The Importance of Context in the Evaluation of Word Embeddings: The Effects of Antonymy and Polysemy

James Fodor
The Centre for Brain, Mind and Markets
The University of Melbourne,
Victoria 3010 Australia
jfodor@student.unimelb.edu.au

Simon De Deyne
School of Psychological Sciences
The University of Melbourne
Victoria 3010 Australia
simon.dedeyne@unimelb.edu.au

Shinsuke Suzuki
Faculty of Social Data Science
Hitotsubashi University
shinsuke.szk@gmail.com

Abstract

Word embeddings are widely used for diverse applications in natural language processing. Despite extensive research, it is unclear when they succeed or fail to capture human judgements of semantic relatedness and similarity. In this study, we examine a range of models and experimental datasets[1], showing that while current embeddings perform reasonably well overall, they are unable to account for human judgements of antonyms and polysemy. We suggest that word embeddings perform poorly in representing polysemy and antonymy because they do not consider the context in which humans make word similarity judgements. In support of this, we further show that incorporating additional context into transformer embeddings using general corpora and lexical dictionaries significantly improves the fit with human judgments. Our results provide insight into two key inadequacies of word embeddings, and highlight the importance of incorporating word context into representations of word meaning when accounting for context-free human similarity judgments.

1 Introduction

Lexical semantics seeks to provide a cognitive explanation of how word meaning is represented and how semantic relations such as hyponymy, antonymy and synonymy are encoded. Vector-space models are one of the dominant approaches to studying lexical semantics. In vector-space models, a word is associated with a vector of real numbers called a *word embedding,* which captures information about word co-occurrences in a document or sentence. Each component of this vector corresponds to an abstract feature in an underlying vector space (Almeida and Xexéo, 2019; Lieto et al., 2017). The meaning of each word is thus represented by the direction of its word embedding in semantic space. (In this paper we use 'word embeddings' loosely, referring to any vector representation of word meaning using real numbers).

Word embedding methods are widely used in natural language processing, where they are utilised by machine learning architectures that have achieved impressive performance on a range of applied language tasks (Devlin et al., 2019; Lenci, 2018; Ranashinghe et al., 2019; Young et al., 2018). Vector-space semantics models also have a natural synergy with neuroimaging techniques that measure patterns of voxel activities in response to linguistic stimuli, thus providing an interface between lexical semantics and cognitive neuroscience (Rodrigues et al., 2018; Wang et al., 2020). It is therefore of considerable interest to evaluate the performance of these methods in modelling word meanings.

Vector-space approaches to semantics hypothesise that many aspects of word meaning, including semantic relationships such as synonymy, antonymy, hyponymy, and logical inference, can be efficiently represented by the relative direction of word embeddings in semantic space (Günther et al., 2019; Clark, 2015). One way to test this hypothesis is to compare the similarity relations

[1] Our code and processed datasets are available at https://github.com/bmmlab/lexical-semantics-eval

between word embeddings with human judgements of word similarity and relatedness (De Deyne et al., 2016; Lenci et al., 2021). A high correlation between the similarity structure of word embeddings and human similarity judgements is evidence that the embeddings successfully encode information about word meaning and semantic relationships between words.

Existing literature evaluating word embeddings against human similarity judgments, however, has typically ignored the implicit context humans use to make these judgements. We hypothesise that this omission is an important factor contributing to the relatively poor performance of word embedding models when evaluated against certain experimental datasets.

In this study we focus on two specific semantic phenomena in which the effects of context are most likely to be apparent: antonymy and polysemy. In the case of antonymy, we hypothesise that humans judge the meaning of a word differently when it is presented in the context of a word opposite in meaning. Likewise, we hypothesise that humans assess the meaning of polysemous words differently than non-polysemous words due to the need to use contextual information to select the relevant sense. We therefore anticipate an investigation into polysemy and antonymy will be important for understanding the limitations of word embeddings resulting from neglecting context.

1.1 Vector-space semantics models

Word embeddings can be constructed using a variety of techniques. Predict-based embeddings are constructed by training a neural network on a word prediction task, such as predicting the next word in a text (Baroni, Dinu, & Kruszewski, 2014). Knowledge-based methods utilise human curated datasets of semantic relations such as WordNet (Pedersen et al., 2004). Transformers are the most recent class of models, which capture context-specific meaning using multilayered attention neural networks trained on very large natural language corpora (Tripathy et al., 2021). Transformers can be used to compute word embeddings which are modified based on the specific usage of the word, and hence are of particular value in assessing the effects of word context.

One of the most common methods for assessing word embeddings is *semantic similarity*. Similarity is sometimes conceptualised as the degree to which two words are interchangeable (Miller and Charles, 1991). Another metric used in the evaluation of word embeddings is *semantic relatedness*. Relatedness refers to the degree to which the words share any type of semantic relation or psychological association (Gladkova et al., 2016; Hadj Taieb et al., 2020). As an example, 'car' and 'van' have high similarity and high relatedness, whereas 'car' and 'wheel' have lower similarity but still high relatedness. See Table 1 in Appendix A for a summary of major word similarity and relatedness datasets.

In most experimental studies, participants are asked to provide judgements about the similarity or relatedness of a set of word pairs, typically measured on an ordinal scale (Hill et al., 2015; Gerz et al., 2016). The averaged ratings are then compared to the cosine similarity of the corresponding word embeddings using a correlation coefficient (Vulić, Ponti, et al., 2020).

Numerous studies have followed this approach to investigate the relationship between human judgements and word embeddings, as summarised in Table 2 in Appendix A. These analyses have typically treated such judgements as non-contextual since word pairs are presented in isolation. However, we argue that this constitutes a failure to consider the implicit context provided by the second word in each word pair. Several studies have found that presenting words within the context of a sentence affects the manner in which humans make semantic judgments (Armendariz et al., 2020; Haber and Poesio, 2021). For example, humans interpret the word 'bank' differently when presented in a sentence about aircraft compared to when presented in a sentence about money (Trott and Bergen, 2021). However, to our knowledge this effect of context on human judgements has not been investigated in experimental datasets consisting solely of word pairs presented in the absence of additional context.

As such, building on previous suggestions (Bloch-Mullins, 2021) we hypothesise that when subjects are presented with two words absent further context, they assess the meaning of each word in the pair based on the implicit context of the other word in the pair. In the present study we investigate this hypothesis by evaluating the ability of word embeddings models to represent the meaning of antonym pairs and polysemous words. These were chosen as inherently relational semantic phenomena where context is most likely to affect human similarity judgements.

1.2 Antonymy

Antonyms are words that are 'opposite in meaning'. They provide a particular challenge for word similarity measures, since words like 'happy' and 'sad' are similar in that they both describe basic emotions, however since they are roughly opposite in meaning, they tend to be given low similarity ratings by humans (Lenci, 2018). It has proven difficult to define precisely what is meant by 'opposite meaning', with different subtypes and variations of antonymy proposed for different contexts or word types (Kotzor, 2021). In this study, we use a broad definition of antonymy by identifying verb pairs with varying degrees of contrasting or opposing meanings.

There are also conflicting views about the relationship between antonymy and similarity. If similarity is defined as the extent to which words are used in similar contexts, antonyms usually are identical in meaning except for the single dimension in which they have opposite values (Etcheverry and Wonsever, 2019). Conversely, if similarity is defined as the extent to which two words can be interchanged without loss of meaning, then antonyms have very low similarity (Kliegr and Zamazal, 2018). In practise, vector-space semantic models tend to give fairly high similarity ratings to both synonyms and antonyms (Nguyen et al., 2016), making it difficult to distinguish between these two relations in such models (Dou et al., 2018).

Various methods have been proposed to improve the representation of antonyms, including training a classifier over a set of word embeddings to distinguish antonyms from synonyms (Ali et al., 2019; Etcheverry and Wonsever, 2019), combining thesaurus or other knowledge-based information with word embeddings (Dou et al., 2018), and modifying standard word embeddings so that antonyms are maximally distant in similarity space (Nguyen et al., 2016; Samenko et al., 2020).

