Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4893–4902
Marseille, 11–16 May 2020

English WordNet Taxonomic Random Walk Pseudo-Corpora

Filip Klubička[1,3], Alfredo Maldonado[2,3], Abhijit Mahalunkar[1], John D. Kelleher[1,3]
[1]Technological University Dublin, [2]Trinity College Dublin, [3]ADAPT Centre
Dublin, Ireland
filip.klubicka@adaptcentre.ie, maldonaa@tcd.ie, abhijit.mahalunkar@mydit.ie, john.d.kelleher@tudublin.ie

Abstract

This is a resource description paper that describes the creation and properties of a set of pseudo-corpora generated artificially from a random walk over the English WordNet taxonomy. Our WordNet taxonomic random walk implementation allows the exploration of different random walk hyperparameters and the generation of a variety of different pseudo-corpora. We find that different combinations of the walk's hyperparameters result in varying statistical properties of the generated pseudo-corpora. We have published a total of 81 pseudo-corpora that we have used in our previous research, but have not exhausted all possible combinations of hyperparameters, which is why we have also published a codebase that allows the generation of additional WordNet taxonomic pseudo-corpora as needed. Ultimately, such pseudo-corpora can be used to train taxonomic word embeddings, as a way of transferring taxonomic knowledge into a word embedding space.

Keywords: WordNet, taxonomy, random walk, language resource, pseudo-corpus, semantic relationship

1. Introduction

Semantic relationships between words or concepts have at least two key dimensions: taxonomic and thematic. Taxonomic relations between concepts are based on a comparison of the concepts' features. Concepts that belong to a common taxonomic category share properties or functions. In contrast, *thematic relations* are formed between concepts performing complementary roles in a common event or theme, which often implies having different, albeit complementary, features and functions (Kacmajor and Kelleher, 2019)[1].

When it comes to language and language resources, as a rule of thumb the two semantic relationships are explicitly encoded in two different kinds of resources: a *natural language corpus* primarily reflects the thematic relationships between words by way of word co-occurrence. Taxonomic relations, on the other hand, are rarely overtly expressed in examples of natural language. Though research has shown that such relationships can be automatically extracted from natural language corpora (Hearst, 1992), they are more accessible and more commonly modeled in the form of *knowledge-engineered language resources* such as knowledge bases, ontologies, taxonomies and similar semantic networks, where relationships are reflected via explicit links between entities (i.e. nodes) in the knowledge graph.

Modelling both kinds of relationships is an important task in building AI with comprehensive natural language understanding abilities, yet most NLP models and systems, especially language models and word/sentence embeddings, solely rely on natural corpora as their main training resource (Mikolov et al., 2013; Salton et al., 2017; Devlin et al., 2018; Peters et al., 2018; Pagliardini et al., 2018).

That said, there have been many efforts to transfer and integrate the taxonomic information encoded in knowledge resources into distributed vector embedding representations of lexical semantics (see Section 2. for details). The approach that we have explored in our work 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 can be generated and used for training word embeddings, in the same way one would train on a natural language corpus. The reasoning behind this approach is that co-occurrence within local contexts in the pseudo-corpus will reflect the connections between words connected in the WordNet graph. In other words, using this approach flattens out WordNet, turning it into a sequential format similar to a natural corpus, where the same implicit connection - co-occurrence - reflects taxonomic relations, rather than thematic.

As such, a WordNet random walk pseudo-corpus can be a valuable way of introducing WordNet structures and knowledge into already existing machine learning pipelines, such as building language models and training word embeddings (Goikoetxea et al., 2016). Naturally, the shape of the underlying knowledge graph (in terms of node connectivity: e.g. tree, fully-connected, radial etc.) will affect the properties of a pseudo-corpus generated via a random walk over the graph, while the types of connections that are traversed will affect the kinds of relations that are encoded in this resource.

We build on previous work on random walks and reimplement the procedure to generate different flavours of WordNet random walk corpora, developing and exploring various combinations of hyperparameters (such as number of restarts, and constraints on direction and minimal sentence length) which we have found control certain properties of the corpora. In this paper we present the WordNet

[1] In the linguistics literature, the concepts of taxonomic and thematic relatedness would roughly correspond to what are respectively called paradigmatic and syntagmatic relations, and there is a nuanced discussion to be had about the extent of the overlap in the terminology. However, as we are focused on resources modeling taxonomic relations exclusively, delving deeper into the differences between these terms falls beyond the scope of this paper, so we lean on the terminology used by Kacmajor and Kelleher (2019).

taxonomic random walk pseudo-corpora that we have generated for the purpose of our own research (Maldonado et al., 2019; Klubička et al., 2019) and provide an analysis of their properties.

We should note that we have constrained our work only to the WordNet **taxonomy**, because: (a) WordNet is one of the most-popular knowledge graphs in use, and (b) in general, the WordNet taxonomy has a well-understood shape (tree-like) which informs the analysis of our results.

The paper is structured as follows: after discussing related work in Section 2., in Section 3. we present the algorithm for generating pseudo-corpora. Section 4. reports on the various statistical properties of the generated resources, while Section 5. points to the published resources.

