Feasibility of BERT Embeddings For Domain-Specific Knowledge Mining

Anonymous ACL submission

Abstract

Extracting information from large corpora of unstructured text using computational methods presents a challenge. Tshitoyan et al. (2019) demonstrated that unsupervised mathematical word-embeddings produced by a static language model could be utilized to uncover 'latent knowledge' within a materials science corpus. The rise of contextualized and massively pre-trained language models like BERT have seen static models becoming surpassed for most NLP tasks. Nevertheless, due to innate architectural and use differences, BERT requires adaptation for knowledge mining. This study tests the suitability of BERT-derived word embeddings for knowledge mining purposes. It utilizes a variation of the approach described by Bommasani et al. (2020) for creating static-equivalent vectors from multiple contextualized word representations. It is conducted using a biomedical corpus, a biomedical BERT variation and validated using domain-specific intrinsic benchmarking tools. Novel, layer-wise BERT performance characteristics are demonstrated. A key finding is that layer-wise intrinsic performance differs for nouns and verbs. Performance also varies according to whether a word of interest belongs to BERT's native vocabulary or is built from sub-word representations: BERT-native representations perform best when extracted from earlier layers, while representations requiring multiple tokens perform best when extracted from the middle-to-latter model layers.

1 Introduction

A vast amount of biomedical knowledge exists as unstructured text within journals, books and abstracts among other formats. This knowledge exists as relationships and connections between described concepts, objects and events. Information extraction from such corpora using supervised methods requires large, manually-labelled datasets. Consequently, these methods do not readily scale.

Recently, Tshitoyan et al. (2019) demonstrated that known and novel relationships between entities described within a materials science corpus could be discovered using unsupervised, high-dimensional word embeddings (Bengio et al., 2003; Collobert and Weston, 2008; Collobert et al., 2011). Here, the authors trained a skip-gram variant of the Word2Vec neural language model (Mikolov et al., 2013) on a corpus of 3.3 million materials science abstracts to produce 200-dimensional embeddings for each word in the corpus vocabulary. Remarkably, when the embeddings representing material names (e.g. 'Bi_2Te_3') were ranked by their cosine similarity to the representation of 'thermoelectric,' several novel thermoelectric conductors were identified. Despite the material name never having appeared alongside, or within the same document as 'thermoelectric,' the direct relationship between the novel material's word representation and 'thermoelectric' was permitted due to indirect relationships between the material's name and other words/phrases such as 'chalcogenide' (chalcogenides are good thermoelectrics) and 'band gap' (which determines thermoelectric properties) within the vector space (Tshitoyan et al., 2019). Venkatakrishnan et al. (2020) subsequently applied the same Word2Vec skip-gram technique to an unstructured text corpus of over 100 million biomedical documents, discovering novel tissue-reservoirs of the ACE2 receptor used by SARS-CoV-2 to infect a host organism.

Both Tshitoyan et al. (2019) and Venkatakrishnan et al. (2020) postulated that context-aware embeddings, such as those from the bidirectional encoder representation from transformers (BERT) model (Devlin et al., 2018) could outperform static models at these tasks. Aside from fundamentally different architecture, BERT produces 'just-in-time' contextualized embeddings from pre-tokenized sequences fed into the model individually. Moreover, unlike static models like Word2Vec

and GloVe (Pennington et al., 2014) which build corpus-specific vocabularies, BERT possesses an innate vocabulary of approximately 30,000 words and handles extra-vocabulary words by decomposition into constituent sub-words. As such, a method of leveraging BERT's unique architecture and training on massive text corpora to ultimately yield word representations capable of use in knowledge mining is lacking. Bommasani et al. (2020) described a method for reducing contextualized word representations to static-equivalents by aggregating contextualized word representations from BERT over a number of contexts: These aggregated representations outperformed static ones in general domain intrinsic benchmarking tasks.

Much like static word representations, BERT-derived equivalents can subsequently be adapted for knowledge discovery by ranking geometric similarity between represented concepts, objects or processes. Nevertheless, as 'latent knowledge' requires physical validation, the quality of novel language model suggestions cannot easily be assessed. Domain-specific intrinsic benchmarks which assess semantic similarity and relatedness between word representations using geometric measures (Chiu et al., 2018) may be utilized as an appropriate surrogate: Higher-fidelity mathematical representations of described reality are expected to approximate human user similarity ratings between concepts and objects. This study subsequently tests the hypothesis of both Tshitoyan et al. (2019) and Venkatakrishnan et al. (2020) that contextual language models yield word representations for knowledge mining that are superior to those produced by static model in a biomedical domain and therefore suitable for knowledge mining. Using a corpus of 500,000 abstracts and full-text articles (Wang et al., 2020), embeddings produced by a series of static models are tested against aggregated contextual representations sampled from the corpus and passed through a biomedically-trained BERT variant, and assessed using domain-specific intrinsic benchmarks.

