Sentence Selection Strategies for Distilling Word Embeddings from BERT

Yixiao Wang, Zied Bouraoui, Luis Espinosa-Anke, Steven Schockaert

Sentence Selection Strategies for Distilling Word Embeddings from BERT

Yixiao Wang[1], Zied Bouraoui[2], Luis Espinosa-Anke[1], Steven Schockaert[1]
[1] Cardiff University, UK, {wangy306, espinosa-ankel, schockaerts1}@cardiff.ac.uk
[2]CRIL-CNRS, Université d'Artois, France, zied.bouraoui@cril.fr

Abstract

Many applications crucially rely on the availability of high-quality word vectors. To learn such representations, several strategies based on language models have been proposed in recent years. While effective, these methods typically rely on a large number of contextualised vectors for each word, which makes them impractical. In this paper, we investigate whether similar results can be obtained when only a few contextualised representations of each word can be used. To this end, we analyze a range of strategies for selecting the most informative sentences. Our results show that with a careful selection strategy, high-quality word vectors can be learned from as few as 5 to 10 sentences.

Keywords: Word Embeddings, Language Models, Natural Language Processing

1. Introduction

The ability to model word meaning *in context* is a central feature of transformer-based language models. Nonetheless, since the introduction of BERT (Devlin et al., 2019), various authors have also explored the possibility of using language models for learning better *static* word embeddings (Ethayarajh, 2019; Bommasani et al., 2020; Vulić et al., 2020; Gan et al., 2020; Li et al., 2021; Gupta and Jaggi, 2021). Distilling static word vectors from language models is useful for several reasons. Some authors have focused on analyzing these vectors to better understand how BERT captures word meaning (Ethayarajh, 2019), and for investigating the social biases of such models (Bommasani et al., 2020). Liu et al. (2020) found that static word vectors can be useful as "anchors" for improving the representations of language models, while Alghanmi et al. (2020) also obtained improved results by combining BERT with static word vectors. Another motivation comes from application settings where efficiency and/or transparency is key (Gupta and Jaggi, 2021). However, our main motivation comes from applications where word meaning has to be modelled in the absence of any sentence context. Examples of such applications include ontology alignment (Kolyvakis et al., 2018), ontology completion (Li et al., 2019), and zero-shot (Socher et al., 2013; Ma et al., 2016) and few-shot (Xing et al., 2019; Li et al., 2020; Yan et al., 2021) learning.

The most common strategy for distilling word vectors from BERT is based on sampling sentences that mention the words of interest. To obtain an embedding of w, we can then simply average the contextualised representations of w across these sentences. Specific strategies differ in (i) which layers of the transformer model are being used and (ii) whether or not the word w is masked. The issue of masking w was studied in particular by Li et al. (2021), who found this to lead to representations that were better suited to predicting semantic properties of nouns. In previous evaluations, a large number of sentences were used for each word, which is computationally expensive. We may thus wonder whether high-quality embeddings can be learned from just a few mentions of each word. A related research question is whether better results are possible by selecting sentences strategically rather than at random. In this paper, we aim to answer these questions by empirically analyzing a range of strategies for selecting mentions of a given word w. On the one hand, this is motivated by the practical desire to distil word vectors from language models in a more efficient way. On the other hand, comparing the effectiveness of different sentence selection strategies can also provide us with insights into how language models acquire knowledge about word meaning.[1]:

2. Related Work

Ethayarajh (2019) was one of the first to distil static word vectors from language models, as a mechanism for probing how models such as BERT and GPT-2 (Radford et al., 2019) capture word meaning. They computed contextualised vectors from different sentences mentioning the target word, and then used the principal component as a static word vector. Most later works used the average rather than the principal component, but both strategies are nearly equivalent given the high anisotropy of contextualised vectors (Ethayarajh, 2019). Bommasani et al. (2020) compared different strategies for pooling contextualised vectors and for getting representations of words that are split into multiple sub-word tokens, among others. They found that the best results were obtained when averaging the representations of the sub-word tokens, and averaging the resulting contextualised word vectors across the available sentences. Furthermore, they found that using 500 sentences for each word led to much better representations than using 10 or 100 words. Another finding from this paper is that the performance of the word vectors can differ quite substantially depending on which

[1]All code and data to replicate our experiments is available at `https://github.com/Activeyixiao/Sentence-Selection-Strategies/`.

layer of the language model is used for obtaining them. Building on this latter insight, Vulić et al. (2020) found that averaging the representations of the first k layers can lead to better results than using the vectors from any individual layer. The optimal value of k depends on the language, task and configuration, but averaging the representations across all layers consistently provided close-to-optimal results. This latter observation was also made by Li et al. (2021), who tried to select k based on validation data. For the word classification benchmarks they considered, using the last layer often performed similar to either averaging the first k layers or selecting a single layer based on the validation data. They also proposed to mask the target word when computing the contextualised vectors. In this case, they only obtained vectors from the final layer, given that the [MASK] token makes the early layers too uninformative. For most classification datasets, they found that masking the target word led to better results. However, on word similarity benchmarks, the vectors obtained with masking under-performed.

