Coconut palms and hand palms: improving similarity ranking by word sense disambiguation

Anouk Visser
email

Rémi de Zoeten
email

Abstract

The abstract will be here.

1 Introduction

2 Related Work

The task of word sense disambiguation is to assign the correct sense to an ambiguous word, whereas word sense discrimination is the task of finding the different senses a word might have (Schütze, 1998). The importance of using co-occurrences for determining the correct sense of a word is already emphasized in (Guthrie et al., 1991) in which the authors propose a method for word sense disambiguation using co-occurrences. More specifically, the authors propose a simple score expressing the relatedness between two words:

$$r(x, y) = \frac{f_x y}{f_x + f_y - f_{xy}} \tag{1}$$

where f_{xy} denotes the frequency of x and y occurring together and f_x and f_y denote the frequency of x, respectively y. In the past years a lot of different methods for word sense disambiguation have been proposed that were either supervised, knowledge-based or unsupervised (Navigli, 2009). Unsupervised mainly focus on word sense discrimination, the majority of unsupervised word sense discrimination method uses some sort of clustering of for example context vectors or individual words that have a similar meaning. An example of an unsupervised method that uses clustering on both first order context vectors and second order context vectors can be found in (Purandare and Pedersen, 2004). Another example of unsupervised word sense discrimination is given in (Dinu and Lapata, 2010) where the authors use the intuition that the meaning of a word can be represented as a distribution over a set of latent senses. Recently (Mikolov et al., 2013a) have released tools for efficiently computing word vectors that capture syntactic and semantic information. The distance between these vectors can be used for identifying linguistic regularities (Mikolov et al., 2013b) and a number of other applications such as word sense discrimination. However, word vector representations suffer from the problem that words may have a number of different meanings that cannot be captured in a single representation. (Reisinger and Mooney, 2010) propose a solution to this problem, they represent a word's meaning by a set of sense specific vectors which are discovered by clustering the contexts in which a word appears. For every context cluster, the authors compute an average vector that can be used to determine the similarity between two words (either in context or isolated). The size of the set of sense specific vectors is set in advance and is the same for every word. (Huang et al., 2012) build upon this work by introducing a new neural network architecture that learns word vectors that also incorporate the global context of a word and can lean multiple vectors for a single word. In addition to this, they present a new dataset of pairs of words in contexts annotated with similarity judgements by human annotators.

3 Datasets

We used the blabla task.
For model training data we used the enwiki8 dataset (http://mattmahoney.net/dc/textdata.html, 23 June 2014) corpus. For our purposes we filtered the corpus for

4 COCONUT

The COCONUT method for disambiguating words is based on two assumptions:

1. the meaning of a word is highly dependent on the words it co-occurs with
2. the co-occurring words that define one meaning of a word are more likely to co-occur with

each other than two words that define two different meanings of the word

Let C be the set of words that co-occur with A, the word we want to disambiguate. COCONUT first constructs and converts the global co-occurrence vectors of the words in C to relatedness vectors. It will then cluster these relatedness vectors in order to determine the two possibly different meanings of A.

4.1 Co-Occurrence Vectors

A global co-occurrence vector contains the frequencies indicating how many times two words co-occur, we convert the absolute frequencies in the global co-occurrence vector to a relatedness score. We use the same function for relatedness as (Guthrie et al., 1991).

Words that are not closely related to A do not contribute to either one of the meanings. Therefore, we will discard the words that have a relatedness score with A that falls in the bottom 50% of all relatedness-scores from C. The terms that are discarded are considered relevant to all meanings of A, we will call this set R.

4.2 Clustering and splitting

Let the set of co-occurrence vectors from the words in C, be called V. After applying k-means clustering on the vectors in V we expect to find two cluster centers that represent the two meanings for A. Note that we are only interested in describing the two meanings of A using the words in C. Therefore, for every vector in V we will discard all words that are not in C. The adjusted vectors can now be used to perform k-means clustering.

The two new co-occurrence vectors for A are initialized with the words in R. As the cluster centers define the different meanings of A, we can look at the words in each cluster to fill the new co-occurrence vectors for A.

COCONUT will split every word in the corpus in order to find two different meanings (we excluded the 75 most frequent words), but not all words are ambiguous. We expect that words that have two distinct meanings will have a greater cluster distance (i.e. a greater distance between the two meanings) than words that do not. We discard all disambiguations for the words that have a cluster distance that falls in the bottom 50% of all cluster distances.

5 Agglomerative Clustering

One way to cluster data is agglomerative clustering. Agglomerative clustering is an iterative bottom up approach to clustering. At the start each data point forms its own cluster and in each iteration the two data points that are closest to each other are merged until there is only a given number of clusters or the smallest inter-cluster distance is larger than a predefined value. Agglomerative clustering can be used to cluster words based on their word vector representation.

6 PALM

In order to disambiguate between multiple meanings of a word, it can be useful to look at the context. PALM is a method for word sense discrimination that trains an SVM for every word that given a context predicts a label that describes the meaning of the word. These labels can be used to relabel a corpus before training a recurrent neural network to obtain multiple vector representations for one word. In this section we describe the PALM method in detail, figure 1 provides an overview of the algorithm.

6.1 Choosing the label

Let W be the word for which we want to train the SVM. PALM starts by extracting all contexts that W appears in. We define 'context' as all words within a window around W (in our experiments we looked five words back and five words ahead). Our aim is to assign a label to each of these contexts that describes the sense of the words best, the collection of labels then represent the different sense a word can have. As we have seen in (Jurgens, 2014) underspecified contexts are often observed. In line with our assumption for the COCONUT baseline (i.e. the co-occurring words that define one meaning of a word are more likely to co-occur with each other than two words that define two different meanings of the word) we expand the context by adding the n (in our experiments we set $n = 5$) most related words to every word in the context (except for W). Finally, a word from the expanded context is selected as a label so that:

$$label = \arg\max_{w} \frac{r(W, w) + \mathit{sim}(W, w)}{2}$$

where $\mathit{sim}(W, w)$ denotes the cosine similarity between W and w.