However, if the goal is to construct a comprehensive representation of word meanings, merely being able to distinguish antonyms from synonyms is insufficient. The fundamental difficulty appears to be that humans judge the similarity of antonyms differently than they judge other words, drawing upon background knowledge about the salient features for which antonyms have opposing values, and *using the context* provided by the presentation of words in a pair to judge the salience of these opposing features (Kotzor, 2021). The goal of the present study is to explore the role of context in more depth, investigating how antonym representation in word embedding models differs from human judgements.

1.3 Polysemy

A word is *polysemous* when it has multiple distinct but related meanings. For example, the verb 'count' can be used either to describe 'calculating using numbers' or 'being included as part of a group'. Vector-space models typically do not directly incorporate polysemy, as the usual approach is to learn a single word embedding vector for each word (Boleda, 2020; Camacho-Collados and Pilehvar, 2018). A major difficulty in incorporating polysemy into vector-space models is that there is no established method for distinguishing or enumerating different senses for a given polysemous word (Emerson, 2020), or in determining how much different senses overlap (Boleda, 2020). WordNet provides one commonly-used set of senses, though these have been criticised as being too finely-grained and lacking any clear structure (Palmer et al., 2007).

Polysemy also presents a problem for evaluating word embeddings, since humans may use the context of the second word in a pair to disambiguate a polysemous word. For instance, when presented with the pair 'bank' and 'river', participants may interpret 'bank' as relating to a riverbank, while when presented with 'bank' and 'loan', they are likely to interpret 'bank' as relating to a financial institution. This differs from word embeddings, which typically represent each word as a fixed vector regardless of which other word it is being compared to. As such, comparisons between human similarity judgements and word embedding similarities may be limited in accuracy by ignoring the contextual effects that affect human judgements.

One potential solution is to replace static word embeddings with *contextual word embeddings*, where instead of being fixed for all uses, word embeddings are dynamically modified based on the context in which they occur (Ethayarajh, 2019; Ranashinghe et al., 2019). Contextual embeddings can be constructed by transformer-based architectures, which have achieved impressive results at sense disambiguation and other investigations of word similarity (Garí Soler and Apidianaki, 2021). However, the highly flexible and contextual nature of transformer embeddings makes it unclear

how exactly these contextual embeddings can be interpreted (Ethayarajh, 2019), and whether it even makes sense to analyse transformer embeddings from two different sentences as existing in the same semantic space (Mickus et al., 2019). Another problem is that contextual embeddings continuously vary in meaning across senses rather than forming discrete clusters, which differs from how polysemy is typically defined (Yenicelik et al., 2020).

An approach adopted by previous studies is to use traditional dictionaries to specify different word senses, combining definitions or example sentences with transformers to produce contextualised word embeddings for each sense (Ruzzetti et al., 2021; Tissier et al., 2017). The present study aims to build on previous research by using example sentences taken from dictionaries to construct word embeddings specialised for a particular context. We use these contextualised embeddings to investigate the extent to which polysemy reduces the ability of word embeddings to account for word similarity and relatedness datasets.

2 Methods

2.1 Analysis of word embeddings

In line with previous work, datasets of similarity and relatedness judgements were used to evaluate word embeddings by computing the Spearman correlation coefficient between human judgements and cosine similarities computed by word embedding models (Baroni et al., 2014). We used Spearman correlation as this is standard practise for evaluating ordinal human judgments of world similarity (Armendariz et al., 2020). See Table 3 in Appendix A for a full description of the embeddings used in this study.

Before computing correlations, the stimuli in the experimental datasets were pre-processed:

- All capitalisation was removed for consistency across datasets.
- Proper nouns were removed, as these have different semantic properties to regular nouns (Boleda et al., 2017).
- Word conjugations were altered to be in simple present infinitive form.
- Spelling was standardised to US spelling.

For the Tr9856 dataset, pre-processing removed so many sentences (mostly due to the presence of many proper nouns) that the modified dataset was renamed to Tr1058 to reflect that this is a small subset of the original dataset. This is indicated in Figure 5 in Appendix A.

For transformer models, decontextualised word embeddings were extracted by passing a single word to the transformer, averaging over multiple tokens when necessary. Contextualised transformer embeddings were computed using ERNIE as explained in Section 3.3. Embeddings were then normalised by dividing by the standard deviation in order to mitigate the problem of 'rogue dimensions', whereby a small number of dimensions account for most of the variation (Timkey and van Schijndel, 2021).

2.2 Verb antonymy

To assess the way antonyms are represented by vector-space semantics models, we manually identified antonym and near-antonym word pairs in the verb datasets, and computed the Spearman correlation between the relevant dataset and word embedding cosine similarities, both with and without these antonym pairs. The purpose of this analysis was to determine whether antonyms are represented differently compared to other word pairs. Verb datasets were chosen for this task as it was observed that the main available noun datasets contained relatively few pairs of antonyms.

2.3 Verb polysemy

To measure the effect of polysemy on semantic similarity judgements, contextual transformer embeddings were reduced to static embeddings using procedures developed previously (Bommasani et al., 2020; Soper and Koenig, 2022). The key idea of this approach is to use a transformer to compute embeddings of the target word in a given sentence context, and then average over multiple sentences to produce a context-sensitive static word embedding. By altering the sentences used to produce the contextual embedding, the resulting static word embeddings can be tailored to particular senses of the target word.

This method was applied using the ERNIE transformer, as it performed similarly to other leading transformer models while also being small and computationally tractable (see Section 3). We produced four distinct contextualised embeddings to test a variety of methods for incorporating contextual information relevant to polysemy. These four methods differ in the amount and quality of contextual information provided, as explained below and summarised in Table 4 of Appendix A.

Note that in order to disentangle the effects of antonymy from those of polysemy, subsequent analyses are performed on the verb datasets with antonyms removed.

The *ERNIE Wikipedia Basic embeddings* were computed from a set of sample sentences, each containing the target word, from a custom Wikipedia corpus of 10,000 articles. These were selected using a Wikipedia list of key articles, in order to provide sentences covering a diverse range of topics while also keeping the corpus a manageable size. The text of each article was imported using a Wikipedia Python API, and then processed to remove image captions, tables, citations, and other metadata. The result was a corpus consisting of 2 million sentences.

Word embeddings were then computed by finding sentences containing each target word within the corpus, up to a maximum of 100 sentences per target word to avoid wasting computational time for very common words. To ensure a match, words in each sentence were lemmatised using the nltk WordNetLemmatizer (Loper and Bird, 2002). Contextualised embeddings were computed for each matching sentence using ERNIE, and the token embeddings of the target word averaged over all sample sentences for that word. A lemmatiser was used to automatically conjugate each word in the sentence as a noun or verb to match the target. In cases in which the target word corresponded to more than one transformer token, the embeddings for each token were averaged.

The *ERNIE Wikipedia Verb embeddings* were computed in the same way, except that words in the sample sentences were now always lemmatised as verbs, thus ensuring the sample sentences reflected cases when the target word was used as a verb. This provides a simple method for controlling for polysemy of words that are used as both nouns and as verbs. A similar approach was taken for nouns, though little gain in performance was observed (see Figure 5 in Appendix A), so subsequent analysis focused only on verb polysemy.

The *ERNIE Dictionary Word embeddings* were calculated from sample sentences extracted for each target word from the Oxford Learner's Dictionary (Turnbull et al., 2010). It was hypothesised that using sentences tailored to providing examples of usage for each word would provide better disambiguation of polysemy than a large collection of assorted Wikipedia sentences. In this case, example sentences were pooled together regardless of the sense they corresponded to.