2 Methods

2.1 Dataset and Text Preprocessing

In response to the COVID-19 pandemic, the Coronavirus Open Research Dataset (CORD-19) was released by governmental and academic institutions. It consists of over 500,000 scholarly articles (with over 200,000 full text articles and preprints) and abstracts pertaining to COVID-19 (Wang et al., 2020)[1]. Corpus metadata was removed and articles aggregated into a single file. All numbers were replaced with a special token and selective lowercasing was performed to preserve abbreviations. For the Word2Vec and GloVe models, common terms and punctuation were removed.

The BERT approach was informed by results of an initial pilot study (see Appendix for preliminary data). Two approaches were adopted, involving extracting n long or short contextual sentence samples from the corpus: Long sequences were created by splitting on periods into constituent sentences. Short context sequences were created by further splitting on commas into constituent phrases. When selecting examples containing each word, the maximum sequence length for both long and short sequences was limited to 512, the maximum sequence length allowed by BERT, assuming each word in the sequence is represented by a single token.

2.2 BERT Approach

BioBERT is a variation of the original BERT model which has been further pre-trained on PubMed abstracts and PubMed Central full-text articles. It outperforms general models at various biomedical NLP tasks (Lee et al., 2020). The open source HuggingFace (Wolf et al., 2020)[2] implementation of BioBERT v1.1 was utilized without any further pre-training or fine-tuning based upon results of the preliminary study (B). Depending on approach, long and short sequences containing words of interest (from the benchmarking vocabulary) were selected. In order to ensure consistency between static and contextual test vocabularies, as BERT is able to use subword pooling for words outside its native vocabulary, only words that were in both the benchmarking vocabulary (see 2.4) and the CORD-19 vocabulary were selected for sampling.

For both long and short sequence approaches, n = 500, 1000 or 5000 samples were extracted from the pre-processed corpus for tokenization. Sequences were selected only if they contained a single instance of the word of interest and were discarded if their pre- or post-tokenized length exceeded 512. Here, for each word w in context c, BERT's tokenizer will either yield a single token or decompose w into k sub-word tokens, where

[1] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[2] https://huggingface.co/

$\{\mathbf{w}_c^1, ..., \mathbf{w}_c^k\} \longmapsto \mathbf{w}_c$. Tokenized sequences were then fed into the model and the sequence representations were extracted from all 13 model layers. For words represented by a single 1x768 representation, this was extracted without further operations. For decomposed words, the arithmetic mean of all $\mathbf{w}_c^k$ was taken to yield a single 1x768 representation from k sub-word representations, per context:

$$\mathbf{w}_c = \text{mean}(\mathbf{w}_c^1, ..., \mathbf{w}_c^k)$$

The arithmetic mean of the n contextual examples of each word w, $\mathbf{w}_{c1}, ..., \mathbf{w}_{cn}$ was then taken. If n examples meeting the inclusion criteria were not available, then the maximum number were taken:

$$\mathbf{w} = \begin{cases} \text{mean}(\mathbf{w}_{c1}, ..., \mathbf{w}_{cn}) & n = 500, 1000, 5000 \\ \text{mean}(\mathbf{w}_{c1}, ..., \mathbf{w}_{c\max(n)}) & n < 500, 1000, 5000 \end{cases}$$

Decision to take arithmetic mean of both sub-word representations and take either n was based on the results of Bommasani et al. (2020), where they found this operation outperformed other possible operations (e.g. max., min., last) for both sub-word pooling and contextual aggregation (see also (Ács et al., 2021)). The present approach also differed from Bommasani et al. (2020) who instead took the representation produced by feeding the word in isolation into the model[3].