2. Related work

Recently there has been an increase in the amount of research on building embeddings from knowledge resources such as WordNet. Prior work shows that embeddings can be enriched with taxonomic knowledge, specialised to better reflect that semantic dimension, or trained from scratch on appropriate taxonomic resources.

Work on enrichment and specialisation tends to focus on the Skip-Gram family of algorithms whereas the approaches taken in research on training embeddings from scratch are more diverse. For example, Faruqui and Dyer (2015) build non-distributional sparse word vectors from knowledge resources, with each dimension representing whether the word belongs to a particular synset, holds a particular taxonomic relation, etc. Another approach is introduced by Nickel and Kiela (2017), who develop Poincaré embeddings that represent the structure of the WordNet taxonomy. This method seeks to encode the semantic structure of a knowledge resource, however it does so in a deterministic manner.

By contrast, 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 pseudo-corpora generated from WordNet random walks. Neither the embedding algorithm nor the objective function is changed in any way. By training on sequences of words that hold taxonomic relations, instead of naturally co-occurring words as in real corpora, the resulting embeddings encode WordNet taxonomic information. A characteristic of random-walk embeddings is that they are of the same "kind" as natural-corpus-trained word embeddings, in the sense that both embeddings are distributional and are trained to satisfy the exact same objective function. If settings and hyperparameters are kept the same, as far as the embedding model is concerned, the only difference between the two sets of vectors is that they were trained on different corpora. This has lead to research that combines WordNet random-walk embeddings with real-corpus embeddings in order to accomplish enrichment or specialisation. For example, Goikoetxea et al. (2016) found that simply concatenating real-corpus word embeddings and WordNet random-walk embeddings gave the best performance on various similarity benchmarks, compared with more sophisticated combination methods. In their work they have also analysed the semantic properties of WordNet random-walk embeddings, and at the time found them 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.

Rather than training word embeddings, Simov et al. (2015) leverage taxonomic knowledge to tackle the task of Word Sense Disambiguation. They pour significant efforts into techniques for enriching the WordNet graph with additional semantic connections (Simov et al., 2016a; Simov et al., 2016b). In their later work, Simov et al. (2017b) build directly on the work of Goikoetxea et al. (2015) and explore how various different varieties of the random walk algorithm impact performance of trained word embeddings, similar to our work on the topic (Klubička et al., 2019). However, rather than constraining the walk to just the taxonomy, they look for additional ways of enriching the graph structure and populating WordNet with as many connections as possible, exploiting all available relationships between WordNet synsets, as well as adding and inferring more from outside resources (Simov et al., 2017a).

3. Resource generation algorithm

Our pseudo-corpus generation process is inspired by the work of Goikoetxea et al. (2015). The core idea of our corpus generation algorithm is that it generates a 'sentence' by performing a random walk over the taxonomic graph of WordNet (Fellbaum, 1998). Each of these random walks begins at a randomly selected synset in the WordNet graph, and each time the random walk reaches a synset, a lemma belonging to the synset is emitted. When the random walk terminates, the sequence of emitted words forms a pseudo-sentence of the pseudo-corpus. This process repeats until a predetermined number of sentences have been generated.

We use three hyper-parameters to control the random walk over the graph: (i) a dampening hyperparamter α, (ii) a directionality hyper-parameter, and (iii) a minimum sentence length hyperparameter.

(i) The dampening factor (α) is used 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 $(1-\alpha)$. Goikoetxea et al. also use a dampening factor and found the best practice is to set it to 0.85. We briefly experimented with slightly higher or lower values, but found it had relatively little impact on pseudo-sentence length when compared to the impact of the other hyperparameters, hence we set ours to 0.85 and did not change it further. While the dampening parameter was introduced by Goikoetxea et al., the directionality hyperparameter, and a minimum sentence length hyperparameter represent extensions that we have introduced ourselves.

(ii) The directionality parameter constrains the permissible directions that the walk can proceed along as it traverses the taxonomic graph (e.g., only up, only down, both). We can do this because we exclusively traverse the

WordNet taxonomy, i.e. hypernym/hyponym connections, which have an inherent directionality to them. This allows us to consider the graph's edges as directed, rather than, as Goikoetxea et al. did, treat them as undirected (due to considering a variety of connections that are not all directional). The motivation for introducing this hyperparameter is that it permits us to explore the relationship between variations in the random walk algorithm, variations in the shape of the underlying graph and the properties of the generated corpora. This relationship will be elaborated on in more detail in Section 4..

(iii) 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 pre-specified length n. As mentioned above, this is necessary because our algorithm constrains the random walk to the taxonomic graph of WordNet. The taxonomic graph is quite sparse – if we only walk along the taxonomic edges, a lot of nodes present in WordNet will end up disconnected, as some synsets are not part of the WordNet taxonomy, but are connected to it via other, non-taxonomic relations. Given that we allow our algorithm to start the random walk anywhere in the graph, the walk often begins, and ends, at a disconnected node. If no minimal sentence length constraint is imposed, this yields many one-word pseudo-sentences that populate the synthesized pseudo-corpus. One-word pseudo-sentences are not at all informative with regards to the word's taxonomic relationship to other words, as these words do not co-occur with other words. To remedy this, we introduce the hyperparameter of minimal sentence length. Most importantly, this can also act as a filtering mechanism that allows us to exclusively traverse the WordNet taxonomy, discarding all words that are not connected to it via a hypernym or hyponym relation. However, the parameter further enables us to generate a corpus of sentences of any minimal length, which allows for a study of different pseudo-corpora properties. More on the hyperparameters will be explained in Section 4.

