BioALBERT: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition

Usman Naseem†, Matloob Khushi†, Vinay Reddy‡, Sakthivel Rajendran§, Imran Razzak¶ and Jinman Kim∥
†‡§¶∥School of Computer Science, University of Sydney, Australia
Email:†‡§¶∥FirstName.SecondName@sydney.edu.au,¶§FirstName.SecondName@deakin.edu.au

Abstract—In recent years, with the growing amount of biomedical documents, coupled with advancement in natural language processing algorithms, the research on biomedical named entity recognition (BioNER) has increased exponentially. However, BioNER research is challenging as NER in the biomedical domain are: (i) often restricted due to limited amount of training data, (ii) an entity can refer to multiple types and concepts depending on its context and, (iii) heavy reliance on acronyms that are sub-domain specific. Existing BioNER approaches often neglect these issues and directly adopt the state-of-the-art (SOTA) models trained in general corpora which often yields unsatisfactory results. We propose biomedical ALBERT (A Lite Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) bioALBERT - an effective domain-specific language model trained on large-scale biomedical corpora designed to capture biomedical context-dependent NER. We adopted a self-supervised loss used in ALBERT that focuses on modelling inter-sentence coherence to better learn context-dependent representations and incorporate parameter reduction techniques to lower memory consumption and increase the training speed in BioNER. In our experiments, BioALBERT outperformed comparative SOTA BioNER models on eight biomedical NER benchmark datasets with four different entity types. We trained four different variants of BioALBERT models which are available for the research community to be used in future research.

I. INTRODUCTION

The growing volume of the published biomedical literature, such as clinical reports [1] and health literacy [2], are fuelling the advancements in the development of text mining algorithms. Biomedical named entity recognition (BioNER) intends to automatically identify biomedical entities such as diseases, chemicals, genes and proteins, etc., from the biomedical literature. So, a crucial step towards this aim is to build better and effective methods which can automatically recognize and extract biomedical entities. BioNER is an essential building block of many downstream text mining applications such as extracting drug-to-drug interactions [3] and disease-treatment [4] relationships. Traditionally, BioNER relies on feature engineering methods (e.g., lexicon-based, rules-based and statistics-based). However, feature engineering is dependent on domain-specific knowledge which does not perform well on BioNER [5].

Deep learning (DL) with its ability to automatically extract features have become common in BioNER recently [6]. For instance, Long Short-Term Memory (LSTM) is usually employed to learn vector representations of each word in a sentence, and then as the input to conditional random fields (CRF) which greatly improved the performance in BioNER [6]. Recently state-of-the-art (SOTA) DL based language models such as Embeddings from Language Models (ELMo) [7], Bidirectional Encoder Representations from Transformers (BERT) [8] and (A Lite Bidirectional Encoder Representations from Transformers (ALBERT) [9] obtained SOTA best performance on many NLP tasks.

Although these models show promising results, but applying them on BioNER has multiple challenges and limitations: (i) a limited amount of training data; (ii) an entity could represent multiple entity types depending on its textual context, i.e., a BRCA1 can be referred as gene name as well as a disease entity depending on its context. Similarly, heart attack and myocardial infarction refer to the same concept and, (iii) the heavy use of acronyms in biomedical texts makes it challenging to identify concepts, i.e., the abbreviation RA may refer to right atrium, rheumatoid arthritis, or one of several other concepts, where the resolution of the abbreviation is, therefore, context-dependent. Therefore, current models in BioNER rely on various context-independent and transformer-based context-dependent language models which are trained on biomedical corpora [10]–[12].