2.3 Static Models

The aggregated embeddings obtained from 2.2 were compared against several static baseline models including 200 and 300-dimensional Word2Vec skip-gram models, and a 300-dimensional GloVe model all trained from scratch on only CORD-19, using default hyperparameters. Additionally, pre-trained 200-dimensional embeddings from BioWordVec (Zhang et al., 2019)[4] were also obtained and used for benchmarking. Briefly, BioWordVec is an open set of static biomedical word vectors trained on a corpus of over 27 million articles, that additionally combine sub-word information from unlabelled biomedical text together with a biomedical controlled vocabulary.

[3] A single word (rather than a sequence) is an 'unnatural' input for BERT, yielding a poorly-performing 'decontextualized' word representation (see (Bommasani et al., 2020) for more detail).

[4] https://github.com/ncbi-nlp/BioWordVec

2.4 Benchmarking

Bio-SimVerb and Bio-SimLex (Chiu et al., 2018) are benchmarking resources for the biomedical domain that offer 988 and 1000 test verb and noun pairs, respectively. These word-pairs have been extracted from 14 open biomedical ontologies and over 14,000 biomedical journals covering over 120 areas of biomedicine. Additionally, some of the test word pairs are from the general domain. These resources address shortcomings of previous biomedical benchmarks such as MayoSRS (Pakhomov et al., 2011) and UMNSRS (Pakhomov et al., 2010) which only test nouns, and fail to distinguish between semantic relatedness and similarity (Chiu et al., 2018). The CORD-19 vocabulary covered 97% of BioSimVerb and 94.43% of BioSimLex test pairs, respectively.

3 Results

3.1 Verb Benchmarks

The left sub-plot of Figure 3 and left column of Table 1 demonstrates the layer-wise performance of n = 500, 1000 and 5000 aggregated contextualized verb representations across all BERT layers. Performance is generally preserved regardless of sequence lengths/number of aggregated contexts. Embeddings extracted and distilled from the 7th and 8th layers performed best for all combinations. Short contexts marginally outperform longer contexts at most layers. The best performing representations for all combinations were extracted from layer 8 and distilled from 1000 contexts, though these representations did not substantially outperform those distilled from other n. In general, representations extracted from the latter 6 layers (with the exception of layer 11) outperform the best-performing static embeddings at verb benchmarking.

3.2 Noun Benchmarks

The right sub-plot of 3 and right column of Table 1 demonstrates layer-wise performance of n = 500, 1000 and 5000 aggregated contextualized noun representations extracted from all BERT layers. Unlike Bio-SimVerb, static models (with the exception of GloVe) outperformed aggregated noun representations from all layers and for all n. The plotted line demonstrates different morphology compared to verb benchmarks: Here, representations distilled from the first 8 BERT layers outperformed those from the latter layers, increasing

