

# Modeling the Evolution of Word Senses with Force-Directed Layouts of Co-occurrence Networks

Robert Schwanhold<sup>1</sup> and Tim Repke<sup>2</sup> and Ralf Krestel<sup>2</sup>

Hasso Plattner Institute

University of Potsdam, Germany

<sup>1</sup> robert.schwanhold@student.hpi.de

<sup>2</sup> fist.lastname@hpi.uni-potsdam.de

## Abstract

Languages evolve over time and the meaning of words can shift. Furthermore, individual words can have multiple senses. However, existing language models often only reflect one word sense per word and do not reflect semantic changes over time. While there are language models that can either model semantic change of words or multiple word senses, none of them cover both aspects simultaneously. We propose a novel force-directed graph layout algorithm to draw a network of frequently co-occurring words. In this way, we are able to use the drawn graph to visualize the evolution of word senses. In addition, we hope that jointly modeling semantic change and multiple senses of words results in improvements for the individual tasks.

## 1 Introduction

Language is dynamic and constantly evolving which leads to changes in the context in which individual words are used and thereby shifting the meaning of words over time. In addition to this semantic change, novel words are introduced or existing words get additional meanings. On the other hand, certain old word meanings can also disappear from active usage in a language. This results in *multiple word senses* per word which in turn can change or shift their *meaning over time*. Current language models typically do not reflect the dynamic and multi-sense aspect of words. There are approaches which tackle one of the aspects, for example, multiple senses (Reisinger and Mooney, 2010) or semantic change (Hamilton et al., 2016).

Static word embeddings, such as word2vec (Mikolov et al., 2013), can only reflect the prevalent meaning a word as it appears in the training data. Contextualized word embeddings, such as BERT (Devlin et al., 2019), circumvent this issue by including the surrounding

words for each usage of the word. However, by using this approach, the representation of a word has to be computed for each time it appears. Furthermore, these models cannot inherently tell which or even how many different senses a word has or how it changed over time.

The boundary between a new word sense and a shift in meaning is blurred. To illustrate this, consider the term “rock”. It has various meanings, e.g., in the context of geology: stone and in the context of music: genre. But those individual meanings are not static. Rock music in the 1960’s is a lot different compared to rock in the 1990’s, for example. Nevertheless, in this case we would argue that the meaning has evolved — the context of usage has changed, and not that there was a new sense added. The problem naturally decomposes into two parts: identifying a sense for a given word in context and tracking the shift in meaning over time.

In this work, we propose a novel data-driven approach that can reflect multiple senses of words as well as how word senses change by jointly modeling different senses over time. We deliberately refrain from defining the senses of a word to be able to also model subtle nuances of different contexts and word usage. To do so, we define a special force-directed graph layout algorithm to align networks of frequently co-occurring words. By modeling words as nodes and connecting co-occurring words via edges, we create a web of language (Dorogovtsev and Mendes, 2001). The algorithm explicitly models multiple word senses by dividing the input data into time slices and duplicating nodes to accommodate changing co-occurrence frequencies. The resulting network layout allows for easy interaction and can be easily explained and understood. This is in contrast to complex embedding models, which function as a black box and are hard to understand intuitively. With this approach, we model the problems of word sense induction and evolution as

a kind of community detection task within a graph. But instead of defining a clustering over the nodes, we propose to visualize the relatedness of words using a force-directed graph layout approach.

## 2 Related Work

Modeling language as a graph has a long tradition (Dorogovtsev and Mendes, 2001; Mihalcea and Radev, 2011; Cong and Liu, 2014; Nastase et al., 2015). We propose to employ word co-occurrence graphs to jointly solve the problems of multiple senses and diachrony. Accordingly, related work can be split into word sense disambiguation, word sense evolution, and approaches that combine both tasks.