A number of strategies which do not rely on averaging contextualised vectors have been proposed as well. For instance, Gan et al. (2020) propose two strategies which use BERT to obtain words that are similar to a target word. This information is compiled into a synthetic co-occurrence matrix, which is then used as input to the GloVe (Pennington et al., 2014) word embedding method. Gupta and Jaggi (2021) use a Word2Vec (Mikolov et al., 2013) inspired model, which uses BERT to obtain a vector encoding of the context of the target word. Their method performed better than averaging strategies, but only when a large number of occurrences of each word were used. A somewhat similar idea was pursued by Wang et al. (2021), who proposed a modification of Skip-gram in which BERT encodings were used to represent contexts.

Beyond work on learning static word embeddings, several authors have investigated which aspects of word meaning are captured by language models. For instance, one line of work has focused on analysing to what extent language models understand the properties of everyday concepts (Forbes et al., 2019; Weir et al., 2020; Li et al., 2021), finding that language models clearly outperform word embedding models such as Word2Vec and GloVe in this respect. Shwartz and Choi (2020) found that language models can to some extent predict properties that are never or rarely stated explicitly, as is often the case for commonsense properties (Gordon and Durme, 2013), but that such models suffer from over-generalization. Comparing the extent to which BERT and FastText (Bojanowski et al., 2017) can act as knowledge bases for answering factual queries (Petroni et al., 2019), Dufter et al. (2021) found that the outperformance of BERT mainly comes from its ability to model the semantic types of candidate answers.

The aforementioned works lend support to the idea that static word vectors induced from language models may have some inherent advantages compared to those from standard word embedding models. As the analysis from Li et al. (2021) revealed, however, different types of word vectors often have complementary strengths. In particular, as argued by Dufter et al. (2021), the ability to learn vectors for a large vocabulary remains an important advantage of traditional word vectors.

3. Distilling Word Embeddings

To obtain the vector representation of a word w, we first sample n sentences $S_1, ..., S_n$ mentioning w. Unless noted otherwise, the source corpus from which these sentences are sampled in our experiments is always Wikipedia, specifically a dump from March 2021. Wikipedia has been used extensively in many areas of NLP, with notable use cases including lexical semantics (Navigli and Velardi, 2010), knowledge extraction and management, or taxonomy learning (Suchanek et al., 2008). For our purposes, moreover, Wikipedia is also a clean resource for encyclopedic information, which, despite its collective nature, undergoes strict editorial revisions, and which has a particular structure that we can exploit. We now discuss the process of sampling our target sentences from Wikipedia. Each of the sentences $S_1, ..., S_n$ is fed through a masked language model such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019). From each sentence S_i we obtain a contextualised vector $\mathbf{w_i}$ using one of the following alternatives:

- MASK: We replace the word w by a single [MASK] token and let $\mathbf{w_i}$ be the representation of this token in the final layer.
- LAST: We feed the original sentence to the language model. In this case, the word w may consist of several sub-word tokens $w_1, ..., w_m$. We let $\mathbf{w_i}$ be the average of the final layer representations of these tokens.
- AVG: Similar as LAST, except that we take the average representation across all layers, rather than only using the final layer.

The final embedding $\mathbf{w}$ of word w is obtained by averaging the contextualised vector $\mathbf{w_1}, ..., \mathbf{w_n}$.

4. Sentence Selection Strategies

We need a strategy for selecting n sentences that mention a given target word w. Our baseline strategy, which we will refer to as RAND, is to randomly sample different sentences from Wikipedia. To avoid poorly structured sentences, we avoid sentences with more than 60 or fewer than 7 words. We now discuss a number of alternative sentence selection strategies, aimed at providing us with more informative sentences. Our hypothesis is that this will allow us to obtain high-quality word vectors from a small number of sentences, which

is essential for scaling up the methods for distilling word embeddings from Section 3. Given this focus on efficiency, we are particularly interested in sentence selection strategies with a low computational overhead. We first consider two strategies that rely on the structure of Wikipedia:

- INTRO: We only sample sentences that occur in the introductory section of a Wikipedia article (regardless of what the article is about). The intuition is that these introductory sections are more likely to contain sentences in which properties of words are mentioned explicitly.

- HOME: If there is a Wikipedia article about w, we select the first n sentences mentioning w from that article. If w does not have a Wikipedia article, we fall back on RAND.

We also analyse a number of strategies that rely on aspects of the sentences themselves:

- POS: We only sample sentences which start with the word w in plural form. The intuition is that such sentences are likely to express generic knowledge about w.

- ENUM: We first select all sentences in which w is preceded or succeeded by a comma or the word 'and'. Then we rank these sentences based on the number of commas, as a simple strategy for prioritizing longer enumerations, and we select the n highest ranked sentences. The intuition is that enumerations can provide us with useful knowledge, capturing the fact that the words in the enumeration have some property in common with w.