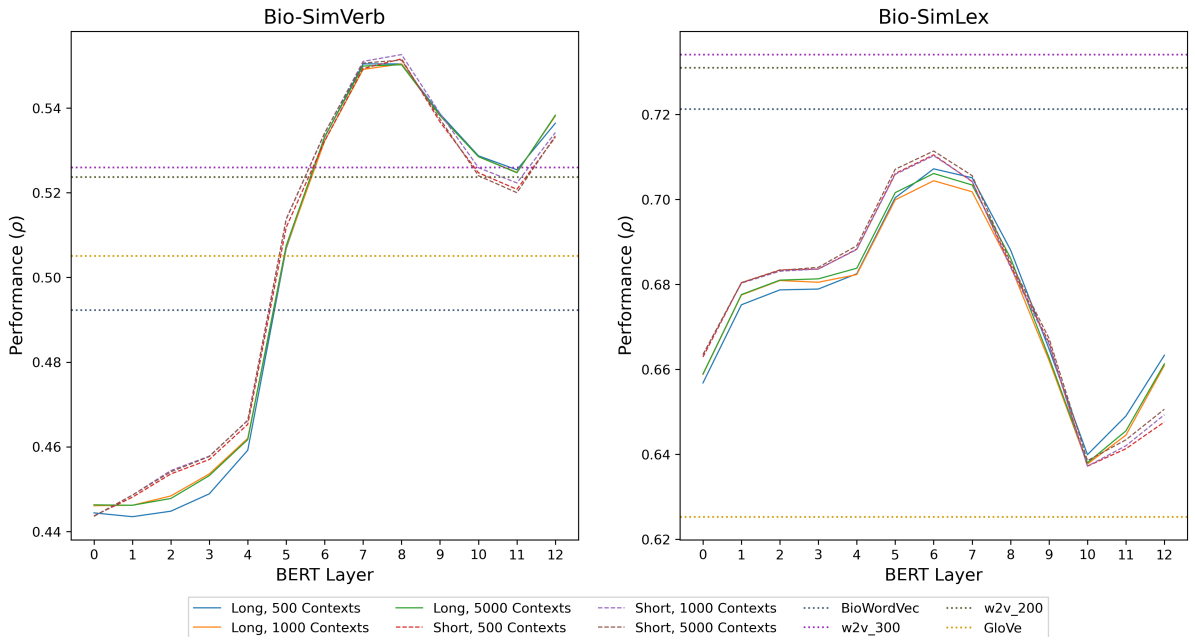

Figure 1: Layer-wise performance of BERT embeddings (0 corresponds to input layer) at both Bio-SimVerb and Bio-SimLex benchmarks. Horizontal dashed lines correspond to performance of static models.

until and then peaking at layer 6 before declining thereafter. Unlike verbal performance, however, embeddings extracted and distilled from shorter sequences demonstrated discernible improvement relative to those from long sequences. The best performance achieved was from $n = 1000$ embeddings, however overall, n made little difference.

Method	**Bio-SimVerb**	**Bio-SimLex**
Long 500	0.5504 (7=8)	0.7072 (6)
Long 1000	0.5504 (8)	0.7044 (6)
Long 5000	0.5502 (8)	0.7061 (6)
Short 500	0.5516 (8)	0.7105 (6)
Short 1000	**0.5526 (8)**	0.7103 (6)
Short 5000	0.5513 (8)	0.7114 (6)
w2v 300	0.5260	**0.7341**
w2v 200	0.5237	0.7310
GloVe 300	0.5051	0.6253
BWV 200	0.4923	0.7213

Table 1: Top performing (Spearman's ρ) distilled BERT embeddings and static embeddings. 'Long/Short n' indicates embeddings distilled from n example sequences of each word. Number in brackets indicates layer. w2v200/300 = Word2Vec 200/300 dimensional embeddings. BWV = BioWordVec 200 dimensional embeddings. Bold entries indicate best overall performance.

3.3 Effect of Sub-Word Pooling

In an attempt to explain observed performance and explore the effects of sub-word pooling, test word-pairs from Bio-SimVerb and Bio-SimLex were separated into two groups using a criteria of whether both words in a respective test pair existed in BioBERT's native vocabulary or not. This yielded test word pairs where both were BERT-native and had a single representation, or where at least one of the words in the pair required a sub-word pooling operation before the n contextual representations could be averaged. Representations from all BERT layers using short contexts and the top performing static model representations (300-dimensional Word2Vec) were then subjected to Spearman's rank testing as per the Bio-Simverb methodology (Chiu et al., 2018), albeit using the modified, separated test pair rankings.

For both verbs and nouns, general and layer-wise native word performance differed to words requiring sub-word pooling prior to context aggregation. Moreover, it is apparent that when test-vocabulary is stratified in this regard, n has little to no bearing upon overall performance of the BERT embeddings. Verbs native to BERT's vocabulary generally outperform those from the top-performing static model at all layers (see left side of Figure 2 and Table 2). Performance declines steadily from layer 0-12. Performance of word representation for

verbs requiring sub-word pooling is overall lower than single-token representations, demonstrating performance increase from layers 0-8. It is only at layers 7-8 where performance slightly exceeds performance of static representations.

Only BERT-native noun representations extracted from the first 4 layers demonstrated superior performance to the corresponding static embeddings. Similar to the BERT-native verbs, performance decreased from layer 0-12, reaching a trough at layer 10, before slightly increasing thereafter. Moreover, sub-word pooled noun representation performance increased substantially from layers 0-6 before declining thereafter. These representations never outperformed the static embeddings (see Figure 2 and Table 2).

Method	Bio-SimVerb	Bio-SimLex
Short 500 (S)	**0.6691 (1)**	0.7255 (1)
Short 500 (M)	0.4603 (8)	0.7417 (6)
Short 1000 (S)	0.6685 (1)	0.7255 (1)
Short 1000 (M)	0.4629 (8)	0.7420 (6)
Short 5000 (S)	0.6688 (1)	0.7256 (1)
Short 5000 (M)	0.4621 (8)	0.7418 (6)
Static (S)	0.5255	0.6959
Static (M)	0.4545	**0.7628**

Table 2: Performance of BERT embeddings aggregated from short contextual examples and with $n = 500, 1000, 5000$. **S** or **M** in brackets indicate whether representations were for words native to BERT i.e. using a single token to represent or those requiring sub-word pooling, respectively. Static representations were from a 300-dimensional Word2Vec model. Bold entries indicate best overall performance.

