Comparative Probing of Lexical Semantics Theories for Cognitive Plausibility and Technological Usefulness

António Branco, João Rodrigues, Małgorzata Salawa,[+2] Ruben Branco, Chakaveh Saedi,[+3]
University of Lisbon NLX Group Faculdade de Ciências Lisbon, Portugal
[2]*AGH University of Science and Technology* Faculty of Computer Science Kraków, Poland
[3]*Macquarie University* Department of Computing Sydney, Australia
antonio.branco@di.fc.ul.pt

Abstract

Lexical semantics theories differ in advocating that the meaning of words is represented as an inference graph, a feature mapping or a vector space, thus raising the question: is it the case that one of these approaches is superior to the others in representing lexical semantics appropriately? Or in its non antagonistic counterpart: could there be a unified account of lexical semantics where these approaches seamlessly emerge as (partial) renderings of (different) aspects of a core semantic knowledge base?

In this paper, we contribute to these research questions with a number of experiments that systematically probe different lexical semantics theories for their levels of cognitive plausibility and of technological usefulness.

The empirical findings obtained from these experiments advance our insight on lexical semantics as the feature-based approach emerges as superior to the other ones, and arguably also move us closer to finding answers to the research questions above.

1 Introduction

Lexical semantics is at the core of language science and technology as the meaning of expressions results from the meaning of their lexical units and the way these are combined. How to represent the meaning of words is a central topic of inquiry and three broad families advocate that lexical semantics is represented as a semantic network (Quillan, 1966), a feature-based mapping (Minsky, 1975; Bobrow and Norman, 1975), or a semantic space (Harris, 1954; Osgood et al., 1957; Miller and Charles, 1991).

In a semantic network approach, the meaning of a lexical unit is represented as a node in a graph whose edges categorically encode different types of semantic relations holding among the units (e.g. hypernymy, meronymy, etc.). In a feature-based model, the semantics of a lexicon is represented by a map where a key is the lexical unit of interest and the respective value is a set of other units denoting characteristics prototypically associated with the denotation of the unit in the key (e.g. color, usage, etc.). Under a semantic space perspective, the meaning of a lexical unit is represented by a vector in a high-dimensional space (aka word embedding), whose components are ultimately based on the frequency of co-occurrence with other units, i.e. on their linguistic contexts of usage.

The motivation for these theories is to be found in their different suitability and success in explaining a range of empirical phenomena in terms of how these are manifest in ordinary language usage and also how they are elicited in laboratory experimentation. These phenomena are related to the acquisition, storage and retrieval of lexical knowledge (e.g. the spread activation effect (Meyer and Schvaneveldt, 1971), the fan effect (Anderson, 1974) among many others) and to the interaction with other cognitive faculties or tasks, as categorization (Estes, 1994), reasoning (Rips, 1975), problem solving (Holyoak and Koh, 1987), learning (Ross, 1984), etc.

These different approaches have inspired a number of initiatives to build repositories of lexical knowledge. Popular representatives are, for semantic networks, WordNet (Fellbaum, 1998), for feature-based models, SWOW (De Deyne et al., 2019), and for semantic spaces, word2vec (Mikolov et al., 2013b) a.o. Different knowledge bases (KBs) are rooted in different empirical sources: WordNet is based on lexical intuitions of human experts; the information in SWOW is evoked from laypersons cued with lexical units; and word2vec reflects the co-occurrence frequency of words in texts.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4004–4019
Barcelona, Spain (Online), December 8-13, 2020

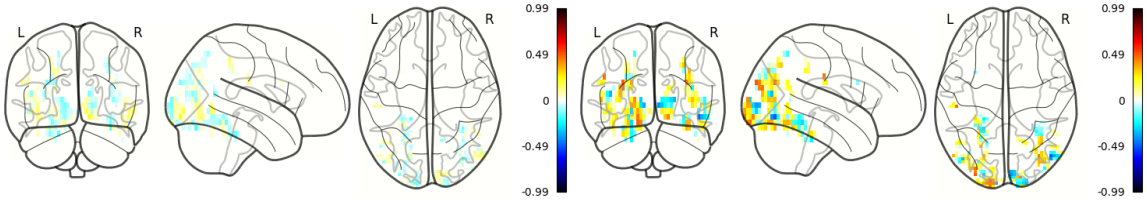

Figure 1: fMRI patterns in Participant 1 for word *eye*: **predicted** (left) with WordNet 60k embeddings via matrix factorization, cf. §3.2.3 below; **observed** (right), as in the data set of (Mitchell et al., 2008).

Against this background, a fundamental research question is: could there be a unified account of lexical semantics such that the above approaches seamlessly emerge as (partial) renderings of (different) aspects of the same core semantic knowledge base? Or in its antagonistic counterpart: is it the case that one of the above approaches is superior to the others in representing lexical semantics appropriately?

In this paper we contribute to these research questions with experiments consisting of two phases. First, different lexical semantic KBs, instantiating different lexical semantic theories, are converted to a common representation. The map-based dataset (SWOW) is converted to a graph-based representation and the two graph-represented datasets (SWOW and WordNet) are then converted to a vector-based representation. Second, to assess the appropriateness of these knowledge bases, the respective word embeddings are evaluated by means of the performance in language processing tasks where they are integrated and where lexical knowledge plays a crucial role.

We resort to the task of predicting brain activation patterns from semantic representations of words, illustrated in Figure 1, to assess the lexical KBs. If a KB k1 is more successful than another k2 in supporting these tasks, this indicates that k1 has higher cognitive plausibility, likely being superior at encoding the meaning of words.

Second, we resort also to the task of determining the semantic similarity between words from their formal semantic representations. Though it may arguably be less well grounded on actual cognitive representation of lexical meaning given the empirical frailty of the similarity scores (Faruqui et al., 2016), this has been a popular task for the intrinsic evaluation of word embeddings.

Third, for extrinsic evaluation, we resort to downstream Natural Language Processing (NLP) tasks.

As reported in the present paper, the findings from these experiments indicate that the feature-based approach emerges as superior to the other approaches to the representation of lexical meaning.

