Computing Science

HOP: A Novel Extrinsic Evaluation Metric for Word Embedding Systems and Taxonomies

Aron Molnar

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed:

Date: December 2, 2022

CS4040 Report

Department of Computing Science
University of Aberdeen
King's College
Aberdeen AB24 3UE

December 2022

HOP: A Novel Extrinsic Evaluation Metric for Word Embedding Systems and Taxonomies

Aron Molnar

Word Count: 3432

1 Introduction

We propose HOP, a novel metric that aims to measure the word meaning capturing power of machine learning systems and human-annotated word taxonomies. HOP utilises an exciting phenomenon found in Wikipedia, which is as follows. Starting from any randomly selected Wikipedia article and continually clicking on one of the first few hyperlinks in the current article, we eventually end up with the article for Philosophy. Apart from two edge cases (cycles and dead-ends), the phenomenon is measurable for all articles. It stems from Wikipedia's nature to describe an entity's category in the first few paragraphs of an article. At every traversal step, we move up one or a few steps in the category (abstraction) chain, eventually ending up at the (arguably) most abstract entity, Philosophy.

In this paper, we evaluate HOP on word2vec embeddings, BERT's contextual embedding layer, the human-annotated taxonomy WordNet, and three baselines. The first baseline is selecting words at random, the second is always selecting the first hyperlinks, and the third is always selecting the second hyperlinks.

We investigate if WordNet's human-annotated semantic relations database is better at capturing word meanings than machine-learned word embeddings like word2vec or BERT's embedding layer. We hypothesise that, as long as we are in the lexical database's vocabulary, WordNet is superior in describing semantic relationships between words compared to trained word embeddings. However, WordNet's vocabulary is limited compared to word embeddings, so if vocabulary is not taken into account, we have found that word2vec produces the best HOP-value, closely followed by WordNet. We also found that both taxonomies and word embeddings perform way better than our proposed baseline word selection strategies, implying that HOP has the potential to become a correct metric of meaning-capturing capacity.

HOP is unique in that it allows us to compare not only word embedding systems but human-annotated taxonomies, yielding a generalized extrinsic meaning-capturing power metric.

The possible impact of the research is four-fold:

1. It provides insight into how well WordNet captures word meanings compared to trained word embeddings.

2. The proposed metric can be re-used to measure the semantic relationship capturing capacity of newly created word embeddings and taxonomies.

3. It allows comparing the concept relation capturing capacity of word embedding methods.

4. In order to achieve a high HOP value, a system must have a considerable vocabulary apart from its meaning-capturing power, which means HOP serves as a great extrinsic evaluation metric for word embeddings and taxonomies.

2 Background and Related Work

One of the central subjects of the research is WordNet: a manually created database of more than 150 000 words with their semantic and syntactic relations. It was introduced in 1995 in [8], and has been immensely influential in the field of (computational) linguistics. Fundamentally, it categorises words along two axes: synonyms and hyponyms. Words are grouped into sets of synonyms called synsets. These synsets are nodes in the WordNet, where vertices represent the hyponym relations.

Word embedding or word representation is an umbrella term that describes the Natural Language Processing task of feature selection, and extraction for words [6]. It is a technique to encode words as real-valued vectors that represent word meanings. Thus, words that approximately mean the same thing are closer together in the multi-dimensional vector space. During the research, we will investigate two different word embedding techniques: word2vec and the embedding layer of the transformer model BERT. word2vec was introduced in 2013 in [7], and it utilises a simple, shallow, two-layer neural network to produce word vectors. It has numerous advantages compared to earlier strategies, such as latent semantic analysis. BERT is an influential transformer model introduced in 2018 in the paper [5]. It is a predecessor to newly released transformer models such as GPT-2 and GPT-3. Transformer models utilise word embedding layers to encode text before they process them. In this research, we evaluate classical and more cutting-edge word embedding techniques to ensure that our results are representative.