4 Discussion

This study demonstrates the feasibility of using BERT-derived word representations for knowledge mining purposes, however their benefit over static representations as used by Tshitoyan et al. (2019) and Venkatakrishnan et al. (2020) is less clear. Overall, this study demonstrates that relatively few, short-sequence contextual word-examples extracted from a corpus can be aggregated and utilized to yield embeddings that can outperform in the case of verbs, or approximate (in the case of nouns) those from the best performing static models trained on entire corpora. Practically this sampling-based approach may offer time and cost savings over training entire static models from scratch, when dealing with large corpora. Moreover, due to the superior performance of the BERT representations for verbs, aggregated contextualized embeddings may even be preferable when mining verb-rich text (e.g. clinical notes).

BERT's pre-training on massive text corpora may be responsible for performance characteristics observed: Evidence is provided by the differing performance of representations innate to BERT's vocabulary compared to representations built from multiple sub-words. It appears that n is less important than whether the word belongs to BERT's innate vocabulary or not: For both nouns and verbs, if the word of interest is BERT-native then it is preferable to utilize representations from the earlier layers and these are superior to static ones, while if a word requires sub-word decomposition then layers 6-8 seem to be optimal. Nevertheless, more work is required to quantify the effect of multiple subwords on performance, as the split vocabulary in this study utilized a relatively imprecise criteria of $k > 1$ for test-pairs where at least one word was non-native to BERT. Moreover, though Bommasani et al. (2020) demonstrated that taking the arithmetic mean of k sub-words was the best performing method on their general-domain intrinsic benchmarking, a later study by Ács et al. (2021) showed that sub-word pooling approach mattered depending on desired downstream NLP tasks. Consequently, further exploration into both k and n parameters should be conducted.

Based on the observed performance, it might also benefit to expand BERT's native vocabulary with domain-specific words prior to conducting pre-training. Though the pilot study showed that further-pretraining was detrimental to performance (see Appendix B), this method utilized BERT's native vocabulary and was only tested using long sequences (which themselves underperform relative to short sequences) before being abandoned. Another consideration is that further pre-training steps are necessary to improve sub-word performance (Liu et al., 2019), which could be important for non-general domains. Moreover, as the benchmarking vocabularies incorporate both general-domain and biomedical-domain word pairs (Chiu et al., 2018), it may also be that the general domain test pairs are contributing disproportionately to performance boosts. Another area for exploration is the comparatively different layer-wise performance for BERT-native words versus extra-vocabulary words, with the former's performance generally decreasing

