

# ABSINTH: A small world approach to word sense induction

Victor Zimmermann<sup>1</sup>, Maja Hoffmann<sup>2</sup>

Leipzig University<sup>1</sup>, Leipzig University of Applied Sciences (HTWK)<sup>2</sup>

victor.zimmermann@uni-leipzig.de

maja\_katharina.hoffmann@stud.htwk-leipzig.de

## Abstract

ABSINTH<sup>1</sup> provides a novel unsupervised graph-based approach to word sense induction. This work combines small world coöccurrence networks with a graph propagation algorithm to induce per-word sense assignment vectors over a lexicon that can be aggregated for classification of whole snippets.

## 1 Introduction

As late as twelve years after publication, the graph-based approach to word sense induction proposed in Véronis (2004) was still cited as 'state-of-the-art' (Tripani and Pelillo 2017, Ustalov et al. 2017) and only recently surpassed by neural substitution-based approaches (Amrami and Goldberg 2018, Amrami and Goldberg 2019). Our goal with this work is to evaluate an approach native to small-world graphs for the word sense induction task. We build on the principles laid out in Hyperlex (Véronis, 2004) with a more dynamic feature set and a graph propagation algorithm previously used for sentiment analysis (Hamilton et al., 2016).

Our system, ABSINTH<sup>1</sup>, provides a simple two-step approach to SemEval-2013 Task 11 (Navigli and Vannella, 2013). To achieve this, we utilise the properties of small world graphs for language (Cancho and Solé, 2001) in general and semantic relations (Newman, 2003) in particular. We extract senses using the root hub algorithm proposed in Véronis (2004).

For word sense disambiguation we use the sense inventory created in previous steps and a graph propagation algorithm to assign each node a sense distribution vector. Lastly, the vectors of each word in a given context are summed up and the context is assigned the sense of the best cumulative weight.

<sup>1</sup> Association Based Semantic Induction Tools for root Hub propagation

Parameter	ABSINTH	Hyperlex	Baseline
Min. context	4	4	4
Min. #nodes	Avg. #nodes	10	9
Min. #edges	Avg. #edges	5	3
Max. weight	0.9	0.9	0.9

Table 1: Minimum context size, minimum number of nodes, minimum number of edges and maximum edge weight for our system, Hyperlex and our Baseline.

In addition to the SemEval scoring methods to evaluate our results we use characteristic path length and global clustering coefficient to evaluate the properties of our coöccurrence graphs.

Our system achieves better results in three out of four metrics than a classifier similar to Hyperlex without label propagation.

## 2 Related Work

Graph-based approaches to word sense induction have been successfully used since the early 2000s (Véronis 2004, Di Marco and Navigli 2013, Amplayo et al. 2019). Véronis proposes the use of root hub detection and minimum spanning trees (Kruskal, 1956) to induce senses and disambiguate search results.

The usefulness of small world graph properties for sense disambiguation has previously been shown in Newman (2003). The term 'small world' was introduced by Travers and Milgram, using it to describe the connectedness of acquaintance networks (Travers and Milgram, 1969). According to their findings, the average path length between two people living in the United States lies around five or six, even though they are selected from a relatively large number of people. The properties of these small world graphs have been formally described in Watts and Strogatz (1998). We show that Hyperlex graphs are indeed small world graphs with the words connected in a similar way to real world

relations between people.

Because of this property, nodes with a high degree (number of outgoing edges) can be selected as so called 'root hubs'. It is assumed that words belonging to a sense are clustered around these root hubs and meaning can be induced by mapping a vocabulary to them.

## 2.1 Coöccurrence graphs & root hub detection

Véronis uses paragraphs including the target string (the word or multi-word expression for which senses are to be induced) from a web corpus as contexts for building coöccurrence graphs. Words in the vocabulary constitute nodes and have an undirected edge when they appear in the same context window. Paragraphs with fewer than 4 words are discarded, further limits on nodes, edges and their weights are introduced (see table 1). The target string is not included in the graph.