As the introduction mentions, our research utilises the ”hop2phil” property of Wikipedia articles. We have conducted a preliminary investigation into the validity of this property and found that it stands for the majority of randomly selected articles. The source code for this investigation is accessible through GitHub[1]. Note that this was only a preliminary inquiry, so the randomly selected starter pages might not have been perfectly distributed uniformly. However, the test shows that only a tiny fraction of traversals lead to either cycles or dead-ends. As this phenomenon is almost trivial from the structure of Wikipedia, no research has been found that thoroughly investigates the dynamics behind it.

However, previous research that utilises the semantic connectivity of Wikipedia articles has been conducted. [9] uses Wikipedia’s categorical system and semantic connectivity to compile a large-scale taxonomy. They have compared their generated database with the two well-established taxonomies, WordNet and ResearchCyc, and found that their Wikipedia-derived taxonomy was competitive with theirs. They reason this might be because they have utilised a well-maintained and already structured knowledge base to feed into their derivation processes.

[1] `https://github.com/Arotte/philhopper`

[10] proposed concept vectorisation methods by making use of Wikipedia's category system. These concept vectors describe what concept belongs to what categories and can be used in many applications, such as information extraction and document classification. The research, however, did not demonstrate concrete applications of their algorithms. Their Vector-based Vector Generation method (VVG) has been used in other research papers ([1], [2]) to calculate the membership scores of a concept.

[11] uses Wikipedia's category and page network to construct a novel semantic similarity metric. As an example, their metric shows a semantic similarity between "train" and "car" of 6.31 but only 0.92 for "stock" and "jaguar" on a scale of 0 to 10. Comparing their metric with other similar metrics, they say they have proved its "reasonableness". These papers show that using Wikipedia's semantic connectivity nature is a reasonable choice as it can provide valuable insight into the semantic similarity of words.

3 Research Question

[3] categorizes word embedding evaluation metrics into two main clusters: extrinsic and intrinsic evaluators. Intrinsic evaluators test representational quality without considering NLP tasks, whereas extrinsic ones evaluate models by measuring their performance in specific NLP tasks. HOP falls into the extrinsic evaluator category as it utilizes a unique real-world phenomenon and takes into account vocabulary size apart from pure word-capturing power. The question is, can HOP compete with other, purely word embedding-focused evaluation metrics? If it can, HOP can become a great generalization of similar metrics as it allows to the comparison of embeddings to taxonomies.

When we only consider words inside WordNet's vocabulary, does WordNet perform better at capturing semantic relationships than trained word embedding models word2vec and BERT's embedding layer? We believe a WordNet-based word selection strategy produces a better HOP value than word embedding-based strategies. However, as HOP aims at being a generalized evaluation metric, we propose that it would be more appropriate if we did not take into account a selection strategy's vocabulary. A strategy with a considerable vocabulary and a high word meaning capturing power would theoretically yield better overall HOP values, whereas a similar strategy with less capturing power would produce poorer values. This seems only natural, as real-world applications also require a strategy to understand a high number of lemmas or words.

If we do not use the intersection of vocabularies of all the evaluated strategies but allow all of them to utilize their whole vocabulary, would WordNet still produce the best HOP values? We believe that word embeddings would produce better values as their vocabulary is superior compared to the taxonomy. Would a system with a slightly smaller meaning capturing power but a bigger vocabulary result in higher HOP values?

When following the first, second or random hyperlinks, do we get a HOP-value higher if we followed a more informed word selection strategy? We believe informed word selection will result in better HOP values.

4 Experimental Design

HOP works in the following way. From a carefully selected set of starter articles, traverse the hyperlink chain until one of the three terminating conditions is reached (cycle, dead-end, and Philosophy article) for each of the word selection strategies. The HOP value of a word selection strategy will be the mean of the hop lengths for each starter article. We select the word with the shortest concept relationship path to 'philosophy' in the WordNet word selection strategy. In the embedding word selection strategy, we select the word whose vector has the smallest cosine similarity metric to the vector of 'philosophy'. Select a set of articles from Wikipedia. For each of these pages, try to traverse to the Philosophy article by hopping from hyperlink to hyperlink. Repeat this process for every starter page and word selection strategy. Not all traversals lead to the target article. We define three possible terminating conditions: a) the target article is reached, b) a cycle is detected, and c) a dead-end is reached. A cycle means the traversal enters an infinite loop, hopping around the same articles. A dead-end can mean a parsing error in the article or an article that does not have enough hyperlinks.