Controlled by these hyper-parameters our random walk algorithm progresses as follows: The random walk starts at a random synset and chooses a lemma corresponding to that synset based on the probabilities in the inverse dictionary (the mapping from synsets to lemmas) provided by WordNet. However, these are expressed as frequencies, rather than explicit probabilities, so we choose one based on the probability distribution derived from the frequency counts. Once the lemma has been emitted, the algorithm stochastically decides whether the walk should be terminated or not, controlled by the hyper-parameter α. Terminating the walk determines the end of the pseudo-sentence, which is then added to the pseudo-corpus and a new random walk is initiated. If the walk is not terminated 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) the dampening factor (α) terminates the process, or (b) there are no more connections to take. We then restart the process and create a new pseudo-sentence. This pseudo-sentence generation process is repeated until we have generated the required number of sentences. One important thing to note is that we allow our algorithm to go back to a node that has already been visited, but we do not allow it to choose a lemma that has already appeared in the sentence we are generating at the time.

As noted above, our pseudo-corpus generation process is based on the work of Goikoetxea et al. (2015), however there are a number of important differences between the two algorithms. First, Goikoetxea et al. performed random walks over the full WordNet knowledge base as an undirected graph of interlinked synsets, making 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 traverse the taxonomic graph of WordNet exclusively. The main motivation behind this decision is that primarily, we are interested in embedding taxonomic relatedness from the generated corpus, and constraining the random walk to the taxonomy is the most explicit way of doing so. 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 full graph quite sparse. These implications have allowed us to further diverge from Goikoetxea et al.'s work and implemented the directionality and minimal sentence length hyperparameters as described above. In addition, as opposed to Goikoetxea et al. who produce multiword terms, such as `Victrola_gramophone`, `natural_glass` and `shatterproof_glass` essentially treating them as words with spaces, in our corpora we divide these terms up into their individual constituent words (e.g. `Victrola gramophone`, `natural glass` and `shatterproof glass`). Though this is not the traditional approach to handle multi-word terms, we do so to make them more compatible for retrofitting with real corpora, which we took advantage of in our related research (Maldonado et al., 2019)[2]. With that in mind, these are examples of typical pseudo-sentences that can be found in our pseudo-corpora, containing only words with taxonomic relations between them:

- *measure musical notation tonality minor mode*
- *decouple tell dissociate differentiate know distinguish*
- *vocalizer castrato vocaliser rapper vocalist caroler*
- *call-back call call-in telephone call trunk call*
- *meeting place facility station first-aid station aid station*

[2] However, our implementation also allows for the option of generating pseudo-sentences where multi-word expressions are not split. It also allows generating sentences that include words found in synsets that are disconnected from the taxonomy, which results in better vocabulary coverage, but ultimately poorer taxonomic representation. We make our implementation publicly available on GitHub (see Section 5.)

4. Resource description and properties

Using the approach outlined in Section 3., we generated taxonomic pseudo-corpora for the following combinations of hyperparameters:

1. **Size**. We define corpus size in terms of the number of pseudo-sentences generated. We generate pseudo-corpora of sizes 1k, 10k, 100k, 500k, 1m, 2m and 3m 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**. Due to the issue of 1-word sentences being generated, we impose a constraint on minimal sentence length. We generate corpora with 1-word, 2-word and 3-word minimum length sentences.

Combining these hyperparameters yielded a total of 63 pseudo-corpora of varying sizes, directions and minimal sentence lengths. Additionally, for a different set of experiments we also generated another 18 corpora without direction or sentence length constraints (i.e. allowing the walk to traverse both directions and generating 1-word sentences). These additional corpora were much larger, upwards of 468 million sentences. We have released all of these corpora to the community; however, due to space constraints and the fact that the larger corpora were generated with constant hyperparameters, in this paper we only discuss statistical data and analyses of the corpus groups up to 3 million sentences. Additionally, because the corpora that contain 1-word sentences by definition contain words found outside the taxonomic graph of WordNet, they are not strictly taxonomic and reflect a graph structure that is not a tree–a distinction that informs the discussion and analysis of our work. As such, they fall outside the scope of our current interest and we thus exclude corpora with 1-word sentences from the below discussion. Still, we have released them together with all other corpora, and their statistics are included in Table 1.

For each pseudo-corpus we measure the following statistics: 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. This data is presented in Table 1.

Token count and sentence length. From Table 1 it is clear that the number of tokens grows with the size in terms of number of pseudo-sentences in a corpus. Interestingly, however, although the average sentence length correlates with absolute number of tokens, it stays constant regardless of the number of sentences, 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, which strongly suggests that there is a common underlying distribution affecting these pseudo-corpora, which is not affected by their size (in terms of pseudo-sentences, i.e. random restarts).