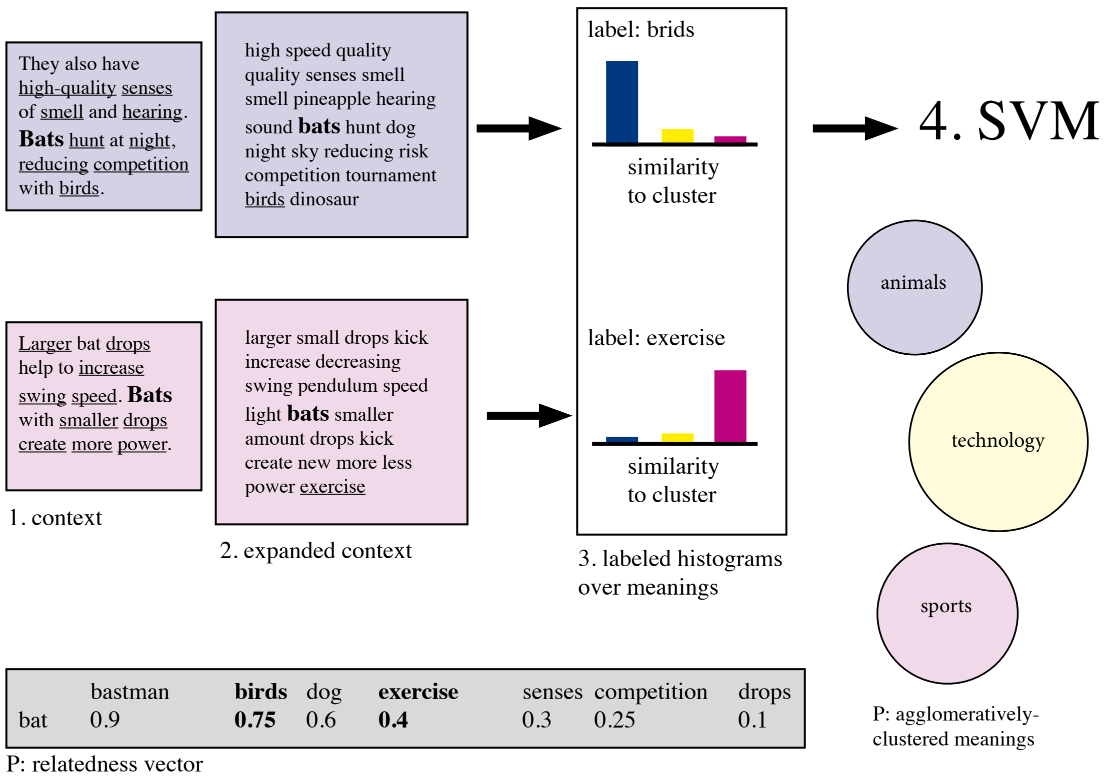

Figure 1: The five steps of the PALM algorithm. P: preprocessing, PALM requires agglomeratively-clusterd latent meanings and relatedness vectors for all words in the corpus. 1. Extract every context word W appears in. 2. Expand the context. 3. Choose a label from the expanded context and create a 'histogram' over the agglomeratively-clusterd latent meanings for every expanded corpus. 4. Train an SVM on these histograms using the labels.

6.2 Probability distribution over agglomeratively-clustered latent meanings

For every word in the expanded context we find the similarity to every agglomeratively-clustered latent meaning, creating an N dimensional vector (where N is the number of agglomeratively-clustered latent meanings that were found). These vectors are accumulated and normalized, so that we have one vector for every context. Although the values denote the accumulated similarity of the words from the expanded context to the meanings, we can interpret the vector as a probability distribution over meanings, given the context (when the context and a meaning are very similar, it is very likely that the context imposes this meaning).

6.3 Label reduction

The last step before training the SVM consists of reducing the number of labels. Although including the second-order context by expanding the context radically decreases the number of new labels, there will still be many labels including a number of labels that describe the same sense of W. We have implemented a method for label reduction that favors labels that were observed most frequently. We iteratively split the labels into two halves: the upper half (containing labels that were seen most frequently) and the lower half (containing labels that were seen least frequently) and merge a label (and all of its vectors) from the lower half (w_l) into a label form the upper half (w_u). w_l and w_u are chosen by:

$$\arg\max_{w_l, w_u} \textit{sim}(w_l, w_u)$$

This process continues until $\textit{sim}(w_l, w_u)$ is below a certain threshold (in our experiments the threshold was 0.5). The reordered labelled data can then be used to train an SVM. The SVM can then disambiguate a word by predicting a the most appropriate label given the probability distribution from an expanded context over the agglomeratively-clustered latent meanings.

7 Experiments

8 Qualitative evaluation

9 Quantitative evaluation

10 Discussion and future work

11 Conclusion

References

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172. Association for Computational Linguistics.

Joe A Guthrie, Louise Guthrie, Yorick Wilks, and Homa Aidinejad. 1991. Subject-dependent co-occurrence and word sense disambiguation. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 146–152. Association for Computational Linguistics.

http://mattmahoney.net/dc/textdata.html. 23 June 2014.

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*, pages 873–882. Association for Computational Linguistics.

David Jurgens. 2014. An analysis of ambiguity in word sense annotations. In *Proceedings of LREC*.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751. Citeseer.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Amruta Purandare and Ted Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning*, volume 72. Boston.

Joseph Reisinger and Raymond J Mooney. 2010. 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.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.