- PMI: For all words that co-occur with w in at least 2 sentences, we compute their Pointwise Mutual Information (PMI) in an offline preprocessing step. This PMI score reflects to what extent these words appear more often in the same sentence than would be expected by chance, given their overall frequency. Given a target word w, we first identify the n words whose PMI score with w is highest. For each of these n related words, we then randomly select one sentence mentioning that word.

Finally, beyond Wikipedia, we consider two external sources for obtaining sentences, because of their focus on generic knowledge.

- DEF: We extract the (primary) definition of w from the English fragment of Wiktionary[2].

- GENERIC: We first select all sentences about w in GenericsKB (Bhakthavatsalam et al., 2020) that originate from a text corpus[3]. We rank the sentences based on their confidence score in GenericsKB and select the top n.

For all strategies, if there are fewer than n sentences that can be selected, we fall back to RAND for the remaining sentences.

5. Experiments

In this section, we empirically compare the sentence selection strategies from Section 4.

Datasets We focus on the problem of predicting semantic properties of words. The reason is that in applications such as zero-shot learning or ontology completion, what matters is whether the word vectors capture particular properties. In particular, following Li et al. (2021), we consider the following four word classification benchmarks: (i) the extension of the McRae feature norms dataset (McRae et al., 2005) that was introduced by Forbes et al. (2019)[4], which focuses on commonsense properties (MCR); (ii) the CSLB Concept Property Norms[5], which also focuses on commonsense properties (CSLB); (iii) WordNet Supersenses[6], which groups nouns into broad categories such as *human* (WNSS); (iv) BabelNet domains[7] (Camacho-Collados and Navigli, 2017), which organises concepts into thematic domains such as *music* (BND). For all datasets, we train a separate binary classifier for each property and we report the (unweighted) average of the F1 scores. To classify words, we feed their word vector directly to a sigmoid classification layer. As a downstream task, we also consider the ontology completion benchmark from Li et al. (2019). In this case, word vectors are used as input features to a graph neural network, whose structure is determined by a given ontology or rule base. In particular, given a rule template such as $\star(x) \wedge \textit{LocatedIn}(x, y) \rightarrow \textit{CapitalCity}(y)$, the task is to predict which concepts can be used for the placeholder $\star$ to make the rule plausible.

Experimental Settings For word classification, we have only considered classes for which sufficient positive examples are available, i.e. at least 10 for MCR, 30 for CSLB, and 100 for WNSS and BND. For MCR, we used the standard training-validation-test split. For the other datasets, we used random splits of 60% for training, 20% for tuning and 20% for testing.
We optimize the network using AdamW with a cross-entropy loss. The batch size and learning rate were

[2] https://www.wiktionary.org

[3] GenericsKB also contains sentences that were generated from knowledge graph triples. Given the short and artificial nature of these sentences, we did not consider them.

