

# Coveting your neighbor’s wife: Using lexical neighborhoods in substitution-based word sense disambiguation

Richard Johansson

Department of Computer Science and Engineering  
University of Gothenburg and Chalmers University of Technology, Sweden  
richard.johansson@gu.se

## Abstract

We explore a simple approach to word sense disambiguation for the case where a graph-structured lexicon of word sense identifiers is available, but no definitions or annotated training examples. The key idea is to consider the *neighborhood* in a lexical graph to generate a set of potential substitutes of the target word, which can then be compared to a set of substitutes suggested by a language model for a given context. We applied the proposed method to the SALDO lexicon for Swedish and used a BERT model to propose contextual substitutes. The system was evaluated on sense-annotated corpora, and despite its simplicity we see a strong improvement over previously proposed models for unsupervised SALDO-based word sense disambiguation.

## 1 Introduction

Probabilistic language models estimate the probability of a word occurring in a given context. This means that for an observed occurrence of a word, a language model can suggest other words – *substitutes* – that could potentially have occurred instead. With a high-quality language model, the set of potential substitutes reflects the *sense* of the word in that specific context. This intuition suggests a simple mechanism for the task of *word sense disambiguation* (WSD) where our goal is to link each occurrence to an item in a fixed sense inventory defined by a lexicon: assuming that the lexicon allows us to generate a set of potential substitutes for each sense, we can then simply compare each of these lists to the one we got from the language model. To disambiguate, we then select the lexicon sense where the substitute set is most similar to the language model’s set of substitutes.

How can we use a lexicon to generate a set of potential substitutes of a given sense? This depends on what information the lexicon represents and how it is structured. In this work, we assume that the lexicon is graph-structured and that proximity in the graph corresponds to substitutability; this assumption allows us to generate a set of potential substitutes of a given sense by considering its *neighborhood* in the graph.

To exemplify, let us assume that we are given the following two occurrences of the Swedish word *ämne* and that we want to associate them with a sense in the SALDO lexicon (Borin et al., 2013):

- (1) *Detta ämne är frätande.*      (2) *Detta ämne kommer att diskuteras senare.*  
‘This substance is corrosive.’      ‘This topic will be discussed later.’

For the first case, the five most probable substitutes suggested by a BERT model are *inhåll* ‘content’, *gift* ‘poison’, *område* ‘area’, *medel* ‘agent’, *föremål* ‘object’; for the second case, they are *område* ‘area’, *problem* ‘problem’, *språk* ‘language’, *tema* ‘theme’, *förslag* ‘proposal’.

We then consider the neighborhoods in the lexicon graph. SALDO defines four senses of *ämne*. Sense 1 corresponds to ‘substance’ and its immediate neighborhood overlaps with the substitute set for the first example: there is an edge in the SALDO graph between sense 1 of *ämne* and sense 1 of *gift*, so we can link the first occurrence to sense 1. Similarly, sense 2 in SALDO corresponds to ‘topic’ and there is an edge between this sense and sense 1 of *tema*, allowing us to disambiguate the second occurrence as well.

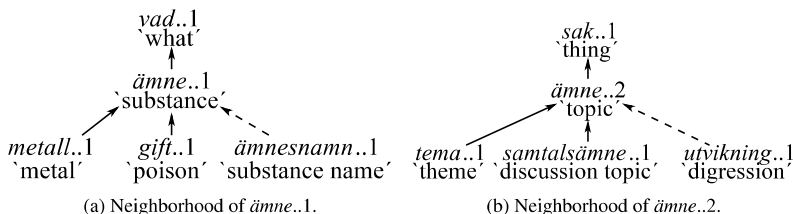


Figure 1: Fragments of SALDO neighborhoods for two of the senses of *ämne*. Primary descriptor edges are drawn as solid arrows and secondary descriptor edges as dashed arrows.

## 2 The SALDO lexicon

The SALDO lexicon (Borin et al., 2013) defines a large sense inventory for Swedish words. While a number of other large-scale lexical resources for Swedish have been developed, SALDO is the largest open resource. It is an extended version of the SAL lexicon (Lönngren, 1989; Borin, 2005) and has been used as a pivot lexicon to define mappings between several lexical-semantic resources in Swedish (Borin, 2010), for instance in the Swedish FrameNet++ project (Borin et al., 2010).

Borin & Forsberg (2009) discuss the conceptual differences between SALDO and WordNet (Fellbaum, 1998). A major difference between these resources is that SALDO tends to use a more coarse-grained sense inventory compared to WordNet. Another fundamental difference is that SALDO does not define typed lexical-semantic relations (e.g. synonymy, is-a, hyponymy) between word senses but instead relies on the notion of *association* (Borin et al., 2013). Association can correspond to several types of lexical-semantic relationships: in many cases, an associated sense can be a synonym or hyperonym, but in other cases it can be e.g. a meronym or be in a predicate–argument relationship.

