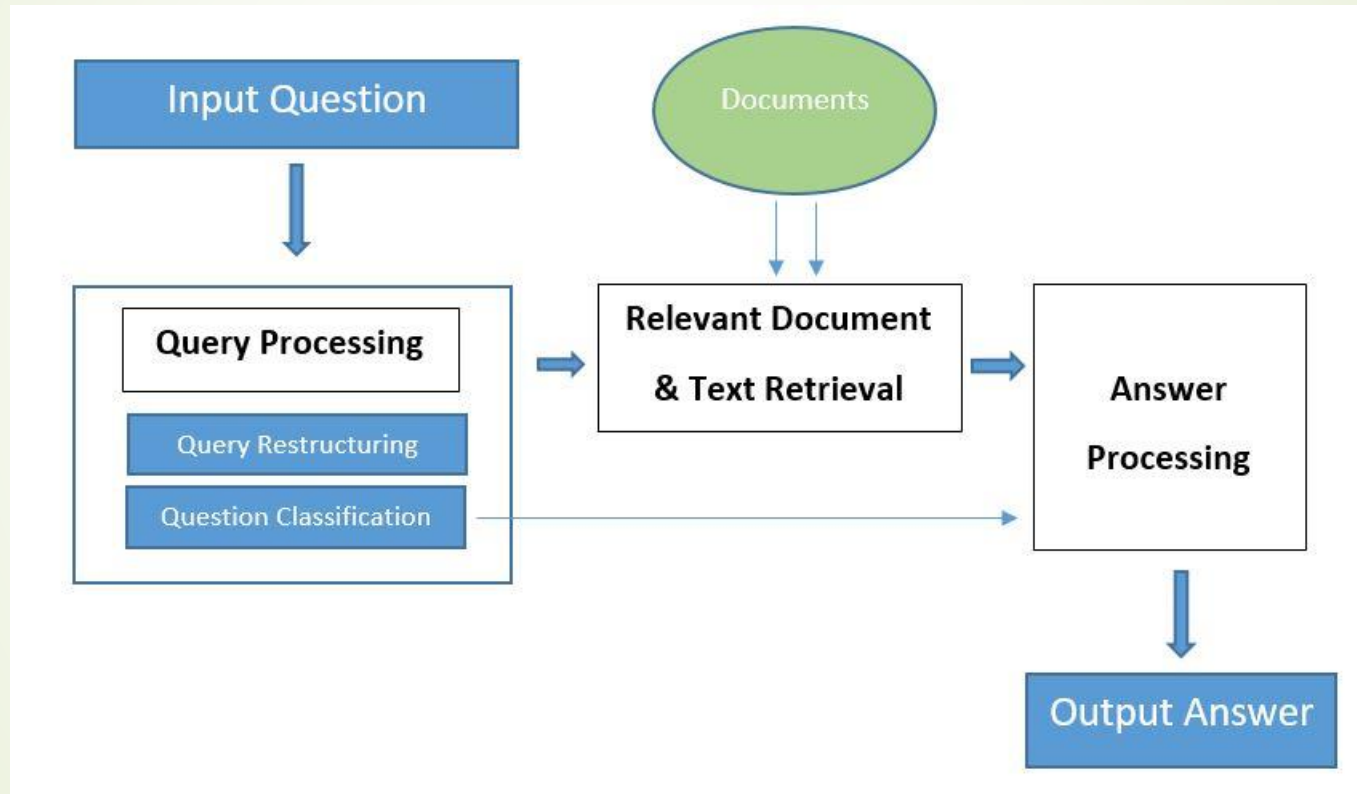
    Eg: Google

  - 2. Domain-Specific QA Systems

    Answer questions pertaining to a specific domain. Can give more detailed answers but restricted to a single domain

    Eg: Medical-domain QA systems (WebMD)

Typical QA Architecture

Stages of QA System

- Question Processing

  Consists of two phases, *query reformation* and *question classification*. Query reformation consists of forming suitable IR/knowledge-base query needed to extract relevant text from available documents/database. Question Classification(QC) consists of assigning the question to one or more of pre-defined classes of questions