Finally, the *ERNIE Dictionary Sense embeddings* were constructed by manually separating example dictionary sentences into up to six different senses for each target word. This was performed by the authors, using the Oxford Learner's Dictionary and Longman Dictionary of Contemporary English Online (Pearson, 2023) as guides. Senses that shared a common grouping or heading in these dictionaries were generally combined, as

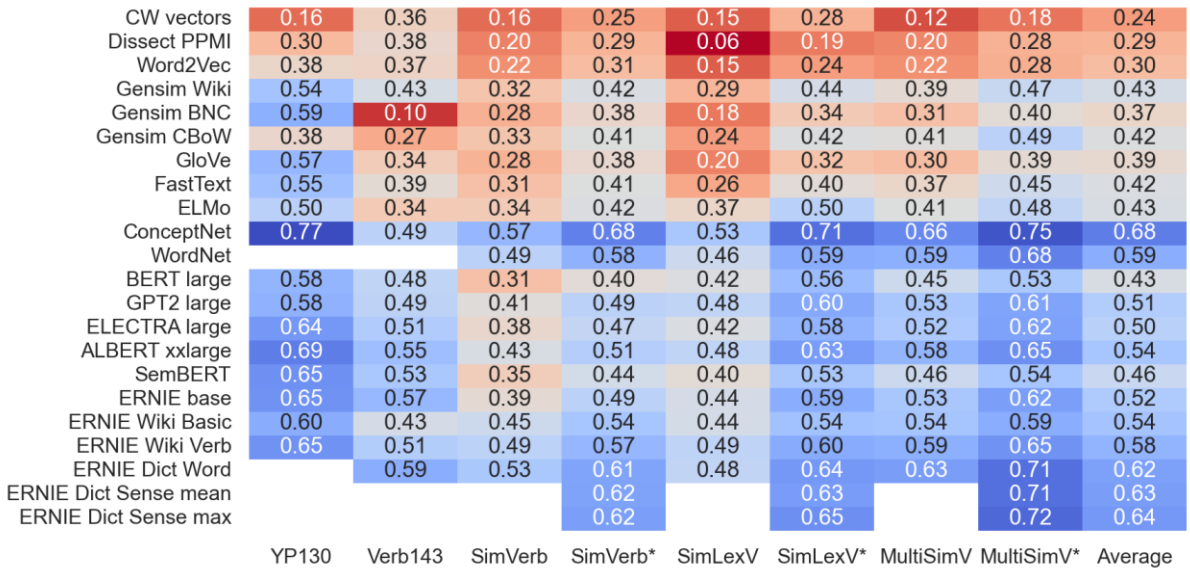

Figure 1: Spearman correlations between embedding models (rows) and verb subsets of experimental datasets (columns). An asterisk denotes exclusion of antonyms from the dataset. SimLexV indicates the SimLexVerb dataset, and likewise for MultiSimV. Average is weighted by dataset size. Note that ERNIE Dict embeddings were only computed for verb datasets with antonyms removed.

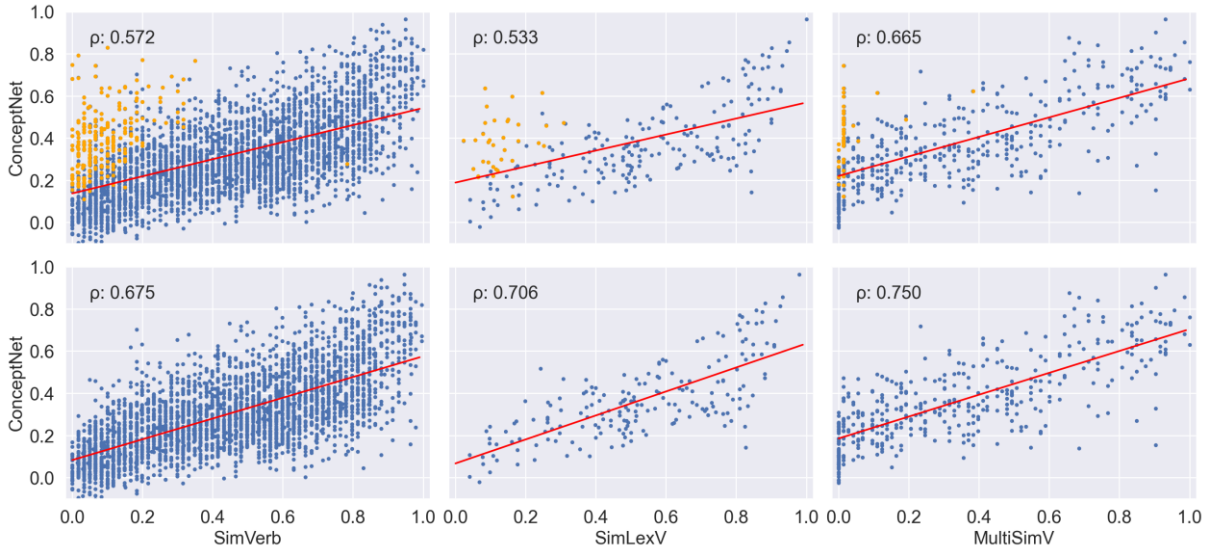

Figure 2: Effects of removing antonyms (shown in orange) from the SimVerb (left), SimLexVerb (centre), and MultiSimVerb (right) datasets, with experimental similarity judgements on the plotted on the horizontal axis against ConceptNet cosine similarities (vertical axis).

were instances where one sense is a subset of another. Rare senses containing few example sentences were excluded to focus on more common uses. We anticipated that manual consolidation of senses would improve the resulting word embeddings by allowing sample sentences to combined from the Oxford, Longman, and Collins Online Dictionaries (Collins, 2023).

Furthermore, while the previous methods pool all senses together, this approach produces embeddings for each individual sense. Such sense embeddings can be compared to the experimental datasets either by taking the average (*mean)* over all senses, or the maximum (*max*) similarity over all pairwise sense comparisons. We consider the maximum pairwise similarity because we hypothesise that participants may be sensitive to the most similar senses of two target words. Both results are shown in Figure 1.

3 Results

3.1 Analysis of word embeddings

To identify the best-performing embeddings to use in subsequent analysis, Spearman correlation coefficients between each word embedding model and the similarity ratings of all verb-based experimental datasets were computed (Figure 1). For comparison, the results for noun datasets are given in Appendix A. For both nouns and verbs, ConceptNet embeddings consistently show higher correlations with human judgements over almost all datasets. Transformers typically perform better than count- and predict-based embedding models, with GPT-2, ALBERT xxlarge, and ERNIE showing the highest correlations. We also observed some clustering of models, with static and contextualised embeddings being more similar to each other than to different types of models, as shown in Figures 5 and 6 in Appendix A.

Given its superior performance, ConceptNet was chosen as the focus of subsequent analysis of antonyms, for which static embeddings are sufficient. ERNIE was selected as a representative transformer for analysis of polysemy, as this required computing contextual embeddings which is not possible with ConceptNet.

3.2 Verb antonymy

Figure 2 shows scatterplots of ConceptNet cosine similarities against three verb-based datasets. The difference between the top and bottom rows of the subplots shows the effect of removing antonyms, which are seen to disproportionally cluster in the top left of the scatterplots. Removal of the antonyms substantially improves the fit between experimental and word embedding similarities, increasing the correlation on the SimVerb dataset from 0.572 to 0.675, from 0.533 to 0.706 on the SimLexVerb dataset, and from 0.665 to 0.750 on the MultiSimVerb dataset. This shows that humans represent the relations between antonyms very differently than do the ConceptNet embed-

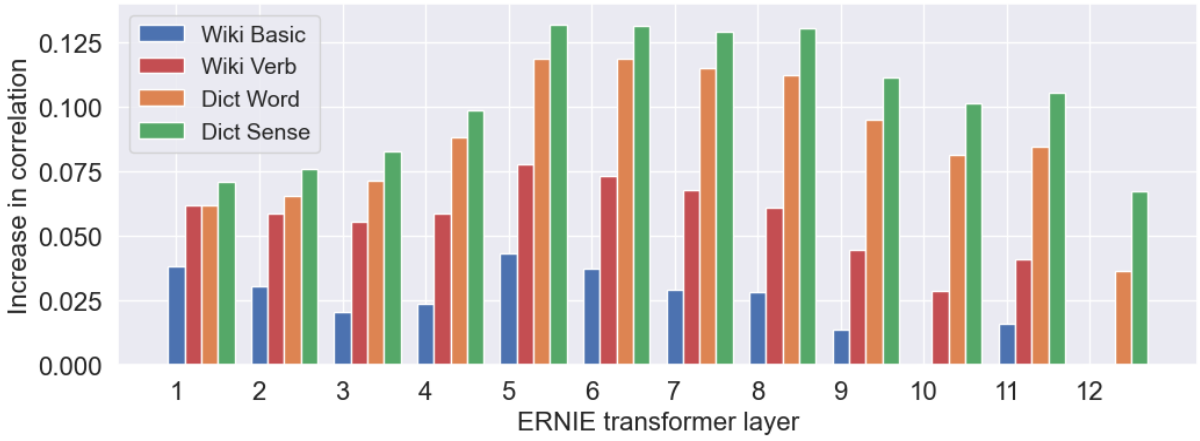

Figure 3: Comparison of the increase in correlation with SimVerb dataset relative to the ERNIE base model for the Wikipedia Basic, Wikipedia Verb, and Dictionary Word, and Dictionary Sense max embeddings. Correlations increase as more specific and fine-grained contextual information is added.

dings. Similar results were observed for ERNIE embeddings, as shown in Figure 8 of Appendix A.