2 Related Work

In (Mitchell et al., 2008), the meaning of each word w was represented by semantic features given by the normalized co-occurrence counts in a big corpus of w with a set of 25 verbs related to basic sensory and motor activities. For each word, the respective fMRI activation level at every voxel is calculated as a weighted sum of each of the 25 semantic features, where the weights are learned by regression to maximum likelihood estimates given the observed fMRI data. Mean accuracy of 0.77 was reported for the 60 words and 9 subjects in the task of predicting brain activation originated by the exposure to words.

In an initial period, different authors focused on different ways to set up the features supporting this task. Jelodar et al. (2010) used the same 25 features but resorted to relatedness measures based on WordNet. As features, Fernandino et al. (2015) used instead 5 sensory-motor experience-based attributes, and the relatedness scores between the stimulus word and the attributes were based on human ratings. Binder et al. (2016), in turn, used 65 attributes with relatedness scores crowdsourced from over 1,700 participants.

As embeddings became popular, authors moved from features to word embeddings. Murphy et al. (2012) found their best results with dependency-based embeddings. Anderson et al. (2017) used word2vec together with a visual model built with a CNN on the Google Images dataset.

Recently, Abnar et al. (2018) evaluated 8 embeddings in predicting fMRI patterns: the experiential embeddings of Binder et al. (2016); the non-distributional feature-based embeddings of Faruqui et al. (2015); and 5 different distributional embeddings, namely word2vec (Mikolov et al., 2013a), fastText

(Bojanowski et al., 2017a), dependency-based word2vec (Levy and Goldberg, 2014), GloVe (Pennington et al., 2014) and LexVec (Salle et al., 2016); as well as the vectors from (Mitchell et al., 2008). The dependency word2vec achieved the best performance among the embeddings, while the Mitchell's et al. (2008) seminal approach with 25 features "is doing slightly better on average" than the other approaches.

In contrast to these papers, the goal here is not to beat the state of the art in brain activation prediction but to probe lexical semantic theories with the help of this task.

3 Common representations

To proceed with such comparative probing, a first step consists in the conversion of the different lexical KBs into the common representation format of semantic vector spaces, which we describe in this section.

3.1 From lexical maps to graphs

In SWOW each word *w* is mapped into a collection of prominent features as these are named by words elicited from laypersons cued with *w* (example in Fig. 2, Appendix). 83,863 participants were cued with 12,217 words to respond with 3 associated words, from which 100 responses were collected per cue, resulting in each cue being associated 300 times (De Deyne et al., 2018). We follow the methodology and data in (De Deyne et al., 2016b) to turn this map into a graph, rendered as an adjacency matrix. In the resulting matrix A_G, with every word displayed in the rows and columns, a cell A_{Gij} contains the count frequency of word i with word j, the accumulated times that j was responded when i was cued.

3.2 From lexical graphs to vectors

To convert graphs into embeddings, we experimented with one outstanding representative from each major family of conversion techniques, viz. based on edge reconstruction, matrix factorisation and random walks.

3.2.1 Edge reconstruction

Conversion techniques based on edge reconstruction support efficient training but ensure optimisation using only local information between nodes that are close to each other in the graph.

They operate on graphs represented as edge lists. An edge is a triple $\langle lhs, rel, rhs \rangle$ where lhs (left-hand side) and rhs are nodes connected by relation rel. The system is trained to recognize triples that are feasible (present in the graph) from the infeasible ones.

As a representative of edge reconstruction techniques, we adopted Semantic Matching Energy (SME) (Bordes et al., 2014) and used its publicly available implementation.[1]

Inference-based With WordNet, the triples were generated this way: for each word w_{lhs} in the vocabulary and for each synset s_{lhs} this word belongs to, a triple is generated for each word w_{rhs} in each synset s_{rhs} (that w_{rhs} belongs to) such that there exists a relation rel between s_{lhs} and s_{rhs}, and both w_{lhs} and w_{rhs} are in the vocabulary.[2]

The models were trained for 500 epochs, with evaluation at every 10 epochs, a learning rate of 0.01 and 200 batches. The remaining parameters were the default ones. The validation and test sets each made up for around 5% of the dataset and the model with the best performance on the validation set was picked. For a fair comparison among conversion methods, the training data is based on the same 60k vocabulary as in the matrix factorisation in Section 3.2.3 below. The vocabulary was selected with the procedure used in (Saedi et al., 2018; Branco et al., 2019), retaining the nodes with the largest number of outgoing edges.

Also for the sake of comparison with the other experiments with cooccurrence-based embeddings in Section 3.3, we chose vectors of dimension 300.

Feature-based With SWOW, the relations were obtained from the associative strength files that were generated by using the publicly available implementation.[3] The strength file is generated for three association types separately (*R1, R2, R3*), which induced three relations that were taken into account as three rel types by the SME method with SWOW (Salawa, 2019).

[1] `http://github.com/glorotxa/SME`.

[2] Data extracted with NLTK `www.nltk.org/_modules/nltk/corpus/reader/wordnet.html`.

[3] `http://github.com/SimonDeDeyne/SWOWEN-2018`

We used the same implementation and methodology to obtain SME models as used for WordNet. We empirically chose a smaller interval between the evaluations (every 5 epochs instead of 10) and a lower learning rate (0.001 instead of 0.01) for better training. We took again a vector size of 300.

3.2.2 Random walk

Another family of graph embedding methods relies on "text" that results from concatenating the words in the nodes that are visited in a random walk through the graph. The word embeddings are obtained from some deep learning techniques over that artificial text. Starting at a random node, at each iteration, a neighbour node is randomly chosen (with a probability α) to be the starting point of the next iteration or stopping the walk (with a probability 1-α) (Goikoetxea et al., 2015).

Differently from the edge reconstruction and matrix factorisation approaches, this technique is effective and accommodates global information about the nodes. However, as it only considers the local context in a path at each iteration, that makes it hard to stumble on an optimal sampling.

Inference-based We used the default Gensim's (Řehůřek and Sojka, 2010) Skip-Gram implementation, with a vector dimension of 300.

For the sake of comparability among KBs, we restricted the original technique to use only the edges among nodes and to ignore the glosses. The random walk was applied to the same WordNet graph (60k vocabulary) used with the edge reconstruction and matrix factorization techniques described above in 3.2.1 and below in 3.2.3.

