An Iterative Approach for Unsupervised Most Frequent Sense Detection using WordNet and Word Embeddings

Kevin Patel and Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
{kevin.patel,pb}@cse.iitb.ac.in

Abstract

Given a word, what is the most frequent sense in which it occurs in a given corpus? Most Frequent Sense (MFS) is a strong baseline for unsupervised word sense disambiguation. If we have large amounts of sense-annotated corpora, MFS can be trivially created. However, sense-annotated corpora are a rarity. In this paper, we propose a method which can compute MFS from raw corpora. Our approach iteratively exploits the semantic congruity among related words in corpus. Our method performs better compared to another similar work.

1 Introduction

Word Sense Disambiguation (WSD) remains to be one of the relatively hard problems in the field of Natural Language Processing. Machine Learning approaches to WSD can be broadly classified into two categories: supervised and unsupervised. Supervised techniques rely on learning patterns from sense-annotated training data. However, such data are hard to come by. SemCor, one of the most common sense-annotated corpus in English language, contains around 700k tokens, 200k of which have been sense-annotated. It is really small as compared to raw corpora such as ukWAC, where the number of tokens is close to 2 billion. On the other hand, unsupervised techniques do not require sense-annotated corpora.

A strong baseline for unsupervised WSD is the Most Frequent Sense (MFS) baseline. While performing sense disambiguation, the baseline completely ignores the context, and simply assigns the most frequent sense to the target word.

Inspite of ignoring context, which is one of the main source of information for performing sense disambiguation, the MFS baseline gives really strong results. This is because of the inherent skew in the sense distribution of the data.

Computing MFS baseline is trivial, if one has access to large amounts of sense-annotated corpora. However, that is not the case as explained earlier. Thus there is a need for uncovering MFS from raw data itself.

Word embeddings collectively refers to the set of language modelling and feature learning techniques, which maps words to real valued vectors (Bengio et al., 2003; Mnih and Hinton, 2007; Collobert and Weston, 2008; Mikolov et al., 2010; Huang et al., 2012; Mikolov et al., 2013a; Mikolov et al., 2013b; Pennington et al., 2014). Do note that most word embedding models only output *one embedding per word*, instead of the ideal case of outputting *one embedding per sense of a word*. Though, some models do exist, which provide one embedding per sense of a word by inferring number of senses either through context clustering approaches (Neelakantan et al., 2015), or by using sense inventory (Chen et al., 2014) . For the rest of this paper, we mean *one embedding per word* models, when we use the phrase word embeddings.

The field of Natural Language Processing is increasingly seeing the use of word embeddings for various problems, and MFS is no exception. Bhingardive et al. (2015) showed that pretrained word embeddings can be used to compute most frequent sense.

In this paper, we propose an iterative approach for extracting most frequent sense of words in a raw corpus. The approach uses word embeddings as an input. Thereby, in order to obtain MFS from some raw corpus, one need to apply the following two steps:

1. Train word embeddings on the raw corpus.
2. Apply our approach on the trained word embeddings.

The key points of this paper are:

- Our work further strengthens the claim by (Bhingardive et al., 2015) that word embeddings indeed capture most frequent sense.
- Our approach outperforms others at the task of MFS extraction.

The rest of the paper is organized as follows: Section 2 describes the related work. Section 3 explains our approach. Section 4.1 details our experimental setup and results. Section 5 provides some error analysis, followed by conclusion and future work.

2 Related Work

Buitelaar and Sacaleanu (2001) present an approach for domain specific sense assignment. They rank GermaNet synsets based on the co-occurrence in domain corpora. Lapata and Brew (2004) acquire predominant sense of verbs. They use Levin's classes as their sense inventory. McCarthy et al. (2007) use a thesaurus automatically constructed from raw textual corpora and the WordNet similarity package to find predominant noun senses automatically. Bhingardive et al. (2015) exploit word embeddings trained on untagged corpora to compute the most frequent sense. Our work is most similar to Bhingardive et al. (2015) owing to our reliance on word embeddings. We therefore evaluate our approach against theirs.

