Project Presentation :

Factoid Question Answer Generation from Wikipedia

Team 28

Vikrant Goyal(201502040), Eavanshi Arora, Nikita Agarwal

Problem Statement

In this project, we consider an automatic Sentence-to-Question-Answer generation task, where given a sentence, the Question Generation (QG) system generates a set of question answer pairs for which the sentence contains or needs answers.

Goals

Goal 1

Our QA generation system should only ask factoid questions i.e questions whose answer is a single entity rather than generating a opinion based question or a description based question.

Goal 2

It should be able to handle the cases of complex sentences too.

Tools Used

- TextBlob
- Nltk
- Python wikipedia library (wikipedia)
- Regex (re)

Types of Questions

- Purpose - the correction of knowledge deficits , the monitoring of common ground
- Type of Information - concept completion questions, “shallow” factual questions
- Source of Information - aims to generate questions for which the source of answer is the literal information in the text
- Length of the Expected Answer - short, usually a single word or short phrase

Process Followed

- Extracting wikipedia articles from the wikipedia library in python and then running a tokenizer on it.
- We have taken only the summary of the articles because it's the only content that is quite useful and would generate good questions.
- Used inbuilt POS tagger for tagging the corpus.
- Now we need to extract the named entities and classify them too. But the inbuilt NER tagger is quite bad as we have experimented on it.

Process Followed

- So we wrote our own grammar for the nltk chunker and parsed each sentence.
- Now extract the essential named entities from each sentence and classify them as location, proper noun and some kind of number or date.
- Based on the above classification we can get the type of question namely where, who, which etc.

Problems Faced

Challenge 1

Semantics of the answer to a question affects the question's form. Mapping answers to "Wh" words and phrases such as who or which is difficult.

Challenge 2

Non-compositionality : A phrase is not just a simple aggregation of the meaning of it's component words. So it can be misleading.
Example : burned to the ground

