Leveraging the Wikipedia Graph for Evaluating Word Embeddings

Joachim Giesen[1] , **Paul Kahlmeyer**[2] , **Frank Nussbaum**[1,2] and **Sina Zarrieß**[3]
[1]Friedrich Schiller University Jena
[2]DLR Institute of Data Science
[3]Bielefeld University

Abstract

Deep learning models for different NLP tasks often rely on pre-trained word embeddings, that is, vector representations of words. Therefore, it is crucial to evaluate pre-trained word embeddings independently of downstream tasks. Such evaluations try to assess whether the geometry induced by a word embedding captures connections made in natural language, such as, analogies, clustering of words, or word similarities. Here, traditionally, similarity is measured by comparison to human judgment. However, explicitly annotating word pairs with similarity scores by surveying humans is expensive. We tackle this problem by formulating a similarity measure that is based on an agent for routing the Wikipedia hyperlink graph. In this graph, word similarities are implicitly encoded by edges between articles. We show on the English Wikipedia that our measure correlates well with a large group of traditional similarity measures, while covering a much larger proportion of words and avoiding explicit human labeling. Moreover, since Wikipedia is available in more than 300 languages, our measure can easily be adapted to other languages, in contrast to traditional similarity measures.

1 Introduction

NLP models based on neural networks require a conversion of words into a vector representation as a first step [Sezerer and Tekir, 2021]. Based on this step, which is called embedding, one can then perform downstream tasks such as text classification, translation, or tagging. NLP models can be learned in an end-to-end fashion if enough data is available. There are however areas, such as neural machine translation [Qi *et al.*, 2018] or text classification [Kim, 2014], where often no large scale training corpora exist, especially in specialized domains. In these cases, pre-trained word embeddings can either be used as off-the-shelf building blocks or as initialization for the first part of a deep learning pipeline [Mandelbaum and Shalev, 2016]. To be generally useful, word embeddings must capture the semantic meaning of words. Therefore, evaluating pre-trained word embeddings in terms of the semantics they capture is an important task.

A large class of metrics for evaluating word embeddings are similarity based: Similar or related words should be close in the embedding space. To set up these metrics, usually relatedness/similarity scores for word pairs are determined by human judgement, see [Bakarov, 2018] and references therein. However, it can be time-consuming and expensive to collect these scores. That may not be a problem for the English language, but NLP models are trained for almost every other language as well. Furthermore, as pointed out by[Faruqui *et al.*, 2016], so far there are no standardized protocols for collecting relatedness/similarity scores, which causes several problems in the evaluation of word embeddings.

The aforementioned difficulties have motivated works that do not rely on human data and survey protocols. For instance, [Torregrossa *et al.*, 2020] set up global metrics that are independent of external data and only use the embedded word vectors. However, even after fine-tuning hyperparameters, so far these global metrics have limited surrogate power for similarity-based metrics, see Figure 1 and Section 4 for details. To address this problem, we propose leveraging relatedness information that is encoded implicitly in Wikipedia graphs. This allows us to bypass the expensive explicit collection of word similarity and relatedness information.

We ground our approach in a theoretical information-foraging framework [Chi *et al.*, 2001] that was developed to model user information needs and actions on the Web. Here, we apply the framework to Wikipedia graphs. We set up an agent that needs to navigate from a given start Wikipedia article to a target article. The agent's strategy is to follow links that maximize relatedness to the target article, chosen from the outgoing links of the current article. Here, the target article corresponds to the agents' information need. Selecting intermediate articles by maximizing relatedness is a simple strategy for satisfying this need. This is consistent with human behavior, as Internet users often greedily follow links that they believe are most likely to meet their information needs.

For the greedy information search strategy, we define the relatedness of (candidate) articles to the target article by the cosine similarity of the embedding of their title words. Therefore, different embeddings lead to different search paths. For a fixed embedding, we calculate the ratio of the shortest path length and the number of Wikipedia pages that the agent visits before the target article is reached (averaged over several start-target pairs). The better the embedding preserves relatedness, the more efficient can the agent navigate towards the target article, yielding a larger value of the metric. We call the metric, which we describe in detail in Section 3, WALES metric (**W**ikipedia **A**gent using **L**ocal **E**mbedding **S**imilarities).

As can be seen in Figure 1, the WALES metric correlates strongly with many relatedness and similarity metrics derived from human judgements. This proves that the implicitly encoded relatedness information in the Wikipedia graph can be used to form a surrogate relatedness metric that does not rely on expensive data collection from humans.