Word selection strategies are applied at every traversal step and determine the next node in the traversal chain. A well-performing strategy would select words semantically closest to the target word 'philosophy', thus minimising the number of hops to reach the target article. An ill-performing strategy would wander around either never finding the target article (maybe ending up in a terminating condition) or finding it with a sub-optimal number of steps. A baseline word selection strategy will be used to compare results. In this strategy, we a) select a hyperlink randomly from the first N links in an article, b) select the first and c) second viable hyperlinks. These strategies would likely be ill-performing. In the WordNet selection strategy, select one with the shortest concept relationship path to the word 'philosophy' from the first N hyperlinks in an article. In the word embedding strategy, we Select a word with the highest cosine similarity to the target word from the first N hyperlinks in an article. This strategy will have two versions. Version A will use word2vec's embedding vectors to determine similarity scores, and Version B will use BERT's word embedding layer.

Figure 1 illustrates the steps and decisions of the BERT word selection strategy during the hopping process. The target Philosophy article is reached with just six hops.

Correctly selecting the set of starter Wikipedia articles is a crucial prerequisite for representative results. The articles have to be uniformly distributed in terms of "abstractiveness". The set has to contain an equal number of articles explaining concrete concepts (like "BMW X1") and more abstract concepts (like "religion"). A possible way could be to measure the number of hops to get to Philosophy by just using the first links in an article. This way, the hop length will be higher for concrete articles and lower for more abstract articles. However, this approach has many edge cases, so some of the pages will have to be selected manually, as there is no efficient and reliable automation available to determine an article's concreteness. A second approach is to gather a large quantity of randomly selected articles with their first one or two paragraphs and use a lemma concreteness database to determine the mean concreteness of every article. This approach uses the database collected in [4]. The number of articles has to be sufficiently high. Considering that the time required to hop between 15

