Extracting Semantics from Random Walks on Wikipedia: Comparing Learning and Counting Methods

Alexander Dallmann
dallmann@informatik.uni-wuerzburg.de
Data Mining and Information Retrieval Group
University of Würzburg

Thomas Niebler
niebler@informatik.uni-wuerzburg.de
Data Mining and Information Retrieval Group
University of Würzburg

Florian Lemmerich
florian.lemmerich@gesis.org
GESIS - Leibniz Institute for the Social Sciences

Andreas Hotho
hotho@informatik.uni-wuerzburg.de
Data Mining and Information Retrieval Group
University of Würzburg and L3S Hannover

Abstract

Semantic relatedness between words has been extracted from a variety of sources. In this ongoing work, we explore and compare several options for determining if semantic relatedness can be extracted from navigation structures in Wikipedia. In that direction, we first investigate the potential of representation learning techniques such as *DeepWalk* in comparison to previously applied methods based on counting co-occurrences. Since both methods are based on (random) paths in the network, we also study different approaches to generate paths from Wikipedia link structure. For this task, we do not only consider the link structure of Wikipedia, but also actual navigation behavior of users. Finally, we analyze if semantics can also be extracted from smaller subsets of the Wikipedia link network. As a result we find that representation learning techniques mostly outperform the investigated co-occurrence counting methods on the Wikipedia network. However, we find that this is not the case for paths sampled from human navigation behavior.

Introduction

The *semantic relatedness* between two concepts describes to what degree the actual meanings of these concepts are related to each other. This information can then be used to enhance tag recommendation, ontology learning or query expansion in search engines. Semantic relatedness has been extracted from a wide variety of sources. Some exemplary sources are unstructured text (Mikolov et al. 2013; Deerwester et al. 1990), ontologies (Budanitsky and Hirst 2006) and tagging data (Cattuto et al. 2008). Recent research showed that semantic information can also be extracted from both biased (West, Pineau, and Precup 2009; Singer et al. 2013) and unbiased human navigation (Niebler et al. 2015) on Wikipedia. In this publication, we explore several new options for improving the extraction of semantic relatedness from Wikipedia navigation structures.

Problem Setting To extract semantic information from Wikipedia navigation data, the predominant approach so far was co-occurrence counting, that is, counting page co-occurrences in paths. In other settings such counting approaches have been shown to be outperformed by learning approaches (Marco Baroni, Georgiana Dinu 2014), such as Word2Vec (Mikolov et al. 2013). Additionally, it is not clear if the task of extracting semantic relatedness requires the investigation of the complete Wikipedia link network, or if significantly smaller networks are sufficient. Furthermore, current methods are based on random walks over the link structure. However, randomly following links on a webpage are a very simplistic model of user behavior, which does not necessarily mimic human users. It is an open issue, if actual navigation paths of human users allows to improve the extraction of semantic relatedness from Wikipedia link structures. Currently available datasets are either biased, because they were created in a game setting, or yield only limited information, such as cumulated transitions instead of actual navigation paths.

Approach In this work, we apply DeepWalk (Perozzi, Al-Rfou', and Skiena 2014), an adaptation of the Word2Vec approach by (Mikolov et al. 2013), on random walks across the Wikipedia link network to learn semantic relations between words. In some aspects navigation data is different to unstructured text, e.g., a walk on Wikipedia only consists of different concepts, whereas unstructured text also contains a grammatical structure as well as different parts of speech. Therefore, we compare the learning approach to counting based measures to see if the claim of (Marco Baroni, Georgiana Dinu 2014), i.e., that word embedding approaches outperform counting methods in semantic challenges, holds on navigation data. We evaluate the learned embeddings on the WS-353 dataset, which contains 353 word pairs with human judgment of semantic relatedness. Furthermore, we propose a PageRank-based approach to restrict the Wikipedia link network to a small, but relevant subset of nodes and links to reduce the time for learning vector embeddings, while maintaining the same performance in our semantic evaluation. Finally, we make use of a dataset of human navigation on the Wikipedia link network to parameterize the random walks in order to simulate human behavior in a more realistic way.

Contributions Our contributions are threefold.

