Universal Biomedical Sentence Embeddings

Ao Liu **Kushagra Pundeer** **Rohan Gangaraju** **Surya Teja** **Tu Vu**
College of Information and Computer Science
University of Massachusetts Amherst
{aoliu, kpundeer, rgangaraju, suryatejadev, tuvu}@cs.umass.edu

Ivana Williams
Chan Zuckberg Initiative
iwilliams@chanzuckerberg.com

Abstract

Pretrained word-level representations learned on large dataset help with the improvement on many NLP tasks. However, using only the word embeddings, models still have the burden to learn the sequence representations. The urge of developing pretrained sentence embeddings increases, especially in biomedical domain, due to their significance in many bio-NLP tasks, such as sentence similarity. Pretrained sentence embeddings allow models to leverage existing contextual knowledge and improve the performances even in a low resource setting. Our goal is to develop a universal biomedical sentence embedding model in order to facilitate researches in the biomedical domain. Due to the effectiveness of multitask learning models in developing sentence embeddings for open-domain NLP as demonstrated in several works, such as GenSen (Subramanian et al., 2018) and Universal Sentence Encoder (Cer et al., 2018), we plan to apply multitasking on biomedical datasets to build a universal sentence encoder model for biomedical domain, with the help of seq2seq models and language models.

1 Introduction

Pretrained word embeddings have driven the success of many Natural Language Processing tasks. Transfer learning is applied on these tasks where the neural models use the pretrained word-level vector representations, which are obtained by unsupervised learning on a large corpus, as the input. It helps in low-resource settings where a large corpus is not available to learn the representations from scratch. While word embeddings can be helpful, models still need to learn relationship between words in context which can be challenging in a low-resource setting. Many recent works such as InferSent (Conneau et al., 2017), Skip-thought Vectors(Kiros et al., 2015b) and Paragram-Phrase XXL (Wieting et al., 2015) have tried to address this issue by learning general purpose sentence representations and show their effectiveness in various NLP tasks. While most NLP techniques in general domain are applied to the biomedical domain, the success of these tasks does not transfer to the biomedical domain without considerable effort and modification (Neumann et al., 2019). Even though there are multiple general purpose sentence embeddings trained on data in public domain, there is only one sentence level representation model for the Biomedical domain (Chen et al., 2018). In this project, we try to develop a universal sentence embedding model for the biomedical domain so that it can accelerate the progress of NLP tasks in biomedical text analysis.

2 Methodology

The objective of this project is to apply multitasking on large biomedical corpus to learn a universal biomedical sentence embedding model. We will explore several models and tasks to find the reasonable ones that we can use to build up our own model.

To facilitate our approach, we want to train our model on a large corpus in biomedical domain in order to have good transferablity. PubMed consists of unlabeled biomedical articles from MEDLINE, life science journals and online books. We obtain a subset of over 5.9 million title-abstract pairs from the publications in the last 5 years. It is one of the largest biomedical dataset that is publicly available. Due to the source and the size of PubMed dataset, it is suitable for our purpose of training a general sentence encoder.

The methodology of our project involves preprocessing, training sentence embeddings and evaluation on downstream tasks.

2.1 Preprocessing

Before we begin training the sentence embeddings, we apply the following preprocessing steps on the PUBMED dataset to clean the dataset and structure it for effective training of the sentence embeddings. We will apply basic preprocessing techniques like lowercasing, tokenization, etc. (using Spacy/AllenNLP). For BERT, we will follow preprocessing techniques similar to bioBERT (Lee et al., 2019) to keep tokens compatible with the original BERT pretrained models.

2.2 Training Sentence Embeddings

The next step after preprocessing is to build and train biomedical sentence embeddings using the sentences from the PubMed dataset. We are presently experimenting with the following models.

2.2.1 Seq2seq Multi-task Learning model

Seq2Seq approach proposed by Britz et al. (2017) has become an effective way to deal with variable length input sequences to variable length output sequences. It directly models the conditional probability of mapping an input sequence of words to output sequence using an encoder-decoder framework.

We intend to use Seq2Seq learning for multi-task learning instead of only focusing on only one task (Luong et al., 2015). The 2 main tasks that we plan on starting with are auto-encoding (Dai and Le, 2015) and skip-thoughts (Kiros et al., 2015a) prediction. Both of these tasks have shown to be more effective when combined in a multi-task learning model when compared to their results independently, as shown in the work by Luong et al. (2015).

Auto-encoders are multi-layer neural network that copy inputs to outputs by first compressing inputs into latent space representation and then reconstructing the output from this representation. Skip-thoughts learns to encode input sentences into fixed dimensional vector representation and then reconstructs surrounding sentences to map sentences that share semantic and syntactic properties into similar vectors. Both these tasks generate an intermediate vector representation which can be used to represent a sentence.