3 Approach

Our approach relies on the semantic congruity of raw text. Consider the following example: Consider the word *cricket* having two senses **sport** and **insect**, and the word *bat* having two senses **sport_instrument** and **reptile**. Then, if in our corpus, we already know that *bat* is in **sport_instrument** sense for most cases, then in order for the corpus to be semantically congruent, the most frequent sense of *cricket* has to be **sport**.

So, in order to find most frequent sense of all words in the vocabulary of the corpus, we start with the word whose sense is already known. So, the approach begins with monosemous words, for which MFS is trivial. Next, it moves on to bisemous words, and uses the monosemous words sense information to detect most frequent sense. Then it moves on to trisemous words, and use the hitherto resolved words for detecting most frequent sense, and so on. Thus the approach **iterates** over the degree of polysemy, and uses the computed MFS of words with degree of polysemy 1 to $n-1$ to compute the MFS of words with degree of polysemy n.

At any point of time, we call the words whose MFS is already established as *tagged words*. For a given word whose MFS is to be computed, we enumerate all senses, and then compute the *vote* for each senses by the top **k** nearest neighbors who are already tagged. The vote is a product of two measures: the cosine similarity (w_i) between the embedding of the current tagged word and the target word, and the wordnet similarity (s_i) between the MFS of current tagged word (which would have been established in the previous iteration), and the sense for which the vote is being computed. The votes are summed over, and the sense with the highest sum is considered to be the Most Frequent Sense of the target word. The basic flow is illustrated in figure 1.

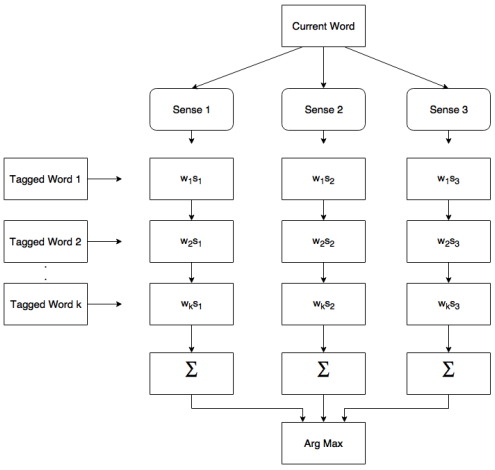

Figure 1: Illustration of our approach

The major parameters in our approach are:

1. K: The number of nearest neighbors who will vote. Through experimentation, we found K=50 to be a reasonable value.
2. WordNet Similarity measure (s_i): We tried all combinations of the six available similarity measures in Princeton WordNet, namely Path similarity, Leacock Chodorow Similarity, Wu Palmer Similarity, Resnik Similarity, Jiang Conrath Similarity, and Lin Similarity.

Our experiments found the average of normalized Wu Palmer and Lin similarity performs slightly better than other combinations.

3. Vector space similarity measure (w_i): We tried both dot and cosine similarity. Dot performed slightly better. In future, we would try other similarity measures such as Tanimoto coefficient.

4 Experiments and Results

4.1 Datasets

We have used the following datasets for our evaluation:

1. SemCor: Sense-annotated corpus, annotated with Princeton WordNet 3.0 senses using WordNet 1.7 to WordNet3.0 mapping by Rada Mihalcea

2. Senseval 2: Sense-annotated corpus, annotated with Princeton WordNet 3.0 senses as above

3. Senseval 3: Sense-annotated corpus, annotated with Princeton WordNet 3.0 senses as above

4.2 Evaluating MFS as solution for WSD

Given that MFS is a strong baseline for unsupervised WSD, a good MFS will give good performance on unsupervised WSD. This is what this experiment evaluates. While in theory, our approach can also use embeddings trained on test corpora directly, we use pretrained word2vec embeddings, as they are crucial to Bhingardive et al. (2015) with whom we are comparing. Table 1 shows the results of using MFS for WSD on Senseval 2 and Senseval 3 only for nouns. We report this noun specific result for comparison with (Bhingardive et al., 2015), who have reported results only for nouns. Here, Bhingardive(reported) and SemCor(reported) are the results as reported in the paper. However, their exact experiment settings are not clear from their paper. Thus we used also computed their results in our setting, which are reported as Bhingardive and SemCor respectively.