Furthermore, the number of tokens also varies largely depending on the other two hyperparameters: directionality and minimum sentence length. Not surprisingly, we see that in corpora with a higher sentence length minimum the number of tokens is consistently larger than in corpora with a lower sentence length minimum. However, most interestingly, both average sentence length and absolute number of tokens are strongly impacted by the hyperparameter of direction. Regardless of the number of sentences, the corpora generated by only walking up the taxonomy create the longest sentences on average and have the largest number of tokens, while exclusively walking down the taxonomy generates the shortest sentences and the lowest number of tokens, and allowing both directions during the walk creates a sort of middle ground where the corpora are slightly larger than only going down, but much smaller than only going up.

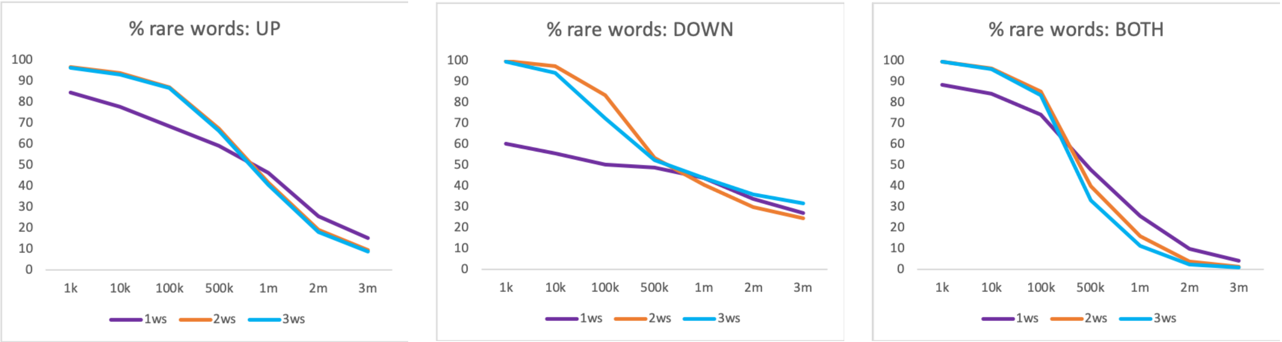
Such behaviour is a direct consequence of the shape of the WordNet taxonomy and the distribution of edges between nodes. The taxonomy is a tree structure with the majority of nodes positioned near the bottom of the tree. Consequently, as there are only a handful of nodes near the top, each time the random walk restarts, it is far more likely to start the random walk at a leaf node somewhere at the bottom of the taxonomy, rather than at the top. Therefore, if the walk is only allowed to go up, on the majority or restarts it will be able to traverse the taxonomy for a 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 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 the algorithm 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.

Repeated sentences. Table 1 also presents statistics on the amount of repetition in the corpora, in terms of identical sentences. We define identical sentences as two sentences whose bags of words contain the same words (effectively disregarding word order). Given that the vocabulary is limited by what can be found in the WordNet, the more we walk the graph, the bigger the chance that the same nodes will be visited, likely via the same paths, and thus identical sentences will be generated. Indeed, looking at Table 1, it is the case that the more sentences there are in the corpora, the more repeated sentences they have. We hypothesised that this would be beneficial for the eventual taxonomic embeddings, as the repetition would reinforce the connections between words, separating information from noise. Our in-depth research on pseudo-corpus sizes has confirmed this hypothesis (Maldonado et al., 2019), but with the caveat that there is a plateau after which growing the size of the random walk pseudo-corpus yields no additional benefits.