1. We compare the performance of concept embeddings and co-occurrence counting on random walks generated on the full Wikipedia link network and find that concept embeddings mostly outperform counting methods in this setting.
2. We implement a scheme to reduce the network size to the k percent most relevant nodes by selecting only the highest ranked nodes according to Pagerank (Brin and Page 1998). We show that the performance of concept embeddings improves on reduced networks due to the reduction in noise.
3. We investigate how incorporating human navigation impacts both algorithms by generating random walks from both the weighted and unweighted network derived from the Clickstream dataset and find that in this case co-ocurrence counting outperforms concept embeddings.

Structure This paper is structured as follows: We first cover related work. After that we give a description and statistics of the datasets used in this paper. In the next section we provide an outline of the methods we use. Then, we describe our experimental setup and show the achieved results. This is followed by a discussion of the findings and a conclusion of this work as well as a collection of ideas for future research.

Related Work

This section covers related work to both deep neural embeddings of graph nodes in vectors as well as semantic relatedness based on the Wikipedia link structure.

Semantic Relatedness on Wikipedia

One of the most well-known works concerning semantic relatedness on Wikipedia was done by (Gabrilovich and Markovitch 2007), where they propose the ESA measure, which calculates semantic relatedness between Wikipedia concepts based on TF-IDF vectors and correlate their resulting relatedness ranking to the WS-353 dataset. However, this method is only based on the article texts and not on any kind of navigation. The potential of Wikipedia's category taxonomy for calculating semantic relatedness is shown by (Strube and Ponzetto 2006) and compared with several baseline approaches using WordNet. They show that methods using the category structure of Wikipedia outperform Google count based methods and a WordNet baseline. However, they obtain the best result using Wikipedia, Google and WordNet in combination. Omitting the Wikipedia category-taxonomy, (Milne and Witten 2008) make use of the static Wikipedia hyperlink structure in order to calculate semantic relatedness, introducing the Wikipedia Link-based Measure. This measure is a combination of a TF-IDF based measure and the Google distance measure, and evaluate a combination of both measures on article-link sets obtained from Wikipedia in comparison to the two methods mentioned above.

(West, Pineau, and Precup 2009) and (Singer et al. 2013) calculated semantic similarities on game navigation from two navigation games on Wikipedia, namely Wikispeedia[1] and the WikiGame[2]. The first uses a probabilistic relatedness measure, which yields results, iff two pages co-occurred in the same path. The second makes use of a sliding window to count co-occurring concepts and this way create a vector representation for each Wikipedia page, so pages can be compared even if they didn't co-occur in a common path. (Niebler et al. 2015) extended the method from (Singer et al. 2013) to unrestrained navigation on Wikipedia. They analysed a real-life dataset with accumulated user navigation on Wikipedia and showed that taking only the binary usage of links (used/not used) into account yielded better results than counting how often a transition has been used, when evaluated for semantic relatedness.

Representation Learning

(Mikolov et al. 2013) presented the Word2Vec approach to calculate continuous word embeddings from raw text. Words are thus associated with points in a feature space, where the spatial distance between these points describes the relation between those words. As an alternative to this approach, the GloVe model does not only focus on local windows, but also takes statistics of the whole corpus into account (Pennington, Socher, and Manning 2014). Based on Word2Vec, Perozzi et al. (Perozzi, Al-Rfou', and Skiena 2014) proposed *DeepWalk*. This method applies Word2Vec on random walks in social networks to learn social representations of entities. In this work, we transfer this technique to learn a vector representation of a Wikipedia article with respect to its position in the Wikipedia network. These vectors can then be used to compute semantic relatedness between articles. (Tang et al. 2015) proposed the LINE algorithm, which addresses the problem of embedding very large information networks with up to millions of nodes into low-dimensional network spaces. LINE also preserves both the local and global properties of the graph, i.e., first-order node proximities (edges) as well as second-order proximities, which are represented by the shared neighborhoods. They perform several experiments on large real-world datasets and outperform DeepWalk as well as SkipGram, while also using less training time. A related approach to our work was presented by (Zhao, Liu, and Sun 2015). They apply DeepWalk on the Chinese Wikipedia and evaluate their findings on a smaller Chinese variant of WS-353. However, there are several key differences to our approach: First, they consider a network not only consisting of Wikipedia pages, but also of Wikipedia categories and even words from the articles. They do not give any configuration parameters and report a Spearman correlation value of about 0.44. Finally, (Marco Baroni, Georgiana Dinu 2014) investigated several relatedness measures for words and showed that prediction models such as Word2Vec and GloVe outperform counting models on word relatedness tasks by a significant margin, when evaluated on WS-353. Still, their experiments have been conducted on unstructured text, in contrast to random walks on Wikipedia.