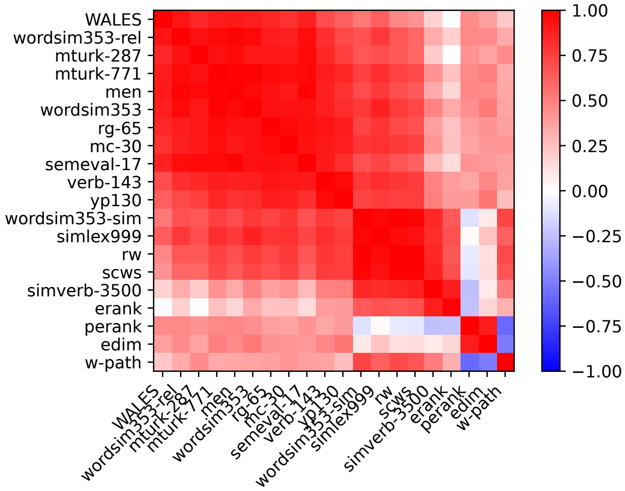

Figure 1: Spearman rank correlation of the proposed WALES metric with other similarity metrics. erank, perank ($p = 2.5$), and edim ($p = 1.0$) are global metrics, and w-path is a baseline metric that is derived directly from shortest paths on Wikipedia.

The performance of word embeddings on information-foraging tasks can be compared with the one of humans. For that purpose, we have scraped a human benchmark data set from the *The Wiki Game* website that is maintained by [Clemesha, 2018]. In this game, humans have to navigate from one Wikipedia article to another while minimizing the number of used links, which is precisely the task that our agents solve. It turns out that for many word embeddings, the performance of our agents is at least on par with the performance of humans.

Finally, we show experimentally that the WALES metric is robust in the sense that it yields consistent rankings of word embeddings for a family of information foraging-strategies, where the greedy strategy that we outlined above is just one member. We also investigate the influence of the distribution of start-target pairs.

2 Related Work

As word embeddings, that is, vector-valued representations of words derived from distributional semantic models, have become an integral part of most modern NLP pipelines, also the need for understanding and evaluating these embeddings has become more important. A thorough and detailed survey of word embedding evaluation methods is given by [Bakarov, 2018] who follows the distinction introduced by [Schnabel *et al.*, 2015] into intrinsic and extrinsic evaluation methods. Extrinsic evaluations measure the performance of word embeddings on some downstream task such as entity recognition or part-of-speech tagging, where the word vectors are used as features. Here, we focus on intrinsic evaluations that test for semantic relationships between words by comparison to human judgement. However, the concept of semantic relations is not clearly defined as long as it is unclear what kind of relationships should be reflected in a word embedding.

[Lastra-Díaz *et al.*, 2019] point out the crucial difference between semantic similarity and semantic relatedness. Semantic similarity can be based on the *is-a* relationship between concepts, whereas semantically related concepts have more general relationships between them. Lastra-Diaz et al. illustrate the difference between semantic similarity and relatedness on the concepts of *car*, *bicycle*, *wheel*, and *fuel*. Cars and bicycles are semantically similar since both are *vehicles*, whereas *wheel* and *car* are in an *is-part-of* (merynomy) relationship. Other classic relationships are hypernymy, hyponymy, antonymy, and synonymy [Murphy, 1986]. Cars and fuel are also clearly semantically related, however, the relationship is not covered by one of the classical relationships. Here, we focus on semantic relatedness and build on semantic relationships reflected in Wikipedia graphs that go beyond the classic relationships.

Classical semantic relationships of words are typically encoded in thesauri, taxonomies, and semantic networks like WordNet or Wikidata. The human knowledge that is encoded in these resources can be used to evaluate word embeddings [Agirre *et al.*, 2009]. However, the majority of evaluations of word embeddings are still by comparison to human judgements on word relationships. Judgements are collected either in laboratory studies or through crowdsourcing on four broad categories of tasks [Liza and Grzes, 2016], namely, direct similarity/relatedness scores, analogy, categorization, and thematic fit. In an analogy task, the goal is to find a term w_2 for a term v_2 such that $w_2 : v_2$ resembles the given relationship $w_1 : v_1$. In categorization tasks, words have to be grouped into clusters, and in thematic fit tasks, it has to be judged how typical a noun is for a verb [Baroni *et al.*, 2014].