[4] https://github.com/mbforbes/physical-commonsense

[5] https://cslb.psychol.cam.ac.uk/propnorms

[6] https://wordnet.princeton.edu/download

[7] http://lcl.uniroma1.it/babeldomains/

		McR				CSLB				WNSS				BND			
		1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
MASK	RAND	44.8	57.0	59.8	61.5	31.0	47.4	51.7	53.8	39.3	53.6	56.0	59.1	28.0	36.2	38.0	40.0
	INTRO	44.0	57.9	58.7	60.7	34.4	47.3	50.8	53.8	41.6	54.2	56.5	57.6	28.3	36.6	38.5	40.0
	HOME	55.3	59.2	61.9	60.0	42.0	50.4	53.2	54.5	45.9	55.2	58.2	58.7	28.9	35.8	37.4	39.1
	POS	42.1	51.6	54.7	56.8	29.5	44.8	47.4	50.1	38.1	51.0	54.9	56.4	28.3	35.7	38.3	39.9
	ENUM	42.9	53.9	55.5	57.1	29.9	43.3	47.8	44.6	41.0	52.5	54.8	56.2	28.3	36.1	39.3	40.0
	PMI	56.8	57.0	59.2	61.6	48.9	46.1	**54.0**	54.4	43.1	55.1	58.3	58.6	29.5	37.6	39.7	41.0
	GENERIC	46.7	52.5	55.4	57.0	33.9	45.7	47.8	50.4	36.4	51.3	55.3	57.8	26.0	34.4	36.7	38.8
	DEF+HOME	56.9	59.9	62.2	**64.1**	49.6	50.4	53.2	**56.0**	49.6	55.2	58.5	59.3	29.2	35.7	37.4	39.1
	DEF+RAND	55.6	58.2	59.2	62.6	48.8	49.8	51.8	55.5	50.3	55.2	57.1	58.6	29.3	35.7	37.9	39.4
LAST	RAND	55.5	59.0	62.3	61.6	46.1	48.3	53.9	53.5	49.4	56.5	58.0	59.0	35.4	42.9	44.7	45.7
	INTRO	53.4	58.7	61.5	59.8	43.3	48.8	50.1	51.8	50.2	**58.3**	58.0	59.1	35.8	42.8	**44.8**	45.6
	HOME	58.3	61.8	62.6	63.0	47.8	48.7	51.8	51.0	52.0	**58.3**	59.1	**59.6**	35.7	42.3	43.9	44.9
	POS	53.7	59.5	58.9	59.8	43.4	52.8	53.2	55.3	45.4	55.4	57.5	58.6	32.6	38.7	40.5	41.5
	ENUM	47.4	59.8	58.1	60.0	41.9	47.8	47.3	52.5	49.5	55.3	57.0	57.4	35.3	42.7	43.7	45.4
	PMI	55.2	60.0	61.8	63.4	43.9	**53.0**	**54.0**	54.4	49.7	57.0	59.1	**59.6**	36.6	42.2	44.6	**45.8**
	GENERIC	54.3	60.7	59.8	61.1	45.1	49.2	51.2	51.3	50.3	57.3	57.9	58.9	36.1	42.3	43.2	44.3
	DEF+HOME	57.0	60.4	61.6	63.0	**50.1**	48.4	52.5	51.7	55.2	**58.3**	**59.6**	59.4	37.2	42.4	44.0	45.1
	DEF+RAND	57.6	60.5	58.8	61.9	**50.1**	49.1	51.5	52.9	55.2	58.0	59.4	59.1	37.2	41.7	44.1	45.5
AVG	RAND	56.5	**62.7**	61.2	60.8	45.4	49.5	50.0	49.1	53.0	56.8	57.7	57.5	39.4	43.5	44.4	45.1
	INTRO	57.4	60.1	58.3	58.8	44.4	49.2	49.2	48.0	52.8	57.9	58.3	58.7	38.7	**43.7**	44.4	44.9
	HOME	**59.4**	60.1	61.1	60.9	47.8	50.3	49.1	48.4	53.7	57.5	58.1	58.5	39.3	43.2	43.6	44.2
	POS	55.2	61.6	60.2	60.2	45.8	50.5	50.6	51.8	47.5	55.0	57.6	58.0	34.8	39.7	41.0	41.4
	ENUM	54.7	59.9	57.8	60.6	43.8	48.1	48.5	48.4	52.1	55.6	56.9	56.8	39.0	42.8	44.0	44.5
	PMI	58.5	61.7	**63.2**	62.2	45.1	50.5	50.5	50.4	53.2	57.8	59.3	58.6	39.5	43.2	44.3	44.6
	GENERIC	58.7	61.0	60.1	61.5	42.3	44.6	46.1	46.1	52.6	56.7	57.6	57.7	39.1	43.1	43.4	44.0
	DEF+HOME	57.9	60.3	61.5	59.7	49.6	49.3	48.1	46.5	**57.6**	57.3	58.7	58.9	**40.4**	43.1	43.7	43.8
	DEF+RAND	58.0	60.2	61.3	60.5	49.7	49.6	50.5	51.1	**57.6**	57.0	58.1	58.1	**40.4**	43.4	44.5	44.8

Table 1: Results for word classification in terms of F1 score. Results were obtained using BERT-base. We report results for 1, 5, 10 and 20 sentences. The best results for a given benchmark and number of sentences are shown in bold. The best results within each embedding strategy (i.e. MASK, LAST, AVG) are underlined.

tuned, with possible values chosen from {4,8,16} and {0.01, 0.005, 0.001, 0.0001} respectively. For ontology completion, we follow the same evaluation methodology as Li et al. (2021), which restricts the evaluation to concept names that appear at least twice in Wikipedia. We use the same hyperparameter settings, and we apply SVD to reduce the dimensionality of the word vectors to 300, as also suggested by Li et al. (2021).

For the pre-trained language models, we used the implementations from `https://github.com/huggingface/transformers`.

Results The results of the word classification experiments are summarized in Table 1. For these results, we used BERT-base-uncased; results for other language models will be discussed below. For the hybrid strategy DEF+HOME we select one sentence using DEF and the remaining sentences using HOME, and similar for DEF+RAND. Our main findings can be summarised as follows. First, compared to MASK, we find that LAST and AVG are far less sensitive to the sentence selection strategy and the number of sentences. Second, the best results are obtained with MASK in the case of MCR and CSLB and with LAST in the case of WNSS and BND. Third, RAND is remarkably competitive, with POS, GENERIC, and ENUM underperforming RAND, while INTRO performs broadly similar. Overall the best results are obtained with PMI, HOME, DEF+HOME and DEF+RAND, all of which clearly outperform RAND. The similar performance of DEF+HOME and DEF+RAND shows that the presence of the definition plays a critical role. Moreover, note that the first sentence selected by HOME is typi-

	Wine	**Econ**	**Olym**	**Tran**	**SUMO**
RAND	16.6	17.2	13.6	8.7	35.2
HOME	18.1	17.9	14.3	9.5	37.9
PMI	16.9	17.6	13.9	8.7	38.6
DEF+HOME	20.1	18.1	16.8	10.0	39.2
BERT-500	23.0	20.0	16.9	11.5	41.4

Table 2: Results for the ontology completion experiment (F1 score). Results were obtained for 20 sentences using BERT-base with the MASK strategy. Wine, Economy, Olympics and Transport are domain-specific ontologies; SUMO is a large open-domain ontology.