3.3 Verb polysemy

Figure 3 shows the results of incorporating contextual information from corpus and dictionary sources by reducing contextual ERNIE embeddings to static embeddings, as outlined in Section 2.3. Relative to the layer 5 ERNIE base embeddings, Wikipedia Basic embeddings increase the correlation with human judgements in the SimVerb dataset by 5 percentage points, Wikipedia Verb embeddings by 8, Dictionary Word embeddings by 12, and the Dictionary Sense max embeddings by 13 percentage points.

We also examined the effect of transformer layers on the correlation with human judgements. Consistent with previous studies (Caucheteux et al., 2021; Timkey and van Schijndel, 2021), the best results are found around the middle layers of the transformer, indicating that later layers progressively incorporate relevant contextual infor-

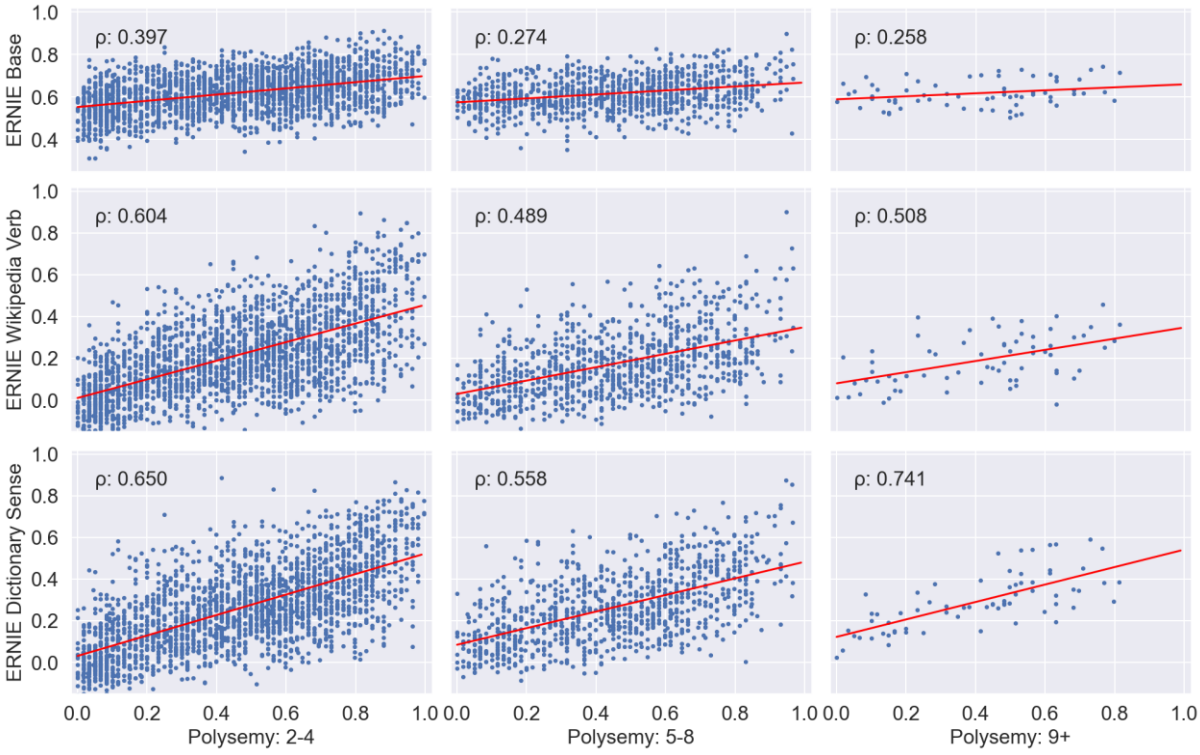

Figure 4: Comparison of the Spearman correlations of the SimVerb dataset with ERNIE Base (top), ERNIE Wikipedia Verb (middle), and ERNIE Dictionary Sense max (bottom) embeddings, split by polysemy score.

mation, but only up to a certain point. Henceforth we discuss results from layer 5.

To further investigate the effect of polysemy on the accuracy of word embeddings, SimVerb word pairs were grouped according to their total polysemy score, defined simply as the sum of the number of senses for both words in each pair. Senses were differentiated for each word during the construction of the ERNIE Dictionary Sense embeddings, as outlined in Section 2.3. As shown in Figure 4, and similarly to the results in Figure 3, the correlation with human ratings increases as more specific and fine-grained contextual information is added, with Wikipedia Verb embeddings showing higher correlations than the base model, and Dictionary Sense embeddings showing higher correlations still.

Furthermore, we found that correlations increase most for highly polysemous word pairs. Relative to the uncontextualised ERNIE base, the ERNIE Dictionary Sense embeddings increase correlations by 0.25 for the least polysemous, 0.29 for moderately polysemous, and 0.48 for the most polysemous word pairs. These results indicate that, while static word embeddings struggle to accurately represent the meaning of highly polysemous words, transformer models which incorporate contextual information perform much better.

4 Discussion

This paper has highlighted significant differences between the manner in which humans and word embedding models represent the meaning of antonyms. While it has long been known that word embeddings perform poorly in predicting antonym similarity judgements (Dou et al., 2018), we have shown the reason for this is that antonyms are given consistently low similarity ratings by humans but moderate to high cosine similarities by embedding models. This effect is consistent across datasets and large in magnitude, reducing correlations by 0.10-0.15, even though antonym or near-antonym word pairs only account for about 10% of each dataset.

Previous research has sought to rectify the low accuracy of word embedding models on antonyms by adding constraints to artificially pull antonyms further apart in semantic space (Mrkšić et al., 2016, Biesialska et al., 2020). However, we argue that this may be inappropriate, because when humans make similarity judgments between words, they may not be performing an analogous task to computing the cosine similarity between the corresponding embeddings. If this is the case, then the failure of word embedding cosine similarities to match human similarity judgments for antonyms should be interpreted as a limitation of the evaluation method, not a flaw of the word embeddings as a model of word meaning.

Relatedly, it has been argued that antonyms should have cosine similarities close to the smallest possible value of -1 (Samenko et al., 2020). In practise, however, negative cosine similarities occur mostly between unrelated words rather than antonyms, with small absolute values (up to around -0.1 for ConceptNet). This is likely because computing cosine similarity averages across all features whether salient or not, thereby computing 'property overlap' (Erk, 2016). Since antonyms share most features in common, this results in a high cosine similarity.

Why then do humans rate antonyms as having very low semantic similarity? One potential explanation is that the salience of the semantic features of a word varies depending on the context in which the word is used. This has been observed for human judgements of noun combination tasks (Bock and Clifton, 2000) and feature verification tasks (Montefinese et al., 2014). Such findings are consistent with our hypothesis that, when assessing the similarity of two antonyms, humans judge the dimension of meaning in which the two words differ as the most salient, and hence rate overall semantic similarity as low. This would also explain earlier findings that humans rate antonyms almost as similar as synonyms when asked to rate features separately, rather than providing an overall similarity score (Crutch et al., 2012).

These considerations highlight the need for a new method which enables more consistent and informative comparison between human similarity judgements and cosine similarities for antonyms. Unfortunately, in this study we were unable to develop such a method. We experimented with simple methods such as providing both words to the ERNIE transformer and extracting the contextualised embeddings of each, but this yielded no useful results. Further improvements will likely require identifying which particular features are most salient for assessment of antonyms, in line with several previous studies (Ali et al., 2019; Nguyen et al., 2016). In addition, our brief treatment of antonymy has not discussed important

issues such as adjectival antonyms or the effects of discourse context on negation (Kruszewski et al., 2016). We leave such considerations for future work.