Milestones Achieved

1. **Questions of the form - Fill In The Blanks.**

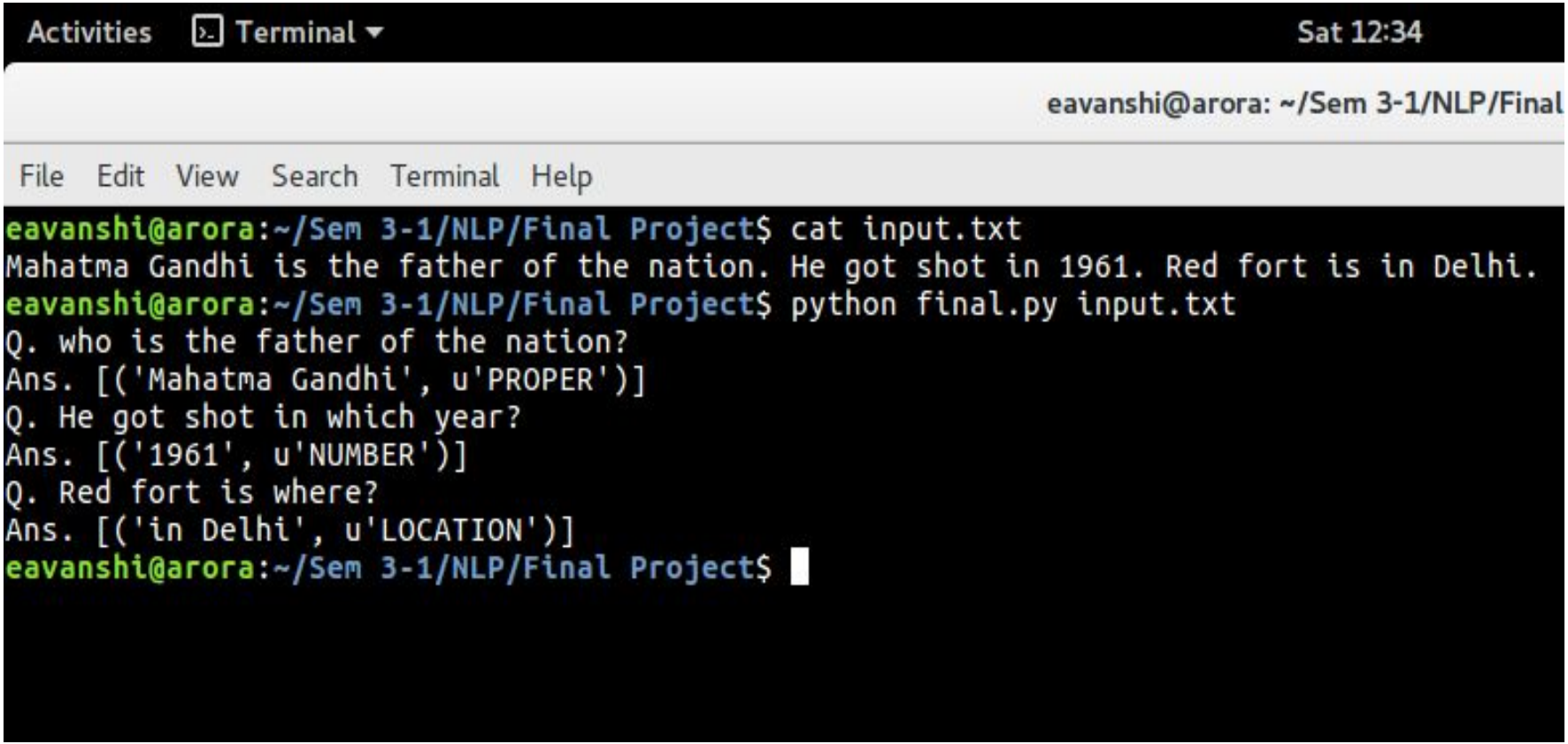
Fill in the blanks type of questions are also quite important and have its applications in the education assessment. We have successfully generated questions of the above mentioned format which are very accurate and can easily deal complex sentences.

```
Ans. [u'Philosophi\xe6 Naturalis Principia Mathematica', u'Mathematical Principles', u'Natural Philosophy', u'1687']
Q. Newton also made pathbreaking contributions to optics, and he shares credit with _____ for developing the infinitesimal calculus.
Ans. [u'Gottfried Wilhelm Leibniz']
Q. Newton's Principia formulated the laws of motion and universal gravitation that dominated scientists' view of the physical universe for the next __
___ centuries.
Ans. [u'three']
Q. By deriving Kepler's laws of planetary motion from his mathematical description of gravity, and using the same principles to account for the trajec
tories of comets, the tides, the precession of the equinoxes, and other phenomena, Newton removed the last doubts about the validity of the heliocentr
ic model of the _____ and demonstrated that the motion of objects on Earth and of celestial bodies could be accounted for by the same principles.
Ans. [u'Solar System']
Q. Newton's theoretical prediction that the Earth is shaped as an oblate spheroid was later vindicated by the geodetic measurements of Maupertuis, La
Condamine, and others, thus convincing most Continental European scientists of the superiority of Newtonian mechanics over the earlier system of Desca
rtes.
Ans. [u'of Maupertuis , La Condamine']
Q. Newton also built the first practical reflecting telescope and developed a sophisticated theory of colour based on the observation that a prism dec
omposes white light into the colours of the visible spectrum.
Ans. []
Q. Newton's work on light was collected in his highly influential book Opticks, first published in _____.
Ans. [u'1704']
Q. He also formulated an empirical law of cooling, made the first theoretical calculation of the speed of sound, and introduced the notion of a Newton
ian fluid.
Ans. []
Q. In addition to his work on calculus, as a mathematician Newton contributed to the study of power series, generalised the binomial theorem to non-in
teger exponents, developed a method for approximating the roots of a function, and classified most of the cubic plane curves.
Ans. []
Q. Newton was a fellow of _____ and the second _____ of Mathematics at the University of Cambridge.
Ans. [u'Trinity College', u'Lucasian Professor']
Q. He was a devout but unorthodox Christian, who privately rejected the doctrine of the Trinity and who, unusually for a member of the Cambridge facul
ty of the day, refused to take holy orders in the Church of England.
Ans. []
Q. Beyond his work on the mathematical sciences, Newton dedicated much of his time to the study of alchemy and biblical chronology, but most of his wo
rk in those areas remained unpublished until long after his death.
Ans. []
Q. Politically and personally tied to the Whig party, Newton served _____ terms _____ for the University of Cambridge, in _____ and _____.
Ans. [u'two brief', u'as Member of Parliament', u'1689\u201390', u'1701\u201302']
Q. He was knighted by _____ in _____ and he spent the last _____ decades of his life in London, serving as Warden (_____) and Master (_____) of the __
___, as well as president of the _____ (_____).
Ans. [u'Queen Anne', u'1705', u'three', u'1696\u20131700', u'1700\u20131727', u'Royal Mint', u'Royal Society', u'1703\u20131727']
eavanshi@arora:~/Sem 3-1/NLP/Final Project$
```

Milestones Achieved

2. Normal QA system for simple questions

We have also made a system which can give out the normal type of questions and would be better only on the simple sentences. It can't handle the case of complex sentences as it would require quite good parsing techniques and extraction of certain clauses from a sentence.


Activities
Terminal
Sat 12:34
eavanshi@arora: ~/Sem 3-1/NLP/Final
File Edit View Search Terminal Help
eavanshi@arora:~/Sem 3-1/NLP/Final Project$ cat input.txt
Mahatma Gandhi is the father of the nation. He got shot in 1961. Red fort is in Delhi.
eavanshi@arora:~/Sem 3-1/NLP/Final Project$ python final.py input.txt
Q. who is the father of the nation?
Ans. [('Mahatma Gandhi', u'PROPER')]
Q. He got shot in which year?
Ans. [('1961', u'NUMBER')]
Q. Red fort is where?
Ans. [('in Delhi', u'LOCATION')]
eavanshi@arora:~/Sem 3-1/NLP/Final Project$

Further Research

Major improvement that can be done is to generate question for Complex sentences, which requires appropriate shallow parsing of text so that we can extract simple sentences or clauses from the sentences for which our system can generate questions easily. Some kind of discourse parsing might also help.