- Passage Retrieval

  Relevant documents or relevant text from those documents, that helps in formation of answer, is retrieved from available documents. QC is also useful in this stage as question category determines the search strategy that needs to be employed to find the most suitable answer(s)

- Answer Processing

  Consists of constructing appropriate answer(s) from the text retrieved in previous stage. This stage also uses QC as it helps in choosing the candidate answer which is most probable to belong to the same class as the question

IBM Watson

- IBM Research undertook a challenge to build a computer system that could compete at the human champion level in real time on the American TV quiz show, Jeopardy

- Meeting the *Jeopardy* Challenge requires advancing and incorporating a variety of QA technologies including parsing, question classification, question decomposition, automatic source acquisition and evaluation, entity and relation detection, logical form generation, and knowledge representation and reasoning

- *Category:* General Science
  Clue: When hit by electrons, a phosphor gives off electromagnetic energy in this form.
  Answer: Light (or Photons)

Question Processing

- Consists of two phases:

   1. Query Reformation

   2. **Question Classification**

   Eg: **Where is 'India gate'?**

   Restructured query: *India gate location*

   Question Category: 'location'

An example

Example(contd.)

Restructured Query: *India Gate, Address*
Question Class: *'Address'*

Example(contd..)

Restructured Query: *India Gate, Coordinates*
Question Class: *'Coordinates'*

Another Example

Answers can be of descriptive type as well

Question Taxonomy

- Also known as 'Question Ontology' or 'Question Taxonomy'

- Pre-defined set of classes that questions are classified into

- Tailored according to the dataset and task at hand

- IBM Watson has 11 pre-defined Question Classes:

    Definition, Fill in the blanks, Abbreviation, Category relation, Puzzle, Verb, Translation, Number, Bond, Multiple choice and Date.

- The classes were tailored towards Jeopardy! Challenge questions

Li and Roth's Taxonomy

Class	#	Class	#
ABBREV.	9	description	7
abb	1	manner	2
exp	8	reason	6
ENTITY	94	**HUMAN**	65
animal	16	group	6
body	2	individual	55
color	10	title	1
creative	0	description	3
currency	6	**LOCATION**	81
dis.med.	2	city	18
event	2	country	3
food	4	mountain	3
instrument	1	other	50
lang	2	state	7
letter	0	**NUMERIC**	113
other	12	code	0
plant	5	count	9
product	4	date	47
religion	0	distance	16
sport	1	money	3
substance	15	order	0
symbol	0	other	12
technique	1	period	8
term	7	percent	3
vehicle	4	speed	6
word	0	temp	5
DESCRIPTION	138	size	0
definition	123	weight	4

Table 1: The distribution of 500 TREC 10 questions over the question hierarchy. Coarse classes (in bold) are followed by their fine class refinements.