[1] `http://www.wikispeedia.net`
[2] `http://www.thewikigame.com`

Datasets

In this section we shortly describe the employed datasets, i.e., the link structure of Wikipedia articles, the Clickstream dataset of real user transitions in Wikipedia, and the WordSimilarity-353 dataset for evaluating semantic relatedness.

Wikipedia Link Network

We downloaded the Wikipedia link network dump from February 2015[3]. We use this dataset to generate random walks across the whole of Wikipedia as well as some subsets of the link network. The subsets are created by taking the top-k ranked nodes and their edges according to Pagerank. All random walks generated from these networks strictly represent the network structure and are unbiased by any actual navigation. Table 4 displays basic statistics for this network and the corresponding subsets.

Clickstream

To represent human navigation behavior, we use a dataset generated from the Wikipedia webserver logs in February 2015 (Wulczyn and Taraborelli 2015). This dataset contains an accumulation of transitions with their respective occurrence counts, i.e., how many users used a particular transition between two Wikipedia pages in the whole month. Transitions with less than 10 occurrences have been removed. For more details on the applied pre-processing, we refer to the website of the dataset. For this paper, we only used the transitions with both source and target pages inside the main namespace of Wikipedia. From these transitions, we build a link network, both in a weighted (edges with transition counts) and unweighted variant (edges without transition counts). For some basic statistics for this network, we refer again to Table 4.

Evaluation Dataset

WordSimilarity-353 (WS-353) (Finkelstein et al. 2001) consists of 353 pairs of English words and names, each with a relatedness value between 0.0 (no relation) and 10.0 (identical meaning). These relatedness values have been generated by 16 raters, denoting the assumed common sense semantic relatedness between two words. Finally, the total rating per pair was calculated as the mean of each of the 16 users' ratings. This way, WS-353 provides a valuable evaluation base for comparing our concept relatedness scores to an established human generated and validated collection of word pairs. In this work, we use a mapping of words to pages, which has also been used in (Singer et al. 2013) and (Niebler et al. 2015). The dataset is freely available for download[4]. The mapping of words to Wikipedia pages can also be downloaded[5].

[3] https://archive.org/details/enwiki-20150205

[4] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html

[5] http://www.thomas-niebler.de/wordpress/?p=88

Methods

In the following section all steps of the applied methodology are presented in detail. For all experiments a directed graph is constructed from the datasets described in the last section. A number of random walks are then performed on each graph and used to i) learn concept embeddings by applying the Word2Vec approach and ii) to construct a vector representation by means of co-occurrence counting. The vectors are then used to compute semantic relatedness scores for word pairs and the results are evaluated against the WS-353 dataset.

Random Walk Generation

Following the intuition of Perozzi, Al-Rfou', and Skiena that a path sampled from a graph can be interpreted as a natural language sentence and each node in a path is identified as a word, paths are sampled from the constructed graphs. In order to capture information about every node, a fixed number c of random walks is executed for every node in the graph. To restrict the generated paths to a maximum length of l_{max} the random walk is stopped when the path length reaches l_{max}. With the exception of the *weighted* graph constructed from the Clickstream dataset, the next edge to follow is chosen uniformly random from the outgoing edges of the current node. In the case of the *weighted* graph the edges are chosen with a probability proportional to the weight of the outgoing edges. Since the investigated graphs are directed it is worth pointing out that only nodes with an out-degree $deg_{out} > 0$ are used to start random walks, effectively eliminating the possibility of paths with a length of $l = 0$.