To overcome the identified limitations, we present Biomedical ALBERT bioALBERT - a context-dependent, fast and effective language model that addresses the shortcomings of recently proposed domain-specific language models. BioALBERT is trained on large biomedical corpora which address the limitation of limited training data. We also innovate in the adoption of cross-layer parameter sharing by learning parameters for the first block and reuse the block in the remaining layers instead of learning unique parameters for each of the layers and sentence-order-prediction (SOP) technique as a measure of coherence loss between sentences. SOP takes two consecutive sentences from training data and creates a random pair from different sentences which helps the model to learn better representations and finally, in BERT based models the size of the embedding was linked to the hidden layer sizes of the transformer blocks. These embeddings are projected directly to the hidden space of the hidden layer whereas in our model we use factorized embedding parameterization.
which decomposes embedding matrix into two small matrices which separate the size of the hidden layers from the size of vocabulary embeddings. This allows for increasing the hidden size without significantly increasing the parameter size of the vocabulary embeddings. BioALBERT is simple and efficient when fine-tuned for BioNER task as compared to other SOTA models. We evaluate our model on eight biomedical named entity recognition benchmark datasets. Our pre-trained BioNER models, along with the source code, will be publicly available.

II. RELATED WORK

A. Language Model

In biomedical text mining research, there is a long history of using shared language representations to capture the semantics of the text. One established trend is a form of word-embeddings [13] that represent syntactic and semantic meaning, and map words into low-dimensional vectors. Similar methods also have been derived for improving embeddings of word sequences by introducing sentence embeddings [14]. These context-independent word embeddings approach such as word2vec [13] were trained on biomedical corpora that contain terms and expressions that are usually not included in a general domain corpus [10]. These methods always require complex neural networks to be effectively used and model context-independent representations.

Another common trend, particularly in recent years, is text representation based on context [7–9]. Unlike traditional word-embeddings, this enables a word to change its meaning depending on the context in which it occurs. Several other works have investigated the usefulness of contextual models in clinical and biomedical domains. Several researchers trained ELMo on biomedical corpora and presented BioELMo and found that BioELMo beats ELMo on BioNER tasks [11, 16]. A pre-trained BioELMo model was released along with their work, allowing more clinical research in the area of BioNER. Beltagy et al. [15] released Scientific BERT (SciBERT), where BERT was trained on the scientific texts. BERT has typically been superior and better than ELMo to non-contextual embeddings.

Si et al. [17], trained the BERT on clinical notes corpora, using complex task-specific models to improve both traditional embedding and ELMo embedding on the 12b2 2010 and 2012 BioNER. Similarly in another study, a new domain-specific language model, BioBERT [12], trained a BERT model on a corpus of biomedical articles from PMC abstracts as well as full texts sourced from PubMed, which gave rise to enhanced performance on BioNER. Peng et al [18] introduced Biomedical Language Understanding Evaluation (BLUE), a collection of resources for evaluating and analysing biomedical natural language representation models. Their study also confirmed that the BERT models pre-trained on PubMed abstracts and clinical notes see better performance than most state-of-the-art models.

Despite this success, BERT has some limitations such as BERT has a huge number of parameters which is the cause for problems like degraded pre-training time, memory management issues and model degradation etc [9]. These issues are very well addressed in ALBERT, which is modified based on the architecture of BERT and proposed by Lan et al. [9]. ALBERT incorporates two-parameter reduction techniques that lift the significant obstacles in scaling pre-trained models. (i) Factorized embedding parameterization - decomposes the large vocabulary embedding matrix into two small matrices, (ii) replaces the NSP loss by SOP loss; and (iii) Cross-layer parameter sharing- prevents the parameter from growing with the depth of the network. These techniques significantly reduce the number of parameters used when compared with BERT without significantly affecting the performance of the model, thus improving parameter-efficiency. An ALBERT configuration similar to BERT-large has 18x fewer parameters and can be trained about 1.7x faster.