Edges with a high association frequency are assigned lower weights using a weighting system described in (Véronis, 2004). Why this weighting algorithm is chosen over a more traditional measure like Dice weights is not further explained, but we expect an algorithm using Dice weights would artificially limit the number of possible neighbours for each node and therefore reduce the number of possible root hubs substantially.

Root hubs are chosen iteratively from the set of graph nodes, limited by the following criteria:

1. the number of neighbours, excluding root hubs and neighbours of root hubs,
2. the mean weight of the candidate's most frequent neighbours, excluding root hubs and neighbours of root hubs.

Additionally, the candidate may not be neighbour to a previously chosen root hub.

Before building the minimum spanning tree, the target string is inserted back into the graph with a distance of 0 to each root hub. This results in the root hubs being selected as the direct children of the target string, allowing the easy mapping of components to a hub.

For disambiguation, Véronis iterates over each node  $v$  in the minimum spanning tree and assigns each a weight vector  $\omega$ :

$$\omega_i = \begin{cases} \frac{1}{1+d(h_i,v)}, & \text{if } v \text{ belongs to component } i, \\ 0 & \text{else.} \end{cases}$$

with  $d(h_i, v)$  being the distance between a root hub  $h_i$  and a node  $v$ .

For a given context, the weight vectors of each token are added up and the sense with the highest cumulative weight is chosen.

We use Véronis' root hub algorithm broadly with more flexible parameters for our corpus. Our disambiguation system still uses Hyperlex' minimum spanning tree as a backup, but fundamentally builds on labelled graph propagation (Hamilton et al., 2016).

## 3 Task Set-up

We evaluate our algorithm on Task 11 of the SemEval-2013 Workshop (Navigli and Vannella, 2013). The aim of the task is to develop a word sense induction (WSI) tool that can be used in web search result clustering. The data is structured as follows:

Each topic is given by a target string. For every topic there is a list of the first hundred internet search results, containing information for the result, namely the URL, title and a text snippet (see table 2).

### 3.1 Corpus

We use an unordered plain-text Wikipedia dump from 2014 as context data to construct the word sense graphs which was not supplied with the shared task. As the sense set used in the task is sourced from Wikipedia as well, using Wikipedia for this purpose satisfies domain and style consistency. Because of soft limits on how many nodes and edges ABSINTH considers, an ordered corpus may favour one sense over another based on if its article randomly fell into our sample.

Additionally we add the titles and snippets of each query to our corpus, since it offers us a guaranteed baseline of around 500 nodes per sense.

## 4 Small World Graphs

Our graphs are so called 'small world graphs'. The connection topography of a small world graph, as described in Watts and Strogatz (1998), lies between a completely random and a completely ordered graph. Therefore small world graphs can be highly clustered, but still have relatively short path lengths between the nodes.

The structural properties of these graphs are defined by characteristic path length  $L(p)$  which measures the average separation between nodes of a graph

ID	47.6
url	<a href="http://us.imdb.com/title/tt0120169/">http://us.imdb.com/title/tt0120169/</a>
title	Soul Food (1997)
snippet	Directed by George Tillman Jr.. With Vanessa Williams, Vivica A. Fox, Nia Long. ...

Table 2: Example dataset entry for 'soul food'.

Target	$L_{sys}$	$C_{sys}$	$L_{rand}$	$C_{rand}$
cool_water	3.675	0.528	6.025	0.030
soul_food	4.664	0.604	4.992	0.022
stephen_king	3.649	0.552	3.791	0.014
the_block	3.905	0.329	3.721	0.006
Average	3.973	0.503	4.632	0.018

Table 3: Characteristic path length (L) and global clustering coefficient (C) for our system and a random graph.

