Computational Linguistics II

Chatbot

Jashn Arora - 2018114006
Tanish Lad - 2018114005

Interim Report

Problem Statement

To Design a Rule-Based Chatbot in Hindi which will interact with the user and answer the user's queries.

Our Approach

After looking through some of the Previous Works on the topic, we have decided to follow this procedure:

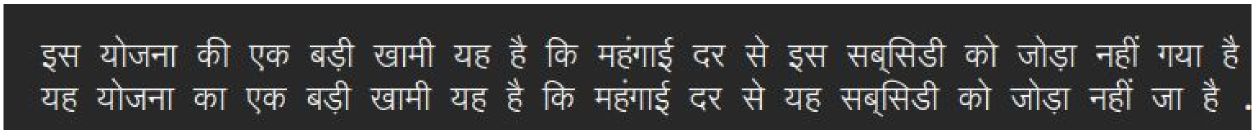
- Upon starting, the Chatbot will first greet the user. This will be done by Brute-Force Approach.
- The Chatbot will ask the user for a query related to our domain (yet to be decided). The user will enter his query.
- We would have already scrapped the data related to our domain, and that would be our dataset. The data will be parsed and each word in every sentence would be present in it's lemmatized form.
- We would then parse and lemmatize each query of the user.

- We would find the most similar sentence from our dataset corresponding to the query using Cosine Similarity Formula (this may be changed if we find a better rule based approach for finding similarity between two sentences). The most similar sentence would be our output for the given query of the user.
- This process will continue until the user signals to exit using a closing command (yet to be decided).

What we have done till this interim submission

- Python code to lemmatize the parsed data. We tested this on sample data and it works successfully. We took some sample sentences, chunked them and gave it as the input to our code. The code will extract the root of each word from the chunked data and store it in an array. The root of each word is present in the chunked output. The code will simply extract root using simple python functions for text manipulation like split and strip.

  Sample Output:

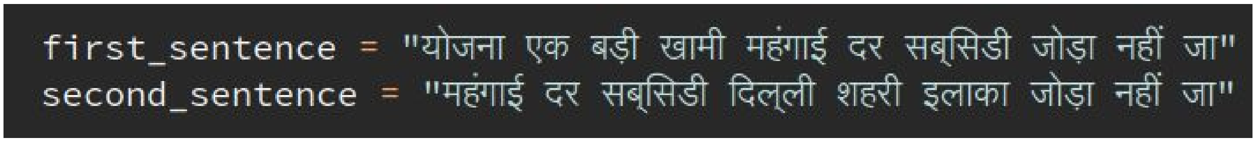
  

  ```
  इस योजना की एक बड़ी खामी यह है कि महंगाई दर से इस सब्सिडी को जोड़ा नहीं गया है
  यह योजना का एक बड़ी खामी यह है कि महंगाई दर से यह सब्सिडी को जोड़ा नहीं जा है .
  ```

  The first sentence corresponds to the actual sentence, and the second sentence corresponds to the lemmatized form for each word of the sentence.

- Python code to find the cosine similarity between two sentences.

  We would first remove the stop words from both the input sentences. Then, using cosine rule from trigonometry, cosine similarity between the sentence vectors will be found. The value would always be greater than or equal to 0 and less than or equal to 1.



```
first_sentence = "योजना एक बड़ी खामी महंगाई दर सब्सिडी जोड़ा नहीं जा"
second_sentence = "महंगाई दर सब्सिडी दिल्ली शहरी इलाका जोड़ा नहीं जा"
```

Upon running our code, the cosine similarity was found to be 0.6324555320336759 which is good because the sentences are quite similar.

What we plan to do till the end

- Selecting a suitable domain like any sport or any topic that has sufficient information available on the web in hindi to scrape.
- Coding a brute force logic to greet the user
- Removing Stop Words from the Sentences
- Successfully simulate a rule based chatbot with a good accuracy.

Challenges we might face

- Finding a way to lemmatize the user query as our current method requires the query to be sent to a Hindi Online Chunker and receive the chunked output so as to lemmatize it. We may change our approach to lemmatize the query if we find a better suitable approach to do it.

- If we find that the cosine similarity formula is not very accurate for some sentences, we would try to find a better rule based approach, and if we do find a better method, we might shift to that method.

Contributions of each person

We both read both the papers together. We both coded all the codes together on a single laptop.

References

[1] Sviatlana Höhn.***A data-driven model of explanations for a chatbot that helps to practice conversation in a foreign language (2017)***

[2] Ananthakrishnan Ramanathan, Durgesh D Rao. ***A Lightweight Stemmer for Hindi (2003)***

[3] Geeks for Geeks. ***https://www.geeksforgeeks.org/python-measure-similarity-between-two-sentences-using-cosine-similarity/***