B. Fine tuning

Fine-tuning has been successfully used to transfer pre-trained weights as initialisation for parameters for various downstream tasks [19]. This improves the efficiency of the

---

Footnotes:

1. https://www.ncbi.nlm.nih.gov/pmc/
2. https://www.ncbi.nlm.nih.gov/pubmed/
TABLE I: Summary of parameters used in the Pre-training

| Name                  | BioALBERT 1.0 | BioALBERT 1.1 |
|-----------------------|--------------|---------------|
| Architecture          | ALBERT Base  | ALBERT Large  |
| Activation function   | GeLU         | GeLU          |
| Number of Attention Heads | 12          | 16            |
| Number of layers      | 12           | 24            |
| Hidden Size           | 768          | 1024          |
| Embedding Size        | 128          | 128           |
| Vocab Size            | 30000        | 30000         |
| Optimizer             | LAMB         | LAMB          |
| Train Batch Size      | 1024         | 256           |
| Eval Batch Size       | 16           | 16            |
| Max Seq Length        | 512          | 512           |
| Max Predictions per Seq | 20          | 20            |
| Learning Rate         | 0.000176     | 0.000062      |
| Training Steps (PubMed) | 200K        | 200K          |
| Training Steps (PMC)  | 270K         | 270K          |
| Warmup Steps          | 3125         | 3125          |

TABLE II: BioALBERT Pre-trained Models

| Model Version | Model | Trained On | # of words | Machine Used | Batch Size | Steps |
|---------------|-------|------------|------------|--------------|------------|-------|
| BioALBERT 1.0 | Base  | PubMed     | 18.8B      | GCP TPU v3-8 | 256        | 50K   |
| BioALBERT 1.1 | Large | PubMed+PMC | 18.8B      | GCP TPU v3-8 | 256        | 50K   |

B. Fine-tuning of BioALBERT

Fine-tuning of BioALBERT on BioNER task is presented in this section. BioNER involves annotating words in a sentence as named-entities. The labelled datasets that were used for this task include four categories representing Disease, Species, Drug/Proteins, and Drugs/Chemicals. The objective is to train and make a prediction on the labels, which are the proper nouns within the domain area. More formally, given an input sentence $S = \{x_1, x_2, ..., x_z\}$, where $x_i$ is the $i$-th word and $z$ represents the length of the sentence. The goal of BioNER is to classify each word in $S$ and assign it to a corresponding label $y \in Y$, where $Y$ is a predefined list of all possible label.

Fine-tuning is simple as compared to the pre-training, and the computational requirements are also not that significant.
TABLE III: Summary of parameters used in fine-tuning

| Summary of All parameters used: (Fine Tuning) | BioNER |
|---------------------------------------------|--------|
| Name                                        | adanw  |
| Optimizer                                   |        |
| Train Batch Size                            | 32     |
| Eval Batch Size                             | 16     |
| Save Checkpoint                             | 200    |
| Max Seq Length                              | 512    |
| Learning Rate                               | 1.00E-05 |
| Training Steps                              | 5336   |
| Warmup Steps                                | 320    |

BioALBERT which takes less physical memory and improvised parameter sharing techniques. The BioNER fine-tuning is trained to learn the word embeddings using the sentence piece tokenization while the BioBERT model was based on the word piece embeddings. For each of the pre-trained models, we constructed a fine-tuning task by using the specific dataset.

The model setup uses the weights of the pre-trained model that are created previously. We have used 1e-5 learning rate, batch size as 32 and lower case texts and finally fine-tuned for 5336 steps. Pre-trained BioALBERT models were pre-trained using TPUv3-8. All of the hyper-parameters used are same as default ALBERT unless stated otherwise. All the tested datasets contain a list of words along with a label B, I, and O where B denotes the beginning of an entity, I stands for inside and is used for all words comprising the entity except the first one, and O means the absence of an entity. For our experiments, we used these datasets as-is and passed to it our pre-trained models for the downstream task. The adanw optimiser was used with evaluation checkpoint so that the model will be evaluated at different time intervals using the holdout development dataset to identify the best model for final predictions. The predictions were performed on the test datasets, and the performance is compared with baseline models that were established previously by calculating the F1 Score, Precision and Recall. The summary of all parameters used in fine-tuning is given in Table III.

