Generation of Questions from Text Using Natural Language Processing

Team 11

Aviral Joshi
01FB15ECS062

Chetna Sureka
01FB15ECS076

Abstract

Natural-language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular it deals with the task programming computers to analyze and process large amounts of natural language data.Challenges in natural-language processing frequently involve speech recognition, natural-language understanding, and natural-language generation.

In this project we have tried to use various NLP techniques to generate relevant questions from any given text that contains phrases/sentences from english language.

Methodology

Software Used

For this task we have used the **Python** programming language, its versatile programming constructs and high level of abstractions allows for faster development time. We have also used the Natural Language Toolkit **(NLTK)** library which has valuable utility for NLP.

Approach

A myriad of of approaches can be used when generating question form a text. The approaches can be sophisticated or simple based on the complexity of the questions that need to be generated. Mentioned below are the approaches that we have tried to generate questions.

1> " ? " are indicative of a question

It is a given that an sentence ending with a question mark is a question and we can use that in our favour to generate questions.

2> Named entity recognition

The **NLTK** library provides functions for extracting named entities from a given passages along with their type such as whether the named entity is a Location, an Organization, or a Person, etc. Questions can then be asked on the character/attributes of the named entity.

3> Hyponyms

Words that are Nouns can be recognized using the NLTK's Parts of Speech Tagging function. Once the nouns are identified questions such as fill in the blanks can be asked by replacing them with a blank. The answers to that blank can the set of all hyponyms for that noun.

4> Looking at Dates and other Numerals in Data

Multiple Choice Questions involving Dates and years was looked at from both the Question formulation side and the Choice generation side. The options were also randomly shuffled.

5> Proper Nouns

Identifying Proper Nouns using Parts of Speech recognition helped us mark the key Proper Nouns from the text and one among many was randomly selected to fill in the blank and closely occurring ones were among the options in the Multiple Choice Answers.

6> True or False

Identifying statements stating facts and making a new version(the incorrect one) based on Semantics of the Language. A random function then either uses the Real version and prints True as answer, or uses the Incorrect version and prints False. This also looks at numbers in the text to suitably modify for generation of a false candidate.

7> Subjective Questions- Definitions

On closely analyzing the way definitions are asked in high schools, using Semantics, questions were generated keeping the definition in text in mind.

Results

By implementing the above mentioned approaches we were able to generate a reasonable question bank. That covered most sub topics of the texts used. The questions were quit generic to what one will find in standard text book for students in school. However these approaches were not able ask questions on the relationships between two named entities, also the questions lacked depth to them, i.e. they were not able to account for and use long term dependencies described in the text to generate questions.

Some Sample Questions Produced :

- 'Who is Augusto Pinochet ?'
- Its economic policies have made _____ intolerable for most of the country's 34m citizens.
- But Mr Maduro may not have to steal the election on the _____ to win it.
- A poll last month by ____ suggests that 38% of the electorate will not vote. What are we talking about here?
- A disn'tpute with holders of its defaulted debt shut off Argentina from international credit markets. True or False?
- A recent survey found that three-quarters of Argentines remain opposed to ____ assistance. What are we talking about here?

A recent survey found that three-quarters of Argentines remain opposed to IMF assistance.

- When Baber made his first forays into India where his dynasty established one of the greatest landpowers of Asia, the Portuguese seapower already controlled the Indian Ocean.True or False?(*True*)
- Define Newton's Third Law of Motion.

Newton's Third Law of Motion is defined as for every action, there is an equal and opposite reaction.

- A Tamil inscription of ____, unfortunately badly damaged, provides evidence for the presence of a South Indian merchants' guild in Sumatra at that time. Which Year did this occur?

  a)1113 b)1107 c)1093 d)1099 e)1088

  1113

Further Enhancements

Better Named Entity Recognition

The named entity recognition (**NER**) is very accurate and obtains an accuracy of 89.71% on the 'schwa.org/projects/resources/wiki/Wikiner' dataset. The stanford NER classifier obtains an accuracy of 92.23%. These NER classifiers do well but other techniques involving the use of Deep Learning have known to perform them.
Replacing the standard NER with one that uses Deep Learning can help improve the accuracy of NER and help in generation of more accurate questions. More specifically Bi-Directional RNN's or LSTM units fare well for NER.

"*Named Entity Recognition with Bidirectional LSTM-CNNs*"
https://arxiv.org/abs/1511.08308

Modelling Relationships Between the Named Entities

To ask deeper questions involving the relationships between 2 or more named entities we must be able to discern the attributes that connect them and whether they are connected positively, negatively, or are neutral.

This can be done by performing some form of sentiment analysis and building a web of interconnections between named entities.
The established links can then be used to ask more sound questions.