Enhanced Biomedical Knowledge Discovery From Unstructured Text Using Contextual Embeddings

Anonymous ACL submission

Abstract

Extracting knowledge from large, unstructured text corpora presents a challenge. Recently, authors have utilized unsupervised, static word embeddings to uncover "latent knowledge" contained within domain-specific scientific corpora. Here semantic-similarity measures between representations of concepts, objects or entities were used to predict relationships, which were later verified using physical methods. Static language models have recently been surpassed at most downstream tasks by massively pre-trained, contextual language models like BERT. Some have postulated that contextualized embeddings potentially yield word representations superior to static ones for knowledge-discovery purposes. In an effort to address this question, two biomedically-trained BERT models (BioBERT, SciBERT) were used to encode n = 500, 1000 or 5000 sentences containing words of interest extracted from a biomedical corpus (Coronavirus Open Research Dataset). The n representations for the words of interest were subsequently extracted and then aggregated to yield static-equivalent word representations. These words belonged to the vocabularies of intrinsic benchmarking tools for the biomedical domain (Bio-SimVerb and Bio-SimLex), which assess quality of word representations using semantic-similarity and relatedness measures. Using intrinsic benchmarking tasks, feasibility of using contextualized word representations for knowledge discovery tasks can be assessed: Word representations that better encode described reality are expected to perform better (i.e. closer to domain experts). As postulated, BERT embeddings outperform static counterparts at both verb and noun benchmarks, however performance varies by model and neither model outperforms static models at both tasks. Moreover, unique performance characteristics are illustrated when task vocabulary is split between BERT-native words and words requiring subword decomposition.

1 Introduction

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

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): When 200-dimensional Word2Vec skip-gram (Mikolov et al., 2013) representations for material names (e.g. 'Bi_2Te_3') were ranked by their cosine similarity to the representation of 'thermoelectric,' several novel thermoelectric conductors were identified and subsequently verified. Despite the material name never having appeared alongside, or within a document containing the word '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 related 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 technique to a corpus of biomedical documents, discovering and validating novel tissue-reservoirs of the ACE2 receptor used by SARS-CoV-2 to invade a host.

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 those

from static models. Nevertheless, a method to adapt models like BERT for this purpose is lacking. Bommasani et al. (2020) described a method for reducing contextualized word representations to static-equivalents by aggregating them over a number of different contexts. These aggregated contextual embeddings outperformed static ones at general domain intrinsic benchmarking tasks (e.g. SimLex-999 (Hill et al., 2015), SimVerb-3500 (Gerz et al., 2016)), suggesting more realistic capture of word syntactic and semantic properties.

BERT-derived embeddings can be used as described by Tshitoyan et al. (2019) for knowledge discovery, by ranking geometric similarity between represented concepts, objects or processes (Figure 1, Figure 2). Nevertheless, as this 'latent knowledge' requires validation, the quality of suggested relationships cannot easily be assessed. For example, Tshitoyan et al. (2019) tested thermoelectric predictions using a mathematical formula, while Venkatakrishnan et al. (2020) utilized a custom-built molecular inference platform to validate postulations. Domain-specific intrinsic benchmarks which assess semantic similarity and relatedness between word representation pairs by comparing them to human-user ratings may be utilized as an appropriate surrogate: Higher-fidelity word representations are expected to better approximate human assessments of word relatedness (and therefore meaning).

This study tests the hypothesis of both Tshitoyan et al. (2019) and Venkatakrishnan et al. (2020) that contextual (BERT) models yield word representations that are superior to those produced by static model, and thus suitable for use in biomedical knowledge discovery. Using a biomedical corpus of 500,000 abstracts, pre-prints 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 processed by two biomedical BERT variants. The contributions of this paper can be summarized as follows:

- A method of utilizing BERT for biomedical knowledge discovery is described and validated. It involves encoding n contextual examples (i.e. sentences) containing vocabulary words, extracting and aggregating their representations. Aggregated representations can then be utilized for knowledge discovery tasks based upon their geometric relationship to other word-representations within the vector space.