IV. EXPERIMENTAL ANALYSIS

In this section, we present the dataset used, baselines and results to show the effectiveness of our model.

A. Datasets

Our model is evaluated on eight biomedical NER benchmark datasets which contain four types of entities and provided by Lee et al. [12]. Table IV shows the statistics of the datasets used. Below we briefly explained each dataset:

- **BC5CDR**: The Bio-creative community challenge for the chemical-disease relation extraction task (BC5CDR) the corpus was made available in a Bio-creative workshop [24]. The two sub-tasks of BC5CDR are identifying: (i) chemical and (ii) disease entities from Medline abstracts.

- **BC4CHEMD**: This dataset is provided by Bio-Creative community challenge IV for the development and evaluation of tools for Chemical NER [25]. BC4CHEMDNER was used for the recognition of chemical compounds and drugs from Pubmed abstracts.

- **NCBI Disease**: To promote disease NER-system research, American National Institutes of Health released the NCBI disease corpus for disease NER-research. The NCBI disease corpus is large-scale and high-quality; it is based on the corpus released by Leaman et al. [26].

- **JNLPBA**: We also used the JNLPBA corpus in our experiments which are provided by Kim et al. [27]. This corpus contains five entity types, including DNA, RNA, Cell Type, Cell Line and Protein.

- **BC2GM**: BC2GM is provided by Ando [28], the state-of-the-art system in the Bio-Creative II gene mention recognition task is a semi-supervised learning method using alternating structure optimization.

- **LINNAEUS**: The LINNAEUS corpus was provided by Gerner et al. [29] which consists of 100 full-text documents from the PMC Open access document set which were randomly selected. All mentions of species terms were manually annotated and normalized to the NCBI taxonomy IDs of the intended species.

- **Species-800**: Species-800, which is also known as S800 [30], is a novel abstract-based manually annotated corpus. S800 comprises 800 PubMed abstracts in which organism mentions were identified and mapped to the corresponding NCBI taxonomy identifiers.

B. Baselines

To assess the performance of the proposed method, an exhaustive comparison is performed with several advanced SOTA methodologies along with their published results. Our model is compared with the following methods.

Yoon et al. [33] introduced CollaboNet, which consists of multiple BiLSTM-CRF models, for BioNER. While existing models were only able to handle datasets with a single entity type, CollaboNet leverages multiple datasets and achieves the highest F1 scores. CollaboNet is built upon multiple single-task NER models (STMs) that send information to each other for more accurate predictions.

Lou et al. [34] proposed a transition-based model for joint disease entity-recognition and normalization, based on transition-based structured prediction framework using structured perceptron with early-update training and beam-search.

3The published results were acquired from respective original publication.
We list the scores of the