and global clustering coefficient  $C(p)$  which measures the cliquishness of a typical neighbourhood. The global clustering coefficient ranges between 0 (for a completely disconnected graph) and 1 (for a highly connected graph). Characteristic path length and global clustering coefficient are calculated as follows:

$$L = \frac{1}{N} \sum_{i=1}^N d_{min}(i, j)$$

$$C = \frac{1}{N} \sum_{i=1}^N \frac{|E(\Gamma(i))|}{\binom{|\Gamma(i)|}{2}},$$

with node count ( $N$ ), the shortest distance between two nodes  $i, j$  ( $d_{min}(i, j)$ ), degree of a node  $i$  ( $|\Gamma(i)|$ ) and proportion of connection between neighbours  $\Gamma(i)$  of a node  $i$  ( $E(\Gamma(i))$ ). To determine whether a graph is indeed a small world graph,  $L(p)$  and  $C(p)$  have to be evaluated against a random connection topography of a graph of the same size.

The random measures are calculated as follows:

$$L_{rand} \sim \log(N)/\log(k)$$

$$C_{rand} \sim 2k/N.$$

A small world graph is defined as follows (Véronis, 2004):

$$L \sim L_{rand}$$

$$C \gg C_{rand}.$$

As can be seen in table 3, our graphs resemble small world graphs, as they feature short average

path lengths, but substantially higher clustering coefficients, compared to what would be expected of random graphs.

Véronis uses these properties mostly for root hub detection. We included a graph propagation system for disambiguation that utilises these graph properties as well.

Because our corpus is much less balanced than Véronis (2004) and our task is more varied<sup>2</sup>, we use a more flexible set of parameters and methods. The task set-up does not support the use of heuristic variables, as some terms are simply too infrequently represented in our corpus to build meaningful graph representations. While setting the euclidean mean of node/edge frequency as a minimum offers a solution to the problem of sparse graphs for less represented terms, more frequent terms seem to over-generate root hubs.

Graph propagation offers a simple method in reducing the total number of senses by essentially merging related root hubs, while retaining the characteristic distribution of senses shown in (Véronis, 2004).

## 5 System

The sense induction works with the properties of small world graphs in mind. The degree of certain nodes makes them ideal root hubs from which a sense distribution can be propagated somewhat organically. The work flow of our system can be roughly translated into induction and disambiguation. The goal of the first task is to produce sensible root hubs. These can be more varied and numerous than in Véronis (2004), as ABSINTH merges and shifts the overlying concepts after initial induction. The root hubs do not themselves carry lexicon definitions of meaning, but provide a structure onto definitions can (hopefully) easily map through propagation.

### 5.1 Word Sense Induction

Induction consists of two steps:

<sup>2</sup>Véronis mostly disambiguates highly polysemous terms and no proper names.

Parameter	ABSINTH	Hyperlex
Min. degree	5	6
Max. mean weight	0.8	0.8

Table 4: Minimum degree and maximum mean weight for root hub detection.

1. Construction and weighting of a coöccurrence graph.
2. Inducing root hubs from this graph.

Our graph is constructed in a straightforward approach, only considering paragraphs including our target string. All nouns and verbs of this sub-corpus are counted, with each coöccurrence within a paragraph being an edge. Stop words are filtered, as is the target string itself, after which every paragraph containing less than four relevant tokens is discarded.

Every node or edge whose frequency falls under a certain threshold (see table 1.) is also discarded. ABSINTH uses the average number of occurrences instead of a heuristic measure, as it is robust enough to deal with over-generation of root hubs and our sub-corpora vary in size too considerably to allow heuristic senses without under-generating root hubs for less frequent targets.

The graph is weighted using the following method from (Véronis, 2004):

$$\omega_{a,b} = 1 - \max[p(A|B), p(B|A)], \quad \text{with}$$

$$p(A|B) = f_{A,B}/f_B \quad \text{and}$$

$$p(B|A) = f_{A,B}/f_A.$$

This weighting method is preferred to a measure like Sørensen-Dice-Weight, as it allows root hubs to have many outgoing edges, while their neighbours can each have a meaningful relation to the root hub without the edge being discarded. We use the algorithm shown in Véronis (2004) to detect root hubs, iteratively choosing hubs by their degree and average weight with their most frequent neighbours (see table 4). We then delete the root hub and its neighbours from the graph before selecting the next hub. After no viable candidates are left, the list of root hubs is returned.