- Using domain-specific intrinsic benchmarking tools (Bio-SimVerb/Bio-SimLex), unique, layer-wise performance differences are shown for verbs and nouns. Moreover, performance for nouns and verbs varies depending upon BERT model used, and layer of the model that representations are extracted from.

- In general, verbs native to BERT's vocabulary drastically outperform those requiring sub-word decomposition. Noun performance benefits from sub-word decomposition.

- Generally, the number of aggregated contexts per word has little effect upon performance. Subsequently, computationally efficient approaches for obtaining and applying these representations in knowledge discovery tasks may be devised.

2 Related Work

2.1 Knowledge Discovery via Semantic Relatedness Measures

Aside from the work of Tshitoyan et al. (2019) and Venkatakrishnan et al. (2020), Voytek and Voytek (2012) described a technique that utilized a co-occurrence algorithm to quantify the relationship and associations between neuroscientific terms and synonyms contained within 3.5 million papers indexed in PubMed. Importantly, the latter authors highlighted that the literature contained "a hidden network of connected facts that, by definition, recapitulate known neuroscientific relationships," almost a decade before the work of Tshitoyan et al. which expounded upon utilization of computational language models to uncover 'latent knowledge' within domain-specific text corpora. Moreover, the work of Venkatakrishnan et al. (2020) successfully demonstrates the scaling-up of this technique, and potential for clinically-meaningful discoveries using a non-trivial portion of the entire digitized biomedical knowledge base (see Figures 1, 2).

2.2 Converting Contextual Word Representations to Static-Equivalents

Bommasani et al. (2020) introduced a technique of converting contextualized word embeddings to static-equivalents for inferential purposes, in

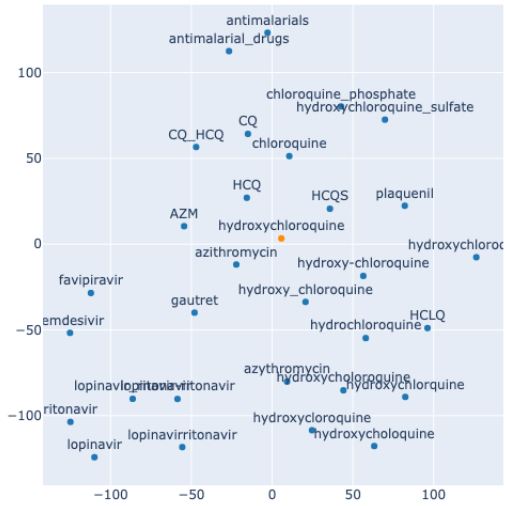

Figure 1: A dimensionality-reduced (T-SNE) plot demonstrating the 30 word-representations closest to 'hydroxychloroquine,' using data derived from the CORD-19 corpus. Apparent are clusters of drugs by type, e.g. antivirals (lower-left quadrant) vs. antiparasitics/antibacterials (middle; lower-right quadrants) which can be subsequently subjected to cluster-analysis using suitable unsupervised techniques. Interestingly the word 'gautret' appears close to the keyword 'hydroxycholoroquine,' as the first clinical trial of this controversial drug's use in treating COVID-19 was authored by Gautret et al. in 2020, indicating the high degree of association between terms in the corpus.

