Utility of second-order co-occurrence for word sense discrimination

Cartisan
Institute of Cognitive Science, University of Osnabrück

Seminar: Ambiguity and polysemy
Session: Word sense disambiguation [Sch98]

March 9, 2014

1 Introduction

Automatic word sense discrimination is a clustering based approach for word sense disambiguation presented by Schütze [Sch98]. It operates on a *Word Space* representation of words that is obtained by representing words as vectors of second-order co-occurrences. Schütze analyses several configurations of the algorithm and comes to the conclusion that the best results are obtained by performing a singular vector decomposition (SVD) before clustering.
SVD is an algebraic method for factor analysis that has been successfully used to uncover latent semantic structures ([DDF+90, p.399]) based on first-order co-occurrence, for a variety of natural language processing tasks like information retrieval or text summarization (cf. [DDF+90], [GL01]).
Based on this observations we make the case that SVD is able to resolve the problems of data-sparseness and robustness that Schütze cites as reasons for employing second-order co-occurrence. Furthermore we question whether the approach taken by Automatic word sense discrimination can be motivated by the Strong Contextual Hypothesis, as claimed by Schütze (cf. [Sch98, p.117]).
For this we give a summary of the proposed framework, introduce the mechanics and rationale behind SVD and use this knowledge to elaborate upon the introduced criticism. We describe and conduct an experiment to examine our claims and finally evaluate the results.

2 Automatic word sense discrimination

This section is used to give a condensed overview of Automatic word sense discrimination. For more details refer to the original paper [Sch98].
The Strong Contextual Hypothesis claims that "two words are semantically similar to the extent that their contextual representations are similar" ([MC91, p.8] according to [Sch98, p.117]). Based on this Schütze hypothesises that two senses of an ambiguous word are similar if their context is similar. Thus the general approach for sense discrimination is to compute a fixed set of context clusters for each ambiguous word during training and discriminate the sense of a new occurrence of a word-token by identifying the cluster closest to it. Second order co-occurences are used to prevent issues with data-sparseness and robustness [Sch98, p.116]. From the different evaluated training approaches the most accurate can be described as following:

1. Derive word vectors for the 20.000 most frequent words (features) in a training set by determining

the number of times each of them co-occurs with each of the 2.000 most frequent words (dimensions) in the same set. A word co-occurs with another if it appears in a distance of maximum 25 words from it, leading to a window-size of 50.

2. Derive context vectors for each context of a word by gathering the word vectors of a context and computing their centroid. Word vectors are weighted according to their inversed document frequency.

3. Derive sense vectors for a word by clustering the context vectors in SVD space using the Buckshot clustering algorithm and computing the centroid of each cluster in Word Space. For that apply Buckshot on the dimensionality reduced left-singular matrix computed by the SVD from the context-vector matrix. Set cluster-number to 10 and reduce to 100 dimensions.

Similarity in the vectors space can be calculated using the cosine-similarity measure.

The discrimination algorithm can be described as follows:

1. Compute context vector for the ambiguous word-token.

2. Retrieve all sense vectors for the word.

3. Assign the context-vector to the nearest sense-vector.

3 SVD

Singular vector decomposition computes the factorization of a $m \times n$ matrix A with $m \geq n$ (without loss of generality)[PTVF92] as cited in [GL01, p.21]:

$$A = U\Sigma V^T \tag{1}$$

where U is a $m \times n$ matrix whose columns are called left-singular vectors, Σ an $n \times n$ diagonal matrix whose non-zero elements are singular values sorted in descending order and V^T a $n \times m$ matrix whose columns are called right-singular vectors. This can be used for an optimal approximation of A in a lower dimensional space. By keeping the k biggest singular values in Σ and setting the rest to zero. The representation can be simplified by deleting the zero rows and columns leading to a matrix Σ'. By inserting Σ' intro eq. 1 U and V^T can also be simplified by deleting the corresponding columns or rows respectively. Thus A' can be obtained:

$$A' = U'\Sigma'V'^T \tag{2}$$

It can be shown that A' is the best k-dimensional approximation of A in a least square sense. Also even in case of a sparse matrix A the singular vectors in U' and V'^T are dense [DM00, p.146].
In semantic analysis approaches like [DDF$^+$90], [GL01] SVD is applied on term-document or term-sentence matrices. The semantic interpretation is that the decomposition captures interrelationships between terms and provides a semantic clustering of both dimensions. Recurring and salient concepts have been shown to be captured by the singular values with the importance degree being represented by the magnitude of the singular value ([BDO$^+$95] according to [GL01, p.21]).
Following this interpretation the dimensionality reduction performed in eq. 2 results in a noise reduction on a conceptual level and thus can increase robustness in distributional semantics.