	BERT-LARGE				RoBERTa-BASE				RoBERTa-LARGE			
	McR	CSLB	WNSS	BND	McR	CSLB	WNSS	BND	McR	CSLB	WNSS	BND
RAND	62.2	55.6	59.4	39.6	59.8	51.6	57.9	39.0	61.3	55.0	59.5	40.3
HOME	63.2	54.8	59.8	39.0	59.3	48.2	58.0	38.4	61.4	53.4	60.3	40.0
PMI	**65.0**	55.4	59.7	**41.3**	**63.6**	**54.0**	**58.7**	**39.8**	62.7	**56.0**	**60.1**	**44.1**
DEF+HOME	62.9	**56.8**	**59.9**	39.1	61.2	50.5	58.5	39.0	**63.1**	53.4	60.0	39.8

Table 3: Comparison of different language models for the word classification benchmarks. Results are reported in terms of F1 score. For all results, we used the MASK strategy with 20 sentences.

	McR	CSLB	WNSS	BND
7-15 words	57.8	51.5	56.8	39.1
15-25 words	59.4	**53.6**	57.9	39.5
25-35 words	**64.0**	50.1	58.1	39.4
35-45 words	60.0	53.3	57.2	**40.1**
45-55 words	60.3	51.0	**58.3**	**40.1**

Table 4: Impact of sentence length on the quality of the resulting word vectors. All results were obtained with BERT-base, using the MASK strategy with 20 sentences.

	McR	CSLB	WNSS	BND
BERT-BASE 500	60.8	51.7	58.3	42.6
BERT-LARGE 500	**62.2**	51.9	**60.2**	43.0
RoBERTa-BASE 500	61.8	49.7	58.2	40.8
RoBERTa-LARGE 500	60.3	54.0	60.0	42.5
SKIP-GRAM	59.6	**54.5**	55.6	**49.1**
CBOW	61.1	50.6	48.4	45.0

Table 5: Comparison with the MASK strategy when using 500 randomly sampled sentences, as well as with static embedding baselines.

cally a definition as well (i.e. the first sentence of the Wikipedia article). For MASK, we clearly see that DEF+HOME outperforms HOME and that DEF+RAND outperforms RAND, while for LAST and AVG the advantage of adding the definition is less obvious.

The results for ontology completion are shown in Table 2. We find that HOME, PMI and DEF+HOME outperform RAND in almost all cases, with DEF+HOME performing particularly well. We furthermore note that these results approach the values that were reported by Li et al. (2021) with 500 randomly selected sentences, which are shown as BERT-500 in Table 2.

Comparison with Other Language Models In Table 3, we present results for BERT-large-uncased, RoBERTa-base and RoBERTa-large, to complement the results for BERT-base-uncased from Table 1. In accordance with the findings for BERT-base, we can see that the PMI strategy is highly effective, consistently outperforming RAND. The HOME and DEF+HOME strategies are somewhat less effective in these cases, especially for the RoBERTa models.

In Table 5 we provide results for vectors that were obtained from 500 randomly sampled sentences using the MASK strategy, covering four language models: BERT-base-uncased, BERT-large-uncased, RoBERTa-base and RoBERTa-large. We find that the results with 20 sentences from Table 3 outperform these vectors (for MCR and CSLB) or are at least competitive with them (for WNSS and BND), thus further illustrating the effectiveness of the sentence selection strategies. Table 5 also shows results for traditional static word vectors that were trained with Word2Vec. In particular, SKIP-GRAM and CBOW vectors were trained on the same Wikipedia dump that we used for sampling sentences (enwiki-20210320). We used a window size of 5 and a minimum frequency threshold of 10. Somewhat surprisingly, perhaps, the best overall results for BD are obtained with the SKIP-GRAM vectors. This provides further evidence for the observation from Li et al. (2021) that BERT-based vectors are particularly suitable for capturing taxonomic properties, while struggling with looser forms of semantic relatedness. For the McRae dataset, CBOW achieves the best results in Table 5, but without outperforming the best configurations from Table 3. The comparatively strong performance of SKIP-GRAM and CBOW for the MCR and CSLB datasets may also be explained by the relatively small size of these datasets, which means that the higher dimensionality of the BERT-based vectors can be sub-optimal.