Current state-of-the-art models to represent words make use of embeddings. Contextualized word embeddings, such as BERT, account for different word senses by computing individual vectors for a word based on its context. Classical, static word embeddings, such as word2vec, use a single vector to represent an individual word. This is problematic because they fail to capture polysemy. Reisinger and Mooney (2010) presented a multi-prototype vector-space model (VSM). The meaning of a word is represented as a set of sense specific vectors. Based on that, Huang et al. (2012) developed a neural network architecture that learns multiple word embeddings per word. However, both of these approaches use a fixed number of clusters, even though different words might have a different number of senses. Brody and Lapata (2009) use a model based on latent Dirichlet allocation (LDA) to solve the word sense induction (WSI) problem. While this approach uses a fixed number of senses across all words, Lau et al. (2012) combine LDA with a varying number of senses per word. However, this approach requires prior knowledge of the number of senses per word. Hierarchical Dirichlet process (HDP), an extension of LDA, can learn the number of topics (or senses in this case) from the data automatically.

Besides the work on detecting word senses, there is also a plethora of work on diachronic modeling of word senses. Kim et al. (2014) separated a text corpus into multiple time slices and trained a model on each time slice to get different word embedding models over time. Diachronic word embeddings were investigated by aligning embeddings trained on consecutive time slices (Hamilton et al., 2016). Bamler and Mandt (2017) developed

the concept of *dynamic* word embeddings. Each document has a timestamp. This allows the word embeddings to change over time. Unlike previous approaches, a single model is used to derive the shifts of word embeddings over time. One advantage of such an approach is that the complete training data can be used for one single model. While these papers focus on shifts of words over time, they do not discover if a word has multiple senses. Spitz and Gertz (2018) use a network to model the co-occurrence of terms in documents. Terms that are co-occurring together are connected by an edge. Topics are discovered by finding edges of frequently co-occurring terms. For each document, the publication time is stored which allows filtering the results by a given time span. Gad et al. (2015) use a layout with multiple vertical line segments to visualize the trends of topics over time. Each vertical line segment corresponds to a time slice. For each time slice, the topic distribution is calculated. Common terms of the underlying topics are grouped together and plotted on the vertical line segments. This visualization shows how different topics split up or converge over time. Very recently, SemEval-2020 (Schlechtweg et al., 2020) featured a task for unsupervised lexical semantic change detection, which has led to a plethora of diachronic approaches.

Mitra et al. (2014) use co-occurrence networks to find changes in word senses over time. They distinguish between four different types of the evolution of language senses: the birth of new sense; splits of a sense; joins of senses; death of senses. Candidate nodes for splits are computed with a distributed thesaurus. For each candidate node, a clustering algorithm is run on the neighborhood graph. Each cluster represents a sense of the term associated with the candidate node. As shown by Ehmüller et al. (2020) however, matching clusters across more than two or three time slices causes problems such as sense shifting when matching partially overlapping clusters. Hu et al. (2019) use deep contextualized embeddings to track the senses of words over time. For each word, the distribution of the senses is calculated on a temporal slice of the corpus. Over time, these distributions show which senses gain or lose importance. While this approach tracks the senses over time, it does not discover them. Instead, the senses are extracted from the Oxford dictionary.

### 3 Force-Direct Graph Layout Algorithm

In this section we describe our force-directed graph layout algorithm for a network of co-occurring words. In this network, each node corresponds to a word in the vocabulary. We first split the corpus in to disjunct sets of documents based on their publication date to create partial corpora across time. For each set, we compute a network of frequently co-occurring words, where the weighted edges represent the frequency of how often words appear in the same context. In our preliminary experiments, we saw promising results by limiting the vocabulary to nouns and using sentences as context windows. In future work, we intend to compare the raw co-occurrence frequencies to more sophisticated measures, such as pointwise mutual information (PMI). We call the sub-networks for individual time periods *period graphs* and edges within each period graph *intra-edges*. We connect nodes representing identical words in neighboring period graphs with *inter-edges*. All edges are undirected.

**Force-Directed Layout.** Our layout algorithm is inspired by traditional force-directed algorithms. Attractive and repulsive forces are applied on nodes based on their edges and on their proximity to other nodes on a two-dimensional canvas. During the layout process, the positions are iteratively updated to minimize the aggregated forces. Traditionally, nodes are allowed to move freely in both dimensions.