[Torregrossa *et al.*, 2020] introduce global metrics as a third class of evaluation methods, next to intrinsic and extrinsic methods. Global metrics are data free in the sense that they do not use data beyond the embedding. Global metrics are basically formulas that are evaluated on the embeddings. Two specific global metrics that have been introduced by Torregrossa et al., called *perank* and *edim*, respectively, are included in the correlation plot in Figure 1. Torregrossa et al. point out that global metrics are inexpensive evaluation methods that avoid the costly and time-consuming collection of human judgements. In this aspect, their work is close to ours. However, while being inexpensive, our approach additionally benefits from the information encoded in Wikipedia graphs, see the experiments in Section 4.

3 The WALES Metric

In this section, we design a family of greedy word-embedding based agents for solving information-foraging tasks on Wikipedia graphs. The agents are derived from the information-foraging framework [Chi *et al.*, 2001]. This framework applies to users that seek information by using proximal cues on a collection of hyperlinked documents.

3.1 Information-Foraging Framework

In a nutshell, the information-foraging framework on n hyperlinked documents and a dictionary D of d words is modeled by a tuple (T, W, K). Here, the (non-symmetric) $(n{\times}n)$-adjacency matrix T describes the hyperlink topology, the non-negative $(d \times n)$-matrix W describes the importance of the words for the documents, and the $(n \times n \times d)$-tensor K describes (proximal) information cues at the links. In a simple model, the information cue at a link is given by a d-dimensional indicator vector over the dictionary. Furthermore, a user's information need is modeled by a query word indicator vector $q \in \mathbb{R}^d$. From the model and the query, the similarity between the query vector and the proximal cues at the links are computed as

$$P_q(i,j) = T(i,j) \cdot \sum_{w \in D} W(w,j) \cdot K(i,j,w) \cdot q(w).$$

Here, $P_q(i,j)$ is the similarity of the query vector q with the target document j, which is linked to from the source document i. The $(n \times n)$-matrix P_q is then normalized such that it becomes stochastic, that is, the entries of each column sum up to 1. Thus, $P_q(i,j)$ is interpreted as the probability that a user with information need q navigates from the i-th to the j-th document. A typical application of the framework is to infer the information need from a user's navigation history [Chi *et al.*, 2001].

3.2 Wikipedia Information-Foraging Agents

Here, we instantiate the information-foraging framework for a Wikipedia like the English Wikipedia. The instantiation of the adjacency matrix T is straightforward. It encodes the hyperlink structure of the Wikipedia. We also use the particularly simple model with $W \equiv 1$. Next, we model the information need by a given target Wikipedia article t. Let w_t be the title word of t. The information need vector $q = q_t \in \mathbb{R}^d$ is the one-hot encoding of w_t. Now it only remains to specify the information cue tensor K in order to be able to compute P_q. We do so by setting

$$K(i,j,w) = \cos(f(w_j), f(w)),$$

where we denote the title of the Wikipedia article j by w_j, and f is a word embedding. Hence, the information cue tensor is independent of the source article i and given by the cosine similarity of the vectors $f(w_j)$ and $f(w)$, that is, of the embedded article titles. Here, the cosine similarity of two vectors a, b of the same dimension is defined by $\cos(a,b) = a^\top b/(\|a\|\|b\|)$. With this instantiation, the unnormalized P_q computes as

$$\begin{aligned} P_q(i,j) &= T(i,j) \cdot K(i,j,w_t) \\ &= T(i,j) \cdot \cos(f(w_j), f(w_t)). \end{aligned}$$

In the direct instantiation of the information-foraging framework, an agent navigates stochastically from article to article, where the transition probabilities are obtained after normalizing the columns of the matrix P_q. In order to derive a deterministic metric on word embeddings, we deviate from the direct instantiation and consider greedy agents that always follow a link with the highest probability. Purely greedy agents, however, may suffer from the problem of getting stuck or cycling on the Wikipedia graph. To address this problem, we move from a Markov chain with stationary transition matrix to a deterministic non-stationary search process that we define in the following.