Impact of Word Frequency While SKIP-GRAM and CBOW were found to be surprisingly competitive in the main experiments, these traditional word embedding models struggle with words that are relatively rare. In such cases, we can expect that the improved ability of language models to model sentence context would become more important. To test this hypothesis, we grouped all the words from the test sets of WNSS and BD into 6 splits, based on their number of occurrences n_{occ} in Wikipedia: $n_{occ} \leq 20$; $20 < n_{occ} <= 50$; $50 < n_{occ} <= 100$; $100 < n_{occ} <= 300$; $300 < n_{occ} <= 500$; $n_{occ} > 500$. The F1 score for each of these cases is reported in Table 6. Note that we did not consider MCR and CSLB for this analysis, as they almost exclusively consist of frequent words. The results in Table 6 clearly show that the BERT-based vectors outperform SKIP-GRAM and CBOW for rare words. For BD, where SKIP-GRAM obtained the best overall results, we can see that the BERT-based vectors outper-

form for words occurring up to 300 times in Wikipedia. In contrast, for the case of BD, for words that occur more than 500 times, CBOW and SKIP-GRAM perform much better than RAND

Impact of Sentence Length When analyzing the disappointing performance of GENERIC, we observed that the sentences from GenericsKB tend to be very short. In Table 4 we therefore analyze the effect of sentence length. While there is no clear overall relationship between sentence length and performance, when selecting (Wikipedia) sentences consisting of 7 to 15 tokens, the performance noticeably drops. Many of the sentences from GenericsKB fall within this range, which suggests that the under-performance of GENERIC may be related to the short length of the selected sentences.

Qualitative Analysis In Tables 7 and 8, we present the top-5 sentences that were selected for the words "banana" and "falcon", for the different strategies considered in this paper. These examples illustrate some of the strengths and weaknesses of these strategies. For example, the 5 RAND sentences for "banana" largely convey information that is irrelevant for learning the meaning of this word. However, since they sometimes use "banana" in an idiosyncratic way, BERT may be able to predict that the masked word is banana, which may result in vectors that behave more like those from the LAST strategy. Meanwhile, the sentences selected using HOME and PMI are typically more informative (e.g. describing a banana's physical appearance; clarifying that banana is an edible plant; etc). The sentences selected using both POS and INTRO appear quite meaningful as well, which is at odds with the relatively poor performance of these methods. The sentences selected using GENERIC seem to be focused on overly specific properties, which may also help to explain the poor performance of this strategy.

6. Conclusions

We have considered the new challenge of distilling high-quality static word embeddings from language models using only a small number of mentions of each word. Based on our analysis, the most effective strategies are to select sentences using PMI and to include a definition of the target word. The success of these strategies makes it possible to use word embeddings obtained from LMs in applications such as ontology completion and zero shot learning with minimal computational overhead.

Acknowledgements

This work has been supported by the Engineering and Physical Sciences Research Council (EP/V025961/1) and HPC resources from GENCI-IDRIS (Grant 2022-[AD011013338]). Zied Bouraoui is supported by ANR CHAIRE IA BE4musIA and FEI INS2I 2022-EMILIE.

7. Bibliographical References

Alghanmi, I., Espinosa Anke, L., and Schockaert, S. (2020). Combining BERT with static word embeddings for categorizing social media. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 28–33, Online, November. Association for Computational Linguistics.

Bhakthavatsalam, S., Anastasiades, C., and Clark, P. (2020). GenericsKB: A knowledge base of generic statements. *CoRR*, abs/2005.00660.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bommasani, R., Davis, K., and Cardie, C. (2020). Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online, July. Association for Computational Linguistics.

Camacho-Collados, J. and Navigli, R. (2017). BabelDomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 223–228, Valencia, Spain, April. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dufter, P., Kassner, N., and Schütze, H. (2021). Static embeddings as efficient knowledge bases? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2353–2363, Online, June. Association for Computational Linguistics.

Ethayarajh, K. (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, November. Association for Computational Linguistics.

Forbes, M., Holtzman, A., and Choi, Y. (2019). Do neural language representations learn physical commonsense? In *CogSci*, pages 1753–1759. cognitivesciencesociety.org.

Gan, L., Teng, Z., Zhang, Y., Zhu, L., Wu, F., and Yang,

WordNet Supersenses						
	0-20	21-50	51-100	101-300	301-500	>500
Rand	**37.7**	**45.9**	**53.4**	**54.8**	**51.7**	**55.9**
CBOW	26.2	37.3	34.1	43.1	44.1	52.3
Skip-gram	30.3	41.8	42.3	47	51.6	54.3
BabelNet Domains						
	0-20	21-50	51-100	101-300	301-500	>500
Rand	**27.8**	**30.3**	**31.5**	**32.6**	34.2	40.1
CBOW	22.1	26.8	22.5	28.4	30.4	**50.8**
Skip-gram	19.6	28.8	31.1	31.8	**37.3**	50.2

Table 6: Breakdown of results for WNSS and BND in function of word frequency (F1 score). For relatively rare words (up to 300 occurrences in Wikipedia), the Rand strategy clearly outperforms the CBOW and Skip-gram baselines. For higher-frequency words, these baselines outperform Rand in the case of BD.

Y. (2020). SemGloVe: Semantic co-occurrences for GloVe from BERT. *CoRR*, abs/2012.15197.

Gordon, J. and Durme, B. V. (2013). Reporting bias and knowledge acquisition. In *AKBC@CIKM*, pages 25–30. ACM.