We restrict this layout as follows. We assign each period graph to equidistant vertically aligned parallel axes, which are ordered from left to right according to their time period. Nodes of each period graph are only allowed to move along their respective axis similar to arc diagrams (Saaty, 1964). All other concepts of traditional force-directed layout algorithms remain the same. As two nodes connected by an intra-edge move further apart on the axis of their respective layer graph, the attractive force grows. Repulsive forces between nodes prevent that all nodes are clustered together. Additionally, we introduce another force to reduce the angle of inter-edges.

Figure 1 illustrates a period graph. Initially, nodes are placed randomly along the axis. As a result, some of the edge span long distances. The positions are then iteratively updated until they converge. As shown in Figure 1b, connected compo-

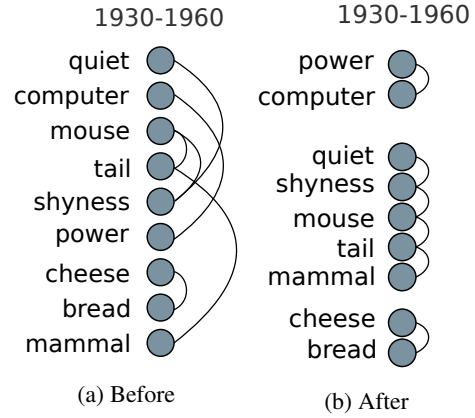


Figure 1: Hand-crafted example of a period graph on initialization and after running our layout algorithm.

nents are clearly separated and all edge lengths are minimal.

Formally, we define the forces between nodes as follows. Let  $V_t$  be the set of nodes of the period graph for time slice  $t$  and  $P_v$  the position along the vertical axis for node  $v$ . The updated position of each node in each period graph in an iteration is defined as

$$P_v := P_v + \psi \left( \alpha F_{intra} + (1 - \alpha) F_{inter} - F_r \right)$$

where  $\psi$  is the learning rate and  $F_{intra}$ ,  $F_{inter}$  and  $F_r$  are the forces between nodes in the current layout. We add  $\alpha$  to balance the attractive forces within and between different period graphs. The forces acting on node  $v$  are defined as

$$F_{intra} := \sum_{u \in N_t(v)} k \times w(\{u, v\}) \times (P_u - P_v)^2.$$

where  $w(\{u, v\})$  is the edge weight and  $N_t(v)$  is the set of nodes directly connected to  $v$  in the current period graph and  $N_{t+1}(v)$  is the set of neighbor nodes of  $v$  from the next period graph. Corresponding nodes in different period graphs are vertically aligned by

$$F_{inter} := \sum_{v' \in N_{t-1}(v) \cup N_{t+1}(v)} \frac{k}{(P_{v'} - P_v)^2}.$$

We use  $k$  as a parameter to control the overall strength of the forces in our system. In physics, this  $k$  is a proportionality constant called *Coulomb's constant* (Gerthsen, 2006). The value of  $k$  is proportional to the electric permittivity of the charged particles in a vacuum. As in other force-directed

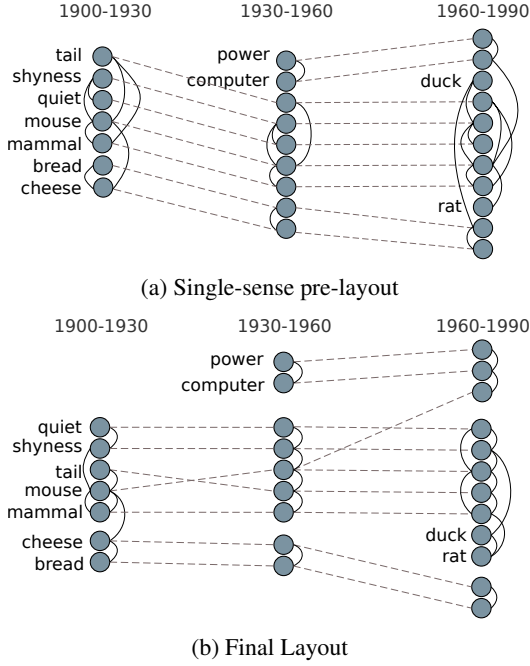


Figure 2: Hand-crafted example to illustrate a resulting layout over three time slices of our proposed approach.

graph layout algorithms, we use a repulsive force to prevent overlapping nodes:

$$F_r := \sum_{u \in V_t} \frac{k}{(P_u - P_v)^2}$$

We limit the calculation of repulsive forces between all pairs of nodes to nodes from the same period graph.

