Synthetic, yet natural: Properties of WordNet random walk corpora and the impact of rare words on embedding performance

Filip Klubička[1,3] **Alfredo Maldonado**[2,3] **Abhijit Mahalunkar**[1] **John D. Kelleher**[1,3]
[1]Technological University Dublin, Ireland
[2]Trinity College Dublin, Ireland
[3]ADAPT Centre, Dublin, Ireland
{filip.klubicka,alfredo.maldonado,john.kelleher}@adaptcentre.ie
abhijit.mahalunkar@mydit.ie

Abstract

Creating word embeddings that reflect semantic relationships encoded in lexical knowledge resources is an open challenge. One approach is to use a random walk over a knowledge graph to generate a pseudo-corpus and use this corpus to train embeddings. However, the effect of the shape of the knowledge graph on the generated pseudo-corpora, and on the resulting word embeddings, has not been studied. To explore this, we use English WordNet, constrained to the taxonomic (tree-like) portion of the graph, as a case study. We investigate the properties of the generated pseudo-corpora, and their impact on the resulting embeddings. We find that the distributions in the psuedo-corpora exhibit properties found in natural corpora, such as Zipf's and Heaps' law, and also observe that the proportion of rare words in a pseudo-corpus affects the performance of its embeddings on word similarity.

1 Introduction

A word embedding model maps the words in a vocabulary to dense low-dimensional vectors, by inferring the relative position of each word in a shared multidimensional semantic space from its context of use in a corpus (Mikolov et al., 2013a; Mikolov et al., 2013b). This approach is founded on the distributional hypothesis (Harris, 1954), which states that words which occur in the same contexts tend to have similar meanings. Such word embeddings are created by training a neural network language model on natural language corpora.

While such embeddings have been shown to perform well on semantic relatedness benchmarks (Baroni et al., 2014; Camacho-Collados and Pilehvar, 2018), training on a natural corpus only models one type of semantic relation between words: thematic (i.e. syntagmatic). On the flip side, taxonomic (i.e. paradigmatic) relations are not explicitly contained in natural language corpora, and as such are not included in those embeddings (Kacmajor and Kelleher, 2019). In fact, research suggests that the best measures of taxonomic similarity and thematic relatedness are different in distributional space (Asr et al., 2018). Furthermore, there are many other kinds of relationships between words and concepts that can be found in knowledge engineered resources, such as knowledge bases, ontologies, taxonomies and other semantic networks.

Modelling these relations is an important task in building AI with comprehensive natural language understanding abilities, and there have been many efforts to bring knowledge graphs into an embedding space (see Section 2 for details). One such approach is the WordNet random walk algorithm (Goikoetxea et al., 2015): by randomly walking the WordNet knowledge graph and choosing words from each synset that has been traversed, a pseudo-corpus is generated and used for training word embeddings. The reasoning is that the distributional hypothesis should also apply in this scenario, in the sense that co-occurrence within local contexts in the pseudo-corpus will reflect the connections between words connected in the WordNet graph.

Naturally, the shape of the underlying knowledge graph (in terms of node connectivity: i.e. tree, fully-connected, radial etc.) affects the properties of a pseudo-corpus generated via a random walk over the graph. Developing a better understanding of the relationship between the shape of a knowledge graph, the properties of the resulting pseudo-corpora, and the properties of the resulting embeddings, has the potential to inform how the walk over a given knowledge graph should be tailored to improve embedding performance.

In this paper we provide an analysis of some

of the properties of pseudo-corpora generated using the random walk method, and examine the impact of these properties on embedding performance. We base this analysis on the WordNet taxonomy, because (a) WordNet is one of the most-popular taxonomies in use, and (b) in general, the WordNet taxonomy has a well-understood shape (tree-like) which informs the analysis of our results. We find that the pseudo-corpora synthesized from the WordNet taxonomy are not as artificial as one might expect - they exhibit properties and regularities also found in natural corpora, following natural language laws such as Heaps' law and Zipf's law. Consequently, we hypothesise that word embeddings trained on such corpora might face the same limitations as those trained on natural corpora would. We explore this notion on the case study of rare (i.e. infrequent) words, which are a known problem for word embeddings (Khodak et al., 2018; Pilehvar and Collier, 2017; Pilehvar et al., 2018).

2 Related work

Research on building embeddings from knowledge resources such as WordNet (Fellbaum, 1998), can be broadly categorised into three approaches: i) enrichment, ii) specialisation, and iii) direct learning from knowledge resources.

Both enrichment and specialisation modify precomputed, corpus-based word embeddings with information from a knowledge resource to either augment them (enrichment) or to fit them onto the specific semantic relation described by that knowledge resource (specialisation). Retrofitting (Faruqui et al., 2015) is an example of enrichment: it modifies corpus-based embeddings by reducing the distance between words that are directly linked in resources like WordNet, MeSH (Yu et al., 2016) and ConceptNet (Speer and Havasi, 2012). In our own recent related work, we have explored the impact of corpus size on vector enrichment (Maldonado et al., 2019).