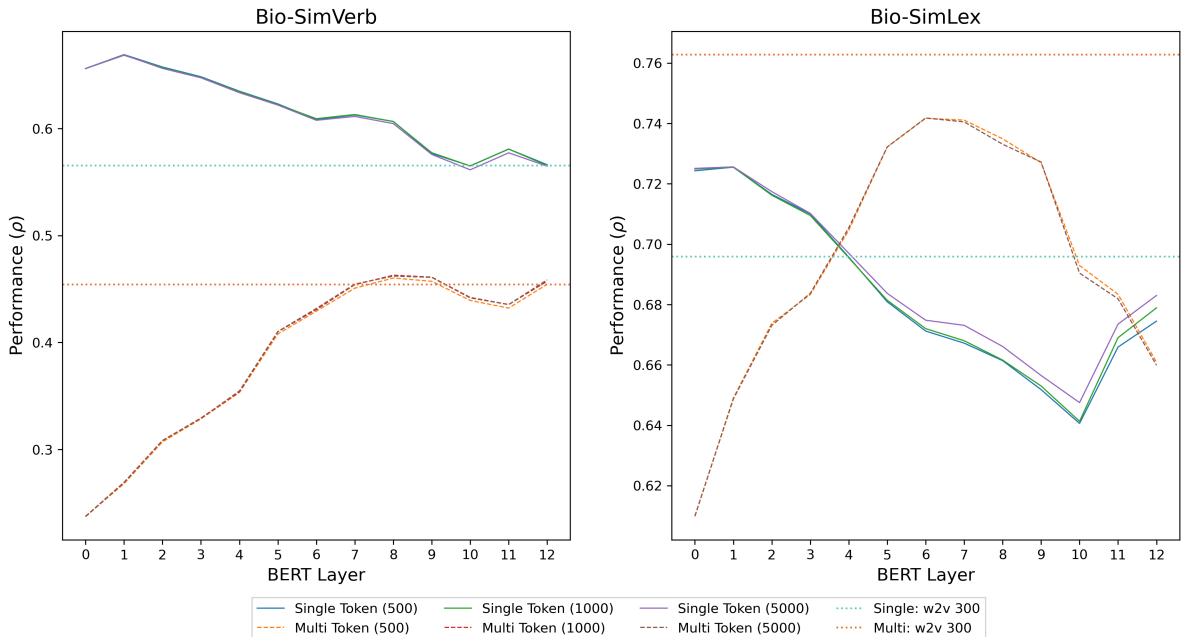

Figure 2: Layer-wise performance of BERT embeddings (0 corresponds to input layer) at both Bio-SimVerb and Bio-SimLex benchmarks. Horizontal dashed lines correspond to performance of static models.

and the latter's increasing.

A working knowledge-mining framework utilizing BERT might consist of first extracting the vocabulary of the corpus upon which mining will be conducted and removing any irrelevant words (e.g. stop words). Then, n samples for each word in the vocabulary may be taken from the corpus and tokenized. As BERT's attention is quadratic to the sequence (Devlin et al., 2018), and representations extracted from short sequences perform better, shorter sample sequences are desirable. Tokenized sequences can then be encoded, and representations extracted, with sub-word pooling performed if necessary. The n contextual examples of each word representation can then be averaged to yield a 1x768 dimensional representation for each word in the corpus vocabulary. It is this collection of vocabulary embeddings that can be subsequently used for mining as per Tshitoyan et al. (2019).

5 Conclusions

This study has successfully demonstrated feasibility of aggregated contextual word representations derived from BERT for biomedical knowledge mining tasks. It has also uncovered several technical and performance-related idiosyncrasies of BERT and BioBERT that require further investigation.

6 Acknowledgements

Thanks to **redacted** and **redacted** for their assistance with this study.

References

Judit Ács, Ákos Kádár, and Andras Kornai. 2021. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, pages 2284–2295. https://doi.org/10.18653/v1/2021.eacl-main.194.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3(Feb):1137–1155.

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 4758–4781.

Billy Chiu, Sampo Pyysalo, Ivan Vulić, and Anna Korhonen. 2018. Bio-simverb and bio-simlex: wide-coverage evaluation sets of word similarity in biomedicine. *BMC bioinformatics* 19(1):1–13.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. pages 160–167.

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(ARTICLE):2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. In *AMIA annual symposium proceedings*. American Medical Informatics Association, volume 2010, page 572.

Serguei VS Pakhomov, Ted Pedersen, Bridget McInnes, Genevieve B Melton, Alexander Ruggieri, and Christopher G Chute. 2011. Towards a framework for developing semantic relatedness reference standards. *Journal of biomedical informatics* 44(2):251–265.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 571(7763):95–98.

AJ Venkatakrishnan, Arjun Puranik, Akash Anand, David Zemmour, Xiang Yao, Xiaoying Wu, Ramakrishna Chilaka, Dariusz K Murakowski, Kristopher Standish, Bharathwaj Raghunathan, et al. 2020. Knowledge synthesis of 100 million biomedical documents augments the deep expression profiling of coronavirus receptors. *Elife* 9:e58040.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv* .

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, pages 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data* 6(1):1–9.

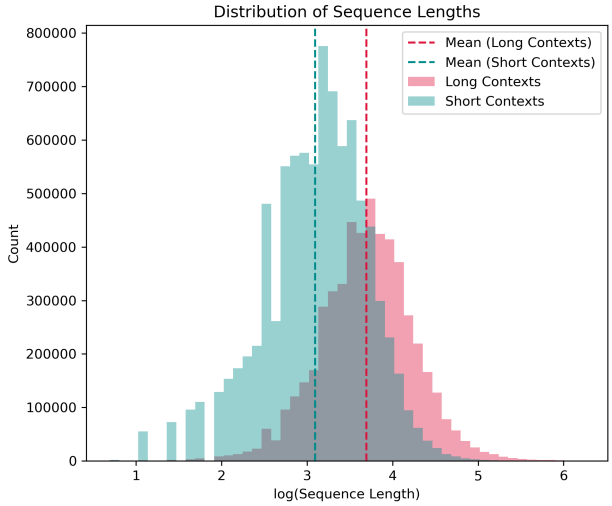

Figure 3: Distributions of log sentence lengths for long and short contextual sequences.