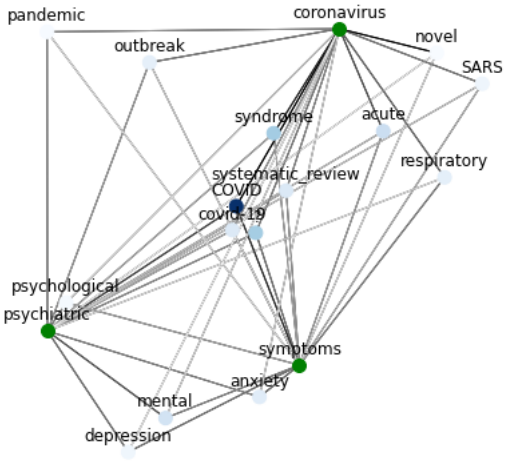

Figure 2: The 15 closest word representations to 'psychiatric,' 'coronavirus,' and 'symptoms,' keywords (green nodes) as derived from models trained on the CORD-19 corpus. Words like 'anxiety,' and 'depression,' among others, appear in the rankings. Pale-blue nodes represent weaker relations to the keywords, while darker-blue nodes represent geometrically-closer (i.e. stronger) relations to the keywords.

an effort to better understand contextualized language models like BERT. Importantly, the static-equivalent embeddings produced by this technique can be utilized in identical ways as those from older Word2Vec or GloVe models, and also outperform static embeddings at various intrinsic benchmarking tasks (Bommasani et al., 2020). Subsequently, novel methods of creating static-equivalents have been described, using continuous bag-of-word approaches (Gupta and Jaggi, 2021), phrases (Wang et al., 2021) and by combining contextual and static embeddings (Hämmerl et al., 2022), for example.

3 Methods

3.1 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 ('<NuM>') and selective lowercasing was performed to preserve abbreviations. For the Word2Vec and GloVe models, common terms and punctuation were removed.

3.2 Overview of Study Approach

The BERT approach was informed by results of an initial pilot study (see Appendix A for preliminary data). The pilot involved comparing contextual representations extracted from either long or short sentence sequences. Long sequences consisted of corpus text split into sentences. Short sequences further decomposed sentences into phrases, by splitting on commas. Due to slightly worse performance of long sequences, and the issue of sequence length often exceeding 512 (the maximum allowable sequence length for BERT), only short sequences were utilized. Two scientifically-specialized BERT models utilized for contextual embedding generation: BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019). Embeddings extracted and aggregated from these models were compared against those extracted from several

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

static models.

3.3 BERT Approach

BioBERT is a variation of BERT which is further pre-trained on PubMed abstracts and PubMed Central full-text articles. It outperforms general models at various downstream 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 (see also Appendix B). SciBERT is another BERT-variant pre-trained on approximately 1.14 million random scientific articles from Semantic Scholar. Approximately 18% of these articles are from the computer science domain, with the remainder from the biomedical domain. It also demonstrates superior performance at downstream biomedical NLP tasks relative to BERT (Beltagy et al., 2019).

n = 500, 1000 or 5000 sentences containing a single instance of the word of interest were sampled and tokenized using either the wordpiece or sentencepiece tokenizer (Kudo and Richardson, 2018) for BioBERT and SciBERT, respectively. Sequences 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 $\{\mathbf{w}_c^1, ..., \mathbf{w}_c^k\} \longmapsto \mathbf{w}_c$. 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 the arithmetic mean of both sub-word representations and n mimicked Bommasani et al. (2020)'s approach, where they found mean-pooling outperformed other possible operations (e.g. max., min., last) for both sub-word pooling and context aggregation (see also Ács et al. (2021)). If n did not meet the threshold, the maximum number of word representations available was aggregated. This approach differed from Bommasani et al. (2020) who instead took the representation produced by the word in isolation[3].

3.4 Static Models

The aggregated embeddings obtained from 3.3 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. Hyperparameters for Word2Vec were a context window of 8, initial learning rate of 0.01, high frequency word downsampling threshold of 0.0001, negative sampling parameter of 15, and ignoring any word with a corpus occurrence frequency of less than 10. All Word2Vec models were trained for 10 epochs. 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.