4 Criticism

Using second-order co-occurrence for context representation is combined with a bigger computational load on training and an increased complexity of the algorithm. Other approaches in distributed semantics that rely on SVD to uncover latent semantic structures (e.g. Latent Semantic Indexing for indexing [DDF$^+$90] and Latent Semantic Analysis [GL01] for summarization) exhibit satisfiable results while still employing first-order co-occurrence.
The reasons brought forward by Schütze for his approach, increased robustness and reduced sparseness, are not elaborated upon. While it is clear that sparseness is reduced by representing a word by the neighbours of it's neighbours it remains unclear weather

this reduction is beneficial. As set out above the discriminative power of the context is motivated by the Strong Contextual Hypothesis. This however does not extend any claims towards second-order context. To the best of our knowledge there are no findings that support the assumption that second-order co-occurrence bear cognitive relevance in semantics. Thus the data brought in to reduce sparseness could as well interfere with the disambiguation process as enhance it. Furthermore performing SVD eludes the problem of data sparseness by shifting the semantic processing to the dimensionality reduced singular vector space that is dense in any case.
It seems not immediately clear what definition of robustness Schütze is employing. Robustness might be taken as a measure of the ability of the algorithm to cope with abnormal data i.e. to not exhibit skews on unusual contextual occurrences. In that case the dimensionality reduction performed by SVD is leading to noise-reduction and an increased robustness.
Summarizing the criticism we claim that employing second-order co-occurrence can not be justified by the Strong Contextual Hypothesis and that SVD can be used in order to deal with data-sparseness and to ensure robustness. Because other approaches in distributional semantics have successfully used first-order context we propose the reimplementation of Automatic word sense discrimination on this basis, resulting in shorter training time and less complexity. Our main claim is thus that the discriminative power of Automatic word sense discrimination is based on SVD and not on the usage of second-order context. Empirical proof can be provided if it can be shown that the proposed alternative approach does not exhibit inferior accuracy in comparison to the original approach.

5 Experiment

In order to assess the formulated criticism Schütze's approach has been reimplemented[1] with slight variations and it's performance was compared to an approach relying on first-order co-occurrence. The performance was measured using a subset of the Senseval2 corpus[2] as provided by Python's NLTK framework. The results show no significant difference in the performance of both approaches, however the precision reported by Schütze could not be replicated. Instead both approaches perform below a baseline performance of assigning every occurrence to the most common sense.
A detailed description of the corpus, the reimplementation and the results is presented below.

5.1 Corpus

The Senseval2 corpus contains 15.000 sentences with 600.000 POS-tagged words. Each sentence in the corpus is representing the context of one of the three disambiguated words *line*, *hard* or *serve*. For each of the sentences a sense-tag is provided.
For this study a subset of the corpus was used, consisting only of sentences comprising one of the two most frequent senses for each ambiguous word. In order to avoid skewed training data sentences have been selected in a way to ensure a ratio of the two senses not worse than $5:6$. The resulting sub-corpus contains 3300 sentences and around 122.000 words. Because Senseval uses sentences as contexts and the proposed algorithms rely on word windows, sentences from different ambiguous words were inserted in alternating order to avoid context blending. 80% of the selected sentences are used for training and 20% for testing.
Using a bigger corpus for training and Senseval solely for testing was considered, however analysis of the Brown Corpus showed that the ambiguous words were to scarce.