However, the number of sentences is not the only factor

size	direction	min.sent.len.	token count	avg.sent.len.	%same sents	vocabulary	%rare words
1k	up	1w/s	4,921	4.92	0.10	2189	84.74
1k	down	1w/s	1,603	1.60	0.50	1425	60.28
1k	both	1w/s	3,378	3.38	0.20	2540	88.62
1k	up	2w/s	7,013	7.01	0.00	2569	96.77
1k	down	2w/s	2,918	2.92	1.00	2280	99.91
1k	both	2w/s	4,691	4.69	0.00	3212	99.47
1k	up	3w/s	7,957	7.96	0.10	2621	96.26
1k	down	3w/s	4,216	4.22	1.70	2895	99.79
1k	both	3w/s	5,519	5.52	0.30	3671	99.48
10k	up	1w/s	48,990	4.90	1.90	12643	77.93
10k	down	1w/s	16,009	1.60	5.87	10810	55.62
10k	both	1w/s	35,085	3.51	2.13	16830	84.34
10k	up	2w/s	70,433	7.04	0.62	12929	93.74
10k	down	2w/s	29,537	2.95	7.18	13943	97.66
10k	both	2w/s	48,022	4.80	0.85	18972	96.37
10k	up	3w/s	80,351	8.04	0.62	13231	93.33
10k	down	3w/s	41,987	4.20	12.40	13857	94.41
10k	both	3w/s	55,988	5.60	0.43	21038	95.91
100k	up	1w/s	492,133	4.92	12.92	51900	68.49
100k	down	1w/s	159,533	1.60	33.03	51412	50.13
100k	both	1w/s	351,970	3.52	13.24	62699	74.28
100k	up	2w/s	705,977	7.06	5.30	44482	87.25
100k	down	2w/s	295,042	2.95	38.56	39999	83.49
100k	both	2w/s	479,014	4.79	6.57	56358	85.43
100k	up	3w/s	804,104	8.04	4.79	44899	86.89
100k	down	3w/s	419,782	4.20	45.70	33118	72.31
100k	both	3w/s	564,113	5.64	3.39	58743	83.68
500k	up	1w/s	2,459,643	4.92	31.66	84842	59.18
500k	down	1w/s	798,474	1.60	68.06	84727	48.95
500k	both	1w/s	1,761,568	3.52	32.71	88707	47.84
500k	up	2w/s	3,515,524	7.03	18.50	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	1w/s	4,924,245	4.92	41.38	90731	46.38
1m	down	1w/s	1,596,776	1.60	79.75	90494	43.93
1m	both	1w/s	3,515,489	3.52	42.32	91958	25.68
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	1w/s	9,828,501	4.91	51.55	92773	25.68
2m	down	1w/s	3,195,186	1.60	87.63	92682	34.02
2m	both	1w/s	7,031,643	3.52	51.29	93119	9.92
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
3m	up	1w/s	14,767,000	4.92	57.37	93,187	15.32
3m	down	1w/s	4,790,103	1.60	90.78	93,140	27.18
3m	both	1w/s	10,554,177	3.52	56.17	93,366	4.35
3m	up	2w/s	21,131,926	7.04	46.67	67,714	9.48
3m	down	2w/s	8,849,429	2.95	89.16	64,416	24.56
3m	both	2w/s	14,402,423	4.80	43.00	67,772	1.41
3m	up	3w/s	24,084,882	8.03	44.78	67,198	8.93
3m	down	3w/s	12,580,624	4.19	88.89	57,499	31.67
3m	both	3w/s	16,918,222	5.64	32.14	67,776	0.82

Table 1: Statistics of generated random walk pseudo-corpora. Statistics are presented in groups based on hyperparameters: we first present size, then minimal sentence length, then direction. Rows presenting data on corpora with a 1-word sentence minimum are shaded cyan, 2-word sentence minimum are shaded magenta and 3-word sentence minimum are shaded orange.



(a) Direction: up

(b) Direction: down

(c) Direction: both

Figure 1: Percentage of rare words plotted against the different sizes of pseudo-corpora. Each graph represents corpora generated in one direction (up, down and both respectively) and displays 3 curves for corpora with a 1-, 2- and 3-word sentence minimum (respectively shaded purple, orange and blue)

controlling the amount of repetition in the corpora: the directionality and minimum sentence length hyperparameters also have a strong impact on the percentage of repeated sentences. Regardless of the number of restarts, when looking at corpora with a 3-words per sentence minimum (shaded orange), the highest percentage of repeated sentences appears in corpora generated by walking down the hierarchy, and allowing both directions generates the lowest percentage, whereas corpora generated going up fall somewhere in the middle. Given that the 'down' corpora have the shortest sentences, as well as the lowest number of words, it is much more likely for their sentences to be the same, as any variation between the sentences generally arises from the random restart, rather than the path of the random walk. Meanwhile, corpora that allow both directions have the most options with regards to the path of the random walk, resulting in high sentence variability and a low percentage of repeated sentences.

Interestingly, the above observation regarding repetition in 3-word sentence minimum corpora does not hold consistently for corpora with a 2-word sentence minimum. Walking down does generate the highest percentage of repeated sentences for both the 2w/s and 3w/s hyperparameter. However, in the 1m 2w/s corpora the lowest percentages of repeated sentences are found in corpora generated from only walking up the taxonomy, and it is only in the 2m corpus that lowest percentage comes from both directions being allowed. This switch between 1m and 2m 2w/s corpora in terms of which direction constraint generates the least number of repeated sentences is peculiar, but given how small the differences are, it is likely that there are confounding effects at play here. We suspect that with the 2w/s corpora allowing both directions makes them more similar to the random walk down, which generates a higher number of short sentences that are then repeated. Once the corpus becomes large enough, this effect is then mitigated and the true effect of the variability comes to the fore. Meanwhile, this effect is not present in the 3w/s corpora because eliminating 2-word sentences compensates for that effect.

Vocabulary. Table 1 also presents statistics on vocabulary size. Naturally, the larger the corpus (both in terms of sentences and tokens), the larger the vocabulary. When comparing the impact of minimal sentence lengths, the vocabulary covered is overall slightly lower in corpora with a 3-word sentence minimum than ones with a 2-word sentence minimum. This difference is small in corpora going up and in both directions, but the difference is quite stark when comparing vocabularies of corpora generated going down (a difference of roughly 8,000-10,000 words). Similarly, when comparing directions, 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 number of tokens and average sentence length. Due to the nature of the random walk going downward the paths are short and there is not much variety, so the vocabulary coverage is significantly lower. Interestingly, allowing for both directions yields a corpus that consistently has almost full coverage, even in the medium-sized corpora, whereas only going up produces a smaller vocabulary in the smaller corpora, but soon catches up as the size increases.