3.5 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 and the general domain. In these tasks, the cosine similarity of word representations from language models are compared to human domain-expert ratings, and subjected to Spearman rank-correlation testing. Bio-SimVerb and Bio-SimLex address shortcomings of previous biomedical benchmarks such as MayoSRS (Pakhomov et al., 2011) and UMNSRS (Pakhomov et al.,

[2] https://huggingface.co/

[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

2010) which only test nouns, and fail to distinguish between semantic relatedness and similarity (Chiu et al., 2018). These tools were used as a surrogate to validate contextualized embeddings use for knowledge discovery: Higher performance of a particular model's word embeddings at noun and verb benchmarks indicates a higher-fidelity mathematical representation of described reality. Consequently, prior to their actual validation, knowledge predictions (i.e. relationships between concepts, objects, entities) can be made with a greater degree of confidence.

4 Results

4.1 Verb Benchmarks

The left sub-plot of Figure 3 and left column of Table 1 demonstrates layer-wise performance of n = 500, 1000 and 5000 aggregated verb representations from BioBERT and SciBERT models. Performance is preserved regardless of sequence lengths/number of aggregated contexts, however it varies by model. SciBERT representations generally underperform compared to both Word2Vec 200 and 300-dimensional embeddings, and compared to BioBERT embeddings taken from the latter 8 layers. BioBERT verb embeddings from the 6th layer onwards outperform both SciBERT and static embeddings. BioBERT and SciBERT performance also differs across layers, as illustrated by curve morphology. Performance for both models reaches a maximum towards the latter layers.

4.2 Noun Benchmarks

The right sub-plot of Figure 3 and right column of Table 1 demonstrates layer-wise performance of n = 500, 1000 and 5000 aggregated contextualized noun representations from BioBERT and SciBERT models. Similar to verbs, n has negligible effect on performance. In contrast, SciBERT noun representations outperform BioBERT and static model representations. Performance for both models generally peaks towards the earlier layers. Again, curve morphology varies for BioBERT and SciBERT.

4.3 Effect of Sub-Word Pooling

To further explain model performance, test word-pairs from Bio-SimVerb and Bio-SimLex were separated into two groups based upon whether both words in a respective test pair existed in BioBERT/SciBERT's native vocabulary or not. This yielded test word pairs where both had a single representation, or where at least one of the words in the pair required sub-word pooling before aggregation. Representations were then subjected to Spearman's rank testing as per Bio-Simverb methodology (Chiu et al., 2018).

Model	Bio-SimVerb	Bio-SimLex
BioBERT 500	0.5516 (8)	0.7105 (6)
BioBERT 1000	**0.5526 (8)**	0.7103 (6)
BioBERT 5000	0.5513 (8)	0.7114 (6)
SciBERT 500	0.5142 (11)	0.7514 (3)
SciBERT 1000	0.5149 (11)	0.7509 (3)
SciBERT 5000	0.5144 (11)	**0.7513 (3)**
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. Performance of BERT model followed by number of aggregated contexts is given in first 6 rows. Number in brackets indicates layer. w2v200/300 = Word2Vec 200/300 dimensional embeddings. BWV = BioWordVec 200 dimensional embeddings. Bold entries indicate best overall performance.

When split using this criteria, model-native verb pairs from BioBERT and SciBERT outperform those requiring subword pooling at all layers, though performance declines from layer 0-12. Embeddings for verbs requiring sub-word decomposition perform better at latter layers, though still underperform. Only model-native BioBERT verb-pairs outperformed the best static embeddings (300-dimensional Word2Vec). Neither model-native, nor multi-token SciBERT verb embeddings outperformed top-performing static embeddings.

In contrast, for both BioBERT and SciBERT, noun embeddings benefitted from subword decomposition, with performance increasing until layer 6 before declining. For both models, native-noun representations performed best when extracted from the early model layers. Subword-decomposed nouns from SciBERT outperformed both model-native noun representations and those from the best-performing static model. Neither native, nor decomposed BioBERT-derived noun representations outperformed 300-dimensional Word2Vec embeddings (see Figure 4 and Table 2).