In this study we also found that polysemy significantly reduces the accuracy of word embeddings in describing the similarity of verbs. The dramatic increase in the correlation of ERNIE embeddings with human judgements when contextual information was incorporated (see Figure 4) is evidence that the quality of the embeddings is significantly impaired by the inability to properly distinguish different word senses. Our results are consistent with a strategy whereby humans assess the similarity of two words using an implicit context that maximises the aspects of meaning they share, ignoring any additional polysemous meanings. This would explain why providing ERNIE with additional information about context, like parts of speech and example sentences, improves the correlation with human judgments.

Our results also highlight the value of using contextual information from lexical dictionaries to augment contextual word embeddings. In particular, ERNIE Dictionary Sense max embeddings increase the correlation by about 5 percentage points for the full SimVerb dataset (excluding antonyms), and about 23 percentage points for the most polysemous word pairs. Similar increases in correlation were observed from the simpler automated method of aggregating all dictionary senses together, as used in the ERNIE Dictionary Word embeddings. We hypothesise that these improvements arise because example dictionary sentences represent common uses of verb, which may reflect the way that humans judge word similarities when asked to judge two words without context.

A different approach to control for the effects of polysemy used in several past studies is to ask participants to judge the similarity of words in the context of a specific sentence, thereby allowing for clearer sense disambiguation (Armendariz et al., 2020; Camacho-Collados and Pilehvar, 2018; Haber and Poesio, 2021). However, it is difficult to ensure that participants do not simply judge the overall similarity of the sentences, or conversely ignore the context and consider the target words in isolation. Furthermore, contextualised word embeddings are more difficult to interpret than static embeddings since they only apply to the word in a specific precise context. Given that a concept is typically defined as a mental representation that is reasonably invariant across contexts (Laurence and Margolis, 1999; Musz and Thompson-Schill, 2018), highly context-specific word embeddings are arguably of less value as cognitive models of concepts. As such, we believe there is also value in incorporating contextual information to improve static embeddings of polysemous words, as we have shown can be done by using example sentences from lexical dictionaries.

In this paper we have focused on ERNIE embeddings, as they showed superior performance over competing models that are purely text-based. The performance of ConceptNet embeddings provide an additional baseline that also incorporates expert linguistic knowledge. The results corroborates previous studies which found that adding expert knowledge can improve the performance of embeddings derived from word co-occurrence statistics (Peters et al., 2019; Xu et al., 2021; Zhang et al., 2020). Nevertheless, transformer models like ERNIE use much larger training corpuses and have more parameters than ConceptNet (Devlin et al., 2019), so the fact that ConceptNet still outperforms all transformer embeddings is a notable finding. However, we do not seek to determine the effect of specific architectural choices or hyperparameters, as such analysis has been conducted in previous studies (Baroni et al., 2014; Lapesa and Evert, 2014; Liu et al., 2021).

5 Conclusion

In this study we have highlighted the problems of ignoring the implicit context in which humans make word similarity judgements. Our results show that word meaning is judged in a context-dependent manner which decontextualised word embeddings struggle to adequately capture. Future work focused on improving embeddings may require better datasets specifically focused on evaluating how humans rate the similarity of different forms of antonyms. Also important is improving the representation of polysemy, which we have shown is possible by combining contextualised embeddings with carefully collated data from dictionaries and other knowledge banks. Our analysis has primarily focused on verbs, and so further work focusing on nouns is also needed. Overall, much work remains to be done to enhance the ability of vector-space semantic models to describe a wide range of semantic phenomena.

References

Muhammad Asif Ali, Yifang Sun, Xiaoling Zhou, Wei Wang, and Xiang Zhao. 2019. 'Antonym-synonym classification based on new sub-space embeddings.' In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6204-11.

Felipe Almeida and Geraldo Xexéo. 2019. 'Word embeddings: A survey', *arXiv preprint arXiv:1901.09069*.

Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, Marko Robnik-Šikonja, Mark Granroth-Wilding, and Kristiina Vaik. 2020. 'CoSimLex: A resource for evaluating graded word similarity in context', In *Proceedings of the 12th International Conference on Language Resources and Evaluation,* pages 5878-86.

Simon Baker, Roi Reichart, and Anna Korhonen. 2014. 'An Unsupervised Model for Instance Level Subcategorization Acquisition.' *EMNLP*, 278-89.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. 'Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors.' In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238-47.

Magdalena Biesialska, Bardia Rafieian, and Marta R. Costa-jussà. 2020. Enhancing Word Embeddings with Knowledge Extracted from Lexical Resources. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 271–278.

Corinne L Bloch-Mullins. 2021. 'Similarity Reimagined (with Implications for a Theory of Concepts)', *Theoria*, 87: 31-68.

Jeannine S Bock and Charles Clifton. 2000. 'The role of salience in conceptual combination', *Memory & Cognition*, 28: 1378-86.

Gemma Boleda. 2020. 'Distributional semantics and linguistic theory', *Annual review of Linguistics*, 6: 213-34.

Gemma Boleda, Abhijeet Gupta, and Sebastian Padó. 2017. 'Instances and concepts in distributional space.' In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 79-*85*.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. 'Interpreting pretrained contextualized representations via reductions to static embeddings.' In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758-81.

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 (Volume 1: Long Papers)*, pages 136-45.

Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2018. 'From word to sense embeddings: A survey on vector representations of meaning', *Journal of Artificial Intelligence Research*, 63: 743-88.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. 'Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity.' Association for Computational Linguistics.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2021. 'Disentangling syntax and semantics in the brain with deep networks.' *International Conference on Machine Learning*, 1336-48. PMLR.

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.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. 'ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators', *ArXiv*, abs/2003.10555.

Stephen Clark. 2015. 'Vector space models of lexical meaning', *The Handbook of Contemporary semantic theory*: 493-522.

Collins. 2023. 'Collins online dictionary'. https://www.collinsdictionary.com/.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. 'Natural language processing (almost) from scratch', *Journal of machine learning research*, 12: 2493−537.

Sebastian J Crutch, Paul Williams, Gerard R Ridgway, and Laura Borgenicht. 2012. 'The role of polarity in antonym and synonym conceptual knowledge: Evidence from stroke aphasia and multidimensional ratings of abstract words', *Neuropsychologia*, 50: 2636-44.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. 'Predicting human similarity judgments with distributional models: The value of word associations.' In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1861-70.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *ArXiv*, abs/1810.04805.

Zehao Dou, Wei Wei, and Xiaojun Wan. 2018. 'Improving word embeddings for antonym detection using thesauri and sentiwordnet.' *CCF international conference on natural language processing and Chinese computing*, 67-79. Springer.

Guy Emerson. 2020. 'What are the Goals of Distributional Semantics?', *arXiv preprint arXiv:2005.02982*.

Katrin Erk. 2016. 'What do you know about an alligator when you know the company it keeps?', *Semantics and Pragmatics*, 9: 17-1-63.

Mathias Etcheverry and Dina Wonsever. 2019. 'Unraveling antonym's word vectors through a siamese-like network.' In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3297-307.

Kawin Ethayarajh. 2019. 'How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings', In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* pages *55-65, Hong Kong, China.*

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. 'Placing search in context: The concept revisited.' In *Proceedings of the 10th international conference on World Wide Web*, pages 406-14.

Aina Garí Soler and Marianna Apidianaki. 2021. 'Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses', *Transactions of the Association for Computational Linguistics*, 9: 825-44.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. 'Simverb-3500: A large-scale evaluation set of verb similarity', In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182.

Fritz Günther, Luca Rinaldi, and Marco Marelli. 2019. 'Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions', *Perspectives on Psychological Science*, 14: 1006-33.

Janosch Haber and Massimo Poesio. 2021. 'Patterns of Lexical Ambiguity in Contextualised Language Models', *arXiv preprint arXiv:2109.13032*.

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 1406-14.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. 'Simlex-999: Evaluating semantic models with (genuine) similarity estimation', *Computational Linguistics*, 41: 665-95.

Tomáš Kliegr and Ondřej Zamazal. 2018. 'Antonyms are similar: Towards paradigmatic association approach to rating similarity in SimLex-999 and WordSim-353', *Data & Knowledge Engineering*, 115: 174-93.

