Midterm 1 - Scope of Work

Background

Kensho is a financial tech company under the umbrella of S&P Global. They develop data analytics platforms and machine learning algorithms to gain insights into financial markets, economic predictions, and trends within human society. One project they have expressed interest in is to use Wikipedia data to aid in NLP tasks. Wikipedia is such an expansive database of information that it can essentially be considered the ground truth of our world. If we can harness that information in the form of a knowledge graph to be used in NLP, then developing insights into our world will be much easier and more accurate.

Problem Statement

Our project's focus is to improve Named Entity Disambiguation (NED), which is an NLP task, using graphical methods. Doing so requires use of Named Entity Recognition (NER), the development and searching of a knowledge graph, and developing deep learning methods that utilize the knowledge graph.

Computing Resource

Currently, all four members are using a personal remote server equipped with Intel Core i7-8750H, 8GB NVIDIA RTX 2070, 32GB RAM. Limited RAM and GPU with only 8GB memory (compared to the knowledge graph) are the main bottlenecks that we are facing when working on knowledge graph data. In addition to this, we are expecting a p2.xlarge to be available to us soon. Nonetheless, we need to design our system while keeping in mind the constraint of limited memory.

Data Sources

The following datasets are related to Wikipedia:

- **enwiki.k_link**: Dataset of hyperlinks on Wikipedia articles that link to other Wikipedia articles.
  - Hyperlink is on source page, while hyperlink links to target page.
- **enwiki.k_plaintext**: Dataset of the text of sections of Wikipedia articles.
- **enwiki.k_raw_anchors**: Dataset of the anchor texts of hyperlinks that link to other Wikipedia articles.
  - An anchor text is the actual clickable text (i.e. the blue text) of the hyperlink.

    - Useful for indicating synonyms/other names of an entity that is not the title of the target Wikipedia page if anchor text != target page title.
- **enwiki.page**: Dataset of Wikipedia pages.
- **enwiki.redirect**: Dataset of redirects.
    - Example: https://en.wikipedia.org/wiki/PRC redirects to https://en.wikipedia.org/wiki/China

The following datasets are related to Wikidata:

- **wikidata.item:** Dataset of Wikidata items.
    - Nodes of the knowledge graph.
- **wikidata.property:** Dataset of Wikidata properties.
    - Edges of the knowledge graph.
- **wikidata.qpq_item_statements**: Dataset of Wikidata triplets.
    - (source, edge, target)

We construct a knowledge graph from **wikidata.qpq_item_statements**. Almost every Wikipedia page has a corresponding Wikidata item. As such, these datasets are essential in producing a training and test set used for NED/NEL, as together they produce data of sentences and their corresponding entities.

Deliverables

The problem can be separated into a 2-step process - named entity recognition (NER) and named entity linking/disambiguation (NED). We utilize the SpaCy NER implementation to identify all the entities within a subset of text extracted from *Wikipedia*. While the exact classification is not necessary, it will be useful in filtering our knowledge graph subsequently, saving on computational resources.

The focus of the project is on NED, where we leverage on *Wikidata* triplets to construct a knowledge graph of all entities mentioned within the text, across a specified local word span. The contextual information of the surrounding words is thus captured within the knowledge graph. In this knowledge graph, there may be several repeated entities. For example, for the text, “**Mount Bromo** is a mountain located in **Java**”, the entity **Java** might have 2 nodes, one referring to the location in Indonesia, and the other referring to the programming language. Using a distance/similarity metric, we can then utilize the knowledge graph to identify which is the **Java** referred to in the text.

We start with the simple approach of computing the number of direct connections to the other entities within the knowledge graph and selecting the entity with the highest degree as the correct entity. This would be a simple baseline for the project. Subsequently, we explore more complicated graph based ranking algorithms and some of the deep learning methods mentioned within the literature review such as Graph Convolutional Network.

We note that by leveraging knowledge graphs in NED, we have essentially identified the relation between entities mentioned within the text. As such, in a typical entity information extraction pipeline, we have already completed both entity tagging, and relation extraction. A natural extension would then be to identify all the coreferences within the text. Due to constraints, we may not be able to do so. However, if time permits, we would explore coreference resolution to complete a full information extraction pipeline that could be used by Kensho on any textual data.

We build our dataset using the *Wikimedia* data. Specifically, given a complete set of text (for example, 3 sentences), we can create several data points for the hyperlinks within the text. Note that each hyperlink is assumed to be a specific entity. Each data point would be defined as (complete text, entity name, correct Wikidata ID, incorrect Wikidata ID). We include incorrect Wikidata IDs in the data set as the disambiguation task requires distinguishing the correct Wikidata ID from the incorrect Wikidata IDs. As such, we need to include incorrect Wikidata IDs for the training of the model. These Wikidata IDs are chosen to not be trivial, and will be based upon a distance/similarity metric such that we choose the Wikidata IDs that are closest to the correct Wikidata IDs.

Having built our model on *Wikimedia* data, we can then explore the generalizability of the model on other textual data, such as the AIDA-CoNLL dataset, which seems to be the standard dataset for NED purposes. Subsequently, if we find finance-specific NED datasets, we would like to explore the applicability of our model on financial data/articles.

Timeline

During phase 3 (model prototyping) from Oct 15th - Nov 8th, we'd like to finish the ideas we proposed on a smaller portion of the dataset for faster iterations. Since we already finished our traditional approach baseline, we'd be mainly experimenting on deep learning approaches for entity linking. By phase 4, we expect to have a thorough comparison between the solutions we proposed earlier. At this stage, we'd focus on retraining the best approach on the full dataset and compare it against the literature results. While doing this, we'd reorganize our deliverables and properly document them to hand them off to Kensho. The article for Towards Data Science will also be slowly refined as we wrap up.

Reference

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the tac 2010 knowledge base population track. In In Third Text Analysis Conference (TAC).

Li Deng and Yang Liu. 2018. Deep Learning in Natural Language Processing (1st ed.). Springer Publishing Company, Incorporated.

Guo, Y., Che, W., Liu, T. and Li, S., 2011, November. A graph-based method for entity linking. In Proceedings of 5th International Joint Conference on Natural Language Processing (pp. 1010-1018).

Cetoli, A., Akbari, M., Bragaglia, S., O'Harney, A.D. and Sloan, M., 2018. Named Entity Disambiguation using Deep Learning on Graphs. arXiv preprint arXiv:1810.09164.

Singh, S., Riedel, S., Martin, B., Zheng, J. and McCallum, A., 2013, October. Joint inference of entities, relations, and coreference. In Proceedings of the 2013 workshop on Automated knowledge base construction (pp. 1-6). ACM.