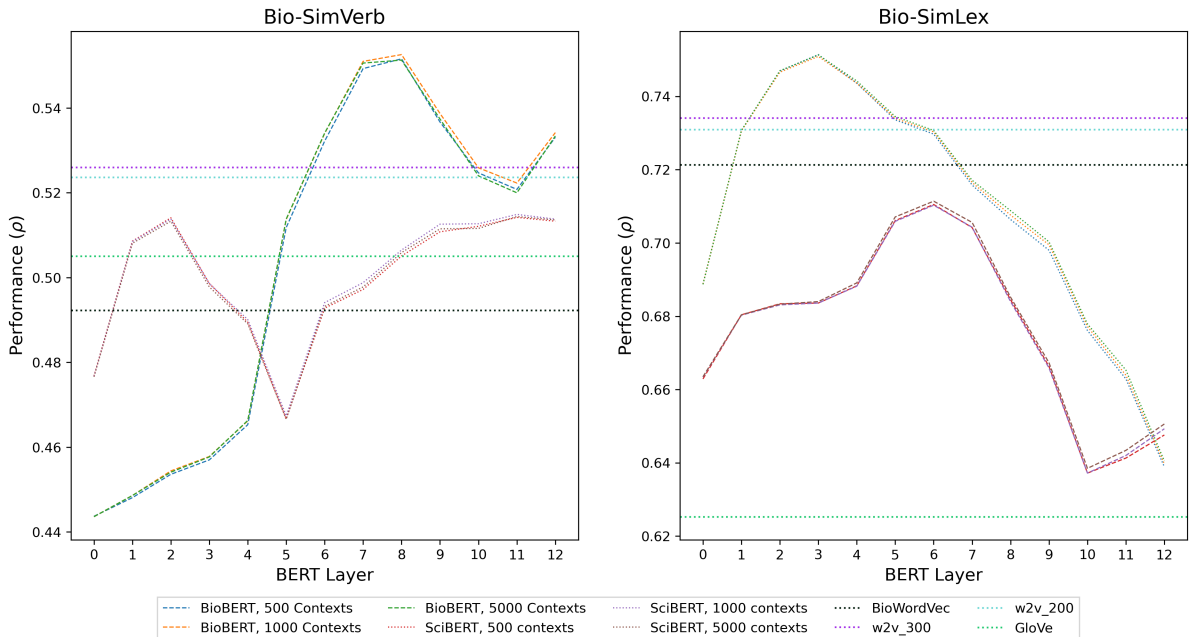

Figure 3: Layer-wise performance of BioBERT and SciBERT embeddings (0 corresponds to input layer) at both Bio-SimVerb and Bio-SimLex benchmarks. Horizontal dashed lines correspond to performance of static embeddings.

