COMP 550: PROGRAMMING ASSIGNMENT 2

JAN TIEGGES NOVEMBER 22, 2023

1 Introduction

This report presents a study on the disambiguation of word meanings using four different methods: Baseline of frequent meanings, Lesk algorithm, BERT model and a bootstrapping approach. The report dives into potential reasons of the results and discusses the limitations of the methods.

2 Datasets

Within the scope of this work, a total of 3 data sets were used. The first data set contains the words to be disambiguated. It is divided into a dev set with 194 samples and a test set with 1450 samples. In total, the dev set contains 128 individual lemmas and the test set 682. The second data set is WordNet, which is used in two of the presented methods to determine the meanings of the words, once by embedding the synsets and once to create a seed set. Finally, the Brown Corpus is used, which serves as an unlabeled data set.

As part of preprocessing, all words are converted to lower case, all punctuation is removed, the stop words removed and lemmas of the words are determined.

3 Methodology

A total of 4 different methods for word disambiguation were examined, which differ fundamentally. The baseline method is the most frequent sense, defined as the most frequent meaning of a word in the WordNet database. Secondly, the Lesk algorithm is used, which determines the meaning of a word based on the meaning of the surrounding words.

The third approach uses BERT, a state-of-the-art pre-trained language model. The pre-training enables them to gain a general understanding of language, which can be used for word disambiguation in a variety of ways. In this work, an approach was chosen that utilizes the generated word or sentence embeddings. More precisely, of all the synsets in question for the word to be disambiguated, the definition and an example of a synset are concatenated and embedded and the cosine similarity to the word embedding, given the context, is calculated. The synset with the highest similarity is then selected as the meaning of the word.

The last method involves the bootstrapping strategy. Here, a small training set is first created by selecting a certain number of lemmas. For each of these lemmas, the definitions, examples and examples of hypernyms, hyponyms and synonyms are then used to create training samples for the most frequent sense. This seed set is then used to train a multinomial naive bayes classifier (NB), which is then applied to the Brown corpus. The samples that have a prediction that is above a certain confidence threshold are then added to the seed set and the NB is trained again. This process is repeated for a certain number of iterations.

For some of the relevant parameters in this context, such as the number of lemmas, the number of iterations or the level of the confidence threshold, a grid search using the development set was carried out to determine the optimal parameters, resulting in the values 50, 3 and 0.7 respectively. Different values were also tested for the smoothing parameter of the NB and whether the prior should be used or not, after which the values 1.0 (Laplace smoothing) and False proved to be the best.

4 Results

The results reveal a clear winner, namely the most frequent sense baseline with an accuracy of 0.6234. Of the other three methods, the BERT model performs best with an accuracy of 0.4110, closely followed by the Lesk algorithm with an accuracy of 0.3393. The bootstrapping method performs worst with an accuracy of 0.0690.

5 Discussion & Conclusion

Even if the results seem surprising at first, it is worth taking a closer look at the predictions of the models. The Lesk algorithm generally uses the overlap of word definitions from a dictionary, with the context in which a word occurs, to disambiguate its meaning. In complex contexts, however, this method is not sufficient. For example, in the context of "Baskonia gives away a victory in Israel", the word "Israel" refers to an ancient kingdom rather than the nation we know today, which the algorithm did not recognize.

The embeddings of the BERT model are better able to recognize these nuances in context, which is why this sample was predicted correctly. On closer inspection of the misclassifications, however, it becomes clear that Bert has particular problems with words with many different meanings (e.g. 'visit'). The reason could be that the embeddings of the different meanings are very similar, and the use of the definition does not sufficiently reflect the meaning of the word. The most frequent sense baseline often solves these cases by simply predicting the most frequent meaning, which is often correct.

The analysis of the bootstrapping approach is more difficult, as it was only trained with 50 of the occurring lemmas. The accuracy of 0.0690 is therefore only meaningful to a limited extent, as it had no chance to disambiguate the non-occurring lemmas and the maximum achievable accuracy was only 0.3062. Another problem is that the lemma to be disambiguated is only attached to the context, but the model does not actually know exactly which lemma it is supposed to disambiguate, but only sees a large context vector. As a result, it sometimes disambiguates words that should not be disambiguated, which shows the limitations of such simple classification models for word sense disambiguation. However, it also shows that bootstrapping is a valid approach to increase the training data.

Further studies could be carried out with more complex models that heavily benefit from the additional training data, such as neural networks. For the BERT model, the construction of the sense embedding also offers plenty of room for experimentation and improvement, and furthermore, a classifier could be trained on top of it.

To summarize, it the experiments highlighted the strengths and limitations of the different methods and showed that the task of word sense disambiguation is not trivial, but can sometimes be solved well with a simple baseline. However, the results also offer plenty of room for further research and improvement of the models.