Feature-based The random walk over SWOW used the same basic setup as used for WordNet.

The SWOW dataset used for edge reconstruction was converted into a graph input for UKB[4] and used the default UKB random walk parameters. To obtain the word embeddings, the default Gensim's Skip-Gram implementation with vectors of dimension 300 was used.

3.2.3 Matrix factorization

A third type of graph embedding method relies on graphs represented by matrices and on their factorization. This is perhaps the family of techniques with the largest number of instances in the literature, which in many cases result from slight variants in the tricks used to weight and condense the nodes in the matrix (Cai et al., 2018).

Matrix factorisation inverts the trade-off found in edge reconstruction. It takes into account the affinity between nodes at the global level of the graph, but at the cost of a large time and space consumption.

Inference-based To convert WordNet, we started by building an adjacency matrix with a size above 155k. Following (Saedi et al., 2018), we resorted to Katz index for the factorization technique, and used the relevant parameters and other options empirically determined there — including, for an affordable computational footprint, the same 60k subgraph, made of words with the largest number of outgoing arcs.

After the Katz procedure, a Positive Point-wise Mutual Information transformation (PMI+) was applied, to reduce the frequency bias (De Deyne et al., 2016a), followed by L2-norm to normalise each line of the matrix output by the Katz procedure, and finally a Principal Component Analysis (PCA) was applied to reduce the dimension of the vectors, set to 300.

Feature-based The adjacency matrix from SWOW was factorised following the same steps.

Due to the small 12k vocabulary available from SWOW, no extraction of a subset was necessary as it formed a dataset computationally manageable.

3.2.4 PMI

In addition to the matrix factorization method used above, in our experiments we resorted also to a streamlined version of it where the computationally highly costly matrix inversion procedure in the Katz index is skipped, remaining only the PMI transformation, followed by an L2 normalization, and PCA to reduce the matrix size to 300.

Feature-based In SWOW, for each pair of cue and associated word, the PMI score was obtained from the number of times they were associated one to the other divided by the product of the number of times each was mentioned normalized by the number of association pairs.

[4] `http://github.com/asoroa/ukb/` (default parameters)

Inference-based In WordNet, for each synset pair related by one edge, the PMI score was obtained as described previously for SWOW, considering a word in a synset as the cue and a word in a synset reached by the edge as an associated word.

3.3 Coocurrence-based vectors

As cooccurrence-based KBs extracted from text, we took the predictive models word2vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2017b), dependency-based word2vec (Levy and Goldberg, 2014) and the contextual embeddings from BERT (Devlin et al., 2019).

4 Brain activity

4.1 Brain activation prediction

The task introduced by Mitchell et al. (2008) consists of predicting the fMRI activation patterns in human subjects from some semantic representation of nouns. To collect the fMRI data used in (Mitchell et al., 2008), 9 participants were randomly shown 60 different noun-picture pairs, each presented 6 times. For each participant, a representative fMRI image for each stimulus was calculated as the mean fMRI response from the 6 repetitions and subtracting the mean of all stimuli. This task consists of mapping an input lexical semantic representation of each word into an approximation of its vector with the activation values for the 3x3x6 mm^3 voxels in the fMRI.

Training and evaluation To obtain the prediction models, we resorted to the implementation by Abnar et al. (2018).[5] The training ran for 1,000 epochs, with a batch size of 29 and a learning rate of 0.001. The loss function was calculated by adding the Huber loss, the mean pairwise squared error and the L2-norm (on weights and bias). Like in previous works, only the 500 most stable voxels were selected. We followed the usual evaluation procedure for this task. Separate models were learned for the 9 participants and evaluated using leave-two-out cross-validation, where the model was asked to predict the fMRI activation for the two held-out words in each iteration. The predictions were matched against the observed activations using cosine similarity over the 500 most stable voxels.

With each word embeddings from Section 3, a different model was trained for this task. Evaluation results are in Table 1 (Appendix). Upon empirical experimentation, from BERT (base-uncased), the concatenation of the last four layers provided the best results, and were used in all experiments.

Discussion The best options with cooccurrence- and feature-based models show a similar level of performance (82.76% and 81.24% of accuracy) and are much better than the best option for the inference-based model (72.76%).

Concerning graph embedding methods, their relative ranking depends on the type of lexical knowledge base to which it is applied. For the inference-based ones (based on WordNet), edge reconstruction is outperformed by matrix factorization, which is on a par or outperformed by random walk. For the feature-based ones (based on SWOW), this ranking appears inverted: the random walk is outperformed by the edge reconstruction — while matrix factorization appears with an outlier score, way below the scores of any other option whatsoever. Hence, graph embedding techniques that take into account the affinity between nodes at the global level of the graph are better for inference-based approaches (likely in line with their supporting of transitivity), while for feature-based approaches (not supporting transitivity), techniques taking into account local information between nodes close to each other are better.

Nevertheless, while showing high variability depending on the knowledge base to which it is applied, it is the much more simple PMI (not originally conceived as a graph embedding method but as a mere measure of association) that when applied to the SWOW-based graph, supports the top performance (81.24%) among all the graph embedding methods.

Focusing on cooccurrence-based models, in turn, these show a good performance in a quite narrow 6 percentage points range, from fastText (76.57%) to dependency (82.76%).

All in all the striking empirical finding in this experiment is the very competitive result of SWOW-PMI (81.24%) — a dataset only with 12k words, crowd sourced from laypersons cued for the simple associative word retrieval task, and converted into embeddings with the simple PMI technique —, against the best

[5] https://github.com/samiraabnar/NeuroSemantics

performing cooccurrence-based KB, dependency word2vec (82.76%) — based on a 1.5B corpus, parsed for syntactic dependencies, with a 175k vocabulary and 900k syntactic contexts.

SWOW reflects the frequency of what types of objects in the world are encountered together in typical situations — rather than the frequency of what words are together in text side by side. Dependency embeddings reflect the frequency of what types of words in the text are encountered together as co-arguments and modifiers in the same predicative structures, which denote typical situations encountered in the world — rather than the frequency of what words are together in text side by side in linear windows of context. These considerations seem to indicate that it is this essential aspect of their design, common to both approaches, that lends them their superiority in this task. Given this is a task on brain activity prediction, they seem to indicate also that, because of this, they have a higher cognitive plausibility to represent lexical semantic knowledge than the other options.