**Representing Multiple Meanings.** Thus far, we described a layout for a graph based on a fixed vocabulary with only one meaning for each word. To reflect multiple senses of a word, we allow the addition of duplicate nodes in a period graph. During the iterative updates of the graph layout, words with multiple senses will cause significantly more stress in the force-directed layout than others. This is due to the fact, that they are associated with different domains, which are likely located far from one another.

We use this to our advantage to discover ambiguous words. First, we run the layout algorithm as described above until it converges. We call the resulting layout our initial layout. In force-based graph drawing algorithms, some nodes induce higher forces on connected or surrounding nodes, causing significant stress in the graph. We identify such nodes duplicate them when the forces of the connecting edges exceed a certain threshold,

which will be determined experimentally. Let node  $v$  be such an ambiguous word, then we split it into two nodes  $v'$  and  $v''$ . The intra-edges that were previously incident to  $v$  are replaced by

$$\forall \hat{v} \in N_t(v) : \begin{cases} (\hat{v}, v'), & \text{if } P_{\hat{v}} > P_v \\ (\hat{v}, v''), & \text{otherwise.} \end{cases}$$

Afterwards, we add inter-edges to connect  $v'$  and  $v''$  to their respective nodes in the previous and following period graphs. This splitting operation can be repeated for the same word again to reflect more than two meanings.

Figure 2 shows an example of the layout before and after adjustment for multiple meanings of words and balancing the forces. Over time, the vocabulary expands and a new meaning of the word “mouse” appears in the context of computers. Note, that in the early days of computing, mice were not used as input devices yet, thus the new sense surfaces only in the last time slice.

## 4 Evaluation Plan

Word sense detection is hard to evaluate given the lack of annotated ground truth data (Usama et al., 2019). General thesauri could be used but only for the period graph for the latest time slice. To our knowledge, there are no established datasets to evaluate both, the multi-sense aspect of a model, as well as the dynamic evolution of senses. Thus, it is necessary to evaluate our approach with respect to both aspects individually and compare results to respective state-of-the-art approaches.

**Evaluation of Word Similarities.** Even though our proposed algorithm focuses on word sense detection, the underlying co-occurrence network can as well be used for other analysis tasks, e.g., word similarity. The vicinity of nodes in a period graph should roughly compare to the neighborhood of vectors in word embeddings trained or fine-tuned on the same set of documents of one time slice.

**Evaluation of the Number of Senses.** The Merriam-Webster dictionary stores metadata for its entries, e.g., a section “First Known Use of . . .”, which lists the year where a sense of that word was first used. Unfortunately, this information does not exist for all entries. However, we can use the existing ones to estimate how well our model performs in finding senses for a specific time period. In addition, manually created thesauri, such as WordNet (Miller, 1995), can also be used.

**Contextualized Word Representations** State-of-the-art embedding models, such as BERT, compute the representation of a word based on the context it appears in. A competitive baseline could be based on contextual word embeddings. Using a pre-trained model, we apply it to each appearance of a word in a corpus. Each meaning of a word should form a cluster of contextual embedding vectors. By doing this for every time slice, we can compare the number of clusters and their similarity neighborhoods to the layout of our graph.

**Qualitative Evaluation of Selected Word Sense Changes.** In a collaboration with digital humanities experts, we developed a use case for a qualitative evaluation by analyzing the different contexts of mentions of natural phenomena in German fiction novels. This allows to qualitatively compare selected parts of our layout to expected changes discussed in relevant literature on digital eco-criticism.