We assume that agents do not know the full Wikipedia graph given by the adjacency matrix T, but memorize the sub-graph that has been revealed to them during their search. Formally, let $v^{(i)}$ be the node visited by the agent at the i-th step. Then, the revealed sub-graph T_i contains only directed edges that emerge from nodes $v^{(j)}, j \leq i$ that have been visited already. The nodes of T_i are given by the nodes adjacent to any edge in T_i. For $\gamma \in [0,1]$, we define an agent based on the following decision rule for selecting a follow-up node:

$$v^* = \underset{v \in T_i \colon \deg(v)=0,\, m_i(v)<\infty}{\arg\max} \cos(f(w_v), f(w_t)) - \gamma m_i(v).$$

This problem maximizes the cosine similarity of the candidate nodes v to the target node t after applying a word embedding f to the corresponding article titles. Moreover, only nodes in T_i with outgoing degree zero ($\deg(v) = 0$) are considered because these are the nodes in T_i that have not been visited before. The parameter γ penalizes the number of links followed, where $m_i(v)$ is the length of a directed shortest path from $v^{(i)}$ to v in the sub-graph T_i if it exists, and $m_i(v) = \infty$ otherwise. The routing of the agents is illustrated in Figure 2 and some real-world examples in Figure 3.

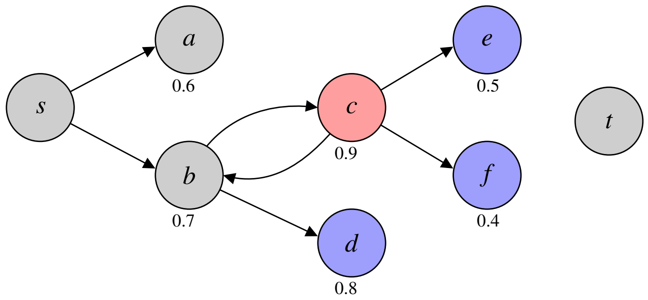

Figure 2: Example for the decision rule on the sub-graph. The agent has followed the path $s \to b \to c$. The cosine similarities with the target node t are noted below the respective nodes. Red: current node, blue: reachable unvisited nodes within the sub-graph. For $\gamma \geq 0.3$, the follow-up node is $v^* = e$ with cost $0.5 - \gamma$, otherwise $v^* = d$ with regularized similarity score $0.8 - 2\gamma > 0.5 - \gamma$.

The decision rule depends on the choice of the embedding f and γ. Here, for $\gamma = 0$, the selected node is the node that maximizes similarity to the target node from among *all* reachable and unvisited nodes in T_i. In contrast, the agent with $\gamma = 1$ always follows a link to a direct neighbor that has not yet been visited–as long as such a neighbor exists. Here, first considering only direct neighbors for the maximization of the cosine similarity allows for fast computation. Only in the rare scenario, where all direct neighbors have already been visited, the agent with $\gamma = 1$ increases its search depth.

3.3 Evaluation of Word Embeddings

For evaluating the performance of an agent, we define an information-foraging task as a pair of a start article s and a target article t, where the latter article represents the information need. We measure the performance of an agent that follows the path $\xi = (v^{(0)} = s, \ldots, v^{(n)})$ for a given task (s, t) by the score

$$L(s, t, \xi) = \frac{m(s, t)}{n}.$$

This score is the ratio of the length $m(s, t)$ of a shortest path from s to t and the length n of the taken path. Here, we assume a strongly connected Wikipedia sub-graph, which ensures that each node is reachable from each node and shortest paths exist. By definition, it holds $0 \leq L(s, t, \xi) \leq 1$, where a value of 1 signifies that the agent found a shortest path.

Now, to evaluate the performance of a word embedding f, we introduce the WALES metric as

$$\text{WALES}_{A,p}(f) = E_{s,t\sim p}[L(s, t, \xi_A(s, t, f))],$$

where p is a probability distribution on the nodes of the Wikipedia graph from which tasks are drawn at random, A denotes the (deterministic) agent, and $\xi_A(s, t, f)$ is the path taken by the agent A using word embedding f.

The choice of the probability distribution p on the nodes of the Wikipedia graph allows to emphasize the importance of certain words for the evaluation. Here, using a uniform task-generating distribution imposes a uniform prior on the words. Since it is generally harder to navigate to nodes with only few incoming edges, another interesting class of task-generating distributions are skewed by putting more weight on nodes with many incoming edges.

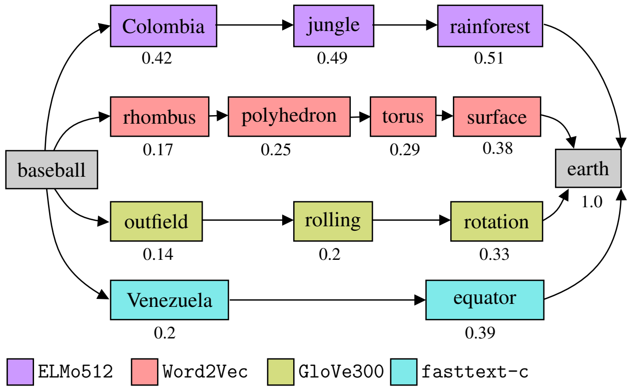

Figure 3: Paths between the Wikipedia articles for *baseball* and *earth* followed by different embeddings. While `ELMo512` and `fasttext-c` navigate over geographic concepts, `Word2Vec` and `GloVe300` use geometric concepts. Cosine similarities to the target *earth* are noted below each node. More examples can be found in the supplement.

