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

Recent years have seen the advent of word embedding models as a popular choice for distributional representation of word meaning. These embedding representations are increasingly being used by the community to improve on NLP tasks such as word sense induction (WSI) and word sense disambiguation (WSD). Existing approaches to solve these problems use hand crafted features to represent and cluster context words. We propose to evaluate current methods and develop a new approach that attempts to learn feature representation of context words using deep learning methods such as convolutional neural networks (CNN) and Long short term memory (LSTM).

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 ambiguities through WSI and WSD has several applications in fields like information retrieval, machine translation and question answering.

2   Related Work

Recently, word embedding models [1] 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 [4] 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 [3] 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 [5] 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[2] 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.

3   Methods

Existing methods use a linear combination of tf-idf weighted word embeddings for creating vectors for context words and apply some sort of clustering algorithm such as K-Means to identify the number of senses of a word. Instead, learning feature representations for context words using deep architectures has been shown to be more efficient in NLU tasks. We plan to replicate the work of Yuan et al [4] for WSD and use it as a baseline for our experiments. We also propose to learn unsupervised feature representations of contexts by training convolutional neural network (CNN) and LSTMs on word embeddings. One-hot encoding of the target words could be used to make predictions. Evaluation of our learned representation could be performed on the below mentioned datasets and compared with the earlier state of the art works on these topics.

4   Dataset Description and Evaluation

SCWS[1]**:** 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[2]**:** 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.

MSH[3]**:** 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.

5   Work division and Timeline

Firstly, we aim to understand and implement the previous work in this area and obtain a baseline on the above datasets. This will be done for the approaches outlined in Yuan et al[4], Neelkantan et al[5] and Trask et al[3] and each team member will take up responsibility for one approach. The aim is to accomplish this work by November 7th, 2016 following which we will work on the above mentioned plan to come up with new approaches, and compare the results with the baselines.

---

[1] http://ai.stanford.edu/ ehhuang/
[2] https://www.cs.york.ac.uk/semeval-2013/task13.html
[3] https://wsd.nlm.nih.gov/collaboration.shtml

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[2] 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.

[3] Andrew Trask, Phil Michalak, and John Liu. sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv preprint arXiv:1511.06388*, 2015.

[4] Dayu Yuan, Ryan Doherty, Julian Richardson, Colin Evans, and Eric Altendorf. Word sense disambiguation with neural language models. *arXiv preprint arXiv:1603.07012*, 2016.

[5] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*, 2015.

[6] Huazheng Wang, Fei Tian, Bin Gao, Jiang Bian, and Tie-Yan Liu. Solving verbal comprehension questions in iq test by knowledge-powered word embedding. *arXiv preprint arXiv:1505.07909*, 2015.

[7] Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. A probabilistic model for learning multi-prototype word embeddings. In *COLING*, pages 151–160, 2014.

[8] Jiwei Li and Dan Jurafsky. Do multi-sense embeddings improve natural language understanding? *arXiv preprint arXiv:1506.01070*, 2015.