Word Sense Induction and Disambiguation using Context Embeddings

Easwaran Ramamurthy
eramamur@andrew.cmu.edu

Devendra Singh Sachan
dsachan@andrew.cmu.edu

Tejus Siddagangaiah
tsiddaga@andrew.cmu.edu

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Problem Definition

Word sense induction (WSI) and word sense disambiguation (WSD) are long standing problems in natural language understanding (NLU) and aim at identifying the correct sense of a polysemic word when used in a sentence. A polysemic word is one which can have multiple meanings when used in different contexts. For example, consider the usage of the word "cold" in the phrases "He has a cold" and "The ice is cold". Specifically, in the first phrase, the usage of the word "cold" points to the common cold and in the second case it points to cold temperature. In WSI, the task is to automatically infer the number of senses of a word while in WSD, the task is to output the correct sense for that word given a fixed set of word senses with their usage. Resolving such disambiguations through WSI and WSD has several applications in fields like information retrieval, machine translation and question answering.

2 Related Work

Recently, word embedding models (Mikolov et al., 2013b) have been gaining popularity in natural language processing tasks like parsing, machine translation and text classification. Such models aim to learn a representation of every word as a vector in a latent space where words that share similar contexts in the corpus are closer to each other. More recently, Yuan et al (Yuan et al 2015) reported state of the art results on a WSD task on the New Oxford American Dictionary (NOAD) using a word embedding representation as input to an LSTM neural network language model followed by a label propagation method on the resulting graph. Trask et al (ICLR 2016) learned embeddings of words and their POS tags to do WSD in which the number of senses of a word is determined by its different POS tags. Neelkantan et al () proposed a Non-Parametric Multi Sense Skip-gram model in which they learn multiple embeddings of polysemous words using context clustering. The approach of Reisinger et al provides a context dependent vector representation of word meaning using clustering for homonymy and polysemy words. We now describe what we propose to achieve through this project building up on the work done in the papers cited above.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

3 Method

Most of these earlier works use a linear combination of tf-idf weighted word embeddings for creating context vectors for words and apply some sort of clustering algorithm such as K-Means to identify the number of senses of a word. We plan to replicate the work of Yuan et al on for WSD tasks and use it as a baseline for our experiments. We will also learn unsupervised feature representation of context by training convolutional neural network (CNN) and LSTM on word embeddings. The prediction will be done using one-hot encoding of the target word. We will evaluate our learned representation on the below mentioned datasets and compare it with the earlier state of the art works on these topics.

4 Dataset Description and Evaluation

- **SCWS:** Stanford's Contextual Word Similarities (SCWS) is a set of 2003 word pairs and their sentential contexts. Each instance in the dataset consists of a pair of words and respective POS tags. The dataset also consists of similarity ratings between the words that are collected from averaging human ratings and also ten specific individual ratings. One possible way to evaluate our model on this dataset is to compute Spearman's rank correlation coefficient between the assigned human ratings and the cosine similarity of the two computed word sense vectors.
- **SemEval 2013, Task-13:** This dataset is drawn from the Open American National Corpus (OANC) which includes text of all genres. It has 50 target lemmas which consists of 20 nouns, 20 verbs and 10 adjectives and has 4664 total instances. It seeks to determine the senses of a word in a fully unsupervised manner. In this, the task is to annotate each instance of a target word with one or more of their senses using either WordNet 3.1 sense inventory or an induced sense inventory. Evaluation can be done using three metrics, namely the Jaccard similarity which can used to measure agreement between senses, positionally-weighted Kendall's Tau similarity to rank senses by their applicability and weighted NDCG to measure agreement with human annotators.
- **SemEval 2015, Task-13:** This dataset consists of tokenized, POS tagged documents from Babelnet in three languages. It contains both named entities and word sense inventories. The task is to annotate all the words with their corresponding senses. To evaluate the performance, we can use precision, recall and F1 metrics.
- **MSH:** The MSH WSD dataset consists of lexical information for medical research publications in PubMed. It is an automatically generated dataset specifically aimed at being a resource for testing WSD algorithms that capture the complexity associated with ambiguous medical terms. It consists of a total of 203 ambiguous words which include 106 ambiguous abbreviations and 88 ambiguous terms and 9 of which are a combination of both. For each ambiguous term or abbreviation, it contains a maximum of 100 labeled instances that include the title and abstract of the publication in MEDLINE and the word sense associated with the ambiguous word. Since the dataset is annotated with labels, simple metrics like accuracy, precision, recall, F1 and auROC can be used to evaluate WSD models on this dataset.

References

Bibliography

[1] Joseph Reisinger and Raymond J Mooney. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics, 2010.