On the other hand, examples of the specialisation approach are PARAGRAM (Wieting et al., 2015), Attract-Repel (Mrkšić et al., 2016), Hypervec (Nguyen et al., 2017) and the work of Nguyen et al. (2016) and Mrkšić et al. (2017) on synonyms and antonyms. Vulić et al. (2018) and Ponti et al. (2018) introduce global specialisation models where vectors for words that are missing in the knowledge resource are also updated.

More related to our work are the approaches to learn directly from knowledge resources. Examples include building non-distributional sparse word vectors from lexical resources (Faruqui and Dyer, 2015), building Poincaré embeddings that represent the structure of the WordNet taxonomy (Nickel and Kiela, 2017) and building embeddings that encode all semantic relationships expressed in a biomedical ontology within a single vector space (Cohen and Widdows, 2017). The latter two methods encode the semantic structure of a knowledge resource in a deterministic manner, while Agirre et al. (2010) follow a stochastic approach based on Personalised PageRank: they compute the probability of reaching a synset from a target word, following a random-walk on a given WordNet relation. Instead of computing random-walk probabilities, Goikoetxea et al. (2015) use an off-the-shelf implementation of the word2vec Skip-Gram algorithm to train embeddings on WordNet random walk pseudo-corpora, changing neither the embedding algorithm nor the objective function[1]. The resulting embeddings encode WordNet taxonomic information rather than natural word co-occurence. An advantage of the embeddings produced by this method is that they can be used as is or can be combined with real-corpus embeddings in order to accomplish enrichment or specialisation (Goikoetxea et al., 2016).

Previous work has analysed semantic properties of word embeddings generated by random walk. Goikoetxea et al. (2016), for example, found WordNet random-walk embeddings to outperform corpus-based word embeddings on the strict semantic similarity (taxonomic similarity) SimLex-999 benchmark (Hill et al., 2015), confirming that they encode taxonomic information better than real-corpus word embeddings. Additionally, other researchers have explored different varieties of the random walk algorithm. Most notably, Simov et al. (2017a) drastically enrich the graph structure by using all available relationships between WordNet synsets, while inferring and adding others from outside resources (Simov et al., 2015; Simov et al., 2017b). However, to the best of our knowledge, there has been no work on analysing the properties of the corpora generated by random-walk processes. In particular, there has been no work on comparing their statistical properties with those of natural corpora.

[1] http://ixa2.si.ehu.eus/ukb/

3 Pseudo-corpora

3.1 Random walk pseudo corpus generation

Our pseudo-corpus generation process is inspired by the work of Goikoetxea et al. (2015). They performed random walks over the full WordNet knowledge base as an undirected graph of interlinked synsets. Their method first chooses a synset at random from the set of all synsets, and then performs a random walk starting from it. They also use a predefined dampening parameter (α) to determine when to stop the walk, so that at each step the walk might move on to a neighbouring synset with probability (α), or might terminate with the probability ($1 - \alpha$). It is usually set to 0.85. Each time the random walk reaches a synset, a lemma belonging to the synset is emitted, using the probabilities in the inverse dictionary. Once the random walk terminates, the sequence of emitted words forms a pseudo-sentence of the pseudo-corpus. The process repeats until a given number of sentences have been generated.

Our pseudo-corpus generation algorithm is similar, however, there are a number of important differences. First, Goikoetxea et al. make use of all available connections in the graph, whereas we only traverse the hypernym/hyponym relationship and ignore non-taxonomic relationship types such as gloss, meronym and antonym relations. This effectively allows us to exclusively traverse WordNet's taxonomic graph, which lets us embed only taxonomic relations. More importantly, this decision is motivated by the fact that we wish to use WordNet's taxonomic graph as a case study of how the underlying structure of a knowledge graph affects the properties of a generated pseudo-corpus. Constraining the random walk to just the taxonomy reduces the graph to a tree shape, which provides an intuitive and transparent understanding of its structure. This restriction to the taxonomic components of the graph has two important implications: (i) it permits us to consider the graph as directed (hypernym/hyponym→up/down), and (ii) it makes the graph quite sparse. The other two significant differences between our algorithm and Goikoetxea et al. are derived from these two implications and are implemented as two new hyperparameters on the algorithm: a directionality and a minimum sentence length parameter.