Gupta, P. and Jaggi, M. (2021). Obtaining better static word embeddings using contextual embedding models. 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 5241–5253, Online, August. Association for Computational Linguistics.

Kolyvakis, P., Kalousis, A., and Kiritsis, D. (2018). DeepAlignment: Unsupervised ontology matching with refined word vectors. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 787–798, New Orleans, Louisiana, June. Association for Computational Linguistics.

Li, N., Bouraoui, Z., and Schockaert, S. (2019). Ontology completion using graph convolutional networks. In *Proceedings of the 18th International Semantic Web Conference*, pages 435–452.

Li, A., Huang, W., Lan, X., Feng, J., Li, Z., and Wang, L. (2020). Boosting few-shot learning with adaptive margin loss. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12573–12581.

Li, N., Bouraoui, Z., Camacho-Collados, J., Anke, L. E., Gu, Q., and Schockaert, S. (2021). Modelling general properties of nouns by selectively averaging contextualised embeddings. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3850–3856.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Liu, Z., Wang, H., Niu, Z.-Y., Wu, H., Che, W., and Liu, T. (2020). Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online, July. Association for Computational Linguistics.

Ma, Y., Cambria, E., and Gao, S. (2016). Label embedding for zero-shot fine-grained named entity typing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 171–180, Osaka, Japan, December. The COLING 2016 Organizing Committee.

McRae et al., K. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37:547–559.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June. Association for Computational Linguistics.

Navigli, R. and Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1318–1327.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? 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 2463–2473, Hong Kong, China, November. Association for Computa-

	banana
RAND	• Born in Puntarenas Province , Lagos ' parent decided to move to Limón where Cristhian went to school and worked in banana plantation • Binding post or banana plug may be used for lower frequency • In India , vegetarian variety may use potato , calabash , paneer , or banana •A later claim suggested that Bubbles had died ; Jackson's press agent Lee Solters quipped to the medium that when Bubbles heard about his demise he went banana ... Like Mark Twain , his death is grossly exaggerated and he 's alive and doing well • At the Royal Variety Performance in 1981 , it was performed in the customary male evening dress by Anita Harris , who brought the house down with the line "I've just had a banana with Lady Diana" in the Buckingham Palace verse of the song
HOME	•A banana is an elongated, edible fruit – botanically a berry – produced by several kinds of large herbaceous flowering plants in the genus "Musa" • In some countries, bananas used for cooking may be called "plantains", distinguishing them from dessert bananas • Almost all modern edible seedless (parthenocarp) bananas come from two wild species – "Musa acuminata" and "Musa balbisiana" • The scientific names of most cultivated bananas are "Musa acuminata", "Musa balbisiana", and "Musa" × "paradisiaca" for the hybrid "Musa acuminata" × "M. balbisiana", depending on their genomic constitution. • They are grown in 135 countries, primarily for their fruit, and to a lesser extent to make fiber, banana wine, and banana beer and as ornamental plants
INTRO	• The area produces citrus, olives, tomatoes and market-garden vegetables, and is one of the few parts of Europe where commercial banana production is possible. • The work, created in an edition of three, consists of a fresh banana taped to a wall with a piece of duct tape • They also sell orange, grape, piña colada, coconut champagne (non-alcoholic), and banana daiquiri (non-alcoholic) fruit drinks • No banana plantation was left unscathed by the hours-long onslaught of strong winds • The crops of highest productivity are plantain, banana, coconut, tomatoes, pepper, eggplant, yucca, rice, beans, maize, "guandules" and sweet potato
PMI	• The common fruits that are used in the preparation include banana, apple, kiwi, strawberry, papaya, pineapple, mango, and soursop • Thus the banana producer and distributor Chiquita produces publicity material for the American market which says that "a plantain is not a banana" • One day Mitchell posted a photo of herself on Twitter next to a bruised banana in response to trolls who had compared her freckles to the overripe fruit • The most important Philippine cooking banana is the saba banana (as well as the very similar cardava banana) • Their meals consist of cooked or steamed rice wrapped in banana or tara or kau leaves that known as "khau how" and boiled vegetables
POS	• Bananas, grown mainly for domestic consumption, amount to a steady annual average crop of 70,000 tons. • Bananas were introduced into the americas in the 16th century by portuguese sailors who came across the fruits in west africa, while engaged in commercial ventures and the slave trade" • Bananas must be transported over long distances from the tropics to world markets • Bananas was edited at the time by the now-legendary horror author r. l. stine • Bananas which are turning yellow emit natural ethylene which is characterized by the emission of sweet scented esters
ENUM	• Crops are, for example, cereals (mainly wheat, barley, rye and triticale), soybeans, banana, rice, coffee, turnips, and red as well as sugar beets • These have included: bacon maple ale and chocolate, peanut butter, and banana ale • There are also wild relatives of jackfruit, mango, cardamom, turmeric and banana • Amelita's signature dish was an organic rib fillet with shaved ham, banana, and hollandaise sauce. • Whereas the larger farming plots are utilized for staple crops, families can choose to grow herbs, flowers and fruit trees (mango, banana, plum, orange, lime) in their personal household garden
GENERIC	• Bananas contain more digestible carbohydrates than any other fruit • Bananas have no fat, cholesterol or sodium • Bananas do contain serotonin • Bananas grow on plants • Bananas contain pectin, a soluble fibre
DEF	• Banana is an elongated curved tropical fruit that grows in bunches and has a creamy flesh and a smooth skin