## 5.2 Word Sense Disambiguation

For allocating contexts to senses, our system uses the graph and list of root hubs built in previous steps. Again, disambiguation is a two step process,

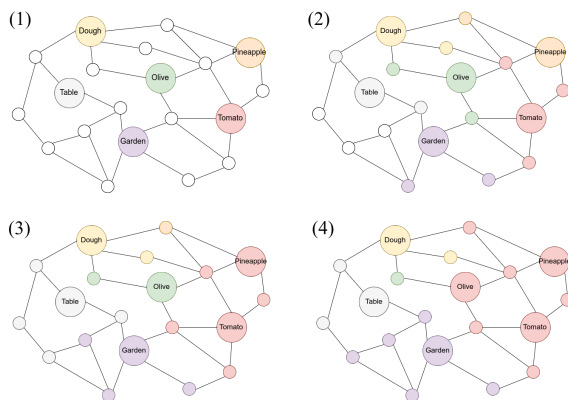


Figure 1: Example of Propagation for the target 'Pizza'.

mirroring the induction process.

First, nodes are labelled according to their 'sense preference' using a propagation algorithm similar to ones used to model voting behaviour (Fowler, 2005) or for sentiment analysis (Newman, 2003). The result is a labelled graph with a sense distribution vector for each node. The best sense of the cumulative vector for a given context is chosen for clustering.

Véronis' algorithm using minimum spanning trees<sup>3</sup> is used as a backup for contexts that could not be matched using the propagation algorithm.

### 5.2.1 Sense Propagation

The goal of our propagation algorithm is to provide an approximation of how indicative a node is for a sense from the root hub inventory. As the sense of a word here is defined by its neighbours, it would follow that whether or not a node is indicative of a sense is also defined by its neighbours. Véronis (2004) offers an algorithm that maps senses to nodes in a binary fashion, but in our understanding a probabilistic distribution would be a more fitting annotation of each node, as this leaves the possibility of a node supporting multiple senses while excluding others, without dividing sense groups.

Our system does not necessarily retain all original root hubs, as they too can be assigned a different sense during iteration (see figure 1). This allows us to over-generate root hubs in earlier steps without much repercussion.

<sup>3</sup>A minimum spanning tree is defined as a sub-graph containing all nodes of the original graph and whose cumulative edge weights are a minimum (Kruskal, 1956).

---

**Algorithm 1** Graph labelling

---

```
1: procedure LABEL_GRAPH
2:    $G \leftarrow$  coöccurrence graph
3:    $H \leftarrow$  list of root hubs
4:    $stable \leftarrow False$ 
5:   for node  $\in G$  do
6:      $node.\omega \leftarrow (\omega_1 \dots \omega_n)$ 
7:      $\omega_1^0 \dots \omega_n^0 \leftarrow 0$ 
8:     if node =  $h \in H$  then
9:        $\omega_h^0 \leftarrow 1$ 
10:     $i \leftarrow 1$ 
11:    while  $stable = False$  do
12:       $stable = True$ 
13:      for node  $\in G, h \in H$  do
14:        for nbr  $\in$  neighbours do
15:          if  $h = \text{argmax}(nbr.\omega)$  then
16:             $\omega_h^i \leftarrow \omega_h^i + (1 - d(node, nbr))$ 
17:             $node.\omega \leftarrow \frac{1}{i+1} \sum_{j=0}^i \omega^j$ 
18:            if  $\text{argmax}(\omega) \neq \text{argmax}(\frac{1}{i} \sum_{j=0}^{i-1} \omega^j)$  then
19:               $stable = False$ 
20:             $i \leftarrow i + 1$ 
return  $G$ 
```

---

Algorithm 1 shows the process in which each node is assigned a sense distribution vector. Notably only the best sense of each neighbour and the weight of their edge<sup>4</sup> ( $d$ ) is considered, not the entire distribution. As our graph is undirected, two conflicting nodes would, should a node’s distribution be based on a neighbour’s own vector, tend to balance each other out, with the graph only reaching a stable state when every connected node features the same distribution, including the same ‘best sense’. This is of course not a desirable outcome.