The directionality parameter constrains the permissible directions that the walk can proceed along as it traverses the tree structure (e.g., only up, only down, both). This hyperparameter permits us to explore the relationship between variations in the random walk algorithm and the number of rare words in the generated corpus (see Subsection 3.2). The minimum sentence length parameter enables us to filter the sentences generated by the random walk algorithm by rejecting any sentence that is shorter than a prespecified length n. The decision to exploit only the taxonomic relations makes the graph quite sparse: a lot of nodes end up disconnected, as some synsets are not part of the WordNet taxonomy, but are connected to it only via non-taxonomic relations. Given that we allow our algorithm to start the random walk anywhere in WordNet, it often begins, and ends, its walk at a disconnected node, which results in a lot of one-word sentences in the synthesized pseudo-corpus. To remedy this, the minimal sentence length hyperparameter disallows generating sentences with only one word, or sentences shorter than the pre-specified value. Section 3.2 contains details on this and other hyperparameters.

In our algorithm[2], the random walk starts at a random synset and chooses a lemma corresponding to that synset based on the probabilities provided by WordNet's inverse mapping from synsets to lemmas. Once the lemma has been emitted, we check if the synset has any hypernym and/or hyponym connections assigned to it (depending on the direction constraint). If it does, we choose one at random with equal probability and continue the walk towards it, choosing a new lemma from the new synset. This process continues until one of two conditions are met: (a) there are no more connections to take, or (b) the process is terminated according to the dampening factor (α). We then restart the process and create a new pseudo-sentence, until we have generated the required number of sentences. Some examples of pseudo-sentences produced by our system:

measure musical notation tonality minor mode

Dutch-processed cocoa powder chocolate milk

[2]Although Goikoetxea et al. provide an implementation of their random walk algorithm, due to the differences outlined above and the special use cases for our research, we have decided to reimplement it in Python and use NLTK's version of WordNet (Bird and Loper, 2004). Our code and generated datasets are being made available online.
`https://github.com/GreenParachute/wordnet-randomwalk-python`

size	direction	min.sent.len.	token count	avg.sent.len.	%same sents	vocabulary	%rare words
500k	up	2w/s	3,515,524	7.03	18.5	64,257	67.35
500k	down	2w/s	1,475,336	2.95	68.56	55,508	53.35
500k	both	2w/s	2,401,498	4.80	20.06	67,049	39.86
500k	up	3w/s	4,011,247	8.02	17.06	63,923	66.48
500k	down	3w/s	2,097,641	4.20	71.01	46,701	52.33
500k	both	3w/s	2,822,171	5.64	12.22	67,353	33.30
1m	up	2w/s	7,041,365	7.04	27.93	66,840	41.84
1m	down	2w/s	2,947,657	2.95	78.57	59,894	40.81
1m	both	2w/s	4,802,354	4.80	28.49	67,647	15.82
1m	up	3w/s	8,032,165	8.03	26.31	66,401	40.52
1m	down	3w/s	4,195,458	4.20	79.46	51,310	43.91
1m	both	3w/s	5,636,469	5.64	18.88	67,683	11.31
2m	up	2w/s	14,079,962	7.04	39.56	67,587	19.32
2m	down	2w/s	5,898,583	2.95	85.91	63,089	30.03
2m	both	2w/s	9,602,490	4.80	37.66	67,756	3.88
2m	up	3w/s	16,061,599	8.03	37.65	67,081	18.20
2m	down	3w/s	8,389,396	4.19	85.92	55,314	35.99
2m	both	3w/s	11,274,757	5.64	26.99	67,757	2.34

Table 1: Statistics of generated random walk corpora

3.2 Pseudo-corpora properties

We controlled the generation of the pseudo-corpora using the following hyperparamters:

1. **Size**. We define corpus size in terms of the number of random restarts, i.e. number of pseudo-sentences generated. We generate pseudo-corpora of sizes 1k, 10k, 100k, 500k, 1m and 2m sentences.

2. **Direction**. As we are only walking the WordNet taxonomy, we define direction as allowing the walk to either only go up the hierarchy, down the hierarchy, or both ways.

3. **Minimum sentence length**. We impose a constraint on minimal sentence length and generate corpora with 2-word and 3-word minimum length sentences.

Combining all the hyperparameters yielded a total of 36 pseudo-corpora of varying sizes, directions and minimal sentence lengths. However, due to space constraints and the fact that the smaller corpora have shown to be too variable to make confident inferences, we only present data and analyses of the three largest corpus groups.

Note that we are not necessarily looking for a combination of hyperparameters that performs best on evaluation tasks, rather we use them as a tool to generate pseudo-corpora with different properties. Following that, for each pseudo-corpus we measure the following statistical properties: total number of tokens, average sentence length (average tokens per sentence), percentage of identical sentences, size of vocabulary, and percentage of rare words in the vocabulary (see Table 1).

From Table 1 it is visible that the number of tokens grows with the size in terms of number of restarts. Interestingly, however, although the average sentence length correlates with absolute number of tokens, it stays constant regardless of the number of restarts, all other things being equal. For example, the average sentence length for the 500k.both.2w/s is 4.8, and the average sentence length for the 2m.both.2w/s corpus is also 4.8 tokens per sentence. This holds for any other analogous combination, further supporting the claim that the underlying graph structure of the corpus is the source of certain word distributions and regularities present in the corpus.