4.2 Further brain activity prediction

CogniVal is a workbench to test word embeddings on a battery of tasks concerning the prediction of cognitive activity related to language processing (Hollenstein et al., 2019). It encompasses 8 tasks for behavioral activity (eye-tracking) and 8 tasks for brain activity (4 with EEG and 4 with fMRI). For a tested word embeddings, CogniVal delivers an aggregated performance score in different tasks.

The 4 fMRI tasks include Mitchell et al.'s (2008) one plus three others, based on the datasets HARRY POTTER (Wehbe et al., 2014), ALICE (Brennan et al., 2016) and PEREIRA (Pereira et al., 2018), which concern the processing of sentences rather than words.

We evaluated all the embeddings used in Section 4.1 above with Mitchell et al.'s (2008) task also on these 3 sentence-rooted fMRI-based tasks. Results are in Table 2 (Appendix).

Discussion A major difference to testing with Mitchell et al.'s (2008) dataset is that, for these 3 sentence-rooted fMRI-based embeddings, while the best cooccurrence-based option is also dependency word2vec — also supporting now the top performance (0.0101; N.B.:lower scores are better) among all options —, the best feature-based option (0.0111, with SWOW-PMI) is closer to the best inference-based option (0.0107, with WordNet60k-PMI) than to the top cooccurrence-based one, and it is even slightly outperformed by this WordNet60k-PMI option.

Another difference is that, while dependency keeps supporting the top performance, the ranking of the other text-based embeddings in the previous experiment (Table 1) is somewhat inverted: for those coming after dependency, the options that were the first two become now the last two and vice-versa (Table 2).

Yet another difference concerns the relative merits of the graph to embeddings conversion methods. While different techniques were better with different KBs in the previous experiment, now there is one method that outperforms the other two in both KBs, inference- and feature-based, namely edge reconstruction. This method takes into account local information between nodes close to each other.

It is worth noting that while the dataset in (Mitchell et al., 2008) was collected with each word being processed in isolation by the subjects, in the other three fMRI datasets in CogniVal, the information was collected for words in the context of sentence processing. As from the experiment with Mitchell et al.'s (2008) task above one noticed that word representations from SWOW and dependency are better at reflecting the frequency of typical situations, this second experiment indicates that SWOW is inferior to dependency at providing such type of cognitively plausible representations when fMRI patterns of words are captured for their occurrence within sentences. Understandably, this should follow from the ways the data in dependency (words in sentences) and SWOW (words in isolation) were collected.

5 Similarity and relatedness

The tasks considered in the third experiment consist of predicting the semantic similarity or relatedness between words in pairs and in seeking to match the gold scores assigned by humans to such test pairs — with the cosine between the predicted vectors mapped into the scale used for gold scores.

The goal in this paper is not to beat the state of the art in similarity prediction tasks but to understand whether when testing embeddings from different empirical sources, these tasks deliver results similar to those delivered by brain activity tasks — our interest is on gaining insight about whether the similarity

tasks, based on simpler and cheaper data, can be used for the same practical purposes of the brain activity tasks, based in much more expensive and hard to get data (Salawa et al., 2019).

For **semantic similarity**, we used SimLex-999 (with 999 pairs) (Hill et al., 2016), WordSim-353-Similarity (203 pairs) (Agirre et al., 2009) and RG1965 (65) (Rubenstein and Goodenough, 1965).

For **semantic relatedness**, WordSim-353-Relatedness (252) (Agirre et al., 2009), MEN (3000) (Bruni et al., 2012) and MTURK-771 (771) (Halawi et al., 2012) were used. The results for WordNet- , text- and SWOW-based embeddings are in Tables 3, 4 and 5 (Appendix), respectively.

Discussion The major difference to the previous experiments is that now the best option is not obtained with cooccurrence-based embeddings but with feature-based ones, whose best results on the 6 tasks (82.72 average; Table 5) are much better than the best results with cooccurrence-based ones (73.10 av.; Table 4), which, in turn, are much better than the best results with inference-based ones (55.30 av.; Table 3).

Another difference concerns the graph to embeddings conversion methods. Now both inference-based and feature-based lexical knowledge repositories are better converted with graph embedding methods that take into account the affinity between nodes at the global level of the graph (matrix factorization or random walk) — while the combination SWOW and PMI still outperforms all the other options in two of the six tasks, and deliver results very close to the top ones in the other four.

Yet another difference is the ranking among the text-based embeddings. fastText, the third in the sentence-based fMRI dataset (Table 2), is now first (Table 4), switching positions with dependency.

This outcome indicates that similarity and relatedness prediction tests are not probing for the same characteristics probed with brain activity prediction tests — cognitive plausibility —, and these two classes of tests for embeddings cannot be interchanged with each other.

Additionally, given the quite different rankings of the cooccurrence-based embeddings in the three experiments above, together with their sharp underperformance vis-a-vis feature-based ones, one wonders what the testing them on similarity tasks — used in the literature to assess their intrinsic quality — is revealing about them, thus empirically reinforcing the issues raised analytically in (Faruqui et al., 2016).

6 Extrinsic evaluation

Results from intrinsic and extrinsic evaluation in NLP are not necessarily aligned with each other as performance increments in intrinsic results may have or not an incremental impact on the intricacies and performance of the larger systems where components happen to be embedded. As intrinsic vs. extrinsic congruence is thus something that has to be determined empirically, it is important to proceed with probing lexical semantics theories with downstream tasks,[6] for their extrinsic evaluation.

For multi-task benchmark, we resort to GLUE platform (Wang et al., 2018), which contains 9 tasks of 3 types, from which we used a subset of 5 tasks with affordable computational footprint: the 2 single-sentence tasks, CoLA and SST-2; 1 of the 3 similarity and paraphrase tasks, namely MRPC; and 2 of the 4 inference tasks, namely RTE and WNLI.