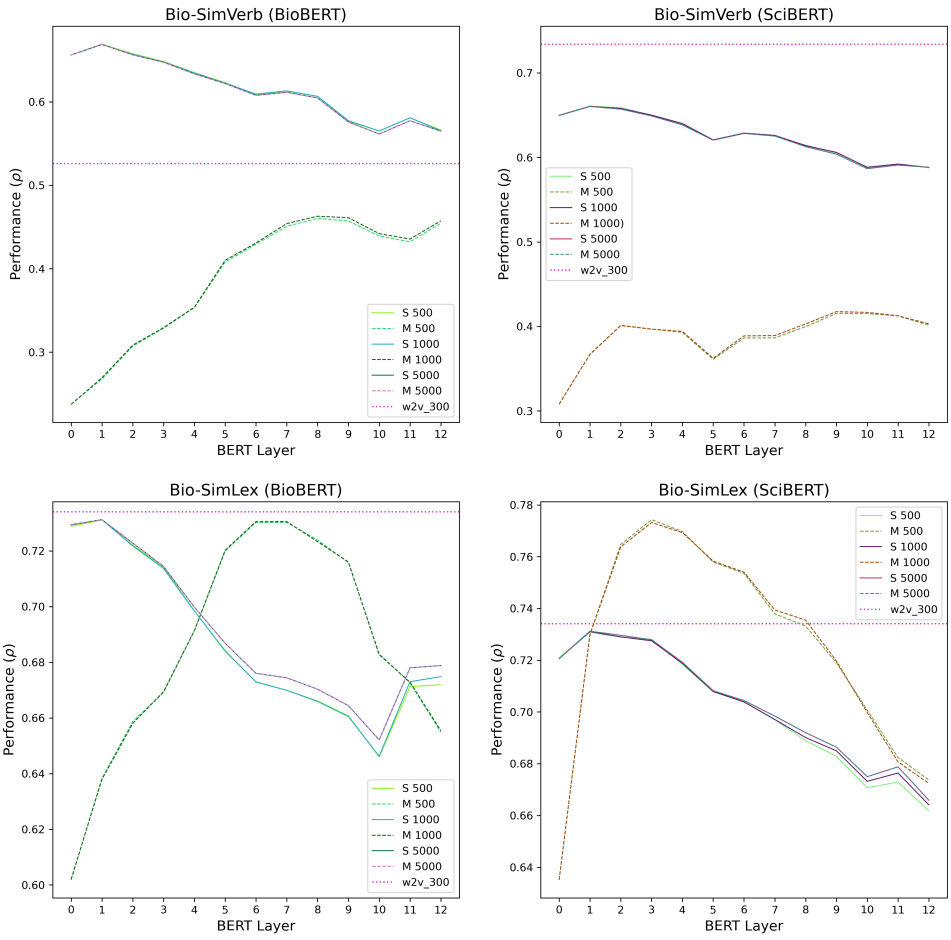

Figure 4: Layer-wise performance of BioBERT and SciBERT embeddings (0 corresponds to input layer) at both Bio-SimVerb and Bio-SimLex benchmarks. Horizontal dashed lines correspond to performance of static models.

Method	Bio-SimVerb	Bio-SimLex
BioBERT 500 (S)	**0.6691 (1)**	0.7255 (1)
BioBERT 1000 (S)	0.6685 (1)	0.7255 (1)
BioBERT 5000 (S)	0.6688 (1)	0.7256 (1)
BioBERT 500 (M)	0.4603 (8)	0.7417 (6)
BioBERT 1000 (M)	0.4629 (8)	0.7420 (6)
BioBERT 5000 (M)	0.4621 (8)	0.7418 (6)
SciBERT 500 (S)	0.6609 (1)	0.7309 (1)
SciBERT 1000 (S)	0.6604 (1)	0.7310 (1)
SciBERT 5000 (S)	0.6603 (1)	0.7312 (1)
SciBERT 500 (M)	0.4155 (9)	**0.7743 (3)**
SciBERT 1000 (M)	0.4175 (9)	0.7732 (3)
SciBERT 5000 (M)	0.4167 (9)	0.7737 (3)
w2v 300 (S)	0.5255	0.6959
w2v 300 (M)	0.4545	0.7341

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.

5 Discussion

Here, the feasibility of BERT-derived word representations for knowledge discovery purposes is illustrated. Static embeddings are outperformed by BERT models, however SciBERT and BioBERT illustrated opposite performance metrics relative to each other depending on whether verbs or nouns are being tested. As n has little bearing on performance, relatively few samples are required to yield embeddings capable of use in knowledge discovery tasks. Postulations of context-aware embeddings being superior to static ones for knowledge discovery may be correct, however a possible caveat is that the correct BERT variant must be chosen depending on word types of interest. Key to this approach is leveraging SciBERT's pre-training on massive and diverse corpora both related and unrelated to the domain of interest (e.g. computer science articles). Here, as few as 500 contextual word representations from a corpus of interest are required to yield aggregated word representations capable of outperforming static ones derived from models which requires training on an entire corpus.