Furthermore, the number of tokens also varies depending on the other two hyperparameters: directionality and minimum sentence length. For example, both average sentence length and absolute number of tokens are sensitive to the direction hyperparameter. Regardless of the number of restarts, corpora generated by only walking up the taxonomy create the longest sentences on average and have the largest number of tokens, while only walking down the taxonomy generates the shortest sentences and the lowest number of tokens.

Such behaviour is a direct consequence of the WordNet taxonomy's structure and the distribution of edges between nodes. The taxonomy is a tree, and as such the vast majority of its nodes are leaf nodes positioned near the bottom. Consequently, each time the random walk restarts, it is far more likely to start somewhere near the bottom

of the taxonomy, rather than at the top. Therefore, if the walk can only go up, on the majority of restarts it will be able to traverse the taxonomy for a large number of nodes before either α kicks in, or it reaches the top and has nowhere to go. Conversely, if the walk is constrained to only move down the taxonomy then on most restarts the walk will only be able to take a few steps before it has nowhere to go and is forced to terminate. Finally, the reason that allowing both directions in the walk generates shorter sentences than going only up is because almost by definition, a synset can have only 1 hypernym, but several hyponyms, so it is more likely to choose a node that is directed downward. In doing so, it behaves more similarly to the algorithm that only goes down and generates shorter sentences than the upward one.

Naturally, the larger the corpus (both in terms of random restarts and tokens), the larger the vocabulary. When comparing the impact of the direction hyperparameter, going down produces corpora with the least WordNet coverage, and going in both directions yields the highest coverage. Again, this is a direct consequence of the structure of the underlying graph. Due to the nature of the random walk going downward the paths are short and there is not much variety, so the vocabulary coverage depends exclusively on the position of the random restarts and is thus significantly lower.

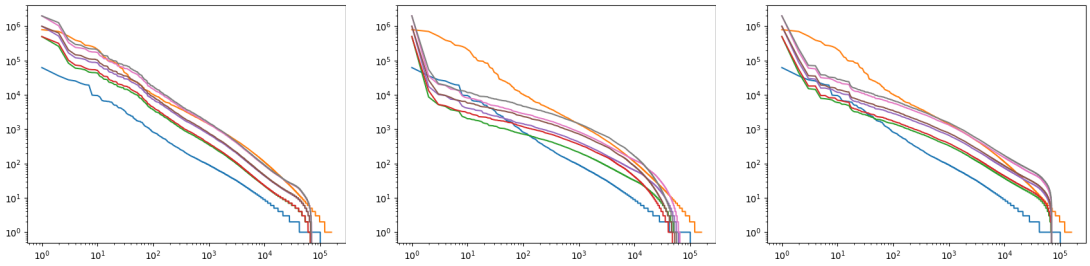
Finally, we look at rare words in the generated corpora. We define a word type as rare if it appears in the corpus less than 10 times. We calculate the percentage of rare words (types/lexemes) versus the full vocabulary. Overall, the percentage of rare words gets smaller as corpus size increases, as more and more words appear over 10 times. However, the hyperparameters seem to have varying effects on this value. For the 500k corpora, the highest percentage of rare words are in corpora generated by only going up, while the lowest percentage are in corpora generated when the walk is allowed to proceed in both directions. All percentages are slightly lower for corpora with a 3-word sentence minimum when compared to corpora with a 2-word sentence minimum. Moving up by one size, corpora with 1m sentences seem to be at a tipping point. Looking at corpora with a 2-word sentence minimum, they follow the percentage of rare words ordering as the 500k corpora of up-down-both, but just barely, and if we look at 3-word sentence minimum corpora the top two rankings switch places. This switch is also apparent in all the 2m-sentence corpora. The percentage of rare words drops off much quicker for corpora generated by only going up compared with corpora generated by only going down. Consequently, even though the up direction generates corpora with the highest percentage of rare words in the smaller sizes, this percentage quickly drops as the corpus size increases. Hence, corpora of 2m sentences generated by only going up have a smaller percentage or rare words compared with the corpora generated by only going down. Likely this is a consequence of the much more drastic increase in absolute number of tokens between the two corpus varieties. The upward corpora consistently have roughly twice as many tokens as the downward corpora, given same number of sentences (i.e. restarts). Overall, the corpus with the smallest percentage of rare words, with only 2.34% rare words in the vocabulary, is the one generated with 2m restarts and allowing the walk to move in both directions. Likely, this is because it is generated from the graph with the most connections, and hence an overall higher coverage; at the size of 2 million sentences, it would have traversed most of the taxonomy several times over, thereby significantly reducing the number of rare words.