SST-2 is a task-based on the Stanford Sentiment Treebank (Socher et al., 2013) consisting of sentences extracted from movie reviews and human annotation of their sentiment. The model trained on these data has to determine the sentiment expressed in input sentences. The **MRPC** task is based on The Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005) with sentence pairs automatically extracted from online news and manually annotated as to whether the sentences in each pair are semantically equivalent. The **RTE** task resorts to The Recognizing Textual Entailment datasets from online news, which result from annual challenges for the task of textual entailment, and gather the data from RTE1 (Dagan et al., 2005), RTE2 (Bar-Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009). This task consists of predicting whether the premise entails the hypothesis in each test item. **WNLI** is a task-based on the Winograd Schema Challenge (Levesque et al., 2012) where each example contains a sentence with a pronoun and a list of admissible antecedents. This task consists of picking the right antecedent. Finally, the **CoLA** task is based on the Corpus of Linguistic Acceptability.

consisting of examples of sentence acceptability judgments taken from books and journal articles on linguistic theory, with each example being a string annotated as to whether it is grammatical.

[6]Examples a.o. of downstream tasks in (Rodrigues et al., 2017; Silveira and Branco, 2012; Costa and Branco, 2012).

6.1 Training and evaluation

The goal here is not to reach the top of GLUE leader board. We are interested rather in the extrinsic evaluation of lexical semantics theories, that is in understanding, under comparable circumstances, if and how the embeddings based on different theories have a different impact on these tasks.

To pursue this goal, we adopted an architecture (Wang et al., 2019) that accommodates pretrained semantic representations, comprising three levels: the input layer, the shared encoder layers and the task-specific model. For the top layer with task-specific information in each downstream task, we used the respective layer from GLUE. This top layer consists of a hidden layer of dimension 512, with the Dropout technique (Srivastava et al., 2014) with $p = 0.2$, and layer normalization (Ba et al., 2016). The final output layer is a softmax layer.

To obtain the middle layer, encoding sentence semantics, we used a 2-layered biLSTM with dimensionality 1024. Instead of random initialization, the sentence encoder is first trained with one of the best performing pretraining tasks reported in (Wang et al., 2019), namely STS-B.[7]

For the input layer, we experimented with each of the pretrained word embeddings discussed above in Sections 4 and 5: for each downstream task, different models were trained with the different pretrained word embeddings, and also with the baseline consisting of embeddings with random vectors. For cross-validation, the framework made use of the original data splitting for training, development and test partitioning. The results are in Table 6 (Appendix) and respective plotting in Figure 3.[8]

Discussion Comparing the overall performance across tasks, a first pattern emerges with a group of two tasks, CoLA and WNLI, performing below half the respective absolute best possible score, and the group with the other three tasks performing above.

In that first group, a large majority of the models with pretrained word embeddings perform below or at best on a par with random embeddings. The two tasks in this group are very hard as they rely on rich information about the grammatical structure of the sentences and on long-distance relations among their expressions: to resolve anaphors and find their antecedents somewhere in the sentence, in WNLI; and to categorically decide sentence membership in the language defined by the grammar, in CoLA. Hence, the signal from the lexical information encoded in pretrained embeddings, from whatever lexical theory or empirical source, has an impact that is null, in CoLa, or even detrimental, in WNLI, being of little value to advance the research questions addressed here.

Turning to the other three tasks, the better is the performance of models for a given task, the more they overperform the random baselines. The top performing datasets are dependency and SWOW-RandomWalk — on a par in MRPC, and with SWOW-RandomWalk just slightly behind dependency in RTE and SST-2.

Concerning feature-based embeddings, these results confirm, also in extrinsic evaluation tasks, their strength found in intrinsic tasks — where SWOW supported the best performing options in semantic similarity and in word-rooted brain activity prediction tests. Given the consistent superiority of SWOW embeddings in intrinsic tasks, and their very competitive performance here in dowstream tasks, these results indicate that SWOW-based embeddings are likely one of the most reliable candidates to be used to enhance NLP downstream tasks for which the contribution of (pre-trained) lexical knowledge is useful.

Concerning text-based embeddings, dependency (top in brain activation prediction) outperforms fastText (top in similarity), thus in line with their relative ranking in the brain task, but inverting that relative ranking in the similarity task. This reinforces the wondering about how useful NLP intrinsic tasks may be to assess the text-based embeddings for their strength in downstream tasks, in line with (Chiu et al., 2016).

Concerning graph embedding methods, taking PMI aside, here edge reconstruction is overperformed by matrix factorization, which is overperformed by random walk. Methods taking into account the affinity between nodes at the global level of the graph seem thus better for extrinsic tasks. Another interesting lesson is that, for each graph embedding method, models using SWOW embeddings overperform models using WordNet embeddings — except for PMI, better now with WordNet than with SWOW.

[7] STS-B, The Semantic Textual Similarity Benchmark (Cer et al., 2017), is a task in GLUE consisting of determining the similarity of two sentences on a continuous scale from 1 to 5.

[8] We repeated all experiments also with the middle layer contributing sentence semantics removed. The same basic outcome patterns in Figure 3 (Appendix) were observed again.

7 Conclusions

Contributions A first major contribution of this paper is *the design of an experimental setup to comparatively probe lexical semantics theories* of all kinds — feature-, inference- and cooccurrence-based — for both their cognitive plausibility and their usefulness in NLP.

This setup consists of converting representative repositories of these theories to a common format and integrating them, under such a common format, in different models for different probing tasks: • To be converted, exemplar techniques from each major graph embedding family (edge reconstruction, matrix factorization, random walk) were used. • To be probed for cognitive plausibility, some of these models addressed fMRI-based brain activation prediction tasks. • To be probed for usefulness in NLP, some other models addressed semantic similarity and relatedness prediction tasks, for intrinsic evaluation, and they were also embedded in downstream NLP tasks, for extrinsic evaluation.

Another major contribution is the empirical results obtained with a systematic application of this probing setup, including the central finding that *the feature-based lexical knowledge base is superior to knowledge bases complying with other lexical semantic theories* in consistently supporting models with top performance across the different probing tasks.

