Project 28: Word Sense Disambiguation Using Wikipedia

The project consists of implementing a new scheme of word disambiguation using Python NTLK and Wikipedia. Typically, a target word like “chair” will have different entries (clickable) that correspond to various senses associated with “chair”. You can also search “chair (disambiguation)” and this should guide you to different senses associated with “chair”. We shall consider initially the sentence S: “I was awarded a chair in computer science”

1) We would like to disambiguate the target word “plant” in sentence S using Wikipedia entries and compare this with WordNet. First, use a simple Lesk implementation to find and display the sense associated with the target word in S.

2) We would like to use Wikipedia as the main lexical database, use either Wikipedia python API or Wikipedia dump (large-scale download) and write a script that identifies the various searches & displays the various entries associated with the query “chair” (or you can also directly query the outcome of the query “chair (disambiguation)”. Output the various senses. Next, calculate the amount of overlap between each sense gloss and the context word of the sentence similarly to simplified Lesk’s approach. Display the sense that yields the highest overlapping.

3) Consider Unified WSD dataset Word Sense Disambiguation: A Unified Evaluation Framework (uniroma1.it), providing 7,253 test instances for 4,363 sense types. Select a small sample of 10 target words to disambiguate from the annotated dataset, and then apply the method in 2) to perform the word sense disambiguation. Next, select another 10 target words at random and repeat the process. Summarize the result in a table in terms of accuracy and comment on the findings, elaborating on the limitation of the Wikipedia-based approach.

4) We want to utilize a more state-of-the-art approach for disambiguation. Study the BERT disambiguation available in GitHub - danlou/bert-disambiguation: Code and CoarseWSD-20 datasets for "Language Models and Word Sense Disambiguation: An Overview and Analysis", and show its application on the same dataset employed in 3). Comment on the robustness of the result and complexity of the data processing pipeline.

5) We want to test the BERT disambiguation model on another dataset, for this purpose, consider the AQUAINT dataset, and restrict only to the first 50 news documents. The

database is available at https://catalog.ldc.upenn.edu/LDC2002T31. Identify a target of your choice and use your script developed in 2) to disambiguate the target word. Elaborate on the result of the BERT embedding model on the AQUAINT dataset.

6) Consider the WikiSim framework available from https://github.com/asajadi/wikisim. Install and Test the above package on your computer and show that you can run simple similarity between pairs, and wordsense disambiguation. Report the evaluation results for the Senseeval-2 dataset. To comprehend the approach you can study their paper "Vector Space Representation of Concepts Using Wikipedia Graph Structure" – Alternative strategies should be discussed with your teacher if failed to execute Wikisim

7) We would like to use the vector representation generated by WikiSim to evaluate the similarity with human judgment on the selected dataset. For this purpose, refer to the above-mentioned paper, take into the general dataset (MC, RG, WS353) for which the ground truth is known, and compute the correlation of the words of each pair of the dataset with that of human judgment. Compare this with word2vec, Glove, and FastText correlation values.

8) We would like to generate a new vector given as a convex combination of WikiSim generated vector and FasText vector. Start with a convex combination of factor 0.5 and compute the new correlation value for the aforementioned dataset. Then start to see various combinations of the convex combination to see if any improvement in the results occurs.

9) Use appropriate literature to comment on the findings. Also, identify any additional input that would allow you to further elucidate any of the preceding, and use appropriate literature of corpus linguistic literature to justify your findings and comment on the obtained results. Finally, comment on the limitations and structural weakness of the data processing pipeline.