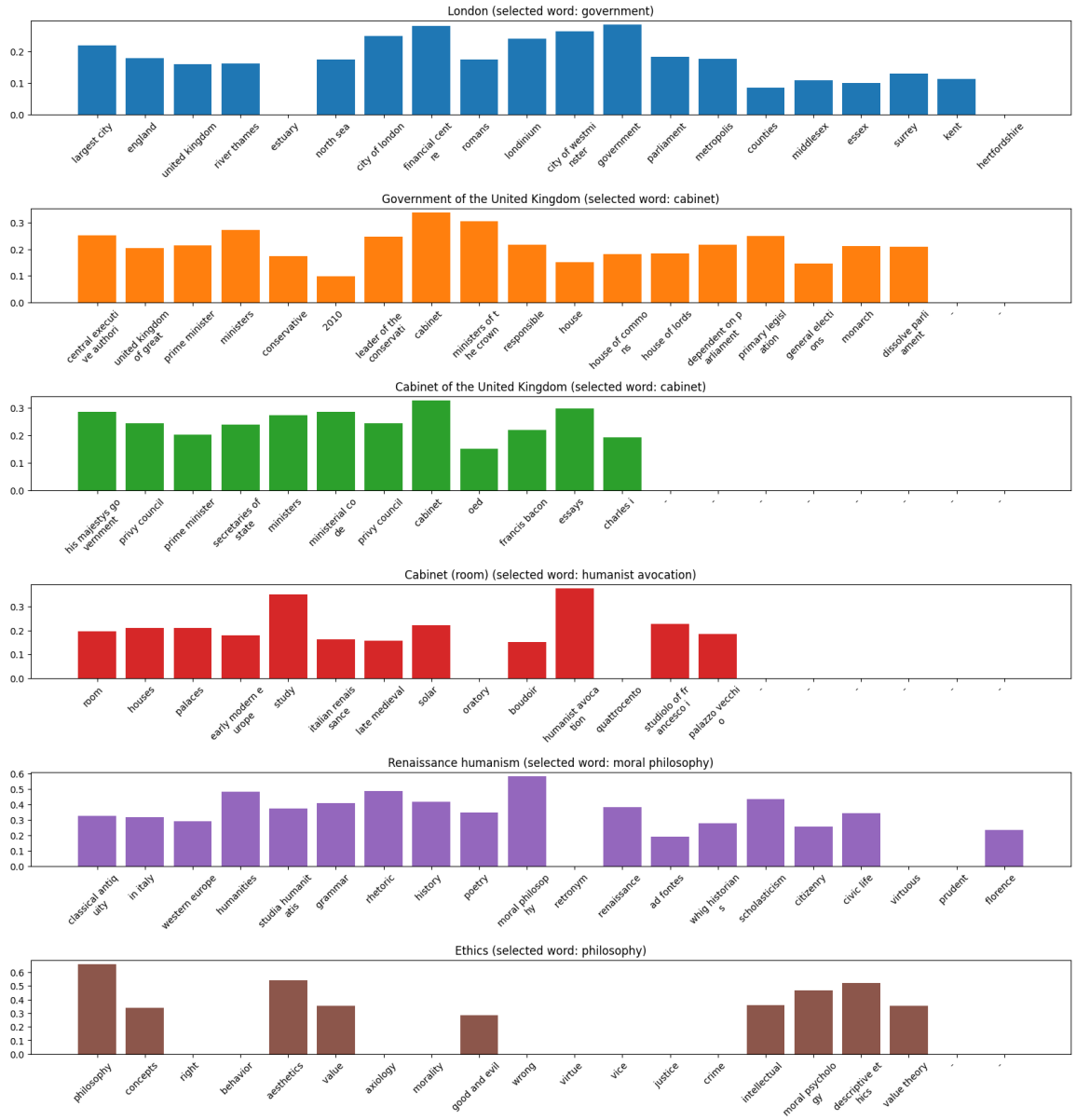

Figure 1: Hopping from London's Wikipedia article to the Philosophy article utilizing the BERT word selection strategy. The target article is reached within just six hops. Each subplot on this diagram is a step in the hopping process, starting from the top. The subplots contain the words that were considered at that step on the X-axis. The Y axis shows the cosine similarity of each word to 'philosophy'. At each step, the hyperlink that has the highest cosine similarity is selected.

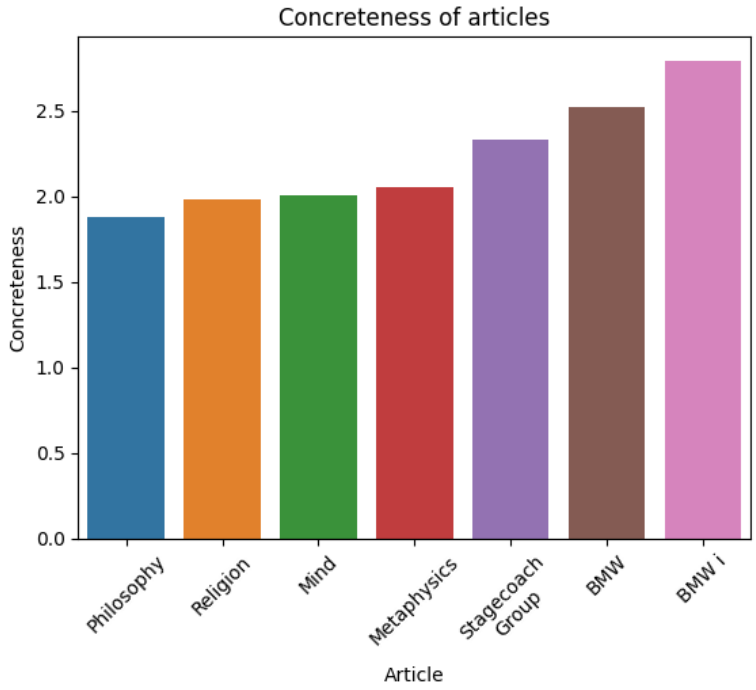

Figure 2: Calculated concreteness score of seven Wikipedia articles.

articles is around 7 seconds (due to a series of GET requests and HTML parsing steps), a reasonable starter page size could be in the thousands. This number can be increased in future experiments when more powerful machines are available.