While each sense could in principle be in an association relationship with many other senses, SALDO explicitly encodes relationships between each sense and its *primary descriptor* (PD): an associated sense that has a more primitive meaning. A few additional relationships are encoded as *secondary descriptors*. SALDO includes no other lexical-semantic information apart from these relations, such as sense definitions or contextual examples. Figure 1 shows the neighborhoods in the SALDO graph around the two senses of *ämne* discussed in the introduction.

## 3 Previous work

Disambiguation systems are implemented in different ways depending on what resources are available. For WordNet-based WSD in English, the most systems tend to use supervised learning because of the availability of moderately large annotated datasets. WordNet is also a fairly rich resource and includes definitions, glosses, as well as several types of labeled sense-to-sense relations. In contrast, SALDO-based WSD is more challenging because of the small quantity of available annotated data and the sparse information in the lexicon. For this reason, most of the WSD systems using SALDO rely on the structure of the lexicon graph only, sometimes in combination with representations learned from unannotated text.

Johansson & Nieto Piña (2015b) proposed a method to align SALDO senses with a word embedding model; this approach naturally leads to a disambiguation mechanism (Johansson & Nieto Piña, 2015a). A tool using this disambiguation method is now integrated in the *Sparv* annotation pipeline (Borin et al., 2016). Nieto Piña & Johansson (2017) used a graph-based regularizer to train word and sense embeddings jointly. Purely graph-based WSD approaches requiring no corpora include graph embeddings using random walks (Nieto Piña & Johansson, 2016b) and personalized PageRank (Agirre & Soroa, 2009).

Nieto Piña & Johansson (2016a) evaluated several WSD systems on all SALDO-annotated corpora that were available at the time. The system by Johansson & Nieto Piña (2015b) was the most effective of those using no training data, but a comparison with a supervised system (on a limited set of target lemmas for which annotated data was available) showed that the unsupervised systems performed relatively poorly.

The idea of disambiguating word senses by using language models to suggest potential substitutes was

first proposed by Başkaya et al. (2013), who applied this approach for WordNet-based WSD as well as for lexicon-free word sense induction (WSI). Subsequent work has mostly focused on WSI: for instance, Amrami & Goldberg (2018) applied a pair of language models to generate substitute sets for WSI.

The same group later used a BERT model for substitute set generation (Amrami & Goldberg, 2019) and this approach is the state of the art in WordNet-based WSI for English as of 2022 (Eyal et al., 2022). The pre-training of BERT (Devlin et al., 2019) involves (among other things) training a *masked language model* (MLM) that tries to predict the identity of a hidden word in a given context, and this aligns perfectly with our goals since a substitute set can then be generated simply by applying the MLM.

## 4 Selecting a SALDO sense for an ambiguous word

Assuming that we are given a context, the position of a word to disambiguate, and a set of SALDO senses to select from, we compute a weighted contextual set of substitute words (§4.1) as well as a weighted word set based on the SALDO neighborhood for each sense (§4.2). We then compute the cosine similarity between the contextual set to each of the SALDO-based sets and select the highest-scoring sense.

### 4.1 Proposing contextual substitutes

We follow the most recent work in substitution-based WSI and apply the MLM of a BERT model. We used the Swedish BERT model published by the Swedish Royal Library (Malmsten et al., 2020). Following Eyal et al. (2022), the MLM is applied in a straightforward manner without masking or modifying the text. We compute the probability distribution at the target position, select the 200 top-scoring items, and exclude inflections of the original target word. The set of potential substitute tokens are weighted proportionally to the probability assigned by the MLM.

While the application of BERT is quite straightforward, the probability distributions are affected by the word piece tokenization. For instance, if a token is followed by a suffix word piece (e.g. `##ar`), the MLM will assign high probabilities mainly to prefixes likely to be followed by this suffix. This likely causes the substitute sets to be of poorer quality for less frequently occurring words and precludes the use of the approach for the disambiguation of multiword expressions. In this work, we simply removed suffix word pieces (starting with `##`) from the set of substitutes; the development of a more systematic approach could potentially be explored in later work.

### 4.2 Extracting neighborhoods from SALDO

We use the neighborhood extraction approach proposed by Nieto Piña & Johansson (2017). For a given SALDO sense, we extract its immediate neighbors in the SALDO graph, following primary and secondary descriptor edges in both directions. Since our goal is to produce a list of words that could potentially be substituted, we only include senses of words of the same grammatical category as the original sense. We repeat the process and add parents, children, and siblings to the set until it has a size of at least 16. Finally, we use the morphological lexicon of SALDO to map every sense to a set of inflected forms, so that e.g. *gift*.1 results in *gift*, *giftet*, ..., *giftens*. The items are assigned weights that depend on the distance in the SALDO graph.