Rare words. Finally, we look at rare words in the generated corpora. We define a word type as rare if it appears in the pseudo-corpus less than 10 times in a sentence with at least one other word in context. The requirement of at least one other word in context for an instance of a word to be counted towards its rare word frequency extends the standard definition of rare words, which generally just considers word occurrences without considering the context of these occurrences. This extension is necessary with our pseudo-corpora because, unlike natural corpora, 1-word sentences occur quite frequently if the random walk traverses a disconnected graph. Instances of words in 1-word sentences should not count towards the word frequencies considered for the definition of rare words for word embedding because these isolated instances provide no contextual information for the word and hence are of no use towards modelling a good taxonomic representation for that word. (Note that for corpora generated with a minimum sentence length hyper-parameter $>$ 1 this definition of rare words becomes simply: words which occur less than 10 times in the pseudo-corpus.)

We calculate the percentage of rare words versus the full vocabulary. Values are presented in Table 1 and their plots in Figure 1.

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 different effects on this value depending on corpus size as well. 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. 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 3m sentences generated by only going up have a smaller percentage or rare words compared with the 3m corpora generated by only going down. This is a consequence of the much more drastic increase in number of tokens between the two corpus varieties. The upward corpora consistently have roughly twice as many tokens as the downward corpora of the same number of sentences. Overall, the corpus with the smallest percentage of rare words, with only 0.82% of rare words in the vocabulary, is the one generated with 3m sentences, a 3 word-sentence minimum 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 3 million sentences, it would have traversed most of the taxonomy several times over, thereby significantly reducing the number of rare words.

These are all properties that arise as a consequence of these corpora being artificially generated. They are all stem from the graph structure of the WordNet taxonomy and from the way the random walk algorithm has traversed this graph. However, we also looked at word distributions and noticed interesting trends that seem to indicate similarities with natural corpora, so we decided to investigate.

4.1. Scaling Linguistic Laws of Natural Languages

The regularities in the frequency of text constituents have been summarized in the form of *linguistic laws* (Gerlach and Altmann, 2014; Altmann and Gerlach, 2016). 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

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

Zipf's Law is approximated by a Zipfian distribution which is related to discrete power law probability distributions. Here, λ is the scaling exponent and it has been found to be ≈ 1.0 for natural languages. In other words, in a natural language corpus, the frequencies of words are inversely proportional to their ranks in the frequency table, i.e. the most frequent word will occur about twice as often as the second most frequent word, three times as often as the third most frequent word, etc.

Heaps' Law is another linguistic law, also a scaling property of language, which describes 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 follows:

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

where the exponent for the Heaps' law for natural languages is found to be $0 < \beta < 1$. In other words, Heaps' law means that as more instances of natural text are gathered, there will be diminishing returns in terms of discovery of the full vocabulary from which the distinct terms are drawn, i.e. as the text gets bigger, there will be less and less new additions to the vocabulary.

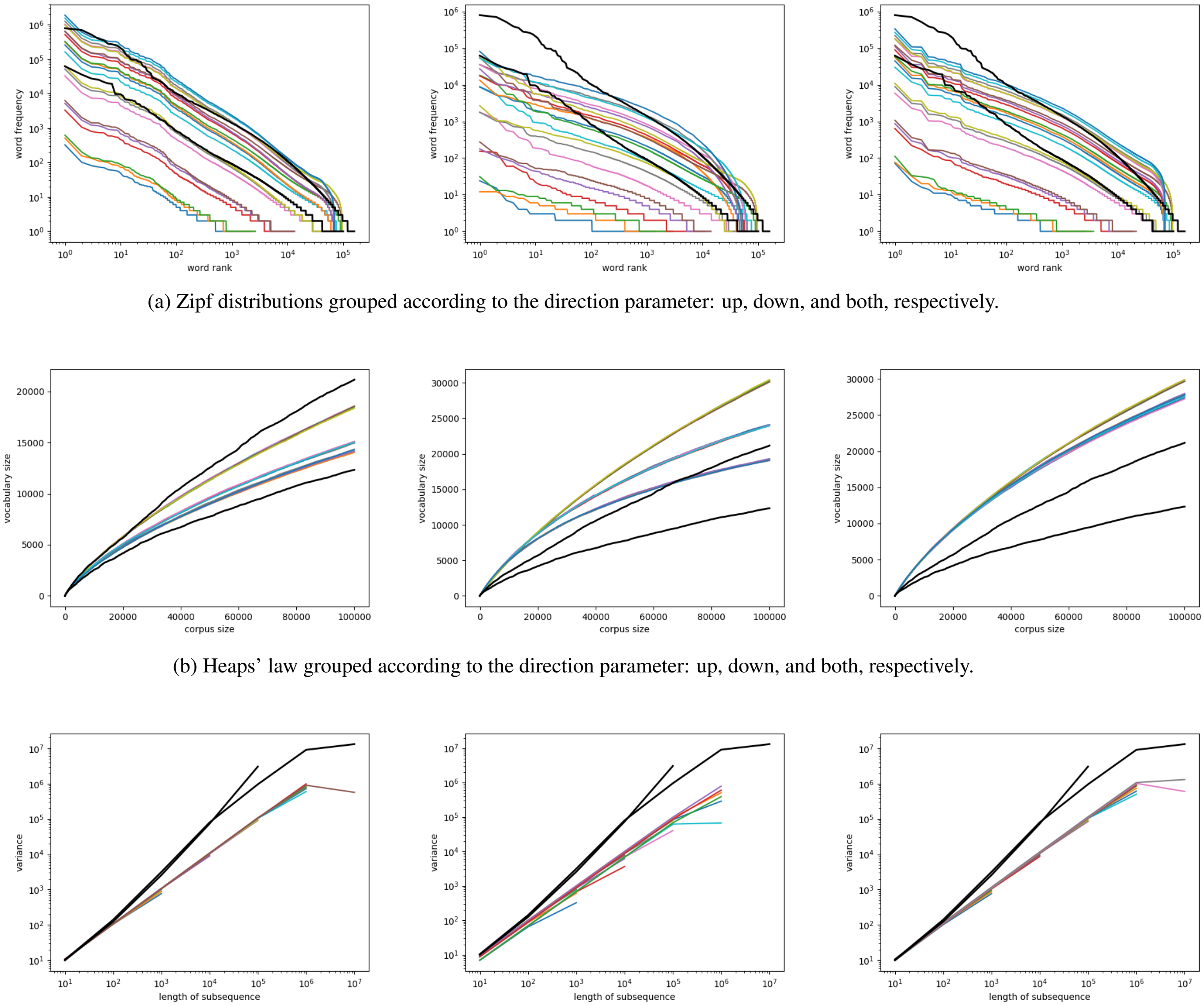
We also consider Ebeling's Law, which studies the growth of variance of individual components (e.g. letters or words in text) in relation to the subsequence length l. Described by Takahashi and Tanaka-Ishii (2019), for a set of words W, let $y(k,l)$ be the number of occurrences of word $w_k \in W$ for all subsequences of length l of the original dataset. Then,