5.2 Reimplementation

Word-sense discrimination has been implemented with the following modifications:

- Instead of Buckshot the clustering algorithm K-means++ is used, because no Buckshot implementations exist for the chosen programming language. The clustering is executed 20 times

[1] `https://github.com/cartisan/worsed`

[2] `http://www.senseval.org/`

	1st order	2nd order	baseline
line	57.25	65.90	58.63
hard	47.72	58.18	60.45
serve	53.63	48.18	57.25
$\overline{x}$	52.87	0.5742	58.74
s	3.93	7.25	1.32

Table 1: Precision of discrimination task in %, for each word. Baseline performance acquired by assigning all word-tokens to most common sense. Last rows show mean and standard-deviation of sample as used for t-test.

with different seeds and the best result in terms of inertia is chosen as final clustering.

- Instead of using cosine similarity to assign word-tokens to senses, the distance function of the K-means++ implementation is used.
- On context creation no word-weighting using the idf-value is performed because the analysed corpus doesn't support the notion of documents.

An alternative discrimination algorithm relying on first-order co-occurrence is implemented in the same way. The only changing point is that no word-vectors are computed and the computation of context-vectors is based on first-order co-occurrence instead.
Before processing the corpus, stop words are removed and stemming takes place. No mentions of stemming can be found in Schütze, however this is necessary because Senseval2 also contains inflected forms of the ambiguous words.
The parameters were set to:

- dimension number: 400
- feature number: 4000
- dimensions in svd space: 100
- context window radius: 25

5.3 Results

Table 1 shows the precisions for each word using either algorithm. The average precision for both algorithms is below average baseline precision. The difference in average precision between both algorithms is not statistically significant ($p = 0.3942$ using independent two-sample t-test).
This seems to back our claim that the discriminative power of word-sense discrimination is not based on second-order co-occurrence. However the low precision of both approaches indicates that the validity of the results should be questioned. There are several possible reasons for our inability to reproduce Schütze's results:

- The discriminative power of words is not taken into account in our implementation which can lead to lower performance.
- The number of training data-points for each word is 880 and lower than in Schütze's experiment.
- The number of samples is low. In special cases Schütze reported results with low precision in his findings as well. For instance the reported accuracy on the word *plant* was 58% [Sch98, p.110]. It is conceivable that the three words analysed in this study are exemplars of such special cases.
- The results reported by Schütze are based predominantly on nouns. In our case a noun, an adjective and a verb were analysed. The only observed above-baseline performance of 66% accuracy was obtained by the second-order co-occurrence approach on discriminating the noun.
- An assessment of the quality of the sense-tags provided by Senseval was not performed. Low quality sense-tags would result in a low precision on valid discrimination decisions.

6 Conclusion

In this term paper we have questioned the utility of second-order co-occurrences in word-sense discrimination. We argued that the justification presented by Schütze does motivate the use of first-order but does not extend to second-order co-occurrence. Further we hypothesized that the discriminative power

of the examined approach can be attributed to the use of SVD.
An attempt to reproduce the results reported by Schütze has failed and thus a subsequent assessment of an alternative approach relying on first-order co-occurrence and SVD was not meaningful. Before any reliable claims on the utility of second-order co-occurrence in the proposed set-up can be made, the validity of the introduced reimplementation has to be verified. For this purpose it would be helpful to obtain the corpus originally used by Schütze. This would clarify compatibility issues and also allow to take into account the discriminative power of words on context-vector creation. Analysis on how the original implementation performed on verbs and adjectives would help further understand the observed results. What in the end would answer all open questions would be a code review of the original implementation. This would allow to identify further differences in implementation and eventually lead to the identification of accuracy bottlenecks.
Following this credo the code of the reimplementation was commented and open sourced itself, in order to be available for public review.

References

[BDO+95] M. W. Berry, S.T. Dumais, G.W. O'Brien, Michael W. Berry, Susan T. Dumais, and Gavin. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573–595, 1995.

[DDF+90] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.

[DM00] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. In *Machine Learning*, pages 143–175, 2000.

[GL01] Yihong Gong and X Liu. Generic text summarization using relevance measure and latent semantic analysis. pages 19–25, 2001.

[MC91] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6, 1991.

[PTVF92] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. Numerical recipes in c: The art of scientific computing. second edition, 1992.

[Sch98] Hinrich Schütze. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–9123, 1998.