Baseline Approach

In (Singer et al. 2013) and (Niebler et al. 2015), two co-occurrence counting approaches to determine semantic relatedness from navigational paths on Wikipedia have been proposed. Both approaches are based on counting co-occurrences of concepts in a window of size w on paths from WikiGame, resulting in co-occurrence vectors for each concept. Here, a window of size $w = 3$ includes the current node and the two successive nodes. These vectors are compared pairwise using the cosine similarity measure, giving a semantic relatedness value of the corresponding concept pair. In (Niebler et al. 2015), this approach has been modified to only consider whether or not two words co-occur instead of the actual number of co-occurrences. This increased evaluation performance on transition datasets, i.e., a window size w of 2.

Model Learning

(Mikolov et al. 2013) presented the Word2Vec approach, which learns low-dimensional word embeddings from a corpus of unstructured text. Such a word embedding is represented by a low-dimensional vector and is learned by training a neural network to guess the correct context in which the words occurr. Technically, Word2Vec uses the Skip-Gram approach, which maximizes the average log probability of all context words with the current word in a sentence. The context is defined by a window of size w, which

includes w words after and before the current word.[6] The probability itself is defined by a softmax function. Using a relatedness measure, e.g., the cosine measure, semantic relatedness between two words can be computed from these vectors. Word2Vec can be adapted to graphs by identifying words with vertices and paths along the edges as sentences (Perozzi, Al-Rfou', and Skiena 2014). In this setting the neural network learns vectors that express the relation between nodes from the graph structure.

Evaluation

To evaluate our results, we correlate the calculated relatedness values with those in our evaluation dataset. This evaluation method has been used often throughout literature, cf. (Budanitsky and Hirst 2006; Gabrilovich and Markovitch 2007; Milne and Witten 2008; Singer et al. 2013; Niebler et al. 2015). We first determine the overlap of word pairs from the evaluation dataset with the experiment dataset, for which we can calculate a relatedness value. This way, we get two rankings on a subset of evaluation pairs: one with human-assigned relatedness scores and the other with our cosine relatedness values. Since an absolute relatedness value is somewhat abstract and may be interpreted differently even by humans, we use the Spearman correlation coefficent ρ to calculate the relationship between those rankings, because this coefficient only considers the relative placement of pairs in a ranking. A high absolute correlation value near 1 means almost perfect correlation, i.e., that the extracted semantic fits well to human intuition whereas a correlation value near 0 means no correlation. If we cannot find a specific word from the evaluation dataset in our experiment datasets, e.g., because it has not been used, we leave out that word pair, since we cannot calculate a relatedness value for it. For evaluation results from a concept embedding model, we denote the corresponding correlation coefficient with $\rho_{concept}$, while results from a co-occurrence counting approach are described by ρ_{coocc}. For the binary evaluation described in (Niebler et al. 2015), we denote the Spearman correlation coefficient as ρ_{binary}.

Experiments

We present the results of the different experiments in this section. First a short motivation and description for each experiment is given. This is followed by the achieved results.

Experimental Setup

In this part we will provide the settings for all parameters and give justifications for our choices. For learning concept embeddings we use the original tool provided by (Mikolov et al. 2013) [7] and train it on the generated random walks. We set the down-sampling threshold to $1e^{-4}$ and the model is trained using the CBOW model and Hierarchical Softmax instead of Skip-Gram because of better computational efficiency. For the random walk generation the length of a random walk is fixed to $l = 20$ transitions, e.g. a walk contains a maximum of 21 nodes. During a random walk it is possible to reach a node with no neighbours. If that happens, the path cannot be extended anymore, but is still kept in the set of generated paths. In order to obtain comparable results, we restrict ourselves to a subset of 276 word pairs from the WS-353 dataset found in all experiment datasets. Another issue is presented by the different notions of a "window" in both the counting method and the concept embedding approach. For a given path $p := (p_1, \ldots, p_n)$ and window size $w = 2$, the concept embedding model considers all nodes in the window $(p_{i-2}, p_{i-1}, p_i, p_{i+1}, p_{i+2})$, whereas the counting method indirectly only takes the window (p_{i-1}, p_i, p_{i+1}) into account. Because of this, a window size of w in the concept embedding model means a window size $w + 1$ in the counting approch. It is worthwhile to note that training a model several times on the same dataset can yield slightly different results, since the concept embedding model uses stochastic gradient descent to speed up training time, which is non-deterministic. A similar problem arises with the generation of random walks: Since the choice of the successing node is *random*, two supposedly similar datasets are not guaranteed to also yield identical evaluation performance.