## 5 Experiments

The largest sense-annotated resource for Swedish was developed in the SemTag project (Järborg, 1999); this covers most of the Stockholm–Umeå corpus (Ejerhed et al., 1992). However, this resource does not use SALDO to define its senses, although SALDO has imported some senses from SemTag lexicon. The Swedish lexical sample of the *SENSEVAL-2* shared task (Kokkinakis et al., 2001) used a subset of the SemTag resource consisting of annotation for 40 ambiguous lemmas. The senses for these lemmas were manually mapped to SALDO by Nieto Piña & Johansson (2016a). Since SALDO uses a coarser division into senses than SemTag, three of the lemmas were not ambiguous after this lexicon mapping and they were removed from the dataset. The only running-text corpus annotated with SALDO senses is *Eukalyptus* (Johansson et al., 2016), which includes texts from eight different domains.

Method	SENSEVAL-2	Eukalyptus
<b>Substitutes</b>	0.6675	0.7020
J & NP (2015)	0.4976	–
Random baseline	0.3557	0.4094
Lowest-sense baseline	0.4952	0.6580
Supervised (BoW)	0.8033	–
Supervised (BERT)	0.9209	–

Table 1: Disambiguation results on the test sets for the different methods.

The instances were preprocessed using the *Sparv* pipeline (Borin et al., 2016). For each word, the pipeline proposes a set of possible SALDO senses, based on the automatically determined morphological analysis and lemmatization. The sense disambiguator chooses one of the candidates from this set.

Unambiguous words are excluded from the experiment, which means that the *practical* accuracy is higher than what we report in the next section, since the majority of the words are unambiguous. We also exclude cases where the annotated sense is a non-compositional reading of a multi-word expression (e.g. *på örat* intended as ‘drunk’, not as ‘on the ear’) or a compositional reading of a compound. After this preprocessing, the SENSEVAL-2 sample consists of a test set of 1,366 instances and a training set of 7,790 instances, and the Eukalyptus set of 12,434 instances.

## 5.1 Results

We evaluated the substitute-based approach proposed in this paper and compare it to a number of trivial and nontrivial baselines. Table 1 shows the disambiguation accuracies on the two test sets. The accuracies are macro-averaged over the 37 lemmas for SENSEVAL-2 and micro-averaged for Eukalyptus.

The most meaningful comparison is with the method by Johansson & Nieto Piña (2015a), which is included in *Sparv*: this system uses a similar setup with a combination of the SALDO graph and a representation model trained in an unsupervised fashion. As we can see, the substitute-based method performs much better on the SENSEVAL-2 test set. Both methods outperform two trivial baselines: random selection, and selecting the sense with the lowest numerical identifier. The substitute-based method also outperforms the lowest-sense baseline on the Eukalyptus set.

For SENSEVAL-2, we also evaluate two straightforward supervised approaches that learn from annotated training examples: a linear SVM using a bag-of-words representation, and a MLP on a BERT representation. Both were implemented as “word experts” that use one classifier per base form. All graph-based methods are strongly outperformed by the supervised models. Practically, the supervised approach cannot be applied to Eukalyptus because of the Zipfian distribution of lemmas to disambiguate.

## 6 Discussion

The proposed method works surprisingly well compared to the baselines despite its simplicity. The method is also quite cheap: in the implementation we have described here, we have only used the graph-based neighborhood, although in the general case it may be possible to exploit other lexical-semantic information to generate more accurate substitute sets. No annotated examples for training are needed.

While the performance is better than previous purely graph-based WSD approaches using SALDO, it is much lower than for supervised models in a lexical sample setting. Obviously, a supervised word expert approach is more difficult to apply in a running-text setting, e.g. in Eukalyptus. Another important practical consideration is the flexibility of the substitute-based method: if we add a new sense to the lexicon and update the edges accordingly, we can *immediately* use the new sense in the disambiguator. The method can therefore be argued to be applicable in an interactive fashion.

This is a first attempt and we see a potential for a more careful consideration of the graph-based substitute set, the contextual substitutes, and the way that these sets are compared. The whole idea hinges on being able to use the lexical resource to suggest potential substitutes. For SALDO, this works less

well in some cases where the neighborhood structure does not correspond well to substitutability. Words referring to professions is one such case; cf. the discussion by Johansson (2014).