3.3 Scaling Linguistic Laws of Natural Languages

The properties described in Subsection 3.2 are a consequence of the corpora being artificially generated from a WordNet's taxonomic graph structure and from the way the random walk algorithm has traversed this graph. However, inspecting word distributions in the corpus showed interesting regularities that seem to indicate similarities with natural corpora. The regularities in the frequency of text constituents have been summarized in the form of *linguistic laws* (Altmann and Gerlach, 2016; Gerlach and Altmann, 2014). Linguistic laws provide insights on the mechanisms of text (language, thought) production. One of the best known linguistic laws is *Zipf's Law* (Zipf, 1949). It states that the frequency, F of the r^{th} most frequent word (i.e. the fraction of times it occurs in a corpus) scales as follows:

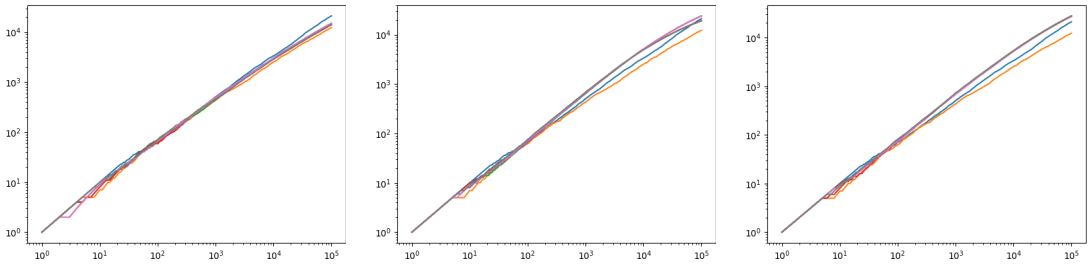
$$F_r \propto r^{-\lambda}, \forall\ r \gg 1 \tag{1}$$

Zipf's Law is approximated by a Zipfian distribution which is related to discrete power law prob-



(a) Direction: up (b) Direction: down (c) Direction: both

Figure 1: Zipf distributions of two natural corpora (shaded blue and orange) and all our pseudo-corpora. We group the three different directions taken by the random walk.



(a) Direction: up (b) Direction: down (c) Direction: both

Figure 2: Heaps' law of two natural corpora (shaded blue and orange) and all our pseudo-corpora. We group the three different directions taken by the random walk.

ability distributions. Here, λ is the scaling exponent and is ≈ 1.0 for natural languages.

Heaps' law is another scaling property and shows how vocabulary grows with text size. Consider n be the length of a text and $v(n)$ be its vocabulary size. Then Heaps' law is formulated as:

$$v(n) \propto n^{\beta}, \forall\ n \gg 1 \tag{2}$$

where the exponent for the Heaps' law is found to be $0 < \beta < 1$ for natural languages.

Here we investigate whether our pseudo-corpora uphold these laws, so as to confirm their naturalness. We employed *Kolmogrov-Smirnov (KS) Distance* to compare the pseudo-corpora against the natural corpora. In our case, we check *KS* distance between the natural and pseudo-corpora for both Zipf's and Heap's law.

Our analysis revealed that the KS distance between our 2 natural corpora is consistent with the distance between the natural and synthetic corpora, indicating consistent variations for both Zipf's and Heaps' law. For both our natural and synthetic corpora, $\lambda \approx 1.1$ and $\beta \approx 0.9$. In this case, it is fair to assume that our pseudo-corpora maintain these properties of natural language. This finding is important because it indicates that embeddings trained on pseudo-corpora will have similar shortcomings to embeddings trained on natural text. For example, past research has highlighted difficulties of learning good embeddings for rare words in natural corpora (Lazaridou et al., 2017; Pilehvar and Collier, 2017).

In addition, in Figures 1 and 2 respectively we also plot Zipf's law and Heaps' law for all our pseudo-corpora, alongside two natural corpora (the Brown corpus (Francis, 1964) and a small chunk of wikitext-2 (Merity et al., 2016)). Though our test of KS distance confirms that all the pseudo-corpora follow Heaps' law and a Zipfian distribution, it is still interesting to note the slight variations in the Zipf curves. Uniformly, the 'up' pseudo-corpora most closely match the natural corpora, the 'down' pseudo-corpora do so to a much lesser degree, and 'both' fall somewhere in the middle. This indicates that the directionality hyperparameter also enables us to simulate slightly different underlying graph structures, in a sense pruning the original graph from the per-

spective of the random walk. These figures reinforce the fact that the nature of the random walk algorithm, the structure of the graph and the paths that are walked have an impact on the resulting pseudo-corpus.

Motivated by these findings, in the next section we will evaluate the performance of a set of embeddings trained on a number of pseudo-corpora and consider the effect of rare words on the performance of these embeddings.

4 Evaluation and analysis

After generating all the corpora, we trained word embeddings and evaluated their performance on the task of word similarity.