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.

Another explanation for performance differences between BioBERT and SciBERT could be their vocabularies: BioBERT utilizes the unmodified BERT vocabulary, while SciBERT utilizes an expanded vocabulary with more scientific terms. A further consideration is that more 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, and SciBERT benefits from having been pre-trained on non-biomedical literature. Another area for exploration is the comparatively different layer-wise performance for BERT-native words versus extra-vocabulary words, with similar characteristics observed for both BioBERT and SciBERT.

A working knowledge-discovery framework utilizing BERT might consist of first extracting the vocabulary of the corpus upon which knowledge discovery 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 discovery as per Tshitoyan et al. (2019), Venkatakrishnan et al. (2020) and Voytek and Voytek (2012).

6 Conclusions

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

7 Limitations

This approach was only tested on a limited-morphology language like English and it is not known if the same results would be seen using corpora consisting of other languages. Moreover, due to resource limitations a maximum number of 5000 contexts were aggregated for the test vocabulary of 4000 words, the process of which took approximately 5 days using a single RTX3080ti, though CPU-based approaches were also trialled which completed the aggregation process for 5000 contexts in approximately the same amount of time using 8 processes.

8 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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text https://doi.org/10.48550/ARXIV.1903.10676.

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* .

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869* .

Prakhar Gupta and Martin Jaggi. 2021. Obtaining better static word embeddings using contextual embedding models. *arXiv preprint arXiv:2106.04302* .

Katharina Hämmerl, Jindřich Libovickỳ, and Alexander Fraser. 2022. Combining static and contextualised multilingual embeddings. *arXiv preprint arXiv:2203.09326* .

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 41(4):665–695.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, Belgium, pages 66–71. https://doi.org/10.18653/v1/D18-2012.

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.

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.

Jessica B Voytek and Bradley Voytek. 2012. Automated cognome construction and semi-automated hypothesis generation. *Journal of neuroscience methods* 208(1):92–100.

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* .

Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-bert: Improved phrase embeddings from bert with an application to corpus exploration. *arXiv preprint arXiv:2109.06304* .

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.

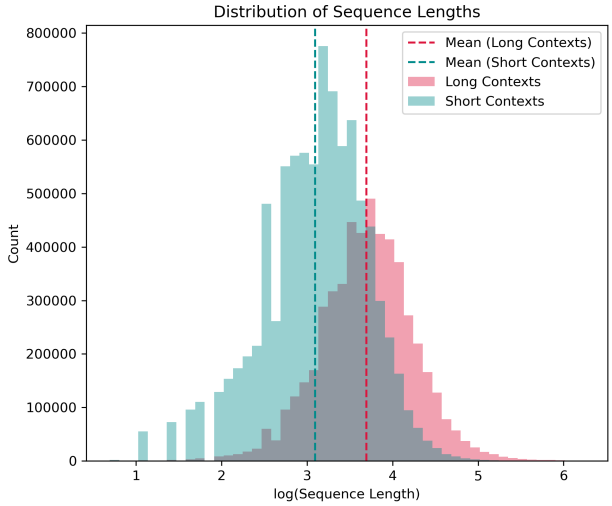

Figure 5: Distributions of log sentence lengths for long and short contextual sequences.