Germán Kruszewski, Denis Paperno, Raffaella Bernardi, and Marco Baroni. 2016. 'There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics.' *Computational Linguistics,* 42: 637-660.

Sandra Kotzor. 2021. *Antonyms in Mind and Brain: Evidence from English and German* (Routledge).

Andrei Kutuzov, Murhaf Fares, Stephan Oepen, and Erik Velldal. 2017. 'Word vectors, reuse, and replicability: Towards a community repository of large-text resources.' In *Proceedings of the 58th Conference on Simulation and Modelling*, pages 271-76. Linköping University Electronic Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. 'ALBERT: A Lite BERT for Self-supervised Learning of Language Representations', *ArXiv*, abs/1909.11942.

Gabriella Lapesa, and Stefan Evert. 2014. 'A large scale evaluation of distributional semantic models: Parameters, interactions and model selection', *Transactions of the Association for Computational Linguistics*, 2: 531-46.

Stephen Laurence, and Eric Margolis. 1999. 'Concepts and cognitive science', *Concepts: core readings*, 3: 81.

Alessandro Lenci. 2018. 'Distributional models of word meaning', *Annual review of Linguistics*, 4: 151-71.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2021. 'A comprehensive comparative evaluation and analysis of Distributional Semantic Models', *arXiv preprint arXiv:2105.09825*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. 'Improving distributional similarity with lessons learned from word embeddings', *Transactions of the Association for Computational Linguistics*, 3: 211-25.

Ran Levy, Liat Ein Dor, Shay Hummel, Ruty Rinott, and Noam Slonim. 2015. 'Tr9856: A multi-word term relatedness benchmark.' 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 419-24.

Antonio Lieto, Antonio Chella, and Marcello Frixione. 2017. 'Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation', *Biologically inspired cognitive architectures*, 19: 1-9.

Liyuan Liu, Jialu Liu, and Jiawei Han. 2021. 'Multi-head or single-head? an empirical comparison for transformer training', *arXiv preprint arXiv:2106.09650*.

Edward Loper and Steven Bird. 2002. 'Nltk: The natural language toolkit', *arXiv preprint cs/0205028*.

Minh-Thang Luong, Richard Socher, and Christopher D Manning. 2013. 'Better word representations with recursive neural networks for morphology.' In *Proceedings of the seventeenth conference on computational natural language learning*, pages 104-13.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees Van Deemter. 2019. 'What do you mean, BERT? Assessing BERT as a Distributional Semantics Model', *arXiv preprint arXiv:1911.05758*.

George A Miller, and Walter G Charles. 1991. 'Contextual correlates of semantic similarity', *Language and cognitive processes*, 6: 1-28.

Nikola Mrkšić, Diarmuid Ó. Séaghdha, Blaise Thomson, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. 'Counter-fitting Word Vectors to Linguistic Constraints.' In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 142-148.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. 'Semantic significance: a new measure of feature salience', *Memory & Cognition*, 42: 355-69.

Elizabeth Musz and Sharon L Thompson-Schill. 2018. 'Finding Concepts in Brain Patterns.' In *The oxford handbook of neurolinguistics*, New York, NY: Oxford University Press.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. 'Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction', In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* pages 454–459.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. 'Making fine-grained and coarse-grained sense distinctions, both manually and automatically', *Natural Language Engineering*, 13: 137-63.

Pearson. 2023. 'Longman Dictionary of Contemporary English Online'. https://www.ldoceonline.com/.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. 'WordNet:: Similarity-Measuring the Relatedness of Concepts.' *AAAI*, 25-29.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. 'Glove: Global vectors for word representation.' In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-43.

Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. 'Knowledge enhanced contextual word representations', In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing,* pages 43–54.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 'Deep Contextualized Word Representations.' *North American Chapter of the Association for Computational Linguistics*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. 'Language Models are Unsupervised Multitask Learners.' *OpenAI blog 1*, no. 8: 9.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. 'A word at a time: computing word relatedness using temporal semantic analysis.' In *Proceedings of the 20th international conference on World wide web*, pages 337-46.

Tharindu Ranashinghe, Constantin Orasan, and Ruslan Mitkov. 2019. 'Enhancing unsupervised sentence similarity methods with deep contextualised word representations.' RANLP.

João Rodrigues, Ruben Branco, Joao Silva, Chakaveh Saedi, and António Branco. 2018. 'Predicting brain activation with WordNet embeddings.' In *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, pages 1-5.

Herbert Rubenstein and John B Goodenough. 1965. 'Contextual correlates of synonymy', *Communications of the ACM*, 8: 627-33.

Elena Sofia Ruzzetti, Leonardo Ranaldi, Michele Mastromattei, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2022. 'Lacking the embedding of

a word? look it up into a traditional dictionary', *2022. Lacking the Embedding of a Word? Look it up into a Traditional Dictionary. In Findings of the Association for Computational Linguistics: ACL 2022,* pages 2651–2662.

Chakaveh Saedi, António Branco, João Rodrigues, and Joao Silva. 2018. 'Wordnet embeddings.' In *Proceedings of the third workshop on representation learning for NLP*, pages 122-31.

Igor Samenko, Alexey Tikhonov, and Ivan P Yamshchikov. 2020. 'Synonyms and antonyms: Embedded conflict', *ArXiv, abs/2004.12835*.

Yong Shi, Yuanchun Zheng, Kun Guo, Wei Li, and Luyao Zhu. 2018. 'Word similarity fails in multiple sense word embedding.' *International Conference on Computational Science*, 489-98. Springer.

Elizabeth Soper and Jean-Pierre Koenig. 2022. 'When Polysemy Matters: Modeling Semantic Categorization with Word Embeddings.' In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 123-31.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. 'Conceptnet 5.5: An open multilingual graph of general knowledge.' *Thirty-first AAAI conference on artificial intelligence*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, and Yuxiang Lu. 2021. 'Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation', *arXiv preprint arXiv:2107.02137*.

William Timkey and Marten van Schijndel. 2021. 'All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality'. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4527–46.

Julien Tissier, Christophe Gravier, and Amaury Habrard. 2017. 'Dict2vec: Learning word embeddings using lexical dictionaries.' In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254-63.

Jatin Karthik Tripathy, Sibi Chakkaravarthy Sethuraman, Meenalosini Vimal Cruz, Anupama Namburu, P Mangalraj, and Vaidehi Vijayakumar. 2021. 'Comprehensive analysis of embeddings and pre-training in NLP', *Computer Science Review*, 42: 100433.

Sean Trott and Benjamin Bergen. 2021. 'RAW-C: Relatedness of Ambiguous Words--in Context (A New Lexical Resource for English)', In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7077-87.

Joanna Turnbull, D Lea, D Parkinson, P Phillips, B Francis, S Webb, V Bull, and M Ashby. 2010. 'Oxford advanced learner's dictionary'. https://www.oxfordlearnersdictionaries.com/.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, and Thierry Poibeau. 2020. 'Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity', *Computational Linguistics*, 46: 847-97.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. 'Probing pretrained language models for lexical semantics', In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pages 7222-40.

Shaonan Wang, Jiajun Zhang, Haiyan Wang, Nan Lin, and Chengqing Zong. 2020. 'Fine-grained neural decoding with distributed word representations', *Information Sciences*, 507: 256-72.

Gijs Wijnholds, and Mehrnoosh Sadrzadeh. 2019. 'Evaluating Composition Models for Verb Phrase Elliptical Sentence Embeddings.' 261-71. Minneapolis, Minnesota: Association for Computational Linguistics.

Wenwen Xu, Mingzhe Fang, Li Yang, Huaxi Jiang, Geng Liang, and Chun Zuo. 2021. 'Enabling Language Representation with Knowledge Graph and Structured Semantic Information.' *2021 International Conference on Computer Communication and Artificial Intelligence (CCAI)*, 91-96. IEEE.

Dongqiang Yang, and David M. W. Powers. 2006. 'Verb similarity on the taxonomy of WordNet.'

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. 'How does BERT capture semantics? A closer look at polysemous words.' In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 156-62.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. 'Recent trends in deep learning based natural language processing', *IEEE Computational Intelligence magazine*, 13: 55-75.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. 'Semantics-aware BERT for language understanding.' In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9628-35.