5 Results

A total of 1177 starter articles were selected. Around a hundred of these were selected manually, ensuring correct concreteness distributions, another hundred or so were selected randomly. The rest of the articles were selected based on their mean concreteness score calculated with the help of the lemma concreteness database from [11]. To calculate a concreteness score for the whole article, the first few sentences were broken into part-of-speech (PoS) tags, and then the concreteness score was retrieved for each tag to calculate their average. If a PoS tag was not present in the database, the Levenshtein [12] distance algorithm was utilized with a threshold of 0.9 to find similar lemmas in the database. Figure 2 demonstrates the calculated mean concreteness score of seven different Wikipedia articles with increasing concreteness. Most of the generated scores make sense. Metaphysics is a branch of Philosophy, thus, it has a higher concreteness score, whereas the concept of the mind is considerably more abstract that the ”BMW i” car model, thus the smaller concreteness score.

The articles were selected such that the concreteness distribution of the final set is uniform and not skewed towards either the more abstract or, the more concrete. Figure 3 displays the distribution of concreteness scores among the selected starter articles. The distribution is not perfectly uniform, it slightly skews towards the less concrete but is in an acceptable interval.

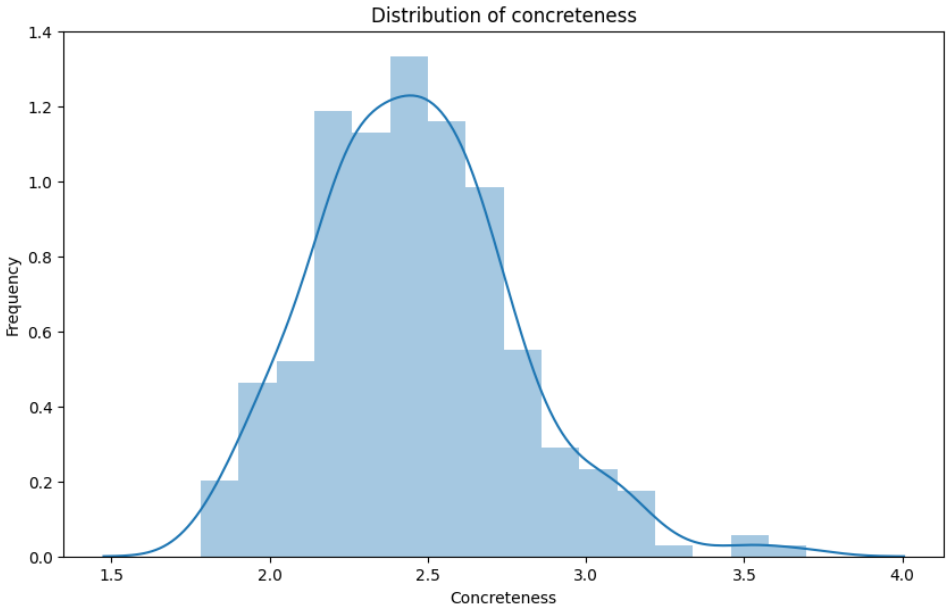

Figure 3: Distribution of concreteness scores among the 1177 starter articles.

Figure 4 aims to convey similar information about the concreteness distribution of the starter articles to Figure 3. Most articles tend to be clustered around the mean, which is ideal.

From the considerable amount of starter articles, not all selection strategies managed to reach the Philosophy article. Table 1 summarizes how many successful traversals happened for each selection strategy. The random strategy only reached the goal three times out of 1177, which is because the maximum hop length was set to 100, and random word selection rarely get to Philosophy in under 100 hops. The BERT embedding selection strategy has only 93 successful traversals because it often got into cycles and sometimes parsing errors and because not all of the 1177 starter articles were tried (but a smaller subset of them with the same concreteness distribution). This was because a typical hop in the case of BERT took between 20 to 35 seconds (due to the compute-intensive contextual embedding aspect).

In general, the most common reason for unsuccessful traversals for both taxonomy-based and embedding-based strategies was infinite cycles. We have tried mitigating this issue with tricks, such as removing the cycle-causing word from the list of words in the current article. However, we have found that removing naturally occurring cycles causes under-representative results, thus, we decided to leave cycles in.