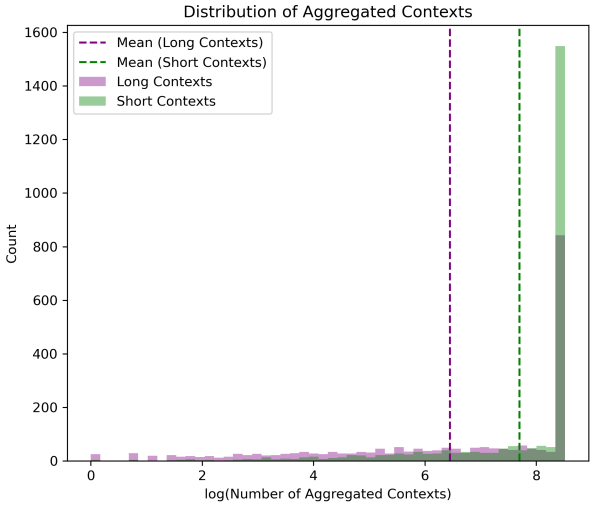

Figure 6: Distributions of log number of aggregated contexts for long and short sequence lengths. There are substantially more examples meeting n = 5000 for short sentences

A Corpus Sampling Characteristics

Figure 3 demonstrates the distribution of log sequence lengths for long and short sequences, respectively. Figure 4 demonstrates the distribution of log number of sequences for long and short sequences, respectively. The sampling criteria was that sequences had a single instance of the word and was <512 words in length. There are 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). 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

The pilot study involved pre-training BioBERT using the entire CORD-19 corpus. 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 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 3.3). Benchmarking was performed as described in 3.5.

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 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

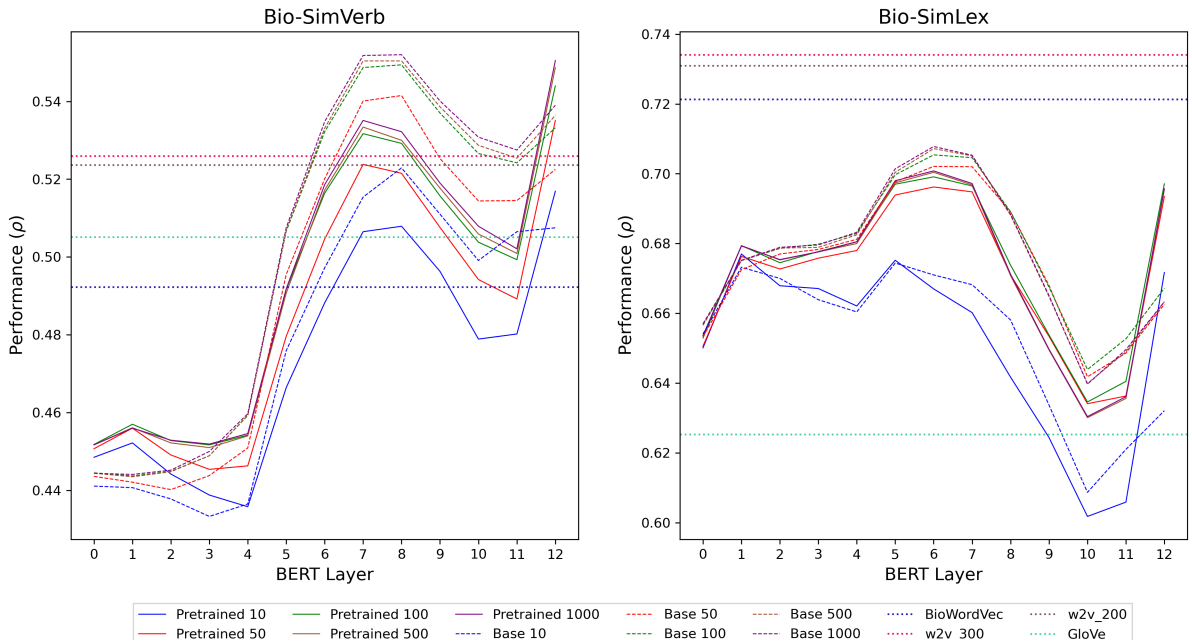

Figure 7: Layer-wise performance of BERT embeddings (0 corresponds to input layer) at both Bio-SimVerb and Bio-SimLex benchmarks. Pretrained n/Base n refer to either the further-pretrained model or the base model, respectively, followed by the n aggregated contexts. Horizontal dashed lines correspond to performance of static models.

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 pretrained 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.