Table 1: Spearman correlations achieved with concept embeddings and both baselines by varying the number of walks started on each node in the network. For concept embeddings the vector size was set to $|v| = 128$. The window size was set to $w = 3$.

walks/node	$\rho_{concept}$	ρ_{cooc}	ρ_{binary}
1	0.676	0.668	0.645
5	0.696	0.676	0.594
10	0.698	0.680	0.582
15	0.696	0.681	0.578
20	0.706	0.682	0.573
25	0.694	0.681	0.573
30	0.705	0.681	0.569

Comparison on the Wikipedia Link Network

Influence of Number of Walks The first experiment concerns the number of walks started from each node to accumulate enough information to learn meaningful concept embeddings. We fix the vector size to $|v| = 128$ and the window size to $w = 3$ and increase the number of walks c for each node. Table 1 shows that concept embeddings and co-occurrence counting give largely the same performance for increasing number of walks. The binary co-occurrence counting instead gives worse performance with increasing c. With the exception of $c = 1$ no apparent dependency between the number of walks and the model performance can be observed. Therefore we decided to set the number of walks $c = 10$ for all further experiments. We also calculated a confidence interval using the bootstrapping method on the spearman correlation values of twenty different random

[6] Note that this includes different nodes than the window of the counting approach in the earlier section.

[7] https://code.google.com/archive/p/word2vec/

Table 2: Spearman correlations for concept embeddings learned with increasing vector sizes on the full Wikipedia network. The window size was set to $w = 3$. As a comparison the co-ocurrence counting method with the same window size achieved a spearman correlation of $\rho_{cooc} = 0.680$.

$\|v\|$	8	16	32	64	128	256
$\rho_{concept}$	0.602	0.647	0.699	0.701	0.698	0.684

Table 3: Spearman correlations for concept embeddings and both baselines when varying the window size on the full Wikipedia network. For concept embeddings the vector size was set to $|v| = 128$.

w	1	3	5	7
$\rho_{concept}$	0.700	0.698	0.675	0.700
ρ_{cooc}	0.664	0.680	0.658	0.638
ρ_{binary}	0.725	0.582	0.511	0.467

walk datasets with $c = 10$. This yielded a mean correlation of 0.699 with a 95% confidence interval of $[0.696\,, 0.703]$.

Influence of Window and Vector Size Next we study the influence of vector size $|v|$ on the results for concept embeddings. For this we set the window size $w = 3$ and train models with different vector sizes $|v|$ on random walks generated for the Wikipedia network. Table 2 shows that the spearman correlation increases with growing vector size and reaches a plateau that extends from $vs = 32$ to $vs = 128$. This matches the results of similar experiments carried out in (Perozzi, Al-Rfou', and Skiena 2014). We analogously chose the vector size $|v| = 128$ for all further experiments.

Both co-occurrence counting and concept embeddings depend on a windows size w that acts as a nodes context. To determine the optimal window size for a fair comparison of the approaches we performed an experiment that studied the impact of increasing window sizes on the outcome. For all approaches we computed models with different window sizes $w \in \{1, 3, 5, 7\}$ and compared the spearman correlation results. The results in Table 3 show that the window size does not greatly impact performance of concept embeddings but that the co-occurrence counting approach performs worse with increasing window size. Furthermore the performance of concept embeddings is slightly better than co-occurrence counting when comparing the spearman correlations for $w = 3$. For $w = 1$ the binary counting approach is able to outperform concept embeddings, but the performance quickly declines for larger window sizes. To facilitate a fair comparison between baselines and concept embeddings we set the window size to $w = 3$ in our experiments.

Effects of Network Reduction to Relevant Nodes

With this experiment we want to study how the network size affects the performance of the different approaches. There are many approaches to reducing a network in size but the most intuitive is to only keep the most relevant nodes of a network. To facilitate this we use the Wikipedia network and compute a ranking for all nodes using Pagerank with 20 iterations. The ranking is then used to create a series of pruned networks using only the top-k percent nodes and corresponding edges. Statistics about the obtained networks can be found in Table 4 and Table 5.