---

**Algorithm 2** Disambiguation w/ labelled graph

---

```
1: procedure DISAMBIGUATE
2:    $S \leftarrow$  context string
3:    $G \leftarrow$  labelled graph
4:    $H \leftarrow$  list of root hubs
5:    $v \leftarrow$  score vector with length  $H$ 
6:   for token  $\in S$  do
7:     if token  $\in G$  then
8:       for  $h \in H$  do
9:          $v_h \leftarrow v_h + token.\omega_h \cdot \frac{1}{1+d(token, h)}$ 
return  $\text{argmax}(v)$ 
```

---

Our disambiguation algorithm (see algorithm 2) uses a score vector with weights for each root hub. For each token in a given context, the sense distribution vector is added to the score vector, with each sense weight adjusted by the distance of the token to the root hub.

<sup>4</sup>We defined the weight of an edge earlier as the inverted coöccurrence probability. As we aim to match the node to the highest score, we chose to invert the measure back for this step. An *argmin* function would work in much the same way as our method.

ABSINTH retains some binding of a sense to a root hub, using the adjustment to counteract a sense straying too far from its root during the propagation step.

### 5.2.2 Minimum Spanning Tree

Contexts that could not be disambiguated using the propagation algorithm are then processed by the algorithm proposed in Véronis (2004). Target string and root hubs are added to the graph with edge weights of 0. A minimum spanning tree is constructed (Kruskal, 1956) and each node assigned a score in a similar way as above:

$$score_{node} = \frac{1}{1 + d(node, roothub)}$$

Again, the scores for each token in a context are accumulated and the best sense is chosen for clustering.

ABSINTH returns this cumulative mapping of our propagation algorithm, supported by Véronis’ components algorithm.

### 5.3 Baseline

We will be comparing our results to different baselines. Firstly we will use singleton and all-in-one clustering. These are not linguistically or even mathematically motivated clustering methods, our Baseline, which is a more naïve approach to graph based word sense induction, features a basic version of Véronis’ algorithm, but using conceptually simple methods and measures. Instead of the root hub selection algorithm detailed above, the baseline simply selects the ten most frequent nodes as root hubs.

The propagation and minimum spanning tree algorithms are replaced by a distance-based scoring measure. Nodes  $v$  are assigned one-hot-vectors based on distance  $d$  to each root hub  $h \in H$ .

$$\omega_i = \begin{cases} 1, & \text{if } h_i = \text{argmax}_{h \in H}(d(h_i, v)), \\ 0 & \text{else.} \end{cases}$$

The final cumulative score vector for a given context of length  $n$  is essentially comprised of the counts of tokens  $w$  corresponding to each sense. The sense with the highest score is selected:

$$sense = \text{argmax}_{h \in H} \left( \sum_{w \in H} \omega_{w_1}, \dots, \omega_{w_n} \right).$$

## 6 Evaluation

We evaluate on the MORESQUE development training set (Navigli and Crisafulli, 2010), consisting of 114 topics and their according search results.

To evaluate the properties of our cooccurrence graph, we use the characteristic path length and the clustering coefficient (see table 3).

### 6.1 Clustering Quality

SemEval-2013 Task 11 evaluates clustering quality on the basis of the following four metrics:

- F<sub>1</sub>-score,
- Rand index
- adjusted Rand index
- Jaccard index.

Additionally, S-recall at  $K$  and S-precision at  $r$  are measured, as well as the average number of clusters and average cluster size.

## 7 Results

System	F <sub>1</sub>	JI	RI	ARI
ABSINTH	<b>55.21</b>	31.73	54.73	6.98
w/o MST	53.57	33.00	<b>56.21</b>	<b>9.08</b>
w/o labelling	50.13	<b>46.20</b>	53.63	5.51
Baseline	49.87	42.52	51.76	3.26
Singletons	<b>68.66</b>	0.00	49.00	-0.07
All-in-one	47.42	<b>51.00</b>	51.00	0.00

Table 5: Results for F<sub>1</sub>-score, Jaccard index (JI), Rand index (RI) and adjusted Rand index (ARI).