Bloom's Taxonomy
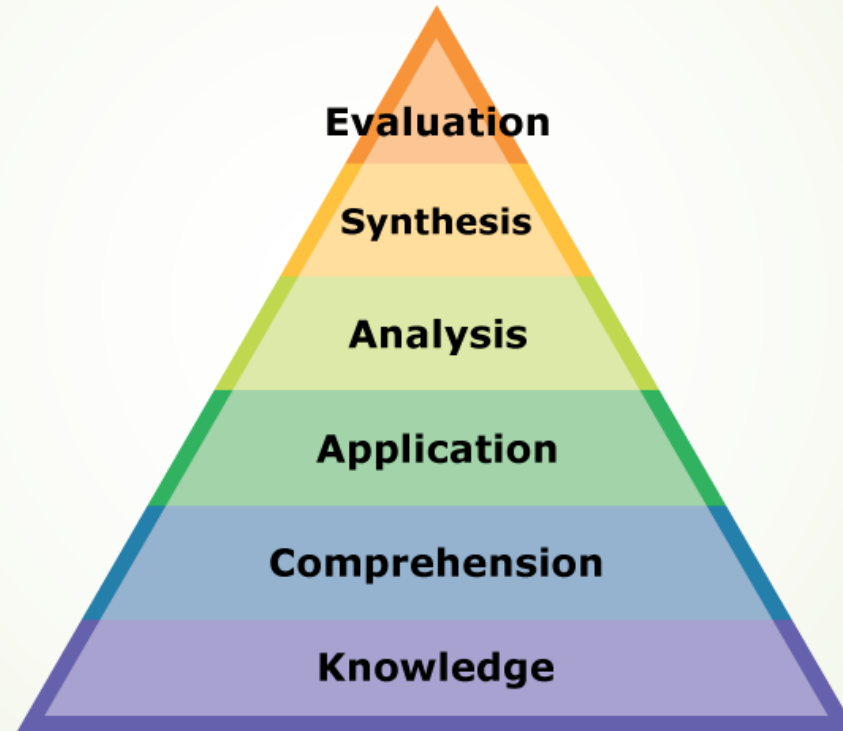
Question Classification Systems

- Essentially two types of Question Classification Systems:
  - 1. Rule-based Systems[1]
  - 2. Learning-based Systems
- Hybrid systems that combine both the approaches also exist

  Eg: IBM Watson

Rule-Based Systems

- A rule based approach consists of hand written rules that are run on the given question to assign it to a pre-defined category

- Such systems do not need any training data

- Eg: (Hull, 1999)

    Set of Question Categories:

            <Person>, <Place>, <Time>, <Money>, <Number>,

            <Quantity>, <Name>, <How>, <What>, <Unknown>

    Mapping from keyword to question category:

            who <Person>, where <Place>, what <What>,

            whose <Person>, when <Time>, which <What>,

            whom <Person>, how <How>, why <Unknown>

Drawbacks of Rule-based Systems

- Lack of thoroughness

- Contradictory rules

- Large set of rules to cater for corner cases

- A small rule set doesn't always do a thorough job and a good enough system often needs large rule set which is difficult to handle.

- Illustration: Rule: Whom - <person>?

  - Works fine: *Whom did the contestant call using the lifeline?*

  - Issue: 'Whom' might also represent organizations

  Whom did Chicago bulls beat in 1992 Championship finals?

Learning-based Systems

- Systems based on machine-learning techniques

- Given labelled data and a set of features, the system learns how to classify the questions into pre-defined categories

- Learning based approaches proposed for QC are mainly supervised learning techniques

- Some of the popular supervised classifiers used are SVM, Maximum Entropy models and language modeling

- Semi-supervised classifiers such as co-learning have also been used effectively.