$$m(l) = \sum_{k=1}^{|W|} m_2(k,l) \propto l^{\eta} \tag{3}$$

$m_2(k,l)$ is the variance of $y(k,l)$. Here, $m(l)$ relates to l with a power-law relationship with exponent η. Ebeling and Pöschel (1994) showed that the Bible has $\eta = 1.69$. In other words, there is a specific relationship between the size of a sequence of natural text and the variance of words that occur in that sequence. It can be seen as describing the variety of words found in a text, which becomes higher as the text size increases.

Taking these natural linguistic laws into account, we test whether our pseudo-corpora uphold such laws, so as to investigate their own naturalness. We have compared the Brown corpus (Francis, 1964) and a relatively small chunk of wikitext-2 (Merity et al., 2016) with all our generated pseudo-corpora. Figures 2a, 2b and 2c display the plots of Zipf's, Heaps' and Ebeling's laws respectively for the two natural corpora as well as all our generated pseudo-corpora. In addition to plotting the individual curves, we employed *Kolmogrov-Smirnov (KS) Distance* to compare the pseudo-corpora against the natural corpora. The Kolmogorov-Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. In our case, we check *KS* distance between the natural and pseudo-corpora for Zipf's, Heap's and Ebeling'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 Zipf's, Heaps' and Ebeling's 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 word representations derived



(a) Zipf distributions grouped according to the direction parameter: up, down, and both, respectively.

(b) Heaps' law grouped according to the direction parameter: up, down, and both, respectively.

(c) Ebeling's law grouped according to the direction parameter: up, down, and both, respectively.

Figure 2: Zipf's, Heaps' and Ebeling's laws of two natural corpora (shaded black) and all our pseudo-corpora. We group the corpora according to the three different directions taken by the random walk.

from taxonomic pseudo-corpora would have similar limitations to representations derived from natural text. For example, previous research has highlighted the difficulties of learning good embeddings for rare words in natural corpora (Lazaridou et al., 2017). And indeed, our own research has confirmed that the presence of rare words in the pseudo-corpora has an impact on embedding performance, just as it would in a natural corpus (Klubička et al., 2019).

Though our test of KS distance confirms that all the pseudo-corpora follow the above distributions, it is still interesting to note the slight variations in the generated plots. Uniformly, the 'up' pseudo-corpora most closely match the natural corpora, the 'down' pseudo-corpora do so to a much lesser degree, while 'both' fall somewhere in the middle. This indicates that the directionality hyperparameter also enables us to simulate slightly different underlying graph structures, accounting for the variation in the statistical distributions. These figures reinforce the fact that the nature of the random walk algorithm, the structure of the graph and the paths that are walked do have an impact on the resulting pseudo-corpus. They might not impact the fact that they reflect scaling laws found in natural language, but they still have an impact on the distributions of the words in the generated text, which can propagate down the line if integrated into various machine learning and language modelling pipelines.

5. Resource publication

Goikoetxea et al. provide an implementation of their pseudo-corpus generation algorithm[3]. However, due to the significant differences our algorithm has introduced, as outlined in Section 3., and the the special use cases required for our research which focused on analysing how the shape of

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

knowledge graph affects the properties of the synthesized corpora, we reimplemented the algorithm using NLTK's Python version of WordNet (Bird and Loper, 2004)[4]. We have also made our random walk code publicly available via GitHub[5], and have included a detailed guide on how to use the provided scripts. In addition to a script for generating pseudo-corpora with varying hyperparameters, there is also a script for calculating basic corpus statistics, and a script for calculating a word similarity score using word embeddings and cosine similarity.

