Bioinformatics, 38(3), 2022, 872–874
https://doi.org/10.1093/bioinformatics/btab702
Advance Access Publication Date: 12 October 2021
Applications Note

Data and text mining

BioVAE: a pre-trained latent variable language model for biomedical text mining

Hai-Long Trieu [1,2,*], Makoto Miwa[1,3] and Sophia Ananiadou[2,4]

[1]Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan, [2]National Centre for Text Mining, Computer Science, University of Manchester 131 Princess Street, M7 1DN, UK, [3]Department of Advanced Science and Technology, Toyota Technological Institute, Nagoya 468-8511, Japan and [4]Alan Turing Institute, 96 Euston Road, London, NW1 2DB, UK

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 14, 2021; revised on September 16, 2021; editorial decision on October 4, 2021; accepted on October 8, 2021

Abstract

Summary: Large-scale pre-trained language models (PLMs) have advanced state-of-the-art (SOTA) performance on various biomedical text mining tasks. The power of such PLMs can be combined with the advantages of deep generative models. These are examples of these combinations. However, they are trained only on general domain text, and biomedical models are still missing. In this work, we describe BioVAE, the first large-scale pre-trained latent variable language model for the biomedical domain, which uses the OPTIMUS framework to train on large volumes of biomedical text. The model shows SOTA performance on several biomedical text mining tasks when compared to existing publicly available biomedical PLMs. In addition, our model can generate more accurate biomedical sentences than the original OPTIMUS output.
Availability and implementation: Our source code and pre-trained models are freely available: https://github.com/aistairc/BioVAE.
Contact: long.trieu@aist.go.jp
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Large-scale pre-trained language models (PLMs) (Beltagy *et al.*, 2019; Lee *et al.*, 2020) have shown state-of-the-art (SOTA) performance on various biomedical text mining tasks. These models provide contextualized representations, learned from large volumes of biomedical text, which then can be easily applied to achieve SOTA on downstream tasks such as named entity recognition (NER), relation extraction (REL) and question answering (QA) (Kim *et al.*, 2019; Lin *et al.*, 2019; Nentidis *et al.*, 2019).

Combining such large-scale PLMs to train latent variables based on deep generative models (DGMs) has been shown to improve representation learning tasks (Bowman *et al.*, 2016; Li *et al.*, 2020). A recent framework called OPTIMUS, has successfully combined BERT-based PLMs (Devlin *et al.*, 2019) and GPT-2 (Radford *et al.*, 2019) with variational autoencoders (VAEs) (Kingma *et al.*, 2013) (a powerful model of DGMs), achieving SOTA in both representation learning and language generation tasks when trained on two million Wikipedia sentences. However, the data distributions between general and biomedical domain are different, which makes it challenging to apply these models directly to biomedical text mining tasks (Lee *et al.*, 2020). In addition, training such large-scale models on a massive amount of biomedical text is costly (Supplementary Appendix SF).

To leverage the advantages of VAE-based PLMs for biomedical text mining, we release BioVAE, the first large-scale pre-trained latent variable language model for biomedical texts. The model is trained using the OPTIMUS framework on 34 million sentences from PubMed articles. We evaluate our BioVAE model on downstream text mining tasks, i.e. NER, REL and QA, and achieve SOTA on most of the tasks when compared with previous powerful biomedical PLMs, i.e. BioBERT (Lee *et al.*, 2020), SciBERT (Beltagy *et al.*, 2019) and PubMedBERT (Gu *et al.*, 2020). For language generation, BioVAE generates more accurate biomedical sentences than the original OPTIMUS output.

2 Approach

OPTIMUS: The OPTIMUS framework (Li *et al.*, 2020) is a large-scale VAE-based language model. VAE defines a joint distribution of observed inputs x and latent variables z with unknown prior distributions $p(z)$. The objective is to maximize the *Evidence Lower Bound* (ELBO):

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)||(p(z)), \quad (1)$$

where $p_\theta(x|z)$ is known as decoder, and $q_\phi(z|x)$ is known as encoder.

Table 1. Results on the text mining test sets

Model	NER			REL	QA
	BC5CDR	NCBI	JNLPBA		
PubMedBERT (Gu *et al.*, 2020)	87.27	79.96	71.82	**85.47**	**75.00**
BioBERT (Lee *et al.*, 2020)	88.85	89.36	77.59	76.68	69.29
SCIBERT (Beltagy *et al.*, 2019)	90.01	88.57	77.28	83.64	72.14
BioVAE ($\beta = 0.0, d_z = 32$)	89.85	88.85	**77.82**	83.68	72.86
BioVAE ($\beta = 0.0, d_z = 768$)	90.10	88.12	77.69	83.05	72.14
BioVAE ($\beta = 0.5, d_z = 32$)	89.69	89.80	77.66	83.54	72.14
BioVAE ($\beta = 0.5, d_z = 768$)	**90.18**	**90.12**	77.57	84.49	72.86

Note: The best scores are in bold, and the scores outperforming the SciBERT baseline are underlined. We report macro F1 scores for NER, micro F1 for REL and accuracy for QA (d_z: latent size).

Table 2. Reconstruction samples generated by OPTIMUS and our BioVAE and corresponding perplexity scores (the lower score is better)

Input	BioVAE	OPTIMUS
Sequence analysis of CDC4/FBXW7 was carried out on gastric **carcinoma** cell lines and xenografts	Sequence analysis of CDC4/FBXW7 was carried out on gastric **cancer** cell lines and xenografts	Electrophysiological studies were performed in the brain and cerebrospinal fluid (CSF)
Perplexity = 1.000	Perplexity = 1.113	Perplexity = 3.534

BioVAE: We used the OPTIMUS framework with the same configurations to train a large-scale VAE language model on biomedical data. We initialize the encoder with the biomedical pre-trained SciBERT (Beltagy *et al.*, 2019) and the decoder with the pre-trained GPT-2 (Radford *et al.*, 2019). We illustrate our model in Supplementary Appendix SA.