Figure 1 is an example of how the BERT word selection strategy reaches the target article in just six hops starting from London's article. When making informed word selections, we considered the first 20 hyperlinks of each article, but some articles did not have that many hyperlinks, which is why some of the bars are missing at the end of the subplots in the figure. Some cosine similarity scores are also missing (or 0.0), which indicates which lemmas were not in the vocabulary of BERT.

There are a considerable amount of hyperlinks in each article which consist of multiple words.

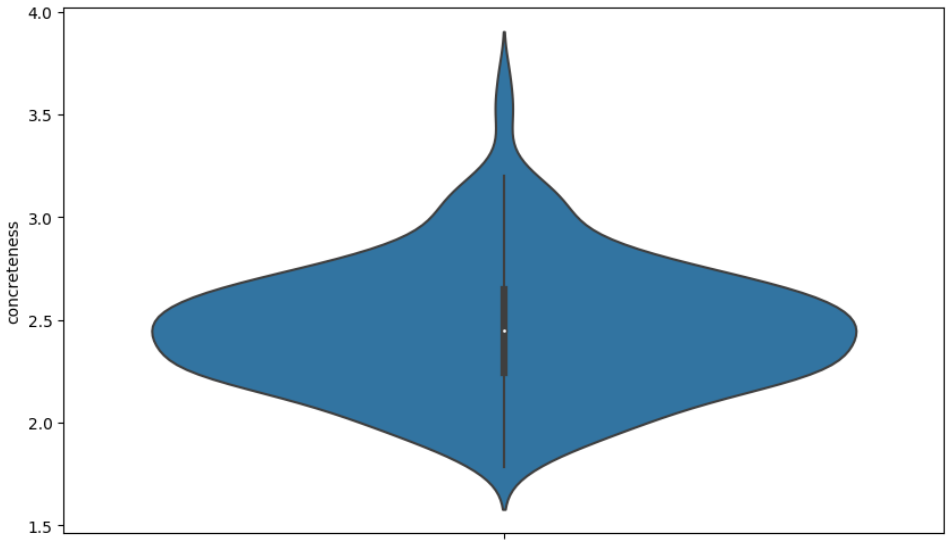

Figure 4: Violin plot concreteness score distribution similar to Figure 3.

Selection	N Reached Phil	Reached Phil (%)
Random	3	0.25
BERT	93	7.90
1st link	295	25.06
2nd link	563	47.83
WordNet	589	50.04
word2vec	611	51.91

Table 1: Traversal success for selection strategies

Selection	HOP-value (Mean Hops)	Std	Count	Min	Max
word2vec	7.27	2.68	611	1.0	19.0
WordNet	7.91	2.96	589	1.0	17.0
BERT	9.12	5.15	93	2.0	30.0
1st link	20.23	7.48	295	1.0	54.0
2nd link	28.39	12.49	563	4.0	51.0
Random	100.00	0.00	3	100.0	100.0

Table 2: Average number of hops to reach Philosophy for each selection strategy combined with min., max. and standard deviation. The average number of hops is the final HOP value for a selection strategy.

This is not a problem for our baseline selection strategies but poses a challenge for our taxonomy and embedding-based strategies. To mitigate this problem, take the mean of the individual word embedding vectors and then compare this mean vector with the target word's embedding vector. In the case of WordNet, a similar averaging algorithm was used.

Table 2 summarizes the measured HOP-values of each selection strategy, and Figure 5 displays the hop length distribution of each word selection strategy. Random selection was excluded as almost all hop lengths, in that case, exceed the maximum allowed hop length (100). Interpretation of the results follows in Section 6.

6 Discussion

We have identified that infinite cycles were not a problem for our baseline selection strategies. Their occurrences were negligible. However, cycles posed a serious limitation on the taxonomy and embedding-based selection strategies. A possible explanation of the observations is the following. Upon reaching an article that is relatively close to Philosophy in terms of concreteness, an embedding strategy might select a hyperlink that is semantically really close to the current article. Because of close semantic similarity, the selected article will almost certainly contain a hyperlink to the previous article. If a hyperlink with a really close semantic similarity to Philosophy is not found, the selection strategy will inevitably select the previous article, leading to an infinite loop.