Table 4: This table gives an overview of all link networks that we use to generate random walks. The weighted *Clickstream* * and the unweighted *Clickstream* $_{uw}$ datasets have the same average outdegree, since for the outdegree of a vertex, the weight of an edge does not matter. *: Please note that for the Clickstream dataset, the numbers give the actually performed transitions from multiple users, which count a single link multiple times.

dataset	$\|V\|$	$\|E\|$	$\emptyset$outdeg
WikiLink-10%	480,150	71,026,160	147.92
WikiLink-20%	960,300	153,751,696	160.11
WikiLink-30%	1,440,450	199,933,460	138.80
WikiLink-40%	1,920,599	231,956,454	120.77
WikiLink-50%	2,400,747	255,128,647	106.27
WikiLink-60%	2,880,883	271,895,395	94.38
WikiLink-70%	3,360,965	285,566,885	84.97
WikiLink-80%	3,841,059	297,133,727	77.36
WikiLink-90%	4,321,158	308,033,742	71.28
WikiLink-full	4,801,501	315,049,408	65.61
Clickstream$_{uw}$	2,255,520	14,362,735	6.37
Clickstream *	2,255,520	1,083,707,336	480.47

As before we perform a series of random walks on the different networks and evaluate the performance of the three models. Table 6 shows the spearman correlations for the different networks and approaches. The results show that the network reduction does not negatively affect the model performance for both concept embeddings and co-occurrence counting and that concept embeddings give the best performance. Indeed it seems that concept embeddings can exploit the reduction and perform slightly better, although a more extensive study would be necessary to validate this. The binary counting approach performs overall significantly worse but definitely profits from the network reduction.

Influence of Human Navigation

The Clickstream dataset consists of transitions between nodes and their weight, e.g. the number of times users have used a specific link between Wikipedia pages. Thus it captures which nodes and links were important to Wikipedia users and we showed in (Niebler et al. 2015) that semantic information can be extracted from the users preferences of certain links. However, the Clickstream dataset only contains transition counts accumulated over a whole month, and explicitly no user-specific navigation paths. To mimic human behavior and thus study more realistic navigation, we created a weighted network from these transition and the corresponding counts and performed both unbiased and biased random walks on this network. When performing biased random walks we selected the next node with a proba-

Table 5: Basic statistics for all random walk datasets generated from Wikipedia link networks. An average path length of less than 21 means that some walks ended with a leaf node, before the max path length could be reached. In the case of Clickstream walks, the average path lengths are significantly shorter than those on the Wikipedia link network. This is due to the low average outdegree of the Clickstream link network, so many paths end with a leaf node, which has no outgoing links.

dataset	#paths	∅len	∅pagefreq
Walks-10%	4,800,410	20.90	208.94
Walks-20%	9,601,410	20.92	209.15
Walks-30%	14,402,200	20.92	209.18
Walks-40%	19,202,710	20.92	209.17
Walks-50%	24,001,980	20.92	209.13
Walks-60%	28,800,120	20.91	209.08
Walks-70%	33,595,690	20.91	209.03
Walks-80%	38,389,140	20.91	208.98
Walks-90%	43,178,910	20.91	208.94
Walks-full	47,975,930	20.91	208.92
$\text{Walks}_{Clickstream}$	14,332,210	13.48	89.09
$\text{Walks}_{Clickstream_{uw}}$	14,332,210	11.31	73.64

bility proportional to the edge's weight. Table 7 and Table 8 show the results for both unbiased and biased random walks. We observe that in both cases, concept embeddings show really bad performance with a difference of at least 0.1 in the case of $w = 3$. In contrast to that, the counting methods can greatly improve their performance, especially on the biased random walks that take the transitions between pages into account.

Discussion

In this paper we set out to study the suitability of concept embeddings learned on the Wikipedia network structure for the task of measuring the semantic relationship of two words and to compare the results to state-of-the-art word co-occurrence counting approaches. Our experiments show mixed results for this specific task and in the case of learning a semantic representation from simulated user navigation data concept embeddings perform worse than the evaluated counting approaches.