Corpus: We train BioVAE on the latest biomedical abstracts from the PubMed 2021 Baseline (https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/). Our data contain 34M sentences (3.35M abstracts).

Settings: We follow the same settings used in OPTIMUS. We set the latent size as 32 and 768, and beta as 0.0 and 0.5. For training on large batch sizes, we used the LAMB optimizer (You *et al.*, 2020). We used 128 GPUs from the AI Bridging Cloud Infrastructure (ABCI https://abci.ai/), which take 3 days to train 34M sentences for one epoch.

3 Results

Tasks: The pre-trained BioVAE model is evaluated on three NER tasks, i.e. BC5CDR (Li *et al.*, 2016), JNLPBA (Kim *et al.*, 2004) and NCBI-disease (Doğan *et al.*, 2014); a REL task, i.e. ChemProt (Kringelum et al., 2016); and a QA task, i.e. BioASQ (Nentidis *et al.*, 2019). We follow the same evaluation settings used in Lee *et al.* (2020) and Beltagy *et al.* (2019).

Fine-tuning: We follow the same settings used in our baseline SciBERT model (Beltagy *et al.*, 2019). The final BERT vectors from the encoder of our pre-trained BioVAE model are fed into a classification layer. The pre-trained model is fine-tuned for 2–5 epochs with a batch size of 32 and a learning rate of 2e-5, similarly to SciBERT's tuning parameters. For QA, we follow the BioBERT settings and evaluation scripts.

Results: Table 1 compares our BioVAE with biomedical pre-trained SciBERT (Beltagy *et al.*, 2019), BioBERT (Lee *et al.*, 2020) and PubMedBERT (Gu *et al.*, 2020) models on the NER, REL and QA tasks. Our baseline is the SciBERT since we use this model to initialize the encoder. BioVAE outperforms the SciBERT on all tasks, i.e. +0.54 F1 (JNLPBA), +0.17 F1 (BC5CDR), +1.55 F1 (for NCBI-disease) and +0.85 F1 (for ChemProt); and +3.57 accuracy (for QA) compared with BioBERT. BioVAE scores are lower than PubMedBERT in REL and QA, but better in NER tasks, and we discuss the reasons in more details in Supplementary Appendix SC.

Text generation: Given an input sequence, our model can reconstruct the input sequence. We compare sentences that have been reconstructed by both our BioVAE and OPTIMUS models in Table 2. The table shows that sentences generated by BioVAE are more accurate than the original OPTIMUS output. Further samples are presented in Supplementary Appendix SB.

4 Conclusion

We have described BioVAE, the first large-scale pre-trained latent variable language model for the biomedical domain. The model is trained using the OPTIMUS framework on large volumes of biomedical text. We achieve SOTA when evaluating the model on text mining tasks such as NER, REL and QA. Our results provide strong evidence that it will be possible to apply the BioVAE model to further biomedical tasks in the future.

Acknowledgement

The authors thank Khoa N. A. Duong for the invaluable support in implementing and evaluating the models.

Funding

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This research is also supported by the Alan Turing Institute and BBSRC, Japan Partnering Award, BB/P025684/1.

Conflict of Interest: none declared.

References

Beltagy,I. *et al.* (2019) SciBERT: a pretrained language model for scientific text. In: *EMNLP-IJCNLP*. ACL, Hong Kong, China, pp. 3606–3611.

Bowman,S. *et al.* (2016) Generating sentences from a continuous space. In: *CONLL, Berlin, Germany*, pp. 10–21.

Devlin,J. *et al.* (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: *NAACL, Minneapolis*, pp. 4171–4186.

Doğan,R.I. *et al.* (2014) NCBI disease corpus: a resource for disease name recognition and concept normalization. *Biomed. Inf.*, **47**, 1–10.

Gu,Y. *et al.* (2020) Domain-specific language model pretraining for biomedical natural language processing. *arXiv*, *preprint arXiv: 2007.15779*.

Kim,D. *et al.* (2019) A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, **7**, 73729–73740.

Kim,J.-D. *et al.* (2004) Introduction to the bio-entity recognition task at JNLPBA. In: *NLPBA/BioNLP*, COLING, Geneva, Switzerland, pp. 70–75.

Kingma,D.P. *et al.* (2014) Auto-encoding variational Bayes. In: *ICLR*, Banff, Canada.

Kringelum,J. *et al.* (2016) Chemprot-3.0: a global chemical biology diseases mapping. *Database*, **2016**, bav123.

Lee,J. *et al.* (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234-1240.

Li,C. *et al.* (2020) Optimus: organizing sentences via pre-trained modeling of a latent space. In: *EMNLP*, ACL. pp. 4678–4699.

Li,J. *et al.* (2016) Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, **2016**, baw068.

Lin,C. *et al.* (2019) A Bert-based universal model for both within-and cross-sentence clinical temporal relation extraction. In: *Clinical NLP Workshop*, pp. 65–71.

Nentidis,A. *et al.* (2019) Results of the seventh edition of the bioasq challenge. In: *ECMLPKDD*, Springer, Würzburg, Germany, pp. 553–568.

Radford,A. *et al.* (2019) Language models are unsupervised multitask learners. *OpenAI Blog*.

You,Y. *et al.* (2020) Large batch optimization for deep learning: training Bert in 76 minutes. In: *ICLR*.