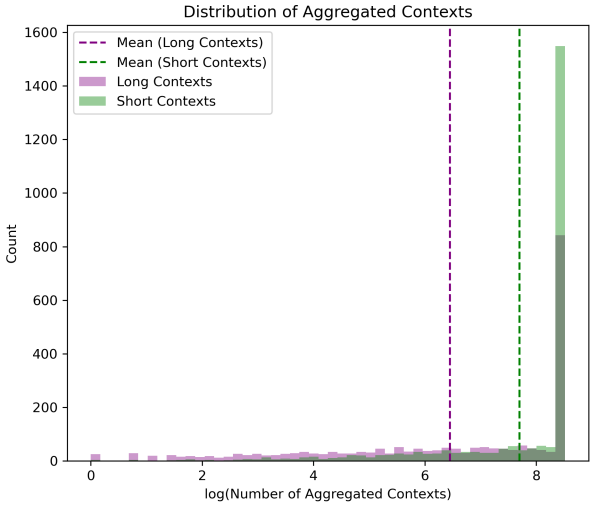

Figure 4: Distributions of log number of aggregated contexts for long and short sequence lengths. There are substantially more examples meeting n = 5000 for short sentences

A Corpus Sampling Characteristics

Figure 2 demonstrates the distribution of log sequence lengths for long and short sequences, respectively. Figure 3 demonstrates the distribution of log number of sequences for long and short sequences, respectively. As one of the sampling criteria was that sequences could only have a single instance of the word and had to be less than 512 words in length, together with the fact that the tokenized sequence could not exceed 512 in length, there are consequently fewer long sequence samples per word compared to short sequence examples. The mean long sequence length was 46.8 words (σ = 29.5) while the mean short sequence length was 26.3 words (σ = 16.7) (2). For long sequences, the mean number of contextual examples per word was 2330.2 (σ = 2194.8). For short sequences, the mean number of contextual examples per word was 3632.2 (σ = 1948.7) (Figure 3).

B Effect of Further Pre-Training on Word Representation Quality

An initial approach attempted was to further pre-train BioBERT using the entire CORD-19 corpus and compare performance of the contextualized word representations at Bio-SimVerb and Bio-SimLex intrinsic benchmarking tasks, with the base BioBERT model and the static baseline models. This approach used only long corpus sequences and the base BioBERT vocabulary (which itself is identical to BERT vocabulary). Pre-training was achieved using the scripts supplied with the TensorFlow implementation of the model (`https://github.com/dmis-lab/biobert`) and involved first creating pre-training data using sentence examples from the corpus, before running further pre-training for 100,000 epochs. Default hyperparameters were used. For this pilot study, n = 10, 50, 100, 500 and 1000. The n selected examples were then all tokenized and passed through either the further-pretrained BioBERT model or the base Bio-BERT model. For either approach, representations corresponding to the word of interest were then extracted wholly (i.e. as a single 1x768 word representation, or k individual sub-word representations) and added to the list of n (explained further in 2.2). Benchmarking was performed as described in 2.4.

For the Bio-SimVerb benchmarks (Left side of Figure 4), there is a clear increase in performance by increasing n from 10 to 1000 contexts. Also apparent is that the representations extracted from the further pre-trained model underperform relative to those extracted from the base model for the same n. Biggest increases in performance are seen going from $n = 10$ to $n = 100$. Increasing n beyond this begins to demonstrate smaller performance boosts. Interestingly, best performing verb embeddings from the further pre-trained model were taken from layer 12 (see 3) while for the base model, performance peaked at embeddings extracted from