Table 7: Example sentences selected for the word *banana*.

	falcon
RAND	• This festival focus on Asayel (local camel) and Majahim (dark skinned camel) , and also feature falcon hunting , Saluki and Arabian horse race , and date packing contest • In a land where mice eat iron, falcons also kidnap children • While in migration , adult are sometimes preyed on by most of the bird-hunting , larger raptor , especially the peregrine falcon •Scottish Wildlife Trust , a charitable organisation , manages the Falls of Clyde site , focusing on the preservation of the endangered or protected wildlife in the ground , such a peregrine falcon , roe deer and badger • The falcon "Ida" come to Pkharmat every morning
HOME	• Adult falcons have thin, tapered wings, which enable them to fly at high speed and change direction rapidly • Fledgling falcons, in their first year of flying, have longer flight feathers, which make their configuration more like that of a general-purpose bird such as a broad-wing. • The falcons are the largest genus in the Falconinae subfamily of Falconidae, which itself also includes another subfamily comprising caracaras and a few other specie • The largest falcon is the gyrfalcon at up to 65 cm in length • As with hawks and owls, falcons exhibit sexual dimorphism, with the females typically larger than the males, thus allowing a wider range of prey species
INTRO	• Peregrine falcons, common kestrels and choughs also nest on the cliffs • Many other versions of this song with motif of falcon drinking water from Vardar were published at the beginning of the 20th century in Macedonia (i.e.) • Common birds are: fantails, kingfishers, tui, kereru, New Zealand falcons • It consists of a golden falcon (Hawk of Quraish) with a disk in the middle, which shows the UAE flag and seven stars representing the seven Emirates of the federation • The school mascot is the falcon and the school colors are scarlet and grey
PMI	• The saker falcon is a large hierofalcon, larger than the lanner falcon and almost as large as gyrfalcon at length with a wingspan of • Because he is so often shown with a falcon, he came to be considered the patron saint of falconry • The island has breeding populations of various raptors: golden eagle, buzzard, peregrine falcon, kestrel, hen harrier and short and long-eared owl • The arrangement is intriguing, because normally the Horus falcon and the hieroglyphs inside the serekh were out of reach and independent of one another • Other birds which can be seen include peregrine falcon, merlin, hen harrier, short-eared owl and ring ouzel
POS	• Falcons of narabedla is a science fiction novel by american writer marion zimmer bradley set in the universe of her darkover series • Falcons rookie lt sam baker was hit in the head in the first half and did not return • Falcons were important in the (formerly often royal) sport of falconry • Falcons defensive end chuck smith questioned the vikings' toughness because of the ease with which they had won during the season • Falcons and cormorants have long been used for hunting and fishing, respectively
ENUM	• The axe did, however, close some country routes including the cuckoo line, the cranleigh line, the steyning line, the new romney branch line and the bexhill west branch line, plus goods yards including deptford wharf and falcon lane • The ford falcon and holden commodore, former chrysler engineers now working for mmal, developed a wider mid-sized car specific to the australian market. • The series features two founding members of the team, ant-man and the wasp, and introduces wonder man, tigra, hawkeye, falcon, vision and scarlet witch • The word perlin is a falconer's term for a cross breed of a peregrine falcon and a merlin • The falcon and the snowman received generally positive notices upon release in 1985 and currently has an 82 percent on rotten tomatoes from 22 critics
GENERIC	• Falcons have long, slim wings which taper to pointed tips • Some falcons eat reptiles • Falcons are small, speedy birds of prey known for their aerial agility • Falcons are birds of prey • Some falcons eat small reptiles
DEF	• Falcon is any bird of the genus Falco, all of which are birds of prey

Table 8: Example sentences selected for the word *falcon*.

tional Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Shwartz, V. and Choi, Y. (2020). Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. Y. (2013). Zero-shot learning through cross-modal transfer. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 935–943.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, 6(3):203–217.

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., and Korhonen, A. (2020). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online, November. Association for Computational Linguistics.

Wang, Y., Cui, L., and Zhang, Y. (2021). Improving skip-gram embeddings using BERT. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:1318–1328.

Weir, N., Poliak, A., and Durme, B. V. (2020). Probing neural language models for human tacit assumptions. In *CogSci*. cognitivesciencesociety.org.

Xing, C., Rostamzadeh, N., Oreshkin, B. N., and Pinheiro, P. O. (2019). Adaptive cross-modal few-shot learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 4848–4858.

Yan, K., Bouraoui, Z., Wang, P., Jameel, S., and Schockaert, S. (2021). Aligning visual prototypes with BERT embeddings for few-shot learning. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 367–375.