We will compare the results of our system to the results of two different versions of itself. The first variant does not use minimum spanning tree for disambiguation. The second is based on the algorithm proposed in Véronis (2004) and uses the same parameters (w/o labelling). It however is not a one-to-one recreation of the original system, as the corpus used is not extracted from the target URLs. We use these two versions for ablation studies.

System	50	60	70	80
ABSINTH	33.99	22.51	<b>17.78</b>	<b>14.51</b>
w/o MST	<b>36.82</b>	<b>22.98</b>	17.18	13.94
w/o labelling	31.73	20.68	15.83	12.57
Baseline	32.75	22.47	15.21	13.96

Table 6: Subtopic precision at recall  $r$  (S-precision@ $r$ ).

ABSINTH outperforms every baseline on the development data, as expected. The three versions of our system vary heavily in F<sub>1</sub>-score and adjusted Rand index. Our system with propagation algorithm and minimum spanning tree as backup performs well on F<sub>1</sub>-score, but lacks in Jaccard index (see table 5). Our recreation of Hyperlex has the best Jaccard index, but is behind every other system in all other measures. Jaccard index may be biased towards fewer larger clusters, as both our system without labelling and all-in-one clustering perform best in this category. Removing the minimum spanning tree as backup boosts adjusted Rand index significantly, with a smaller bump in Rand index.

System	# cl	ACS
Gold standard	3.98	19.83
ABSINTH	5.39	22.99
w/o MST	4.82	20.61
w/o labelling	1.46	74.81
Baseline	4.54	33.69

Table 7: Average number of clusters (# cl.) and average cluster size (ACS).

The gold standard features a smaller number of clusters with a high average cluster size, which would indicate that the development data may not be an entirely accurate representation of most sense distributions, as other sets have shown to have different distributions (Navigli and Vannella, 2013). We expect better efficacy for Rand index and adjusted Rand index on a different dataset.

We are hesitant to remove Véronis’ components algorithm as backup, as the influence of the minimum spanning tree is only minimal, but it supports our system with a tried and tested approach which may outweigh the efficacy gain indicated on the development set.

The low average cluster count may also have affected the remarkably high efficacy of all-in-one clustering, outperforming every other system in Jaccard index and Rand index by a large margin. We expect this measure to drop significantly when testing on datasets with higher cluster counts.

In terms of precision (see table 6) and recall (see table 8), our full system and our system without minimum spanning tree perform about the same, which is expected due to the small influence the minimum spanning tree has on the results. In both metrics, ABSINTH without label propagation and

dynamic limits trails behind every other version of our system, as well as the baseline.

Across the board, adjusted Rand index has been the most stable measure of the system’s efficacy, with the other measures being more susceptible to changes in cluster size and count. While accurate prediction of number of senses is certainly an important part of the task, we felt overall clustering quality had to be optimised before any reasonable approach in this direction could be taken.

System	5	10	20	40
ABSINTH	51.58	<b>70.32</b>	<b>78.21</b>	<b>88.44</b>
w/o MST	53.46	69.52	77.83	88.21
w/o labelling	<b>55.99</b>	65.77	73.75	84.69
Baseline	55.14	66.25	76.18	87.41

Table 8: Subtopic recall at rank  $K$  (S-recall@ $K$ )

## 8 Conclusion

The similarity of coöccurrence networks and human relations in small world graphs lead to a broad spectrum of possible approaches to optimising a system that had been tried and tested for over a decade. Our system produced solid results on the development data despite the age of the basic components.

Hyperlex has proved to be a very robust baseline on which to build on. Using graph-based algorithms on top of the networks built by Hyperlex could open up interesting avenues for further research and improvement in (non-neural) word sense induction.