4 Experiments

We evaluate the WALES metric along three dimensions, namely, robustness with respect to the involved hyperparameters, comparison to baseline methods, and, of course, correlation with metrics based on collected human similarity judgements from the literature. Here, we use a sub-graph of the English Wikipedia hyperlink graph collected by the Stanford Network Analysis Project (SNAP) [Boldi *et al.*, 2011]. Since WALES is a routing task, we further restrict the graph to the largest strongly connected component of the original graph. The resulting graph has $n = 38\,609$ of the original $4\,203\,323$ nodes, making it easier to handle and preventing impossible routing tasks. We demonstrate WALES on pretrained word embeddings for Google's `Word2Vec` [Mikolov *et al.*, 2013], Stanford's `GloVe` [Pennington *et al.*, 2014], Facebook's `fasttext` [Mikolov *et al.*, 2018], and AllenNLPs `ELMo` [Peters *et al.*, 2018] (for details and versions see the supplement). As can be seen in Figure 3, different embeddings lead to distinctly different behaviors of the information foraging agents. Note that even though some of these embeddings were trained on the English Wikipedia, this does not pose a problem as we are using the topology of the hyperlink graph instead of the text corpus. Also, our claims about robustness and correlation with human judgments of the WALES metric are not affected by training on the Wikipedia.

4.1 Robustness of the WALES Metric

The WALES metric depends on the choice of the distribution of start-target pairs and the decision rule for the agents. In this section, we show that the WALES metric yields consistent rankings for the considered embeddings when distribution and decision rule are varied. However, first we show that the expected value can be replaced by empirical averages in practice.

Practical Calculation of the WALES Metric. By the law of large numbers, the expectation in the WALES metric is well approximated by the empirical average

$$\text{WALES}_{A,p}(f) \approx \frac{1}{k} \sum_{i=1}^{k} L(s_i, t_i, \xi_A(s_i, t_i, f))$$

over a sufficiently large number of tasks (s_i, t_i), $i = 1, \ldots, k$ that are drawn i.i.d. from the task-generating distribution p.

As a preliminary experiment, we determined a size for the task sets for which the empirical average is close to the expected value of the WALES metric. For that, we computed 95% confidence intervals, averaged over 50 randomly generated task sets. Table 1 shows the results for a uniform task-generating distribution and the `fasttext-c` embedding, see the supplement for additional embeddings and distributions. A good approximation of the expected value is obtained for $k = 1\,000$ tasks. Further increasing the size of the task sets is not justified since the computational cost increases linearly in the number of tasks (for $1\,000$ tasks, good embeddings require less than 3 seconds, see the supplement for details). Therefore, from now on, we use task sets of size $1\,000$.

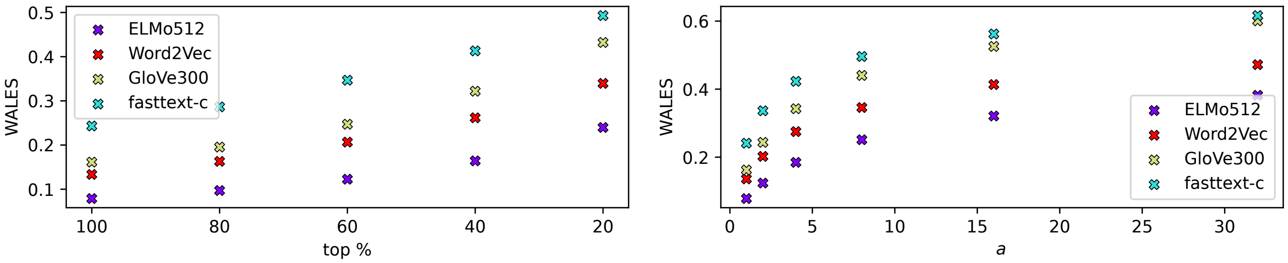

Figure 4: Left: Evaluation of word embeddings by the WALES metric with tasks from a uniform distribution over the top percent of nodes, sorted ascendingly by number of incoming edges. Right: Evaluation of word embeddings by the WALES metric with tasks that consist of nodes from a power law. Task sets of size 1 000 ensure that the expectation in the WALES metric is recovered reliably.

k	10	50	100	500	1000
mean	0.23	0.24	0.24	0.24	0.24
CI	0.046	0.022	0.015	0.008	0.006

Table 1: Mean and size of 95% confidence intervals (CI) of the WALES metric for an increasing number of uniformly selected tasks k and the `fasttext-c` embedding, averaging over 50 trials.