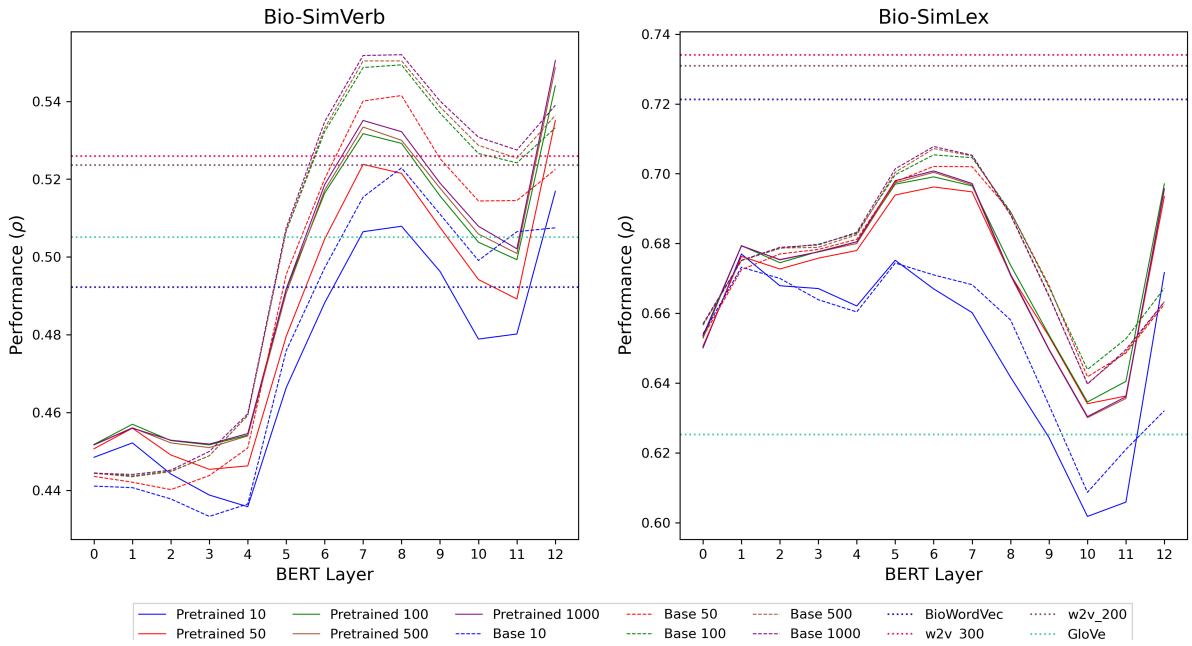

Figure 5: Layer-wise performance of BERT embeddings (0 corresponds to input layer) at both Bio-SimVerb and Bio-SimLex benchmarks. Pretrained n/Base n refer to either the further-pretrained model or the base model, respectively, followed by the n aggregated contexts. Horizontal dashed lines correspond to performance of static models.

layer 8. In some cases, embeddings taken from layer 12 of the further pre-trained model almost reached peak performance from embeddings taken from layer 8 of the base model.

For the Bio-SimLex benchmarks (Right side of Figure 4), though there was a general performance increase between representations extracted from the further-pretrained model and the base BioBERT model, it was less pronounced as it was for the verb benchmarks, with performance for the first 6 layers approximately equal before diverging thereafter. Moreover, a substantial boost is seen going from $n = 10$ to $n = 50$, becoming less pronounced as n increases. Again, performance for the representations extracted from a further pre-trained model demonstrate a trough following their maximum performance at layer 8, but increase substantially thereafter going from layer 11 to 12, though without reaching their layer 6 peak. This characteristic was not observed with the base model representations. Finally, representations from either the further-pretrained or base models did not outperform either Word2Vec 200 or 300 dimensional representations, or the BioWordVec representations.

Method	**Bio-SimVerb**	**Bio-SimLex**
Pre-Trained 10	0.5169 (12)	0.6770 (1)
Pre-Trained 50	0.5351 (12)	0.6991 (6)
Pre-Trained 500	0.5440 (12)	0.7004 (6)
Pre-Trained 1000	0.5487 (12)	0.7008 (6)
Base 10	0.5229 (8)	0.6744 (5)
Base 50	0.5415 (8)	0.7054 (6)
Base 500	0.5494 (8)	0.7072 (6)
Base 1000	**0.5504** (8)	0.7078 (6)
w2v 300	0.5260	**0.7341**
w2v 200	0.5237	0.7310
GloVe 300	0.5051	0.6253
BWV 200	0.4923	0.7213

Table 3: Top performing (Spearman's ρ) distilled BERT embeddings and static embeddings from pilot study. 'Pre-Trained/Base n' indicates embeddings extracted from n examples taken from the distilled pre-trained or base model, respectively. Number in brackets indicates layer. w2v200/300 = Word2Vec 200/300 dimensional embeddings. BWV = BioWordVec 200 dimensional embeddings. Bold entries indicate best overall performance.