A Additional Figures and Tables

Model Name	Number Word Pairs	Part of Speech	Data Type	Citation
RG65	65	Nouns	Similarity	Rubenstein and Goodenough (1965)
WordSim-353	353	Nouns	Relatedness	Finkelstein et al. (2001)
SimLex-999	999	Mixed	Similarity	Hill et al. (2015)
YP-130	130	Verbs	Similarity	Yang and Powers (2006)
Verb-143	143	Verbs	Similarity	Baker et al. (2014)
Multi-SimLex	1,888	Mixed	Similarity	Vulić, Baker, et al. (2020)
SimVerb-3500	3,500	Verbs	Similarity	Gerz et al. (2016)
MEN	3,000	Nouns	Relatedness	Bruni et al. (2012)
MTurk-287	287	Nouns	Relatedness	Radinsky et al. (2011)
MTurk-771	771	Nouns	Relatedness	Halawi et al. (2012)
Tr9856	9,856	Nouns	Relatedness	Levy, Dor, et al. (2015)
SemEval-2017	500	Nouns	Relatedness	Camacho-Collados et al. (2017)
Stanford-RW	2,034	Mixed	Similarity	Luong et al. (2013)

Table 1: Summary of word similarity and relatedness experimental datasets.

Models Tested	WS353	SL999	MEN	MT287	MT771	RW	SV	Citation
PMI model, Skip-gram, GloVe	.71	.43	.78	.69		.51		Levy, Goldberg, et al. (2015)
PMI model, CBOW	.79	.43	.79	.78	.71			De Deyne et al. (2016)
Skip-gram	.70	.34	.73	.66	.61	.40		Chiu et al. (2016)
Count-based, CBOW, GloVe, FastText	.70	.40	.78					Wijnholds and Sadrzadeh (2019)
BERT, GPT-2, RoBERTa, XLNet, DistilBERT	.72	.55					.45	Bommasani et al. (2020)
LSA, LDA, CBOW, skip-gram, GloVe, RI, FastText, BERT	.71	.49	.79	.71		.48	.41	Lenci et al. (2021)

Table 2: Summary of previous analyses of word embedding models, showing the highest Spearman correlation recorded by each paper for each analysed dataset. WS: WordSim, SL: SimLex, MT: MTurk, RW: Stanford-RW, SV: SimVerb.

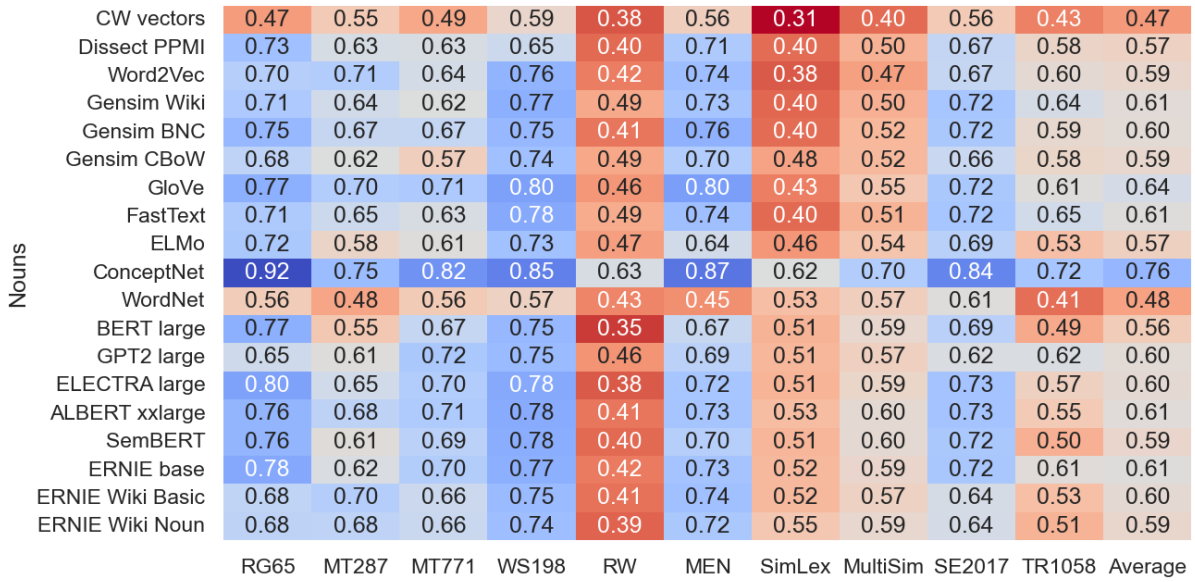

Figure 5: Spearman correlations between embedding models (rows) and noun-based experimental datasets (columns).

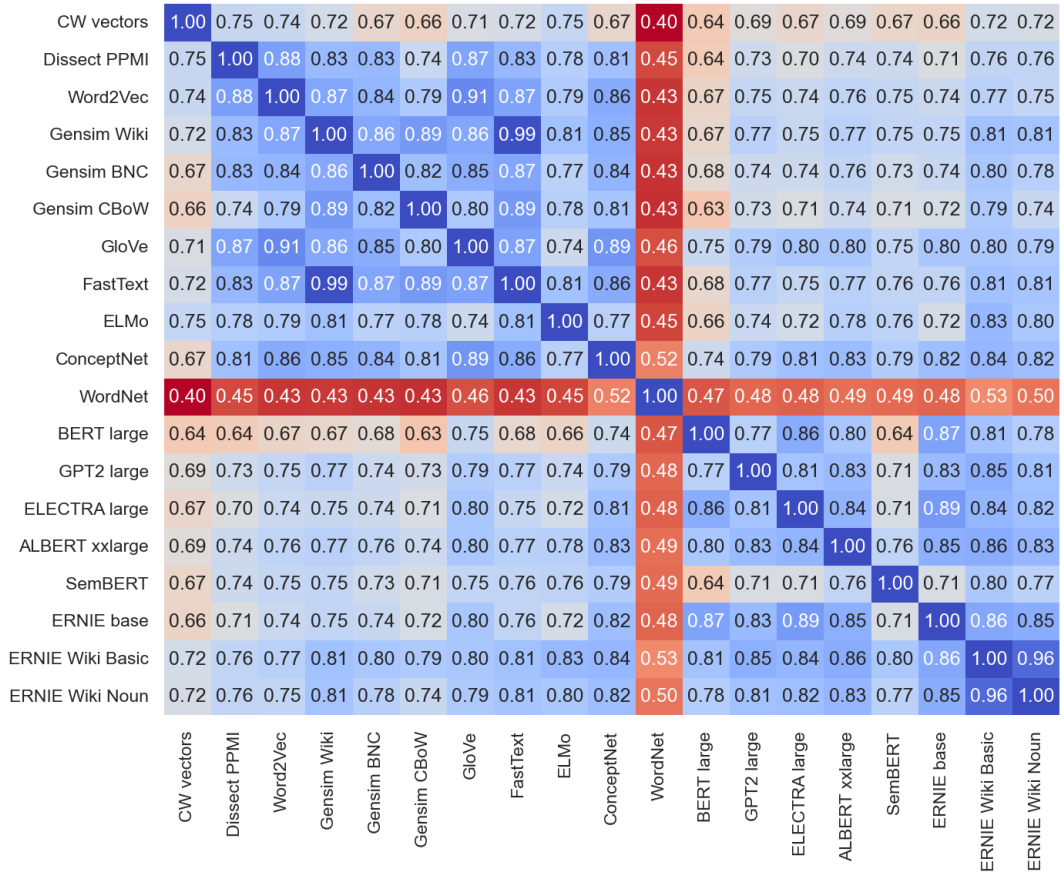

Figure 6: Correlation matrices of all models computed over the vocabulary of the MEN noun dataset.