We conclude that the concreteness metric-based starter article selection strategy was ideal. As Figure 2 demonstrates, the metric performs well in determining the abstractiveness of Wikipedia articles, and the distribution of the concreteness metric scores is near-uniform.

The HOP value of the random selection strategy is 100 with 0 standard deviations, meaning that all traversals took more than 100 steps (after 100, the hopping process is terminated). This proves that it is impossible to achieve a small HOP value by randomly following hyperlinks, resulting in the fact that the HOP metric has some validity.

Always following the second hyperlink produced the second-largest HOP value of 28.39 with the highest standard deviation of 12.49. Figure 5 provides some insight into the reason behind the high standard deviation. The number of hops does not cluster around the mean. Instead, there is clustering

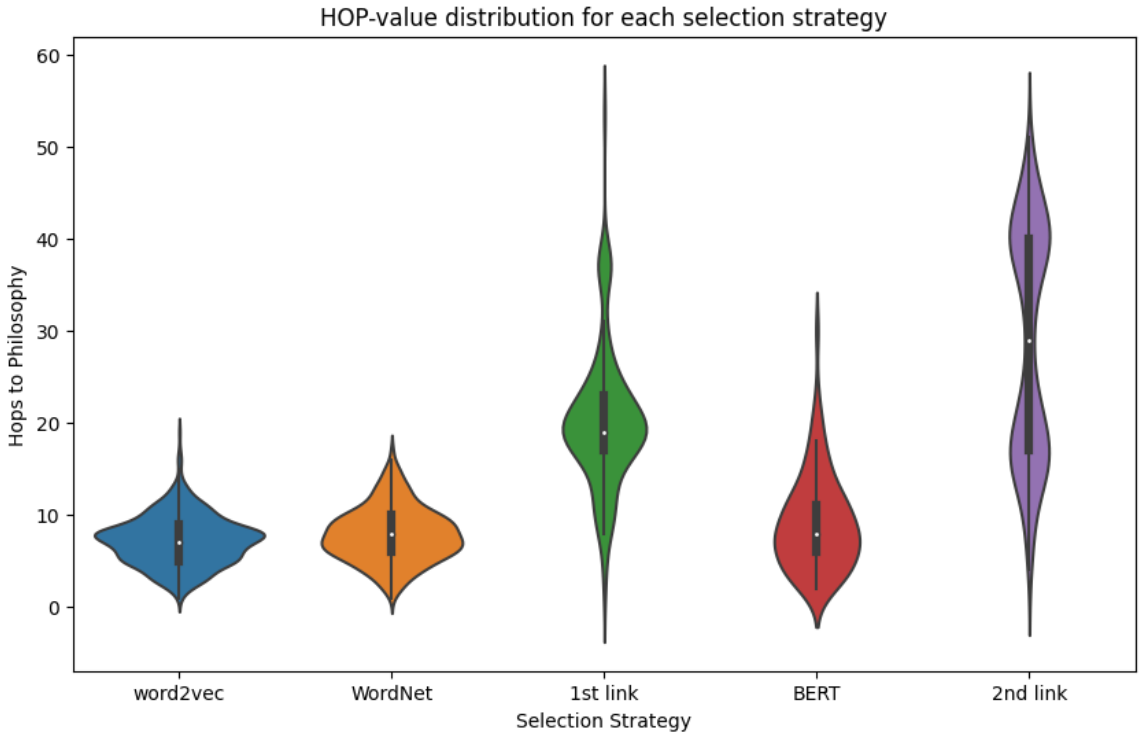

Figure 5: Violin plot summarizing the hop length distribution of each word selection strategy. Random selection is excluded as it would skew the diagram too much without providing much insight.

at around 20 and around 45 hop lengths. We believe this was caused by the near-uniform distribution of the starter article concreteness scores. Following the first hyperlinks through the traversal process produces a smaller HOP value than the second link approach. The HOP value for the first link strategy is 20.23, around eight hops less than the second link approach, with a considerably smaller standard deviation. The violin plot shows a small clustering around hop length 40, which somewhat resembles the second link strategy. The trend here suggests that the further away we are from the first hyperlink, the longer it takes to reach the target article. This is a good result as Wikipedia articles tend to define and categorise entities in the first few words.