4.1 Training

We trained our embeddings using the 2017 version of Pytorch SGNS, a publicly available implementation[3] of the skip-gram with negative sampling (SGNS) algorithm, introduced by Mikolov et al. (2013a). We only made minor data-handling optimisations – the objective function is not modified in any way.

The vectors were computed with SGNS using a window of five words on both sides of a sliding focus word, without crossing sentence boundaries. Twenty words were randomly selected from the vocabulary based on their frequency as part of the negative sampling step of the training. The frequencies in this weighting were smoothed by raising them to the power of $\frac{3}{4}$ before dividing by the total. All vectors produced by the SGNS system had 300 dimensions and trained for 30 epochs. We train separate embeddings on each combination of the three hyperparameters and report scores from the best performing epoch.

4.2 Evaluation

We evaluate the performance of our embeddings on five different benchmarks: the similarity-focused SimLex-999 (Hill et al., 2015); the English test set from the SemEval 2017 Task 2 challenge (Camacho-Collados et al., 2017) (henceforth referred to as SemEval-17); the relatedness dataset WS-353 (Finkelstein et al., 2002); and the Princeton evocation benchmark (Boyd-Graber et al., 2006). However, we suspect none of these benchmarks are ideally suited to the task at hand, as they are all based on human judgements on an often broad idea of word similarity, yet we are specifically modelling taxonomic relations. For this reason, in addition to the above benchmarks, we develop a novel test set, inspired by the work of (Pedersen et al., 2004)[4]: we take the word pairs from SimLex, and replace the human similarity judgements with a WordNet similarity measure (based on the distances in the graph). We refer to this benchmark as WordNet-paths. This serves as a sanity check and an appropriate test set for our taxonomic embedding model.

As is common practice, we evaluate our model by computing a Spearman correlation score between the cosine similarity of the word vectors from our model and the scores in our benchmarks. Table 2 presents the results alongside the percentage of rare words in a given benchmark.

4.3 Discussion

The aim of this experiment is not to beat state of the art scores, but rather to investigate different WordNet taxonomic structures generated by the random walk hyperparameters and their impact on rare words and performance of word embeddings trained on the pseudo-corpora. We hypothesise that the direction constraint of the random walk has an effect on the percentage of rare words in the resulting corpus, which in turn affect the performance of the trained embeddings.

With that in mind, we look at Table 2. Our highest correlation scores come from the WordNet-paths benchmark, which is not surprising as this benchmark reflects most accurately what our models have learned – taxonomic relations in WordNet. The highest overall score comes from the largest corpus, but looking at the different groups of different-sized corpora, the best performing model is always the one allowing both directions in the random walk, which generates the lowest percentage of rare words. Our hypothesis is clearly confirmed on this benchmark, where all the best scores come from corpora with the lowest percentage of rare words, while the lowest scores come from corpora with the highest percentage of rare words in two out of six cases.

In contrast with WordNet-paths, our worst performance is achieved on the evocation benchmark. This is to be expected, as the evocation benchmark models a relationship between words that is very

[3] https://github.com/theeluwin/pytorch-sgns

[4] http://wn-similarity.sourceforge.net

corpus	simlex %rare	simlex score	ws353 %rare	ws353 score	semeval %rare	semeval score	evoc %rare	evoc score	wn-paths %rare	wn-paths score
500k-up-2w/s	2.63	39.03	8.01	39.24	11.81	37.23	5.26	7.93	2.63	52.89
500k-down-2w/s	2.53	19.22	6.86	21.23	10.47	20.46	3.72	4.46	2.53	41.86
500k-both-2w/s	1.14	32.56	2.97	42.76	4.83	38.12	1.31	9.87	1.14	56.31
500k-up-3w/s	2.92	37.07	7.09	34.65	11.60	35.70	4.71	8.61	2.92	50.60
500k-down-3w/s	2.97	31.26	8.70	33.34	10.06	27.51	5.26	4.13	2.97	49.12
500k-both-3w/s	1.04	34.84	2.75	45.53	4.72	40.36	1.10	10.61	1.04	57.00
1m-up-2w/s	1.24	41.73	3.20	43.34	5.85	39.56	2.08	8.61	1.24	53.44
1m-down-2w/s	1.09	30.46	3.43	41.69	6.26	35.09	2.08	6.90	1.09	47.56
1m-both-2w/s	0.50	40.55	0.92	48.25	1.75	40.93	0.44	11.14	0.50	57.60
1m-up-3w/s	1.19	42.28	2.75	39.75	5.85	40.51	2.19	9.75	1.19	54.15
1m-down-3w/s	1.93	36.37	5.03	42.65	8.11	36.19	4.05	5.48	1.93	51.15
1m-both-3w/s	0.35	42.13	0.69	46.59	1.33	39.16	0.33	10.93	0.35	57.73
2m-up-2w/s	0.59	42.58	1.14	44.38	2.77	39.61	0.77	8.63	0.59	53.52
2m-down-2w/s	0.69	34.87	1.14	41.79	4.00	36.75	0.99	5.62	0.69	47.67
2m-both-2w/s	0.15	43.28	0.46	47.03	0.41	40.48	0.22	10.95	0.15	58.00
2m-up-3w/s	0.50	43.40	1.14	43.97	2.46	39.71	0.77	9.65	0.50	54.01
2m-down-3w/s	1.04	36.80	3.43	44.29	5.44	35.17	2.41	4.85	1.04	49.47
2m-both-3w/s	0.05	43.28	0.46	47.51	0.31	40.35	0.22	11.14	0.05	56.55