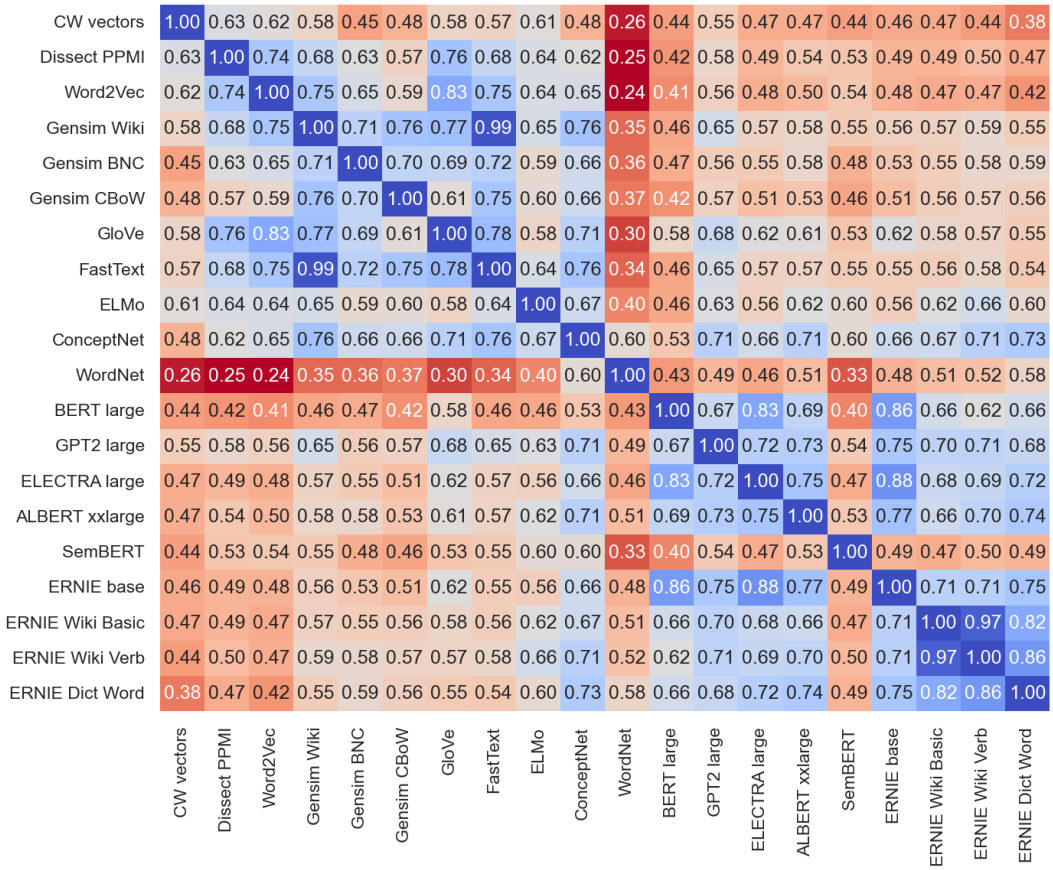

Figure 7: Correlation matrices of all models computed over the vocabulary of the SimVerb verb dataset.

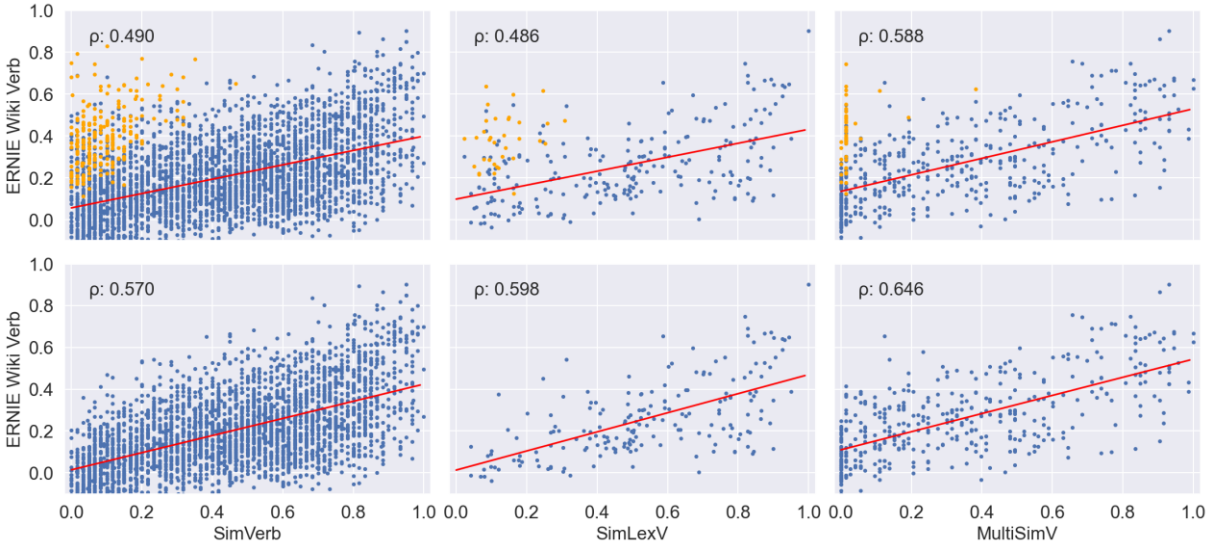

Figure 8: Effects of removing antonyms (shown in orange) from the SimVerb (left), SimLexVerb (centre), and MultiSimVerb (right) datasets.

Model Name	Type	Explanation	Citation
CW vectors	Count	Regression model trained over Wikipedia corpus.	Collobert et al. (2011)
Dissect PPMI	Count	Trained using Positive Point-wise mutual information (PPMI) over ukWaC, Wikipedia, and the British National Corpus.	Baroni et al. (2014)
Word2Vec skipgram	Predict	Skipgram model trained over Wikipedia.	Kutuzov et al. (2017)
Gensim Wiki	Predict	Skipgram model trained over Wikipedia and Gigaword corpus.	Kutuzov et al. (2017)
Gensim BNC	Predict	Skipgram model trained over British National Corpus.	Kutuzov et al. (2017)
Genism CBoW	Predict	Continuous Bag of Words (CBoW) model trained over Gigaword corpus.	Kutuzov et al. (2017)
GloVe	Predict	Custom regression model trained over 840 billion token corpus from the Common Crawl.	Pennington et al. (2014)
FastText	Predict	Skipgram model trained over Wikipedia and Gigaword corpus.	Kutuzov et al. (2017)
ELMo	Predict	A 94 million parameter bidirectional Long Short Term Memory (LSTM) trained over a 30 million word corpus.	Peters et al. (2018)
ConceptNet	Knowledge	ConceptNet relations are encoded into vectors by applying PPMI to the relation adjacency matrix, plus extra information from GloVe and word2vec.	Speer et al. (2017)
WordNet	Knowledge	WordNet relations encoded into vectors by counting number of intermediate nodes.	Saedi et al. (2018)
BERT large	Transformer	A 340 million parameter transformer model trained on a 3.3 billion token corpus from Wikipedia and BooksCorpus.	Devlin et al. (2019)
GPT2 large	Transformer	A 1.5 billion parameter transformer model trained on a web corpus of 8 million documents.	Radford et al. (2019)
ELECTRA large	Transformer	A 335 million parameter transformer model trained on a 33 billion token web corpus.	Clark et al. (2020)
ALBERT xxlarge	Transformer	A 233 million parameter transformer model trained based on BERT.	Lan et al. (2020)
SemBERT	Transformer	A 240 million parameter transformer model based on BERT and incorporating semantic role labelling.	Zhang et al. (2020)
ERNIE	Transformer	A 10 billion parameter transformer model trained on a corpus of plain text and knowledge graphs.	Sun et al. (2021)

Table 3: Summary of word embedding models used in this paper.

Model Name	Explanation	Specificity of Context
ERNIE base	No context provided.	None
ERNIE Wiki Basic	Context provided from a corpus of Wikipedia articles, with words matched using automatic lemmatisation.	Least
ERNIE Wiki Verb	Context provided from a corpus of Wikipedia articles, and only matching words conjugated as verbs. This should avoid matching cases where verbs as used as nouns.	Less
ERNIE Dictionary Word	Context provided by example sentences extracted automatically from Oxford Online Dictionary. This should provide higher-quality and more relevant use cases representative of the words.	More
ERNIE Dictionary Sense	Context provided by a curated set of example sentences separated by sense from the Oxford, Longman, and Collins Online dictionaries.	Most

Table 4: Summary of ERNIE embeddings constructed in the paper, and with an indication of how fine-grained is the context incorporated into the embeddings.