Concerning the other, incidental empirical findings: • As to graph embedding techniques, it was possible to understand that *graph embedding methods taking into account the affinity between nodes at the global level of the graph tend to better serve downstream tasks*. • As to intrinsic vs. extrinsic NLP evaluation tasks, there emerged not an alignment between a superior performance in the former and a superior performance in the latter, hence *intrinsic performance of a lexical knowledge base is an unreliable predictor of its extrinsic performance*. • As to text-based embeddings, top performance models in different tasks are supported by different embedding methods, hence *the performance of a given method for text-based embeddings in one task is an unreliable predictor of its performance in other tasks*.

Discussion The overall superior performance of the feature-based SWOW lexical knowledge base is the striking empirical result of the present study. This knowledge repository supported the top results, without a close second, in semantic similarity and relatedness prediction. It was on a par to top results or a close second in extrinsic tasks where the lexical signal has an impact. It was a close second in word-rooted brain activity prediction. As expected, it was not shining only in sentence-rooted brain activity prediction.

It only adds to its outstanding record that it is uncertain how comparability can be fairly ensured between SWOW and its alternatives. Dependency embeddings are extracted from a 1.5B corpus, with a 175k vocabulary (word2vec/100B, fastText/600B, GloVe/840B). SWOW has 12k words linked to each other whose links were crowdsourced from 83k laypersons cued 3 times for basic word association.

It also adds to the overwhelming performance of SWOW that it comprises just 12k words while the lexicon of an adult is estimated to consist of over 40k words (Brysbaert et al., 2016) — which additionally emphasizes its yet untapped potential and the promise of extended versions with larger numbers of words.

To explain its strength, it is tempting to attribute it to SWOW's higher cognitive plausibility. This bears the underlying assumptions — epistemologically non trivial — that better performance at the word-rooted brain activity prediction correlates with higher cognitive plausibility of lexical semantic knowledge repositories, and higher cognitive plausibility correlates with better performance of NLP systems where these repositories are embedded.

Future work Returning to one of our driving questions: could there be a unified account of lexical semantics such that the three families of approaches to lexical meaning seamlessly emerge as (partial) renderings of (different) aspects of the same core semantic knowledge base? The findings reported in this paper make a positive answer to it increasingly attractive, with feature-based lexical knowledge being a good candidate to that core knowledge repository.

It is worth noting though that one has a grasp on how to obtain cooccurrence-based lexicons from feature-based ones (De Deyne et al., 2016b and the present paper); and also from inference-based ones (Saedi et al., 2018 and the present paper) and vice-versa (Tarouti and Kalita, 2016). But whether it is possible to go from text-embeddings like fastText to feature-based lexicons like SWOW, or from SWOW to ontologies like WordNet remain open questions. Seeking to address these challenges is future work that should help to further progress towards finding an answer to the research question above.

Acknowledgements

The research reported here was partially supported by PORTULAN CLARIN—Research Infrastructure for the Science and Technology of Language, funded by Lisboa 2020, Alentejo 2020 and FCT—Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016.

Reproduction

To support the reproduction of research results (Branco et al., 2017), the embeddings are available from `https://hdl.handle.net/21.11129/0000-000D-C67B-A`

References

Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. 2018. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018) , pages 57–66*.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of The Annual Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies Conference (NAACL-HLT2009)*, pages 19–27. Association for Computational Linguistics.

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association of Computational Linguistics*, 5(1):17–30.

John Robert Anderson. 1974. Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 6(4):451–474.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, and Danilo Giampiccolo. 2006. The second PASCAL recognising textual entailment challenge. *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, 01.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In *TAC*.

Jeffrey R. Binder, Lisa L. Conant, Colin J. Humphries, Leonardo Fernandino, Stephen B. Simons, Mario Aguilar, and Rutvik H. Desai. 2016. Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, 33(3-4):130–174.

Daniel G. Bobrow and Donald Arthur Norman. 1975. Some principles of memory schemata. In *Representation and Understanding: Studies in Cognitive Science*, page 131–149. Elsevier.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017a. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017b. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2014. A semantic matching energy function for learning with multi-relational data. *Machine Learning*, 94(2):233–259.

António Branco, Kevin Bretonnel Cohen, Piek Vossen, Nancy Ide, and Nicoletta Calzolari. 2017. Replicability and reproducibility of research results for human language technology: introducing an lre special section. *Language Resources and Evaluation*, 51:1–5.

Ruben Branco, João António Rodrigues, Chakaveh Saedi, and António Branco. 2019. Assessing wordnets with wordnet embeddings. In *Proceedings of Global Wordnet Conference (GWC2019)*, pages 253–259.

Jonathan R Brennan, Edward P Stabler, Sarah E Van Wagenen, Wen-Ming Luh, and John T Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, 157:81–94.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL2012)*, pages 136–145. Association for Computational Linguistics.

Marc Brysbaert, Michaël Stevens, Pawel Mandera, and Emmaneul Keuleers. 2016. How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, 7(1116).

Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *SemEval@ACL*, pages 1–14. The Association for Computer Linguistics.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6, Berlin, Germany, August. Association for Computational Linguistics.

Francisco Costa and António Branco. 2012. Aspectual type and temporal relation classification. In *Proceedings of 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL2012)*, pages 266–275.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. pages 177–190, 01.

Simon De Deyne, Daniel J Navarro, Amy Perfors, and Gert Storms. 2016a. Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General*, 145(9):1228.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016b. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING2016)*, pages 1861–1870.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, pages 1–20.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, 51(3):987–1006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

William K Estes. 1994. *Classification and Cognition*. Oxford University Press.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 464–469.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany, August. Association for Computational Linguistics.

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

Leonardo Fernandino, Colin J. Humphries, Mark S. Seidenberg, William L. Gross, Lisa L. Conant, and Jeffrey R. Binder. 2015. Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. *Neuropsychologia*, 76:17–26.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, 06.

Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. Random walks and neural network language models on knowledge bases. In *Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 1434–1439.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1406–1414. ACM.

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

Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41:665–695.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. Cognival: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 538–549.

Keith J Holyoak and Kyunghee Koh. 1987. Surface and structural similarity in analogical transfer. *Memory & Cognition*, 15(4):332–340.