| Type        | Metrics | SOTA          | BioBERT v1.0 | BioBERT v1.0 | BioBERTV1.1 | BioALBERT 1.0 | BioALBERT 1.0 | BioALBERT 1.1 | BioALBERT 1.1 |
|-------------|---------|---------------|--------------|--------------|-------------|--------------|--------------|--------------|--------------|
|             |         | (PubMed)      | (PMC)        | (PubMed)     | (PMC)       | (PubMed)     | (PMC)        | (PubMed)     | (PMC)        |
| Disease     | NCBI    | Base          | Base         | 88.30        | 86.76       | 86.16        | 89.04        | 88.22        | **97.45**    | **96.84**    | 97.18        | 97.38        |
| Drug/Chem   | BC5CDR  | p             | 89.30        | 89.02        | 89.12       | 89.06        | 89.71        | **95.89**    | 95.61        | **97.18**    | 95.85        |
| Drug/Protein| BC5CDR  | R             | 89.31        | 92.76        | 92.63       | 91.61        | 92.26        | 93.97        | **98.26**    | 98.06        | 95.62        | 95.68        |
| Species     | BC4CHEMD| p             | 92.29        | 91.77        | 91.65       | 92.23        | 92.80        | 97.76        | 97.71        | **97.78**    | 97.61        |
| Species     | JNLPBA  | p             | **91.14**    | **91.26**    | **91.41**   | **91.41**    | **92.50**    | **93.95**    | **96.25**    | **96.24**    | **96.22**    |
| Species     | LINNAEUS| p             | **81.81**    | **81.72**    | **82.86**   | **83.76**    | **84.72**    | **97.86**    | **98.46**    | **97.40**    | **97.05**    |
| Species     | Species-800| p          | **81.57**    | **83.35**    | **84.21**   | **85.93**    | **85.12**    | **94.87**    | **94.27**    | **95.72**    | **94.78**    |
|             |         | R             | **81.69**    | **82.54**    | **83.51**   | **84.40**    | **84.72**    | **96.34**    | **96.02**    | **96.97**    | **96.33**    |
|             |         | F             | **78.58**    | **76.83**    | **76.57**   | **77.39**    | **77.49**    | **84.72**    | **83.42**    | **84.01**    | **83.53**    |
|             |         | p             | **92.80**    | **91.85**    | **91.82**   | **92.88**    | **93.77**    | **95.95**    | **99.98**    | **99.38**    | **99.92**    |
|             |         | R             | **94.29**    | **84.72**    | **85.48**   | **86.11**    | **85.56**    | **89.43**    | **91.94**    | **97.41**    | **99.20**    |
|             |         | F             | **97.54**    | **88.11**    | **88.88**   | **89.81**    | **88.24**    | **99.71**    | **99.72**    | **99.72**    | **99.72**    |
|             |         | p             | **74.31**    | **70.60**    | **71.54**   | **72.84**    | **73.80**    | **90.17**    | **98.93**    | **98.10**    | **98.75**    |
|             |         | R             | **75.25**    | **78.20**    | **79.11**   | **80.24**    | **81.56**    | **98.24**    | **98.14**    | **98.32**    | **98.59**    |
|             |         | F             | **74.98**    | **73.08**    | **73.09**   | **74.21**    | **74.06**    | **98.76**    | **98.39**    | **99.02**    | **98.72**    |

Notes: Precision (P), Recall (R) and F1 (F) scores on each dataset are reported. The best scores are in **bold** and the second best scores are underlined.

Table V presents the performance of all the variants of BioALBERT and contrasts them to baseline methodologies.

Our model outperforms all other methods on all eight datasets and entity types. For, (i) Disease-type datasets, BioALBERT improved the performance by 7.47% and 10.63% for NCBI-Disease and BC5CDR-Disease datasets respectively; (ii) Drug/Chem type datasets increase in performance by 4.61% and 3.89% for BC5CDR-Chem and BC4CHEMD datasets respectively; (iii) Gene/Protein type datasets increase in performance by 12.25% and 6.42% for BC2GM and JNLPBA datasets respectively and; (iv) Species type datasets increase in performance by 6.19% and 23.71% respectively is observed.

We have performed multiple comparisons to analyse the effectiveness of BioALBERT. Fig. 2 presents, the performance comparison of the same versions (trained on same corpora and for the same number of steps) of both BioALBERT and BioBERT. We can see that in Fig. (2a), we compared BioBERT v1.0-base model which is trained on PubMed with BioALBERT 1.0-base model trained on Pubmed and similarly in Fig. (2b), we compared BioBERT v1.0-base model trained on Pubmed and PMC with BioALBERT 1.0-base model trained on PubMed and PMC biomedical corpora. In both cases, BioALBERT outperformed BioBERT on all eight datasets. We also compared the training time of BioALBERT with BioBERT. We found that all models BioALBERT beat BioBERT with a considerable margin. The run time statistics of both pre-trained models are given in Fig. 3.