Table 2: Results for all embeddings trained on various corpora, showing Spearman correlation scores for best epoch per corpus trained on, as well as the percentage of rare words in a given benchmark. Cells shaded green represent the lowest percentage of rare words and the highest Spearman score obtained in the given group of embeddings on a given benchmark. Cells shaded red represent the highest percentage of rare words and the lowest Spearman score on the given group.

different in nature from the purely taxonomic relationship that we model here. This, together with the fact that our best correlation scores come from the WordNet paths benchmark, confirms that our embeddings do indeed reflect a purely taxonomic understanding of words. Yet in spite of the correlation scores being so low, our hypothesis holds here as well – in each group of comparable embeddings, the highest score comes from pseudo-corpora that traversed both directions, and generated the least rare words. The lowest scores stem from corpora with the highest percentage of rare words in five out of six cases.

As expected, we achieve much higher correlations scores on the remaining three benchmarks. Though the highest scores are achieved on WS-353, the overall performances between benchmarks are comparable insofar as they all model word similarity and relatedness. Our hypothesis holds just as consistently when examining the results on SemEval-17 and WS-353, where five out of six times and six out of six times respectively, the best performing model stems from a corpus that yields the lowest percentage of rare words, while the inverse holds four out of six times.

SimLex-999 seems to be somewhat of an outlier among these benchmarks. This is peculiar because, though it is more similarity-focused, the nature of the relations should not be that different from the one in WS-353 and SemEval-17. Our hypothesis still holds in the larger corpora (2m-2w/s, 2m-3w/s and 1m-3w/s), but in the smaller ones the lowest percentage of rare words is produced by the corpora allowing both directions, yet the highest scores actually come from the corpora produced going up. Given that the inconsistencies happen in the smaller corpora, it is possible that this is just an unlucky sample, or that the interplay of confounding factors has a stronger effect in the smaller corpora and negatively affects the performance of the corpora allowing both directions.

Overall, the distribution of best-worst models is fairly consistent across the 5 benchmarks. The best models are those going in both directions, and 2-word sentence minimum models are usually slightly outperformed by 3-word sentence models, though the differences are marginal. Unsurprisingly, models allowing both directions also consistently produce the lowest percentage of rare words. From this, it seems, also follows that more often than not those models have the best scores.

5 Conclusion

In our work we expand our understanding of the random walk algorithm, in terms of the relationship between the structure of the underlying knowledge graph, the properties of the pseudo-corpora generated from the graph, and the performance of the embeddings trained on these pseduo-

corpora. We use the WordNet taxonomy as a case study for our work. We find that all our pseudo-corpora resemble natural corpora at a statistical level. We attribute these properties to the underlying tree structure of the graph from which the pseudo-corpora are built. We also train word embeddings on these corpora to study the impact of these properties on the embedding performance on word similarity evaluation tasks. Our evaluations confirm a successful modelling of taxonomic relations, and on most benchmarks our data supports the hypothesis that the ratio of rare words in a pseudo-corpus affects embedding performance.

Understanding the properties of the pseudo-corpora generated from a knowledge graph structure can inform how the random walk should be designed and run for any graph. E.g. knowing that a tree-like graph structure results in pseudo-corpora exhibiting Zipfian properties is useful as it highlights the presence of rare words in the corpora. As the vocabulary of the lexical resource is finite, the problem of rare words within the generated pseudo-corpora can be addressed by ensuring that the pseudo-corpus is large enough so that even the relatively rare words appear frequently enough to learn adequate embeddings. This perspective helps in answering questions such as: *how large should a pseudo-corpus be?*

Though this might seem obvious, an important takeaway is that the properties of any pseudo-corpus generated from a knowledge graph will be affected by the properties of that graph–its structure and node connectivity will be reflected in the generated corpora, thus impacting the resulting embeddings. We do not claim that any graph structure will exhibit the exact properties we found, but rather that this kind of analysis should be considered when using a random walk algorithm.