Impact of the Task-Generating Distribution. Here, we investigate how the choice of the distribution for the start-target pairs influences the rankings of the embeddings. For this experiment, we sample nodes dependent on their number of incoming links (in-degree). This characteristic provides some control over the difficulty of the routing tasks. This is because intuitively, target nodes with a larger number of incoming links can be reached more easily.

To facilitate the sampling, we sort the n nodes of our Wikipedia sub-graph such that they have ascending in-degrees. As a first class of task-generating distributions, we consider power laws, where nodes for the tasks are generated as $\lfloor nx \rfloor \in \{0, \dots, n-1\}$ with $x \sim ax^{a-1} \in [0,1]$. Here, we use $a = 1, 2, 4, 8, 16, 32$ as hyperparameters. As a second class, we consider uniform distributions on the top b percent of nodes with the largest in-degrees, where $b = 100$ amounts to the uniform distribution over all tasks.

We respectively drew 1 000 tasks (start-target pairs of Wikipedia articles) from each distribution. As we have seen, task sets of this size approximate the expected value in the WALES metric well. The results of the WALES metric using greedy agents with $\gamma = 1$ can be found in Figure 4. WALES yields consistent rankings of the considered embeddings for all considered distributions, namely, from best to worst, `fasttext`, `GloVe`, `Word2Vec`, `ELMo`.

In the supplementary material, we show that the ranking for different `fasttext` embeddings is also stable, where the `fasttext-c` embeddings consistently outperformed the `fasttext-w` embeddings.

Impact of the Decision Rule. The rankings obtained from the WALES metric are also robust with respect to the choice of $\gamma \in [0,1]$, see Figure 5. For roughly $\gamma \geq 0.1$, the penalty for taking detours is already large enough such that in practice, agents behave like the decision rule with $\gamma = 1$. As the rankings are stable for all $\gamma \in [0,1]$, we suggest using the greedy strategy with $\gamma = 1$. It allows for the fastest implementation as most of the time only outgoing links from the current article are considered.

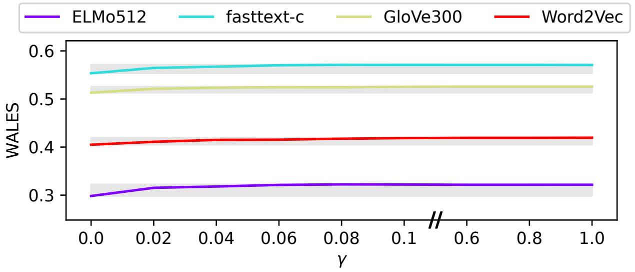

Figure 5: WALES metric on 1 000 tasks from a power law ($a = 16$) for agents with decision rules parameterized by $\gamma \in [0,1]$.

4.2 Baselines

In this section, we present two baselines. The first facilitates interpretability of the WALES metric. The second is an alternative metric that also builds on the Wikipedia graph, but does not measure relatedness as well as the WALES metric.

Human Baseline: The Wiki Game. Interestingly, there is a game called *The Wiki Game* in which information-foraging tasks as we used to set up the WALES metric play a central role. In each round of the game, players receive a challenge given by a start and a target Wikipedia article. Then, they have to navigate from the start to the target Wikipedia article, following as few links as possible. *The Wiki Game* is implemented online [Clemesha, 2018], where for each challenge there is a time limit of 120 seconds.

With friendly permission, we scraped human benchmark data from the *The Wiki Game* website. For this, we collected a data set of $k = 1\,000$ challenges (start-target pairs) that appeared on the website. During the collection, we skipped start-target pairs for which at least one of the start and target articles was not a node of our strongly connected Wikipedia sub-graph. Moreover, we skipped start-target pairs for which no successful human attempts were recorded on the website.

