Independent Study Report

Pranav Kumar Sivakumar

12.22.2018
Advisor: Prof. James H. Martin
Lead Researcher: Shafiuddin Rehan Ahmed

Objective

Assist in collecting huge amounts of data from Wikipedia, organizing them and developing models along the pipeline for performing ranking and prediction of the most suitable Wikipedia links for the obtained resolved named entities in a given text. Also, generate the test data to support the bigger data pipeline.

INTRODUCTION AND BACKGROUND

Disambiguation:

Disambiguation is word-sense disambiguation, the process of identifying which meaning of a word is used in the context.

Coreference Resolution:

Coreference resolution is the task of finding all expressions that refer to the same entity in a text. It is generally an important step for much higher level NLP tasks that involve natural language understanding, such as document summarization, question answering, and information extraction.

Named Entity Recognition:

The named entity recognition model identifies named entities (people, locations, organizations, and miscellaneous) in the input text.

Wikification:

The process of adding wiki syntax to text in a wiki platform, or converting HTML to wiki markup. In our context, It is the task of identifying and linking expressions in a text to their referent Wikipedia pages.

TECHNOLOGY USED

Languages involved: Python, MySQL, Java

Libraries: AllenNLP**,** mysqlDB, Wikipedia, BeautifulSoup, Flask

PROCEDURE

Since I was working on the test-data side, the following procedure is a part of the whole pipeline.

1. Scrape data from a list of links to Wikipedia pages that are related to disambiguation. For eg: This link. Extract a paragraph for each of the titles from their respective pages.
2. Apply coreference resolution using the trained model from AllenNLP (which is based on End to End Coreference Resolution)and change all the entities of the same cluster to the first referred one.
3. Apply the fine-grained Named Entity Resolution model (which uses a biLSTM with CRF layer and ELMo embeddings) from AllenNLP to obtain all the referred entities in the text.
4. Create an API if needed to be used remotely.
5. Match and collect around *n* (Say 15) relevant titles for each of the entity from the constructed database using Full-text search, assign pos/neg labels.
6. Generate the data by creating multiple negative examples for each of the entities alongside their correct title for testing the trained machine learning model that ranks the related titles and finally predicts the most relevant title for the corresponding entity.
7. Test the accuracy of the model developed.

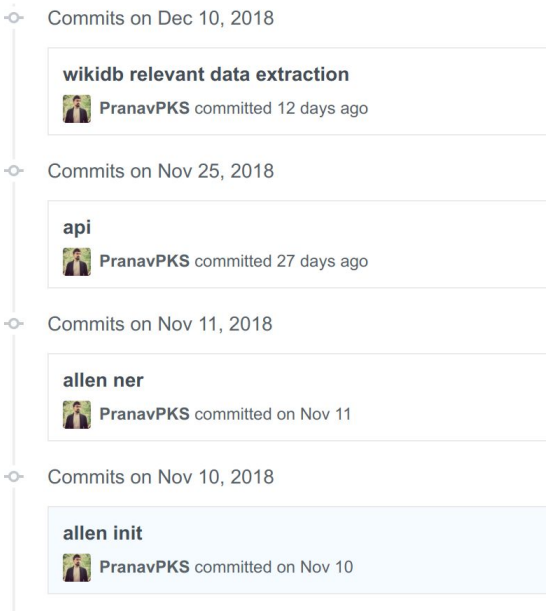
DATA COLLECTED

```
mysql> show tables;
+-------------------+
| Tables_in_wiki_db |
+-------------------+
| Category          |
| MetaData          |
| Page              |
| PageMapLine       |
| Page_idb          |
| category_inlinks  |
| category_outlinks |
| category_pages    |
| inlink_count      |
| links             |
| page              |
| page_categories   |
| page_inlinks      |
| page_outlinks     |
| page_redirects    |
| page_views        |
+-------------------+
16 rows in set (0.02 sec)
```

The data is huge (~40GB) to be delineated in this report. It basically contains the entirety of Wikipedia from the contents of its pages to the analytics of the pages. Analytics mainly refers to the page views, number of inlinks (number of pages this page has been referred) and number of outlinks (number of links in this page). Since these attributes represent the popularity of the pages, they are mainly required to compute the necessary features to train the machine learning model. These were collected with the help of the standard Wikipedia API.

GITHUB REPO

All the project files and work progress were continuously tracked using this Github repository link: https://github.com/PranavPKS/wiki-hop-data. Only the data that can be open-sourced were uploaded. Only if a component is completely ready it had been pushed to the repository.



RESULTS AND CONCLUSION

Since the project is still ongoing, the accuracies that we were measuring on the test data were around 95%. The wikification process will run like the following,

The raw text can be any paragraph like this one,

“Anthony Endrey was a Hungarian-Australian lawyer and author. He was a Queen's Counsel and Master of the Supreme Court in Victoria, Australia, and a member of the Victorian Bar. Endrey was born in Hungary, and graduated Doctor of Law from the University of Budapest. He was a research assistant at Friedricks-Wilhelm University in Berlin. He served in the Royal Hungarian Army during World War II, and was a prisoner of war in the Soviet Union until his release in 1945. He emigrated to Australia in 1949, and in order to practice in Australia he studied law at the University of Tasmania.”

It will have all its coreference resolved, entities recognized and its most associated Wikipedia links assigned for each of the entities. Please download and check this >>> HTML file for reference. The image is attached here,

Anthony Endrey was a Hungarian - Australian lawyer and author . Anthony Endrey was a Queen 's Counsel and Master of the Supreme Court in Victoria , Australia , and a member of the Victorian Bar . Anthony Endrey was born in Hungary , and graduated Doctor of Law from the University of Budapest . Anthony Endrey was a research assistant at Friedricks - Wilhelm University in Berlin . Anthony Endrey served in the Royal Hungarian Army during World War II , and was a prisoner of war in the Soviet Union until Anthony Endrey release in 1945 . Anthony Endrey emigrated to Victoria , Australia in 1949 , and in order to practice in Victoria , Australia Anthony Endrey studied law at the University of Budapest .

REFERENCESS

1. Ratinov, Lev, et al. "Local and global algorithms for disambiguation to Wikipedia." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.
2. Lee, Kenton et al. “End-to-end Neural Coreference Resolution.” EMNLP (2017).
3. Peters, Matthew E., et al. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018).