As far as future work, there are several exciting avenues that can be explored. Most immediately, it would be important to examine whether the natural properties and rare word percentages in the pseudo-corpora hold when applied to more dense graph structures with connections beyond the WordNet taxonomy, such as WordNet gloss relations, polysemy, antonymy, meronymy, etc. Going further, one could apply the random walk to other knowledge bases to see if the regularities hold there as well. Additionally, combining pseudo-corpora from different knowledge bases, or simply enriching one graph with connections from another, adding additional thematic relations from other knowledge bases. Certainly, this would make the problem more complex, and would render the directionality parameter moot, as a lot of those connections do not have an inherent directionality to them. But this is definitely the next step in improving scores and increasing coverage.

Going even further, it would be beneficial to explore the application of both these taxonomic embeddings, as well as more complex knowledge graph embeddings, on tasks other than word similarity, such as hypernym prediction (which are better suited to exploiting taxonomic knowledge) or perhaps using them to tackle the problem of type and token identification of multi-word expressions.

Acknowledgements

This research was supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Eneko Agirre, Montse Cuadros, German Rigau, and Aitor Soroa. 2010. Exploring knowledge bases for similarity. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'10)*.

Eduardo G. Altmann and Martin Gerlach, 2016. *Statistical Laws in Linguistics*, pages 7–26. Springer International Publishing, Cham.

Fatemeh Torabi Asr, Robert Zinkov, and Michael Jones. 2018. Querying word embeddings for similarity and relatedness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 675–684.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, MD.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. In *Proceedings of the third international WordNet conference*. Citeseer.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver.

Trevor Cohen and Dominic Widdows. 2017. Embedding of semantic predications. *Journal of Biomedical Informatics*, 68:150–166.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional Word Vector Representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 464–469, Beijing.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1606–1615, Denver, CO.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Winthrop Nelson Francis. 1964. A standard sample of present-day english for use with digital computers.

Martin Gerlach and Eduardo Altmann. 2014. Scaling laws and fluctuations in the statistics of word frequencies. *New Journal of Physics*, 16:113010, 11.

Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. Random Walks and Neural Network Language Models on Knowledge Bases. In *Human Language Technologies: The 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1434–1439, Denver, CO.

Josu Goikoetxea, Eneko Agirre, and Aitor Soroa. 2016. Single or multiple? combining word representations independently learned from text and wordnet. In *AAAI*.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

Magdalena Kacmajor and John D. Kelleher. 2019. Capturing and measuring thematic relatedness. *Language Resources and Evaluation*.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*.

Angeliki Lazaridou, Marco Marelli, and Marco Baroni. 2017. Multimodal word meaning induction from minimal exposure to natural text. *Cognitive science*, 41:677–705.

Alfredo Maldonado, Filip Klubička, and John D. Kelleher. 2019. Size matters: The impact of training size in taxonomically-enriched word embeddings. *Open Computer Science*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12, Scottsdale, AZ.

Tomas Mikolov, Ilya Stutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the Twenty-Seventh Annual Conference on Neural Information Processing Systems (NIPS) In Advances in Neural Information Processing Systems 26*, pages 3111–3119, Lake Tahoe, NV.

Nikola Mrkšić, Diarmuid O Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. *arXiv preprint arXiv:1603.00892*.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 454–459, Berlin.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical Embeddings for Hypernymy Detection and Directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 233–243, Copenhagen.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré Embeddings for Learning Hierarchical Representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6338–6347. Curran Associates, Inc., Long Beach, CA.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Nigel Collier. 2017. Inducing embeddings for rare and unseen words by leveraging lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 388–393.

Mohammad Taher Pilehvar, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge rare word dataset-a reliable benchmark for infrequent word representation models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1391–1401.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293.

Kiril Simov, Alexander Popov, and Petya Osenova. 2015. Improving word sense disambiguation with linguistic knowledge from a sense annotated treebank. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 596–603.

Kiril Simov, Petya Osenova, and Alexander Popov. 2017a. Comparison of word embeddings from different knowledge graphs. In *International Conference on Language, Data and Knowledge*, pages 213–221. Springer.

Kiril Ivanov Simov, Svetla Boytcheva, and Petya Osenova. 2017b. Towards lexical chains for knowledge-graph-based word embeddings. In *RANLP*, pages 679–685.

Robert Speer and Catherine Havasi. 2012. Representing General Relational Knowledge in ConceptNet 5. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679—-3686, Istanbul.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-Specialisation: Retrofitting Vectors of Words Unseen in Lexical Resources. In *Proceedings of NAACL-HLT 2018*, pages 516–527, New Orleans, LA.

John Wieting, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Dan Roth. 2015. From Paraphrase Database to Compositional Paraphrase Model and Back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Zhiguo Yu, Trevor Cohen, Elmer V. Bernstam, Todd R. Johnson, and Byron C. Wallace. 2016. Retrofitting Word Vectors of MeSH Terms to Improve Semantic Similarity Measures. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 43–51, Austin, TX.

George. K. Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*, volume 47. 01.