## 5 Conclusion

In this paper, we proposed a novel approach for a multi-sense time-sensitive word similarity model. As it is based on a force-directed graph layout of aligned co-occurrence networks, it allows direct and intuitive interpretation as opposed to most black box embedding models. In future work, we are developing the model further and plan to perform an extensive evaluation as discussed in Section 4. To this end, we will compare our model to existing state-of-the-art language models for word sense disambiguation and evolution, as well as to community detection methods working on graphs.

## References

Robert Bamler and Stephan Mandt. 2017. [Dynamic word embeddings](#). In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389. JMLR Inc. and Microtome Publishing.

Samuel Brody and Mirella Lapata. 2009. [Bayesian word sense induction](#). In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 103–111. Association for Computational Linguistics.

Jin Cong and Haitao Liu. 2014. Approaching human language with complex networks. *Physics of Life Reviews*, 11(4):598–618.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of](#)

[deep bidirectional transformers for language understanding](#). In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186. Association for Computational Linguistics.

Sergey N Dorogovtsev and José Fernando F Mendes. 2001. [Language as an evolving word web](#). *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1485):2603–2606.

Jan Ehmüller, Lasse Kohlmeyer, Holly McKee, Daniel Paeschke, Tim Repke, Ralf Krestel, and Felix Naumann. 2020. [Sense tree: Discovery of new word senses with graph-based scoring](#). In *Proceedings of the Conference "Lernen, Wissen, Daten, Analysen" (LWDA)*, volume 2738 of *CEUR Workshop Proceedings*, pages 246–257. CEUR-WS.org.

Samah Gad, Waqas Javed, Sohaib Ghani, Niklas Elmqvist, E. Thomas Ewing, Keith N. Hampton, and Naren Ramakrishnan. 2015. [Themedelta: Dynamic segmentations over temporal topic models](#). *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 21(5):672–685.

Christian Gerthsen. 2006. *Gerthsen Physik*. Springer-Verlag.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. [Diachronic word embeddings reveal statistical laws of semantic change](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1489–1501. Association for Computational Linguistics.

Renfen Hu, Shen Li, and Shichen Liang. 2019. [Diachronic sense modeling with deep contextualized word embeddings: An ecological view](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3899–3908. Association for Computational Linguistics.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 873–882. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. [Temporal analysis of language through neural language models](#). In *Proceedings of the Workshop on Language Technologies and Computational Social Science@ACL*, pages 61–65. Association for Computational Linguistics.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. [Word sense induction for novel sense detection](#). In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 591–601. Association for Computational Linguistics.

- Rada Mihalcea and Dragomir Radev. 2011. *Graph-based natural language processing and information retrieval*. Cambridge University Press.
- Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119.
- George A. Miller. 1995. [Wordnet: A lexical database for english](#). *Communications of the ACM*, 38(11):39–41.
- Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. [That’s sick dude!: Automatic identification of word sense change across different timescales](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1020–1029. Association for Computational Linguistics.
- Vivi Nastase, Rada Mihalcea, and Dragomir R Radev. 2015. [A survey of graphs in natural language processing](#). *Natural Language Engineering*, 21(5):665–698.
- Joseph Reisinger and Raymond J. Mooney. 2010. [Multi-prototype vector-space models of word meaning](#). In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 109–117. Association for Computational Linguistics.
- Thomas L Saaty. 1964. [The minimum number of intersections in complete graphs](#). *Proceedings of the National Academy of Sciences of the United States of America*, 52(3):688.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [Semeval-2020 task 1: Unsupervised lexical semantic change detection](#). In *Proceedings of the Workshop on Semantic Evaluation*, pages 1–23.
- Andreas Spitz and Michael Gertz. 2018. [Entity-centric topic extraction and exploration: A network-based approach](#). In *Proceedings of the European Conference on Information Retrieval (ECIR)*, volume 10772 of *Lecture Notes in Computer Science*, pages 3–15. Springer.
- Muhammad Usama, Junaid Qadir, Aunn Raza, Hunain Arif, Kok-Lim Alvin Yau, Yehia Elkhatib, Amir Hussain, and Ala I. Al-Fuqaha. 2019. [Unsupervised machine learning for networking: Techniques, applications and research challenges](#). *IEEE Access*, 7:65579–65615.