As far as our corpora, we have published all resources related to our research on Arrow@TUDublin[6], which is Technological University Dublin's official archive and data repository. This includes an archive of all 81 pseudo-corpora that were generated for our research[7]. They are published in the form of a compressed archive of text files, and once extracted each individual pseudo-corpus can be used with our statistics script, or as input for any word embedding system.

Additionally, we have also used the data repository as an archive for our taxonomic word embeddings, which we trained on the above pseudo-corpora (with some exceptions). This includes a total of 72 pre-trained taxonomic word embedding models that were trained for the purposes of our research (Maldonado et al., 2019; Klubička et al., 2019) [8].

6. Conclusion

The original motivation and distinctive element of our work was to explore how the shape of the knowledge graph affected the properties of the generated pseudo-corpora. It was this motivation that led us to look into a taxonomic graph, in turn developing the specialised taxonomic random walk algorithm. Using the algorithm to create all these corpora allowed us to train taxonomic embeddings and look into the impact that the properties of the different corpora have on their performance.

When looking into the corpora properties, 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 Zipf's, Heaps' and Ebeling's law. We also find that changing hyperparameters of the random walk–and thus the shape of the graph–can heavily impact statistical properties of the generated pseudo-corpora, such as vocabulary size, sentence length, amount of repetition, and percentage of rare words.

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. The PRECISE4Q project (`https://precise4q.eu`) is funded under the EU's H2020 programme grant No. 777107.

7. Bibliographical References

Agirre, E., Cuadros, M., Rigau, G., and Soroa, A. (2010). Exploring knowledge bases for similarity. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'10)*.

Altmann, E. G. and Gerlach, M., (2016). *Statistical Laws in Linguistics*, pages 7–26. Springer International Publishing, Cham.

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

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ebeling, W. and Pöschel, T. (1994). Entropy and long-range correlations in literary english. *EPL (Europhysics Letters)*, 26(4):241.

Faruqui, M. and Dyer, C. (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.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Francis, W. N. (1964). A standard sample of present-day english for use with digital computers.

Gerlach, M. and Altmann, E. (2014). Scaling laws and fluctuations in the statistics of word frequencies. *New Journal of Physics*, 16:113010, 11.

Goikoetxea, J., Soroa, A., and Agirre, E. (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.

Goikoetxea, J., Agirre, E., and Soroa, A. (2016). Single or multiple? combining word representations independently learned from text and wordnet. In *AAAI*.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

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

Kacmajor, M. and Kelleher, J. D. (2019). Capturing and measuring thematic relatedness. *Language Resources and Evaluation*.

Klubička, F., Maldonado, A., and Kelleher, J. (2019). Synthetic, yet natural: Properties of wordnet random walk corpora and the impact of rare words on embedding performance. In *Proceedings of GWC2019: 10th Global WordNet Conference*.

[4] `http://www.nltk.org`

[5] `https://github.com/GreenParachute/wordnet-randomwalk-python`

[6] `https://arrow.dit.ie`

[7] `https://arrow.dit.ie/datas/9/`

[8] `https://arrow.dit.ie/datas/12/`

Lazaridou, A., Marelli, M., and Baroni, M. (2017). Multi-modal word meaning induction from minimal exposure to natural text. *Cognitive science*, 41:677–705.

Maldonado, A., Klubička, F., and Kelleher, J. D. (2019). Size matters: The impact of training size in taxonomically-enriched word embeddings. *Open Computer Science*.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Mikolov, T., Stutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). 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.

Nickel, M. and Kiela, D. (2017). Poincaré Embeddings for Learning Hierarchical Representations. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 6338–6347. Curran Associates, Inc., Long Beach, CA.

Pagliardini, M., Gupta, P., and Jaggi, M. (2018). Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of NAACL-HLT*, pages 528–540.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, abs/1802.05365.

Salton, G., Ross, R., and Kelleher, J. (2017). Attentive language models. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 441–450.

Simov, K., Popov, A., and Osenova, P. (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.

Simov, K., Popov, A., and Osenova, P. (2016a). The role of the wordnet relations in the knowledge-based word sense disambiguation task. In *Proceedings of Eighth Global WordNet Conference*, pages 391–398.

Simov, K. I., Osenova, P., and Popov, A. (2016b). Using context information for knowledge-based word sense disambiguation. In *AIMSA*.

Simov, K., Osenova, P., and Popov, A. (2017a). Comparison of word embeddings from different knowledge graphs. In *International Conference on Language, Data and Knowledge*, pages 213–221. Springer.

Simov, K. I., Boytcheva, S., and Osenova, P. (2017b). Towards lexical chains for knowledge-graph-based word embeddings. In *RANLP*, pages 679–685.

Takahashi, S. and Tanaka-Ishii, K. (2019). Evaluating computational language models with scaling properties of natural language. *Computational Linguistics*, 45(3):481–513.

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