We calculate the human benchmark on the *The Wiki Game* data set by averaging over task scores $m(s,t)/\mathrm{agr}(s,t)$, where the aggregation function $\mathrm{agr}(s,t)$ calculates either the average, median, or best (=shortest) path length from all successful human attempts for the challenge (s, t) from the *The Wiki Game* data set. This way, we obtained three baselines: *average*, *median*, and *best*. Note that *The Wiki Game* is implemented on the full Wikipedia graph, consequently, we used shortest paths $m(s, t)$ from the full Wikipedia, which we obtained from the online service by [Wenger, 2018].

The results of the WALES metric for different embeddings on the *The Wiki Game* data set are shown in Figure 6. Here, the `fasttext` embeddings consistently outperform the human baseline. Note that the human benchmark likely overestimates the average human performance because failed human attempts for solving a challenge are not recorded publicly on the *The Wiki Game* Website. The fact that `fasttext` and `GloVe` embeddings still beat the human baseline is a good indicator for the quality of these embeddings.

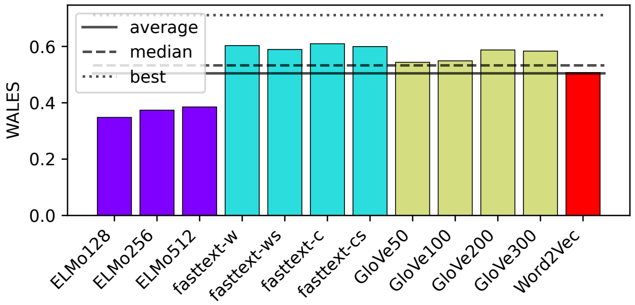

Figure 6: WALES metric ($\gamma = 1$) for different embeddings on 1 000 tasks with human benchmarks collected from [Clemesha, 2018].

Observe that the values of the WALES metric in Figure 6 on the *The Wiki Game* data set are close to the ones for the power law with $a = 32$ in Figure 4. Indeed, the distribution of the tasks in the *The Wiki Game* data set is well approximated by a power law with $a = 32$, see the supplement.

Wikipedia Shortest Path Baseline. One of the earliest, but still influential, measures of word similarity is by [Rada *et al.*, 1989] who define semantic distance as the length of the shortest path between concepts in a taxonomy. In this spirit, we use the (negative) minimal distance between two articles in the Wikipedia graph as a proximal measure of relatedness for title words of Wikipedia articles. Formally, we define a baseline metric *w-path* for word embeddings as the Spearman rank correlation of the negative shortest paths $-m(i, j)$ and the cosine similarity $\cos(w_i, w_j)$ of embedded title words, computed over 1 000 pairs of Wikipedia articles (i, j). Note that traditional similarity metrics are computed analogously, using human-provided similarity scores instead of the negative shortest path distances. However, the *w-path* metric does not correlate strongly with these traditional metrics, as can be seen in Figure 1, and thus should not be used to replace them.

4.3 Comparison to Other Similarity Metrics

Here, we compare the WALES metric with traditional similarity metrics from [Bakarov, 2018]. These metrics are named after data sets with human similarity/relatedness judgements. Note again that a traditional similarity metric is computed as the Spearman rank correlation of the cosine similarities of word pairs in the embedding space with the corresponding human judgements, over all word pairs from the metric-specific data set. A first observation is that the WALES metric involves a larger vocabulary than most traditional similarity metrics, see the supplement. Here, we study in detail how WALES correlates with traditional similarity metrics.

[Torregrossa *et al.*, 2020] observed that most (traditional) similarity metrics strongly correlate–indicating that these metrics measure a similar concept, despite the fairly different protocols for collecting human similarity/relatedness scores. Here, we check how WALES correlates with the traditional similarity metrics from [Bakarov, 2018]. For that, we computed Spearman rank correlations after evaluating all pretrained word embeddings with each considered metric.

The results are shown in Figure 1. There are two main groups with strongly positive correlation. The WALES metric is in the first group, which contains metrics that focus on measuring semantic relatedness, among them *Wordsim353-rel*, *mturk*, and *men*. WALES measures semantic relatedness well since links in Wikipedia strongly result from relatedness.

All metrics from the second group focus on semantic similarity rather than relatedness. As we noted above, WALES primarily measures relatedness. Therefore, WALES correlates to a lesser extent with the measures from the second group. Additionally, some data sets for metrics from the second group measure semantic similarity of verbs and adjectives (for example, *simverb-3500*, *RW*). However, the Wikipedia graph has a strong noun bias [Hoenen, 2016].