More generally, we may want to develop methods that align token representations from a language model with a representation of the graph. One might use an embedding of the SALDO graph, either a purely graph-based embedding (Nieto Piña & Johansson, 2016b) or one based on a combination of the graph and a corpus (Johansson & Nieto Piña, 2015b; Nieto Piña & Johansson, 2017). It may then be possible to build a mapping of the BERT-based representation into the space of the embedded graph.

## Acknowledgements

This work builds on resources developed in projects running over several decades at Språkbanken and Uppsala University: the development of the SALDO lexicon and the sense-annotated corpora in the SemTag and Eukalyptus projects. I was funded by the projects *Interpreting and Grounding Pre-trained Representations for NLP* and *Representation Learning for Conversational AI*, both funded by WASP.

## References

- Eneko Agirre & Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athens.
- Asaf Amrami & Yoav Goldberg. 2018. Word Sense Induction with Neural biLM and Symmetric Patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4867, Brussels. Association for Computational Linguistics.
- Asaf Amrami & Yoav Goldberg. 2019. Towards better substitution-based word sense induction. arXiv preprint 1905.12598, <https://arxiv.org/pdf/1905.12598.pdf>.
- Osman Başkaya, Enis Sert, Volkan Cirik, & Deniz Yuret. 2013. AI-KU: Using Substitute Vectors and Co-Occurrence Modeling For Word Sense Induction and Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 300–306, Atlanta, Georgia. Association for Computational Linguistics.
- Lars Borin & Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. NEALT Proceedings Series*, volume 7.
- Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, & Dimitrios Kokkinakis. 2010. The Past Meets the Present in the Swedish FrameNet++. In *Proceedings of EURALEX*.
- Lars Borin, Markus Forsberg, & Lennart Lönngrén. 2013. SALDO: a touch of yin to WordNet’s yang. *Language Resources and Evaluation*, 47(4):1191–1211.
- Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, & Anne Schumacher. 2016. Sparv: Språkbanken’s corpus annotation pipeline infrastructure. In *Swedish Language Technology Conference*, Umeå.
- Lars Borin. 2005. Mannen är faderns mormor: Svenskt associationslexikon reinkarnerat. *LexicoNordica*, 12:39–54.
- Lars Borin. 2010. Med Zipf mot framtiden - en integrerad lexikonresurs för svensk språkteknologi. *LexicoNordica*, 17.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Eva Ejerhed, Gunnel Källgren, Ola Wennstedt, & Magnus Åström. 1992. The linguistic annotation system of the Stockholm-Umeå corpus project – description and guidelines. Technical report, Department of Linguistics, Umeå University.
- Matan Eyal, Shoval Sadde, Hillel Taub-Tabib, & Yoav Goldberg. 2022. Large Scale Substitution-based Word Sense Induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4738–4752, Dublin. Association for Computational Linguistics.

- Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press.
- Richard Johansson & Luis Nieto Piña. 2015a. Combining Relational and Distributional Knowledge for Word Sense Disambiguation. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 69–78, Vilnius. Linköping University Electronic Press.
- Richard Johansson & Luis Nieto Piña. 2015b. Embedding a Semantic Network in a Word Space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1428–1433, Denver.
- Richard Johansson, Yvonne Adesam, Gerlof Bouma, & Karin Hedberg. 2016. A Multi-domain Corpus of Swedish Word Sense Annotation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 3019–3022, Portorož.
- Richard Johansson. 2014. Automatic Expansion of the Swedish FrameNet Lexicon. *Constructions and Frames*, 6(1):92–113.
- Jerker Järborg. 1999. Lexikon i konfrontation. Technical report, University of Gothenburg. Research Reports from the Department of Swedish, Språkdata, GU-ISS-99-6.
- Dimitrios Kokkinakis, Jerker Järborg, & Yvonne Cederholm. 2001. SENSEVAL-2: The Swedish Framework. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 45–48, Toulouse.
- Lennart Lönngren. 1989. Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi. Technical report, Uppsala University. Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi.
- Martin Malmsten, Love Börjesson, & Chris Haffenden. 2020. Playing with words at the National Library of Sweden – Making a Swedish BERT. arXiv preprint 2007.01658, <https://arxiv.org/pdf/2007.01658.pdf>.
- Luis Nieto Piña & Richard Johansson. 2016a. Benchmarking word sense disambiguation systems for Swedish. In *Swedish Language Technology Conference*, Umeå.
- Luis Nieto Piña & Richard Johansson. 2016b. Embedding Senses for Efficient Graph-based Word Sense Disambiguation. In *Proceedings of the 2016 Workshop on Graph-based Methods for Natural Language Processing*, pages 2710–2715, San Diego.
- Luis Nieto Piña & Richard Johansson. 2017. Training Word Sense Embeddings with Lexicon-based Regularization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 284–294, Taipei.