Comparison on the Wikipedia Link Network If trained on random walks performed on the full Wikipedia network, concept embeddings are able to outperform both counting baselines for most parameter choices. Only for the smallest possible windows size $w = 1$ the binary counting approach is able to outperform concept embeddings. A point in favor of concept embeddings is their robustness towards increasing window size which negatively affects the performance of the cooccurrence counting approaches. Overall concept embeddings seem to tolerate bad parameter choices better than counting approaches. One potential explanation might be that concept embeddings are better able to capture latent meanings induced by the context. It is noteworthy that the

Table 6: Spearman correlations for networks pruned to the top k-percent nodes according to Pagerank. For concept embeddings we set $w = 3$ and $|v| = 128$.

top(%)	$\rho_{concept}$	ρ_{coocc}	ρ_{binary}
10	0.711	0.669	0.627
20	0.719	0.676	0.626
30	0.700	0.674	0.619
40	0.716	0.681	0.610
50	0.716	0.677	0.605
60	0.706	0.678	0.594
70	0.709	0.681	0.595
80	0.699	0.677	0.588
90	0.709	0.682	0.587
100	0.698	0.680	0.582

Table 7: This table shows the spearman correlation for concept embeddings and both baselines on unbiased random walks based on the Clickstream data.

w	1	3	5	7
ρ_{coocc}	0.721	0.729	0.719	0.697
ρ_{binary}	0.740	0.733	0.701	0.675
$\rho_{concept}$	0.576	0.628	0.624	0.624

spearman values for the binary co-occurrence counting approach are slightly different compared to the values obtained in (Niebler et al. 2015). This might be due to different datasets being used. To verify this the effects of changes in network structure on the different approaches need to be studied.

Effects of Network Reduction to Relevant Nodes When reducing the network size by retaining only relevant nodes and their edges according to Pagerank we can see that concept embeddings perform slightly better with decreasing network size and overall give the best performance. The binary version of the co-occurrence counting scheme can profit from the reduction while the normal co-occurrence counting scheme is negatively affected. This might be because we gradually remove less relevant and therefore less connected nodes from the network as can be seen in Table 4. The binary counting approach then reduces the contribution of more relevant nodes and at the same time increases the contribution from less relevant nodes which may decrease performance for larger networks. This indicates that for concept embeddings only a limited part of the network around the relevant nodes contributes to their representation and that in fact the quality of the representations can increase when shrinking the network. It also significantly speeds up computation, since a lot less data needs to be processed to obtain high quality representations.

Influence of Human Navigation In our third experiment we used the Clickstream dataset that captures how often links were used in Wikipedia and thus denotes real user preferences in link selection. The intuition is that user behavior

Table 8: This table shows the spearman correlation for concept embeddings and both baselines on biased random walks based on the Clickstream data.

w	1	3	5	7
ρ_{coocc}	0.701	0.756	0.759	0.755
ρ_{binary}	0.741	0.752	0.721	0.694
$\rho_{concept}$	0.533	0.516	0.506	0.494

contains information that can be used to create better semantic representations. We find that this holds for both baselines which give significantly better values compared to the results from the Wikipedia link networks. However concept embeddings perform worse on this kind of network compared to counting approaches as shown in Tables 7 and 8.

In (Niebler et al. 2015), we already showed that human navigation outperforms the static Wikipedia link network with a window size of $w = 2$ (only transitions) regarding semantic evaluation performance. We now extended that comparison and show that this still holds for greater window sizes. Without a doubt, more context in the form of random walks improves the results substantially. Unfortunately, we don't know if this hypothesis still holds on real human navigation data on Wikipedia.

Somehow concept embeddings fail to capture the information conveyed by the users navigation patterns. A possible cause for this can be found in Table 4 and Table 5. The Tables show that nodes in the network derived from Clickstream data have a significantly lower average out-degree. Also the average path length of a sampled random walk is much lower than the paths samples from the Wikipedia link network, i.e., paths very often end with leaf nodes, i.e., the last node has no outgoing links and thus the path ends there despite not having reached maximum length. This means that individual nodes in the network are less tightly connected and might not capture as much context information as the original Wikipedia link network does. Finally the average page frequency for a random walk shows that nodes occur less often in the Clickstream induced random walks than in the WikiLink random walks, thus further limiting the amount of exploitable information for concept embeddings.