Also observe in Figure 1 that the global metrics *erank*, *perank*, and *edim* do not correlate strongly with any of the other similarity metrics, even after considering different hyperparameters [Torregrossa *et al.*, 2020]. Hence, WALES is more suitable as a surrogate metric (for semantic relatedness) that does not rely on human annotations.

5 Conclusion

In this work, we proposed the information-foraging based WALES metric for assessing the ability of word embeddings to preserve semantic relatedness. In contrast to traditional similarity/relatedness metrics, WALES has a *simple* set-up that does not rely on explicit human ratings of word similarities. Instead, WALES exploits implicitly encoded relational knowledge from Wikipedia graphs. Since Wikipedia is available in many languages, WALES is *portable* to other languages, whereas traditional metrics are tied to a single language. WALES is *precise* as it correlates strongly with traditional measures for semantic relatedness. WALES is *robust* because it yields consistent rankings of word embeddings over its hyperparameter configurations. For practical future evaluations, we collected a human benchmark data set on which the performance of the information-foraging agents for different embeddings is *comparable* to the one of humans.

Acknowledgements

This work was supported by the Carl Zeiss Foundation within the project “Interactive Inference”. We thank Alex Clemesha for the *The Wiki Game* data.

References

[Agirre *et al.*, 2009] Eneko Agirre, Enrique Alfonseca, Keith B. Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of the Conference on Human Language Technologies (NAACL-HLT)*, pages 19–27, 2009.

[Bakarov, 2018] Amir Bakarov. A Survey of Word Embeddings Evaluation Methods. *CoRR*, abs/1801.09536, 2018.

[Baroni *et al.*, 2014] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 238–247, 2014.

[Boldi *et al.*, 2011] Paolo Boldi, Marco Rosa, Massimo Santini, and Sebastiano Vigna. Layered label propagation: a multiresolution coordinate-free ordering for compressing social networks. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 587–596, 2011.

[Chi *et al.*, 2001] Ed Huai-hsin Chi, Peter Pirolli, Kim Chen, and James E. Pitkow. Using information scent to model user information needs and actions and the web. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 490–497, 2001.

[Clemesha, 2018] Alex Clemesha. The Wiki Game. https://www.thewikigame.com, 2018. Accessed: 2022-01-13.

[Faruqui *et al.*, 2016] Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. In *Proceedings of the Workshop on Evaluating Vector-Space Representation for NLP (RepEval)*, pages 30–35, 2016.

[Hoenen, 2016] Armin Hoenen. Wikipedia Titles As Noun Tag Predictors. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2016.

[Kim, 2014] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

[Lastra-Díaz *et al.*, 2019] Juan J. Lastra-Díaz, Josu Goikoetxea, Mohamed Ali Hadj Taieb, Ana García-Serrano, Mohamed Ben Aouicha, and Eneko Agirre. A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence*, 85:645–665, 2019.

[Liza and Grzes, 2016] Farhana Ferdousi Liza and Marek Grzes. An Improved Crowdsourcing Based Evaluation Technique for Word Embedding Methods. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP (RepEval@ACL)*, pages 55–61, 2016.

[Mandelbaum and Shalev, 2016] Amit Mandelbaum and Adi Shalev. Word embeddings and their use in sentence classification tasks. *CoRR*, abs/1610.08229, 2016.

[Mikolov *et al.*, 2013] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR) Workshop Track*, 2013.

[Mikolov *et al.*, 2018] Tomás Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2018.

[Murphy, 1986] M. Lynne Murphy. *Semantic Relations and the Lexicon*. Cambridge University Press, 1986.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[Peters *et al.*, 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the Conference on Human Language Technologies (NAACL-HLT)*, pages 2227–2237, 2018.

[Qi *et al.*, 2018] Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation? In *Proceedings of the Conference on Human Language Technologies (NAACL-HLT)*, pages 529–535, 2018.

[Rada *et al.*, 1989] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

[Schnabel *et al.*, 2015] Tobias Schnabel, Igor Labutov, David M. Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 298–307, 2015.

[Sezerer and Tekir, 2021] Erhan Sezerer and Selma Tekir. A Survey On Neural Word Embeddings. *CoRR*, abs/2110.01804, 2021.

[Torregrossa *et al.*, 2020] François Torregrossa, Vincent Claveau, Nihel Kooli, Guillaume Gravier, and Robin Allesiardo. On the Correlation of Word Embedding Evaluation Metrics. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 4789–4797, 2020.

[Wenger, 2018] Jacob Wenger. Six degrees of wikipedia. https://www.sixdegreesofwikipedia.com/, 2018. Accessed: 2022-01-13.