We start with one-to-many seq2seq network which consists of one encoder and multiple decoders. A sequence of words is provided as the input and we use the same model to generate both the same sequence of words using one decoder and next sequence of words for skip-thoughts objective using another decoder.

We will train deep LSTM models using 4 layers with 1000-dimensional cells and embeddings on the PUBMED dataset. We will use mini-batch size of 128 and uniformly initialize the parameters in $[-0.06, 0.06]$. A dropout with probability of 0.2 will be applied over the vertical connections (Pham et al., 2013). We will make use of stochastic gradient descent (SGD) and reverse the input sequences. Finally we will employ a simple finetuning schedule where after every x epochs we cut the learning rate by half for every y epochs. The values of x and y also referred as finetuning start and finetuning cycle along with number of training epochs will vary from task to task and will be experimentally determined.

2.2.2 Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018)

BERT is a language representation model that achieved state-of-the-art results on many NLP tasks, including tasks that use sentence embeddings, for open-domain NLP. However, the performance of language representation models depends heavily on the data on which the model is pretrained on, and the biomedical domain has a large corpus of words not found in open-domain corpus. Due to this problem, we plan to train sentence embeddings using BERT for the biomedical domain also. The BERT model consists of a multi-layered transformer network and bi-directional attention mechanism, and is pretrained by optimizing two objectives: masked language model and next sentence prediction. BERT uses a single or a pair of sentences as the input, and marks the start of each input using the token `[CLS]`. We can obtain the sentence embedding for a given sentence using BERT by passing it through the trained model and returning the final layer output of the `[CLS]` token used at the start of that sentence.

In this project, we intend to use two pretrained BERT models: one from the original work (Devlin et al., 2018) and the other from bioBERT (which is a fine tuned version of the original BERT model using Pubmed and PMC datasets) (Lee et al., 2019). Further, since our objective is to learn universal sentence embeddings, we intend to fine tune both the pretrained BERT models using multi-task learning with tasks like named entity recognition

and fake sentence detection, using the available PubMed data.

2.2.3 Fake Sentence Detection (Ranjan et al., 2018)

Now that we have auto-encoder-based Seq2Seq model and language model to help building our own model, we also want to see if there are other models or tasks that can facilitate the learning of sentence embeddings. InferSent (Conneau et al., 2017) proposed that natural language inference (NLI) task can help build good sentence embeddings. Recent researches (Subramanian et al., 2018; Cer et al., 2018) adopt their idea and significantly improve the quality of sentence embeddings with multitasking. However, there is no competitive replacement of the SNLI dataset (Bowman et al., 2015) they used in their papers in the biomedical domain; that is, no NLI dataset exists for biomedical domain. Thus, we have to find an appropriate alternative. The fake sentence detection task first generates fake sentences by shuffling or dropping words in the original sentences and then builds a classifier to predict whether the input is a fake sentence or not. The reason we think it could be a suitable task is that it allows our model to learn more syntactic and semantic information than that are need for sentence embeddings.

3 Evaluation

The final step of our methodology is to evaluate the trained models on downstream biomedical NLP tasks. Since semantic sentence similarity task is one of the most vital tasks in the biomedical domain, and has applications in tasks like duplicate sentence identification in diagnosis reports, identifying cohorts among patients based on extent of similarity in scan reports, diagnosis, etc. The first downstream task we are considering is the semantic sentence similarity task using BIOSSES dataset. BIOSSES contains 100 sentence pairs selected from TAC2 Biomedical Summarization Track Training Data Set. Five human experts score the similarity of sentence pairs in the range $[0, 4]$ following SemEval 2012 Task 6 Guideline. The mean of the scores assigned by the 5 experts is taken as the gold standard. We use the Pearson coefficient (which indicates the extent to which two variables are linearly related) to evaluate the cosine similarity scores obtained by our model with the gold standard normalized to $[0, 1]$.

Further, we plan to evaluate our sentence embeddings on probing tasks proposed by Conneau et al. (2018). The probing tasks tend to evaluate sentence embeddings based on their capabilities of predicting following information with a simple output layer:

1. surface informations: the length of the sentence and the information of the original words
2. syntactic information: word orders and hierarchical structures
3. semantic information: tense and the number of subjects or objects, etc.

We will evaluate our model based on the accuracies for these probing tasks.

References

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2018. Biosentvec: creating sentence embeddings for biomedical texts. *arXiv preprint arXiv:1810.09302*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. *CoRR*, abs/1511.01432.

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*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015a. Skip-thought vectors. *CoRR*, abs/1506.06726.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015b. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing.

Vu Pham, Christopher Kermorvant, and Jérôme Louradour. 2013. Dropout improves recurrent neural networks for handwriting recognition. *CoRR*, abs/1312.4569.

Viresh Ranjan, Heeyoung Kwon, Niranjan Balasubramanian, and Minh Hoai. 2018. Fake sentence detection as a training task for sentence encoding. *arXiv preprint arXiv:1808.03840*.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.