Conclusion and Future Work

In this ongoing work, we compared different approaches to extract semantic relatedness from random walks on networks of Wikipedia articles. In that regard, we applied a method based on DeepWalk (Perozzi, Al-Rfou', and Skiena 2014) and the co-occurrence counting approach from (Niebler et al. 2015) on random walks generated from the Wikipedia link network, subsets of this network, a network of human navigation without any navigation weights and a weighted network of human navigation. We evaluated our results on a semantic relatedness dataset with human intuitions of relatedness. Our results show, that concept embeddings mostly outperform the co-occurrence counting approach except on random walks generated from human navigation data.

Future work includes comparison of the cosine similarity with Euclidean distance and an extension of our experiments on other Wikipedia datasets, such as the Simple English or German Wikipedia. We plan to extend our current experimental setting and compare concept embeddings obtained using GloVe, LINE and DeepWalk and evaluate our findings on more datasets for better validation. We also want to find out the best possible combination of parameters, so we will investigate the effect of the length of random walks on semantic relatedness. As the random walks generated from Clickstream data seemed to yield really good results with the co-occurrence counting method, we also want to investigate if the same still holds when taking other human navigation data into account, e.g., from the WikiGame or even, if possible, from Wikipedia log data. Finally, a combination of both the Clickstream weights and the Wikipedia link network would probably yield very interesting results, because this could counter the possible negative effects which we talked about in the discussion when evaluating the concept embedding model on Clickstream random walks.

Acknowledgements

This work was partially funded by the DFG German Science Fund research project "PoSTs II".

References

Brin, S., and Page, L. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Comput. Netw. ISDN Syst.* 30(1-7):107–117.

Budanitsky, A., and Hirst, G. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguists* 32(1):13–47.

Cattuto, C.; Benz, D.; Hotho, A.; and Stumme, G. 2008. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In *Proc. of the 7th International Conference on The Semantic Web*, volume 5318, 615–631. Springer-Verlag.

Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T.; and Harshman, R. 1990. Indexing by Latent Semantic Analysis. *Journal of the American society for information science* 41(6):391–407.

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; and Ruppin, E. 2001. Placing Search in Context: The concept revisited. In *Proc. of the 10th international conference on World Wide Web*. ACM.

Gabrilovich, E., and Markovitch, S. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proc. of the 20th IJCAI*, IJCAI'07, 1606–1611. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Marco Baroni, Georgiana Dinu, G. K. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proc. of the Conference* 1:238–247.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and

Phrases and their Compositionality. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc. 3111–3119.

Milne, D., and Witten, I. H. 2008. An Effective, Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *Proc. of the Conference on Artificial Intelligence*, AAAI '08.

Niebler, T.; Schlör, D.; Becker, M.; and Hotho, A. 2015. Extracting Semantics from Unconstrained Navigation on Wikipedia. *KI-Künstliche Intelligenz* 1–6.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*, volume 14.

Perozzi, B.; Al-Rfou', R.; and Skiena, S. 2014. DeepWalk: Online Learning of Social Representations. In Macskassy, S. A.; Perlich, C.; Leskovec, J.; 0010, W. W.; and Ghani, R., eds., *KDD*, 701–710. ACM.

Singer, P.; Niebler, T.; Strohmaier, M.; and Hotho, A. 2013. Computing Semantic Relatedness from Human Navigational Paths: A Case Study on Wikipedia. *International Journal on Semantic Web and Information Systems (IJSWIS)* 9(4):41–70.

Strube, M., and Ponzetto, S. P. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proc. of the National Conference on Artificial Intelligence*, volume 21, 1419. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; and Mei, Q. 2015. LINE: Large-scale Information Network Embedding. In *Proc. of the 24th International Conference on World Wide Web*, WWW '15, 1067–1077. New York, NY, USA: ACM.

West, R.; Pineau, J.; and Precup, D. 2009. Wikispeedia: an online game for inferring semantic distances between concepts. In *Proc. of the 21st international jont conference on Artifical intelligence*, IJCAI'09, 1598–1603. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Wulczyn, E., and Taraborelli, D. 2015. Wikipedia Clickstream.

Zhao, Y.; Liu, Z.; and Sun, M. 2015. Representation Learning for Measuring Entity Relatedness with Rich Information. In *Proc. of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 1412–1418. AAAI Press.