**D. Discussions**

BioALBERT gives better performance and addresses the previously mentioned challenges in the biomedical domain. We attribute this to the BioALBERT built on top of the decoding. In another study, Lou et al. [34] proposed a neural network approach, i.e., attention-based bidirectional Long Short-Term Memory with a conditional random field layer (Att-BiLSTM-CRF), to document level chemical NER. The method leverages document-level global information obtained by attention mechanism to enforce tagging consistency across multiple instances of the same token in a document. It achieves better performances with little feature engineering.

Xu et al. [31] proposed a novel dictionary-based and document-level attention mechanism with a deep neural network NER method, named as DABLC. DABLC tags the consistency of multiple instances in a document at the document level and combines an external disease dictionary that is constructed with five disease resources containing a rich collection of disease entities. The authors adopted the efficient exact string matching method for dictionary matching; this method can effectively and accurately match the disease names.

Lee et al. [12] introduced BioBERT, which is a pre-trained language model for biomedical text mining. Authors showed that pre-training BERT on biomedical corpora is crucial in applying it to the biomedical domain. BioBERT outperforms previous models on biomedical text mining tasks such as NER, RE and QA. We compare BioALBERT with both BioBERT (v1.0 and V1.1) models and other SOTA models used in BioNER task. We selected those methods because they are the SOTA, and based on the conducted meta-analysis exhibit the highest performance among the techniques so far developed.

**C. Results**

Table V presents the performance of all the variants of BioALBERT and contrasts them to baseline methodologies.
transformer-based language model that learns contextual relations between words (or subwords) in a text. As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the transformer encoder reads the entire sequence of words at once. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word) and address the issue of contextual representation.

Our model addresses the shortcomings of BERT based biomedical models. At first, BioALBERT uses cross-layer parameter sharing and reduces 110 million parameters of 12-layer BERT-base model to 31 million parameters while keeping the same number of layers and hidden units by learning parameters for the first block and reuse the block in the remaining 11 layers. Secondly, our model uses the SOP, which takes two segments from the training corpus that appear consecutively and constructs a random pair of segments from different documents. This enables the model to learn about discourse-level coherence characteristics from a finer-grained distinction and leads to better learning representation in downstream tasks. Thirdly, our model uses factorized embedding parameterization in which a smaller size layer vocabulary and hidden layer to decompose the embedding matrix into two small matrices which reduce the number of parameters between vocabulary and the first hidden layer whereas in BERT based biomedical models embedding size is equal to the size of the hidden layer. Furthermore, finally, our model is trained on massive biomedical corpora to be effective on BioNER to address the issue of the shift of word distribution from general domain corpora to biomedical corpora. All these, when combined, address all the issues associated with BioNER earlier. As our model offers a consistent improvement over all other methods for all tested datasets, we can conclude that it is a robust solution for BioNER.

To extend our analysis, we analysed the performance of different pre-trained models of BioALBERT. We found out that the performance of all BioALBERT models are almost equally good, but BioALBERT 1.1-large trained on PubMed works better than others (shown in Fig. 3). BioALBERT 1.1-large model, which is trained on PubMed with dup-factor as five performs better on most of the datasets. This shows that the relevance of duplication data in NLP tasks.

V. CONCLUSION

In this study, we presented BioALBERT, which is a pre-trained language model for biomedical named entity recognition. We presented four different variants of BioALBERT models which are trained on huge biomedical corpora for a different number of steps. We showed that training ALBERT on biomedical corpora is a crucial step in applying it to BioNER. As future works, we plan to pre-trained other versions which include hybrid of general and biomedical corpora of ALBERT on biomedical corpora with more training steps and fine-tune on biomedical text mining task. We also plan to fine-tune BioALBERT on other text mining tasks to show the effectiveness of our model.

REFERENCES

[1] Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics, 17(01):128–144, 2008.