Ahmad Babaeian Jelodar, Mehrdad Alizadeh, and Shahram Khadivi. 2010. WordNet based features for predicting brain activity associated with meanings of nouns. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 18–26. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 302–308.

David E Meyer and Roger W Schvaneveldt. 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2):227.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *In the ICLR Workshop Papers.*

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

George Miller and Walter Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6:1–28.

Marvin Minsky. 1975. A framework for representing knowledge. In *Psychology of Computer Vision*. McGraw-Hill.

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 114–123. Association for Computational Linguistics.

Charles E Osgood, George J Suci, and Percy H Tannenbaum. 1957. The measurement of meaning. *Urbana: University of Illinois Press*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.

M Ross Quillan. 1966. Semantic memory. Technical report, Bolt Beranek and Newman Inc., Cambridge MA.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. European Language Resources Association.

Lance J Rips. 1975. Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, 14(6):665–681.

João António Rodrigues, Chakaveh Saedi, Vladislav Maraev, João Silva, and António Branco. 2017. Ways of asking and replying in duplicate question detection. In *Proceedings of 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 261–270.

Brian H Ross. 1984. Remindings and their effects in learning a cognitive skill. *Cognitive Psychology*, 16(3):371–416.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Chakaveh Saedi, António Branco, João António Rodrigues, and João Silva. 2018. WordNet embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 122–131, Melbourne, Australia, July. Association for Computational Linguistics.

Małgorzata Salawa, António Branco, Ruben Branco, João António Rodrigues, and Chakaveh Saedi. 2019. Whom to learn from? graph- vs. text-based word embeddings. In *Proceedings of Recent Advances in Natural Language Processing (RANLP2019)*, pages 1041–1051.

Małgorzata Salawa. 2019. *Word Embeddings from Lexical Ontologies: A comparative study*. AGH University of Science and Technology of Kraków, MA Dissertation.

Alexandre Salle, Marco Idiart, and Aline Villavicencio. 2016. Matrix factorization using window sampling and negative sampling for improved word representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, volume 2, pages 419–424. Association for Computational Linguistics.

Sara Silveira and António Branco. 2012. Combining a double clustering approach with sentence simplification to produce highly informative multi-document summaries. In *Proceedings of IEEE International Conference on Information Reuse and Integration (IEEE-IRI2012)*, pages 482–489.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Feras Tarouti and Jugal Kalita. 2016. Enhancing automatic wordnet construction using word embeddings. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 30–34.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy, July. Association for Computational Linguistics.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar, October. Association for Computational Linguistics.

Appendix

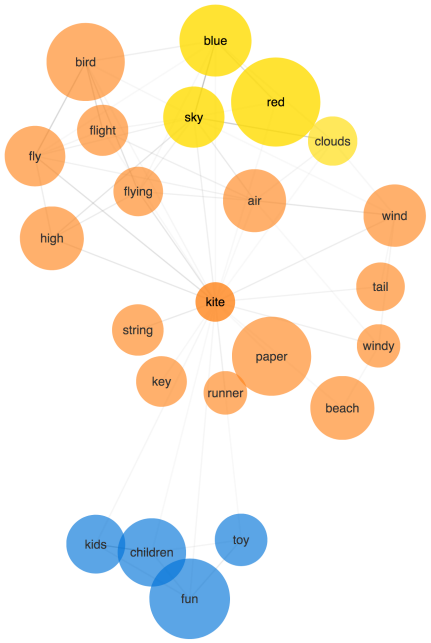

Figure 2: Visualization of the cue *kite* and its associated words in the SWOW lexical knowledge base. Source: `https://smallworldofwords.org/en/project/explore`

	fastText	word2vec	BERT	GloVe	depend
Cooccurrence-based	76.57	77.78	77.98	78.59	**82.76**
	Edge	Factor.	Walk	PMI	
Inference-based (60k)	61.08	**69.42**	69.37	62.72	
Inference-based (120k)			**72.76**	72.54	
Feature-based	76.40	54.16	73.65	**81.24**	

Table 1: **Intrinsic evaluation**: Performance with knowledge bases (rows) under different embedding techniques (columns) in terms of accuracy in **predicting brain activation** with the **word-rooted fMRI-based data set** of (Mitchell et al., 2008) (higher is better).

	BERT	GloVe	fastText	word2vec	depend
Cooccurrence-based	0.0362	0.0256	0.0189	0.0143	**0.0101**
	Edge	Factor.	Walk	PMI	
Inference-based (60k)	0.0117	0.0153	0.0520	**0.0107**	
Inference-based (120k)			0.0280	**0.0113**	
Feature-based	0.0121	2.5005	0.0288	**0.0111**	

Table 2: **Intrinsic evaluation**: Performance with knowledge bases (rows) under different embedding techniques (columns) in terms of accuracy in **predicting brain activation** with the three **sentence-rooted fMRI-based CogniVal datasets** (Hollenstein et al., 2019) as mean squared error (**lower is better**).

	PMI	Edge	Factor.	Walk
Similarity				
Simlex-999	28.26	39.63±1.55	49.90	**50.93**±0.15
WordSim-353Sim	42.61	54.93±2.31	50.80	**67.40**±0.30
RG1965	26.38	57.70±4.84	57.00	**77.50**±0.95
Relatedness				
WordSim-353Rel	16.69	26.20±4.10	**30.90**	28.43±0.76
MEN	29.18	39.67±2.55	45.00	**52.17**±0.70
MTurk-771	37.22	42.40±1.25	52.80	**52.90**±0.50

Table 3: **Intrinsic evaluation**: Performance with **WordNet** 60k under graph to embedding conversion techniques (columns) over test sets for **semantic similarity and relatedness prediction** (rows) in Spearman's correlation coefficient (higher is better), with three runs averaged where relevant. The coverage of the test sets with WordNet 60k: 100% of Simlex-999; 100% of WordSim-353 S; 98.0% of RG1965; 97.6% of WordSim-353 R; 83.4% of MEN; 99.9% of MTurk-771.

