Project Title: **Anusaaraka**

Sanjana Goel
Computer Science & Engineering
Indira Gandhi Delhi Technical University for Women(IGDTUW), New Delhi
Ph: +91-8630214743

Duration: May 27, 2019- July 20, 2019

1.1 Task 1

Objective

To familiarise with the basic terminologies used in NLP, a brief introduction to Statistical and Neural language modelling and to draw a comparison between Context counting and context predicting semantic vectors.

Tools and Utilities

Python, Jupyter Notebook, Sublime text

Procedure

Result (*Click here*)

- Successfully implemented the 2 context counting algorithms- Pointwise Mutual Exclusion (PMI) and Single Value Decomposition(SDV).
- Documented a paper on NLP terminologies, Statistical and Neural language modelling.
- Drafted a comparison between Context counting and context predicting techniques.

Conclusion

Based on my findings and considerations, I would certainly suggest the use of Distributional Semantics Models(DSMs) for theoretical/practical applications to go for predict models.

References

Referred the following research papers and webLinks:

https://machinelearningmastery.com/statistical-language-modeling-and-neural-language-models/
https://machinelearningmastery.com/what-are-word-embeddings/
https://nlp.stanford.edu/pubs/glove.pdf
https://www.aclweb.org/anthology/P14-1023
https://www.aclweb.org/anthology/P13-4006
https://www.kaggle.com/gabrielaltay/word-vectors-from-pmi-matrix

1.2 Task 2

Objective

To automate the Collins ENG-HINDI Dictionary using Selenium WebDriver as the automation tool and web scraper and Google Chrome as the web browser.

Tools and Utilities

Python, Sublime Text, Selenium WebDriver

Procedure (Find the README and source code *here*)

- Install Selenium WebDriver
- Download the latest version of ChromeDriver from *here*
- Extract and paste the driver in your cd
- Add the input-output text file and python script in the cd
- Run the following command: python script_name.py <input_file> <output_file>

Result

ENG-HINDI word translation is automated with simultaneous data collection.

Conclusion

- Selenium is a wonderful open source automation tool which enables record and playback for testing web applications and can run multiple scripts across various browsers. (*Read more about Selenium*)
- On an average, manual search on Collins Web Dictionary takes about 35 seconds per word. However, by automating this search process using Selenium, the search time is reduced to approximately 5 seconds per word. Thus, enhancing the quality and productivity of web search.

References

Important links:

https://www.collinsdictionary.com/dictionary/english-hindi
https://automatetheboringstuff.com/
https://www.pluralsight.com/guides/web-scraping-with-selenium
https://towardsdatascience.com/web-scraping-using-selenium-python-8a60f4cf40ab

1.3 Task 3

Objective

To make improvements in Harshit Sir's module to automate Bing translator using Selenium WebDriver as the automation tool and web scraper and Firefox as the web browser.

Tools and Utilities

Python, Selenium WebDriver

Procedure

- Install Selenium WebDriver
- Download the latest version of GeckoDriver.
- Extract and paste the driver in your cd
- Add the input-output text file and python script in the cd
- Run the following command: python script_name.py <input_file> <output_file>

Result

ENG-HINDI word translation using Bing translator is automated

Conclusion

- Selenium is a wonderful open source automation tool which enables record and playback for testing web applications and can run multiple scripts across various browsers. (*Read more about Selenium*)

References

Important links:

https://automatetheboringstuff.com/
https://www.pluralsight.com/guides/web-scraping-with-selenium
https://towardsdatascience.com/web-scraping-using-selenium-python-8a60f4cf40ab

1.4 Task 4

Objective

To observe and tabulate the word-alignment outputs produced by Anusaaraka ie to check if the words(phrases) are properly aligned while ENG-HIN translation. It is an important supporting task for most methods of SMT.

Tools and Utilities

OCR, Champollion, Font Converter

Procedure (Find the README *here*)

- Carefully observe the English and Hindi sentences. Then, observe the following: Translation Issues, Anusaaraka Issues, ENG-HINDI construction change, Necessary/ Unnecessary paraphrasing, Pth layer observation, solution by parser, solution by the improvement in dictionary

Observation

- A- English Sentence
- K- Anusaaraka layer
- L- ENG-HINDI phrases according to English Hindi phrasal translator tool
- M- HINDI-ENG phrases according to Hindi English phrasal translator tool
- N- Parser layer
- O- Scoring on the basis of L, M, N layers
- P- Manual translation

Click *here* to find the word alignment observations of Class 11, Geography(NCERT) chapter 2.

References

Important links:

https://docs.google.com/spreadsheets/d/16S7nZKTHcfFs9gws_kXGKclP8mxfJ7WqeuTJiZkF__A/edit#gid=1839226296