In addition to this, we also report the performance on all parts of speech, in table 2. Here, Bhingardive(reported) is the result with the parameter configuration for their approach as reported in their paper. We also tried out different parameter settings for their algorithm, and Bhingardive(optimal) is the best result obtained with optimal parameter setting. It is clear that our approach outperforms both their reported approach and the one with empirically obtained optimal parameters.

Method	**Senseval2**	**Senseval3**
Bhingardive(reported)	52.34	43.28
SemCor(reported)	59.88	65.72
Bhingardive	48.27	36.67
Iterative	63.2	56.72
SemCor	67.61	71.06

Table 1: Accuracy of WSD using MFS (Nouns)

Method	**Senseval2**	**Senseval3**
Bhingardive(reported)	37.79	26.79
Bhingardive(optimal)	43.51	33.78
Iterative	48.1	40.4
SemCor	60.03	60.98

Table 2: Accuracy of WSD using MFS (All Parts of Speech)

4.3 Evaluating MFS as classification task

Another way to evaluate our approach was to learn MFS from pretrained word embeddings which were trained on large corpora, and compare it with WordNet First Sense (WFS). Table 3 shows how our approach fares against Bhingardive et al. (2015)'s when both the approaches are applied on pretrained word2vec embeddings (trained on Google News Dataset with billions of tokens and released by them).

A similar evaluation can also be done by using true MFS obtained from frequencies in sense-annotated corpora. Tables 4 show the results for the same.

5 Discussion

Even though our approach performs better than Bhingardive et al. (2015), we are not able to cross SemCor and WFS results. The following are the reasons for the same:

- There are words for which WFS doesn't give *proper* dominant sense. Consider the following examples:
  - *tiger* - an audacious person

Method	Nouns	Adjectives	Adverbs	Verbs	Total
Bhingardive	43.93	81.79	46.55	37.84	58.75
Iterative	48.27	80.77	46.55	44.32	61.07

Table 3: Percentage match between predicted MFS and WFS

	Nouns (49.20)	Verbs (26.44)	Adjectives (19.22)	Adverbs (5.14)	Total
Bhingardive	29.18	25.57	26.00	33.50	27.83
Iterative	35.46	31.90	30.43	47.78	34.19

Table 4: Percentage match between predicted MFS and true SemCor MFS. Note that numbers in column headers indicate what percent of total words belong to that part of speech

  - *life* - characteristic state or mode of living (social life, city life, real life)
  - *option* - right to buy or sell property at an agreed price
  - *flavor* - general atmosphere of place or situation
  - *season* - period of year marked by special events

- In some cases, the tagged words actually rank very low in order for them to make a significant impact. For instance, while detecting MFS for a bisemous word, it may happen that the first monosemous neighbour actually ranks 1101, *i.e.* a 1000 polysemous words are closer than this monosemous word. Thus in such cases, the monosemous word may not be the one who can influence the MFS.

6 Conclusion

In this paper, we proposed an iterative approach for unsupervised most frequent sense detection in raw corpus. The approach uses word embeddings. Our results bears similar trends to those of Bhingardive et al. (2015), thereby strengthening the claim that word embeddings do indeed capture most frequent sense. Through 2 different categories of experiments, we established that our method is better than theirs. Since there are no language specific restrictions, we believe that our approach should be easily applicable to other languages. In the future, we would like to experimentally validate this claim.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

Sudha Bhingardive, Dhirendra Singh, Rudramurthy V, Hanumant Redkar, and Pushpak Bhattacharyya. 2015. Unsupervised most frequent sense detection using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1238–1243, Denver, Colorado, May–June. Association for Computational Linguistics.

Paul Buitelaar and Bogdan Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proceedings of the WordNet and Other Lexical Resources: Applications, Extensions and Customizations. NAACL Workshop*, Pittsburgh. o.A.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–73.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of pre-

dominant word senses. *Computational Linguistics*, 33(4):553–590.

Tomáš Mikolov, Martin Karafiát, Luk Burget, Jan Cernock, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, pages 1045–1048.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Andriy Mnih and Geoffrey E. Hinton. 2007. Three new graphical models for statistical language modelling. In Zoubin Ghahramani, editor, *ICML*, volume 227 of *ACM International Conference Proceeding Series*, pages 641–648. ACM.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Conference on Empirical Methods in Natural Language Processing, 2014*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12.