In general, our baseline selection strategies perform considerably worse than the taxonomy and embedding-based strategy, which is an ideal result. It confirms our hypothesis that an informed word selection strategy produces lower HOP values.

The BERT-based strategy gives the worse HOP value of 9.12 out of the three, which is an unexpected result. We have hypothesised that BERT would perform better given its superiority in vocabulary and size compared to word2vec. This poor performance might stem from BERT's contextual embedding aspect. During experimentation, we only considered the first two paragraphs of the current article to receive word embeddings from the Transformer. It might have produced more accurate word embeddings and thus performed better if we fed all the previous articles and the target article's first two (or maybe the first three or four) paragraphs.

The best-performing strategy was word2vec, with a HOP value of 7.27, closely followed by WordNet with 7.91. This result is not surprising as we did no justice to WordNet - we did not eliminate words that were not in the taxonomy's vocabulary. The results are interesting because they show that a simple neural network-based word embedding can compensate for its poor meaning by capturing power with an extensive vocabulary. However, the HOP value closeness for both approaches shows that WordNet is considerably better at capturing word meanings.

7 Conclusion & Future Work

In this research, we have proposed HOP, a novel extrinsic evaluation metric for word embeddings and human-annotated taxonomies. Current irrelevant evaluation metrics focus primarily on measuring the performance of word embedding models, but we have shown that creating a more general metric is possible and viable. It allows comparison between different embedding models and human-annotated databases of lemma and word meanings.

We have evaluated HOP on three different baseline selection strategies and found that an informed selection strategy produces considerably better HOP values, meaning that HOP has the potential to become an efficient metric of meaning-capturing power.

We have identified that WordNet gives a slightly worse HOP value than word2vec. Still, they are their values. They are close, meaning that, although word2vec vocabulary is more extensive, WordNet performs at least as well - and probably better - at capturing word and lemma meanings. BERT unexpectedly produced a worse HOP value than word2vec, but this might be because of its contextual embedding aspect, which requires further consideration.

8 Reflective Analysis

Infinite cycles were a considerable bottleneck, and further research is required to mitigate them. Increasing HOP values with increasing n-th hyperlink selections suggest and trend that might be worthy of further examination. We might improve BERT's performance by feeding it more context in the future.

Experience shows that the random selection strategy we used as a baseline is unnecessary as it usually reaches the target after a high amount of hops and does not provide the relevant insights as a baseline should. We hypothesised that when only considering the intersection of vocabularies of all evaluated methods, WordNet would be superior. This assumption was not explicitly measured, so further research should explore the direction.

Generating BERT embeddings was performed on the CPU, making the hopping process really slow. This is part of the reason for having a relatively small amount of target articles reached. Future research should utilize the GPU for Transformer embedding generation.

9 Source Code

1. The whole source code of the project is accessible at `https://github.com/Arotte/HOP`.
2. A lightweight, first-link hopper script is accessible at `https://github.com/Arotte/philhopper`.

References

[1] Mehdi Allahyari and Krys Kochut. Automatic topic labeling using ontology-based topic models. pages 259–264, 12 2015.

[2] Mehdi Allahyari, Krys Kochut, and Maciej Janik. Ontology-based text classification into dynamically defined topics. pages 273–278, 06 2014.

[3] Amir Bakarov. A survey of word embeddings evaluation methods, 2018.

[4] Marc Brysbaert, Amy Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46, 10 2013.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[6] Yang Li and Tao Yang. *Word Embedding for Understanding Natural Language: A Survey*, pages 83–104. Springer International Publishing, Cham, 2018.

[7] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[8] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995.

[9] Simone Paolo Ponzetto and Michael Strube. Deriving a large scale taxonomy from wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, page 1440–1445. AAAI Press, 2007.

[10] Masumi Shirakawa, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. Concept vector extraction from wikipedia category network. In *Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication*, ICUIMC '09, page 71–79, New York, NY, USA, 2009. Association for Computing Machinery.

[11] Feiyue Ye, Feng Zhang, Xiangfeng Luo, and Lingyu Xu. Research on measuring semantic correlation based on the wikipedia hyperlink network. In *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*, pages 309–314, June 2013.

[12] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095, 2007.