Midterm 1 - Progress Report

Due to limited computing resources, we experiment with our models on a small sample of 5000 Wikipedia page introductions. We used Wikipedia page introductions since they usually have the highest density of anchor links. Subsequently, we will expand to the full dataset after we have finalized our models. Currently, we have completed the baseline model. The results are as detailed below.

In the NER stage of our base model, we saw that NER was able to identify 85% of the entities in our sample dataset. Note that we went with a less strict approach: As long as there is any overlap in an identified entity and an actual entity, we consider this correct. We then discarded all the entities which were not detected by NER for subsequent stages.

In the candidate selection stage, we were able to correctly include the correct entity in the candidates list 53% of the time. Currently, we are only doing a form of exact matching, where we account for stop words, capitalization, and punctuation. As such, there seems to be a lot of work for improvement in this step. Because of cascading errors, we report different evaluation metrics - at the overall pipeline result and at the performance of each individual component independent of others, by readjusting for the errors from the previous input component.

In the entity linking stage, we only retrieve direct relations between all the candidate wikidata items and then return the top 10 highest degree centrality wikidata items for each identified entity. Out of the NER identified entities, we see the top 1 item is the correct entity 26% of the time, the top 5 items contain the correct entity 40% of the time, and the top 10 items contain the correct entity 47% of the time. This translates to 48%, 74%, and 86% if we were to evaluate NEL independent of candidate selection stage, i.e. accounting only for cases the candidate selection is able to generate the correct entities. The evaluation of the top 10 items and adjusting for candidate selection gives us a rough upper bound of our models' performance, i.e. assuming that NEL is significantly improved along with a perfect candidate selection we should expect about 86% recall.

To evaluate the relative performance of our pipeline, we compare against End-to-End Neural Entity Linking by Kolitsas et al. (CoNLL 2018), a state-of-art model on Gerbil benchmark. Using this pretrained model off-the-shelf, the top 1 item contains the correct entity 47% of the time on this sample dataset, compared to the 26% from our pipeline. Note that this value is adjusted by only accounting for the entities that the model is able to detect. Without the readjustment, this model only obtains 32.5%. This indicates that there's a lot of room for improvement for our model in terms of cascading errors. In addition, the way Wikipedia entities are annotated with references is not consistent with how we perform NER leading to low recall on this dataset when accounting for the overall performance without readjusting anything.

The baseline model gives us a good idea of how a naive method performs on entity disambiguation. In particular, we note that the baseline model performs badly mainly because of cascading errors, particularly at the candidate selection stage. Subsequently, we will begin exploration of deep-learning approaches, using text and graphical embeddings.