Support Vector Machine (SVM)
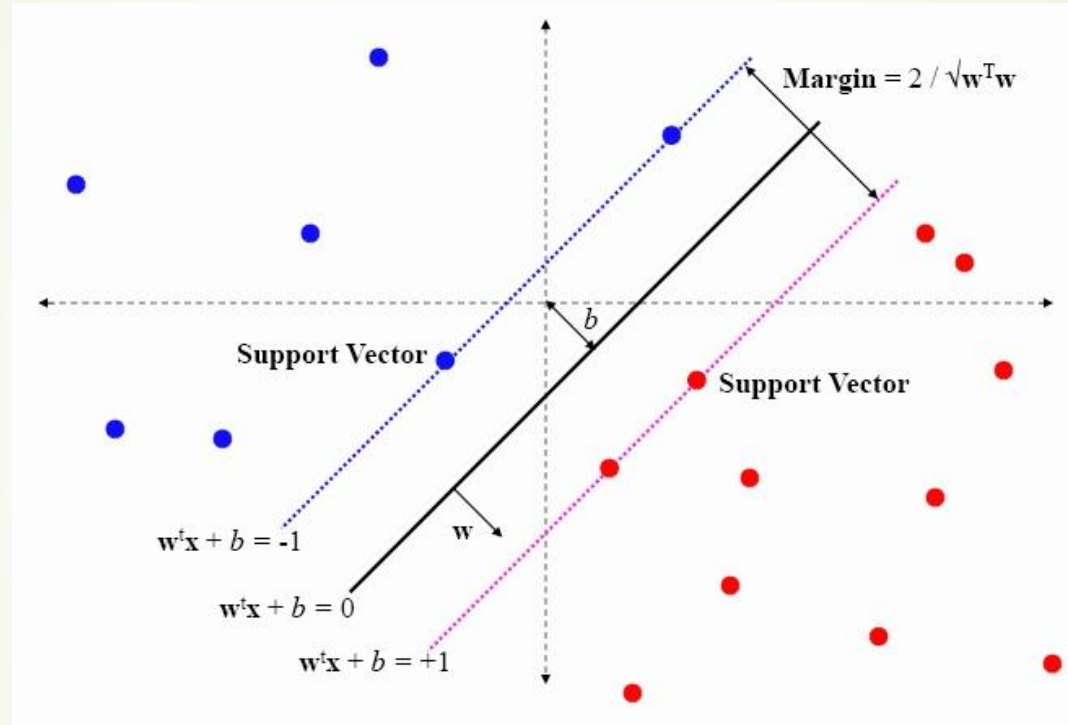
SVM tries to find a hyper plane which has maximum margin between separating classes.

Maximum-Entropy Models

Probability that a sample x_i belongs to class y_i is calculated as:

$$p(y_j | x_i, \lambda) = \frac{1}{Z(x_i | \lambda)} exp \sum_{k=1}^{n} \lambda_k f_k(x_i, y_j)$$

- f_k is feature indicator function which is usually a binary-valued function defined for each feature. λ_k is weight parameter which specifies the importance of $f_k(x_i, y_i)$ in prediction and $Z(x_i | \lambda)$ is the normalization function

- To learn parameters λ_k, the model tries to maximize log-likelihood LL, defined as follows:

$$LL = \sum_{i} log \frac{exp \sum_{k=1}^{N} \lambda_k f_k(x_i, y_i)}{\sum_{j} exp \sum_{k=1}^{N} \lambda_k f_k(x_i, y_i)}$$

Language-Modeling

- Base idea of language modeling is that every word in the text is viewed as being generated by a language.

- Each question can be viewed as a document. Probability of a question belonging to language of a given class 'c' can be computed as:

$$p(x|c) = p(w_1|c)p(w_2|c,w_1) \ldots p(w_n|c, w_1, \ldots, w_{n-1})$$

- As learning all the probabilities needs quite a large amount of data, unigram assumption can be made i.e., probability of each word is only dependent on the previous word. This reduces the equation to:

$$p(x|c) = p(w_1|c)p(w_2|c, w_1) \ldots p(w_n|c, w_{n-1})$$

- Most probable class can be determined using Bayes rule:

$$c = \text{argmax } p(x|c) \, p(c)$$

where $p(c)$ is a prior probability that can be assigned to the classes or can be taken as equal for all classes.

Semi-Supervised Learning Models

- Semi-supervised learning methods such as co-training have also been used to construct QC systems successfully.

- Co-training is a method of training two classifiers simultaneously.

- Given a set of labeled and unlabeled data, both classifiers are trained on labeled data and unlabeled data is marked by both the classifiers.

- Top results with high confidence from each classifier is fed to the other classifier for training. This process is repeated again.

Hybrid Approach

- Question Classification systems have also been constructed using hybrid approach which uses both rule-based and learning-based classifiers.

- IBM Watson is a very good example of such system. The detection in Watson is mostly rule based, which includes regular expressions patterns to detect the question class. On top of which, a logistic classifier is employed to get the best possible class.

- IBM Watson has 11 pre-defined Question Classes:

  Definition, Fill in the blanks, Abbreviation, Category relation, Puzzle, Verb, Translation, Number, Bond, Multiple choice and Date.