	BERT	GloVe	depend	w2vec	fastText
Similarity					
Simlex-999	27.65	37.52	44.56	43.61	**49.24**
WordSim-353Sim	58.66	62.98	75.88	74.08	**79.74**
RG1965	55.70	65.77	71.11	74.77	**81.31**
Relatedness					
WordSim-353Rel	37.72	57.09	49.23	60.97	**71.33**
MEN	51.27	61.78	67.65	69.89	**80.87**
MTurk-771	44.43	63.07	62.23	65.69	**76.13**

Table 4: **Intrinsic evaluation**: Performance of cooccurrence-based word embeddings **BERT, GloVe, dependency embeddings, word2vec and fastText** (columns) over test sets for **semantic similarity and relatedness prediction** (rows) in Spearman's correlation coefficient (higher is better).

	Edge	Walk	Factor.	PMI
Similarity				
Simlex-999	54.13±6.20	**69.33**±0.06	67.80	68.54
WordSim-353Sim	77.07±4.76	84.53±0.06	**85.00**	83.73
RG1965	83.50±4.50	90.23±0.49	**92.90**	92.48
Relatedness				
WordSim-353Rel	70.70±3.68	77.73±0.23	**79.30**	78.50
MEN	78.50±3.90	84.27±0.06	87.20	**87.38**
MTurk-771	74.77±4.21	81.10±0.17	80.90	**82.39**

Table 5: **Intrinsic evaluation**: Performance with **SWOW** under graph to embedding conversion techniques (columns) over test sets for **semantic similarity and relatedness prediction** (rows) in Spearman's correlation coefficient (higher is better), with three runs averaged where relevant.The coverage of the test sets with SWOW: 99.6% of Simlex-999; 90.6% of WordSim-353 S; 83.1% of RG1965; 87.3% of WordSim-353 R; 89.4% of MEN; 93.3% of MTurk-771.

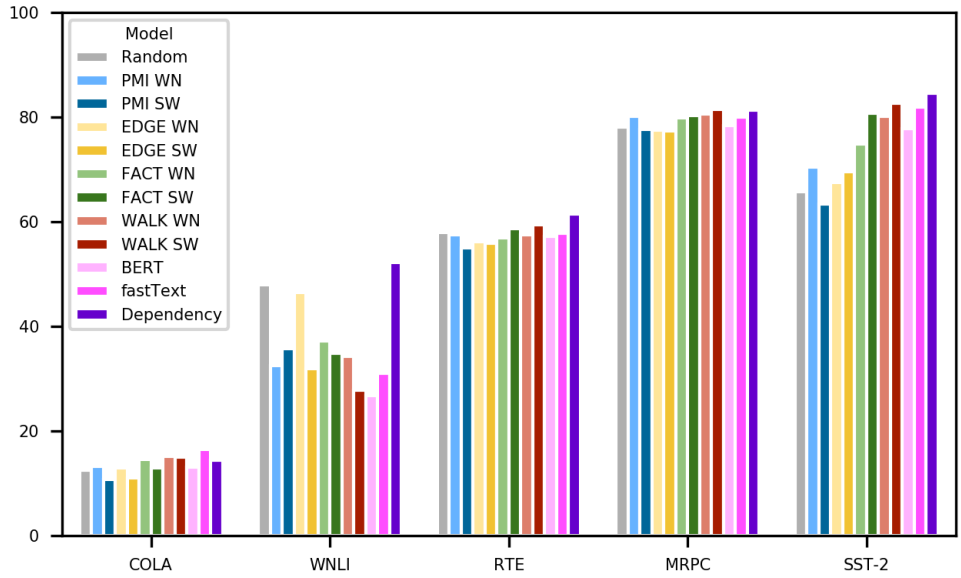

Figure 3: **Extrinsic evaluation**: Performance of models with different input layer word embeddings (bars) over the five different GLUE **downstream tasks** (groups), measured in different evaluation metrics projected to a common range [0, 100]. The reported scores are the average of three runs. The source data for this graph is in Table 6.

	CoLA	WNLI	RTE	MRPC	SST-2
Random	12.37 ± 0.65	47.87 ± 8.80	57.90 ± 2.35	78.02 ± 0.78	65.70 ± 0.95
PMI WN	13.17 ± 2.83	32.40 ± 7.58	57.40 ± 0.99	80.12 ± 0.19	70.40 ± 0.59
PMI SW	10.70 ± 2.35	35.70 ± 8.15	54.97 ± 2.73	77.63 ± 0.47	63.30 ± 1.08
Edge WN	12.93 ± 2.11	46.47 ± 13.56	56.20 ± 2.71	77.47 ± 0.80	67.47 ± 0.31
Edge SW	10.97 ± 1.53	31.93 ± 4.28	55.83 ± 1.77	77.33 ± 0.78	69.50 ± 0.90
Fact WN	14.43 ± 1.85	37.10 ± 7.74	56.93 ± 1.17	79.82 ± 0.89	74.90 ± 0.52
Fact SW	12.93 ± 2.01	34.77 ± 7.76	58.60 ± 2.71	80.23 ± 1.03	80.70 ± 0.92
Walk WN	15.13 ± 1.95	34.27 ± 17.21	57.53 ± 1.81	80.62 ± 0.43	80.20 ± 0.26
Walk SW	14.87 ± 0.15	27.70 ± 3.58	59.43 ± 1.37	**81.50** ± 1.26	82.63 ± 1.12
BERT	13.03 ± 1.97	26.77 ± 3.06	57.17 ± 0.46	78.38 ± 0.24	77.80 ± 0.78
fastText	**16.43** ± 2.25	30.97 ± 6.40	57.73 ± 1.56	79.92 ± 0.48	81.87 ± 0.76
Dependency	14.40 ± 0.78	**52.10** ± 7.27	**61.47** ± 2.56	81.30 ± 0.23	**84.57** ± 0.58

Table 6: **Extrinsic evaluation**: Performance over five GLUE **downstream tasks** (columns) of models with different input layer word embeddings (rows). For the task CoLA, performance is measured with Matthews correlation. For MRPC, an average of accuracy and F1 is reported. For the remaining tasks, accuracy is reported. The evaluation scores were projected to a [0-100] common scale (higher is better), with bold denoting top results. Each score is the average of the results from three runs with the random seeds 1147, 1256 and 1179. To enhance the readability of eventual data patterns, the content of this table is rendered in Figure 3.