Small world graphs, not really a native field of computational linguistic research, have proven themselves quite apt in modelling semantic relations. Even though the graphs built were useful and stable, better results could be obtained by using various sources instead of the Wikipedia corpus. Especially proper names of obscure bands and other pop culture references have posed a challenge to our system which could have been solved with a less information- and more entertainment-based corpus.

As graphs tend to explode with a larger prominence of the target string in the context corpus (see figure 2), parameters such as minimum number of neighbours should be tied to a dependent variable in future work.  $\log(\Gamma(i)) \cdot \Gamma(i)$  was tested, but still

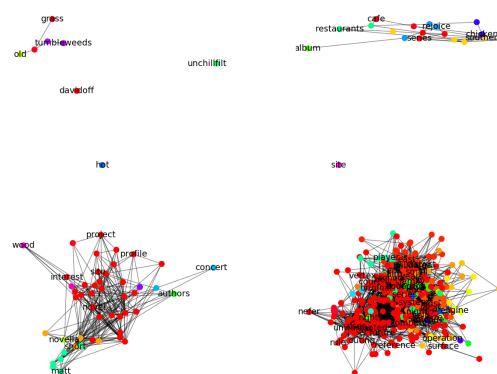


Figure 2: Graphs of different sizes.<sup>6</sup>

performed worse than the heuristic measure<sup>5</sup>.

This small study hints towards the small world property of semantic graph networks opening up a larger world of established tools and methods from intersecting fields of research that can be appropriated and employed for semantic modelling tasks.

## References

- Reinald Kim Amplayo, Seung-won Hwang, and Min Song. 2019. Autosense model for word sense induction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6212–6219.
- Asaf Amrami and 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.
- Asaf Amrami and Yoav Goldberg. 2019. Towards better substitution-based word sense induction. *arXiv preprint arXiv:1905.12598*.
- Ramon Ferrer i Cancho and Richard V. Solé. 2001. *The small world of human language*. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1482):2261–2265.
- Antonio Di Marco and Roberto Navigli. 2013. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754.
- James H Fowler. 2005. Turnout in a small world. *The social logic of politics: Personal networks as contexts for political behavior*, 269.

<sup>5</sup>We lowered the heuristic minimum number of neighbours from 6 to 5 for our system based on limited tests on a subset of the development data, to some minimal improvements.

<sup>6</sup>From top left to bottom right: cool\_water, soul\_food, stephen\_king, the\_block

- William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. [Inducing domain-specific sentiment lexicons from unlabeled corpora](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 595–605.
- Joseph B. Kruskal. 1956. [On the shortest spanning subtree of a graph and the traveling salesman problem](#). *Proceedings of the American Mathematical Society*, 7(1):48–50.
- Roberto Navigli and Giuseppe Crisafulli. 2010. [Inducing word senses to improve web search result clustering](#). In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 116–126.
- Roberto Navigli and Daniele Vannella. 2013. [SemEval-2013 Task 11: Word sense induction and disambiguation within an end-user application](#). In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14-15, 2013*, pages 193–201. The Association for Computer Linguistics.
- M. E. J. Newman. 2003. The structure and function of complex networks. *SIAM REVIEW*, 45:167–256.
- Jeffrey Travers and Stanley Milgram. 1969. An experimental study of the small world problem. *SOCIOLOGY*, 32(4):425–443.
- Rocco Tripodi and Marcello Pelillo. 2017. [A game-theoretic approach to word sense disambiguation](#). *Computational Linguistics*, 43(1):31–70.
- Dmitry Ustalov, Alexander Panchenko, and Chris Biemann. 2017. [Watset: Automatic induction of synsets from a graph of synonyms](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1579–1590, Vancouver, Canada. Association for Computational Linguistics.
- Jean Véronis. 2004. [Hyperlex: lexical cartography for information retrieval](#). *Computer Speech & Language*, 18(3):223–252.
- Duncan J. Watts and Steven H. Strogatz. 1998. [Collective dynamics of 'small-world' networks](#). *Nature*, 393(6684):440–442.