Features in Question Classification

- Pivotal part of using a classifier is construction of feature vector using optimal set of features.

- A simple feature vector can be constructed as:

$$x = (w_1, w_2, \ldots, w_n)$$

  where w_i is frequency of word i in question x. This would be a very sparse feature vector

- A simple modification can be done by dropping words with zero frequency from the vector

- Various other features that provide much deeper information about a question are used, in practice

Syntactic Features

- Syntactic features consist of structural aspects of the given question such as parts of speech(POS) tags and head words.

- Successful POS taggers exist which can give POS tags with high accuracy (~96%) such as Stanford NLP POS tagger.

- A head word is usually defined as most informative word in the sentence.

- Extracting head word of a sentence is a challenging problem and requires construction of parse tree of the question based on a set of grammar rules.

- Probabilistic Context Free Grammars(PCFGs) can be used for such purpose.

Head Word Example

What year did the Titanic sink?
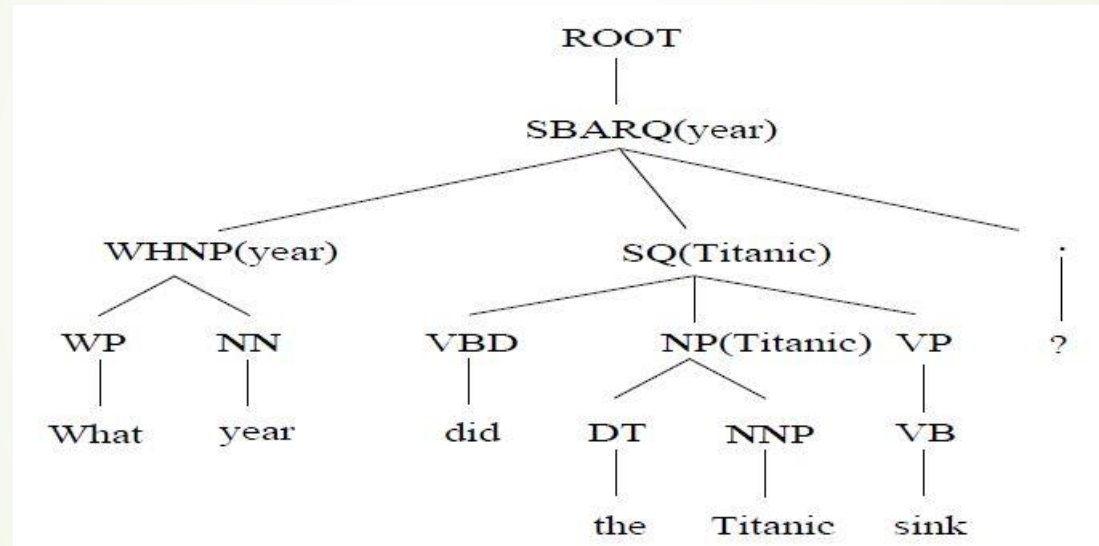
Head Word: **year**

Semantic Features

- Semantic features are extracted based on meaning of the words in the question

    Eg: Hypernyms and named entities

- Hypernym is word which denotes a higher level semantic concept to the given word

  Eg: 'animal' is a hypernym of 'cat'

- *Wordnet* can be used to find hypernyms of given words

- Named entity is a well-known place, person or event, approximately a proper noun present in the question. Named entity recognition(NER) is a well researched area in NLP with lot of existing systems which achieve high accuracy

Evaluation

- Performance metric for any QC system would be accuracy of the system, i.e.

$$\text{Accuracy} = \frac{\text{Number of correctly classified questions}}{\text{Total number of input questions}}$$

- Standard IR metrics such as precision and recall also can be looked at for a given question category.

$$\text{Precision} = \frac{\text{Number of correctly classified questions as a given category}}{\text{Total number of input questions labeled as that category}}$$

$$\text{Recall} = \frac{\text{Number of correctly classified questions as a given category}}{\text{Total no. of questions actually belonging that category in input data}}$$

Questions?