Biomedical Named-Entity Recognition by Hierarchically Fusing BioBERT Representations and Deep Contextual-Level Word-Embedding

Usman Naseem*, Katarzyna Musial†, Peter Eklund‡ and Mukesh Prasad§
*†§School of Computer Science, University of Technology Sydney, Australia
‡School of Information Technology, Deakin University, Australia
Email:{*usman.naseem@student,†katarzyna.musial-gabrys@,§mukesh.prsad@}.uts.edu.au,‡peter.eklund@deakin.edu.au

***Abstract*—Text mining in the biomedical domain is increasingly important as the volume of biomedical documents increases. Thanks to advances in natural language processing (NLP), extracting valuable information from the biomedical literature is gaining popularity among researchers, and deep learning has enabled the development of effective biomedical text mining models. However, directly applying advancements in NLP to biomedical sources often yields unsatisfactory results, due to a word distribution drift from the general language domain corpora to specific biomedical corpora, and this drift introduces linguistic ambiguities. To overcome these challenges, this paper presents a novel method for biomedical named entity-recognition (BioNER) through hierarchically fusing representations from BioBERT, which is trained on biomedical corpora and Deep contextual-level word embeddings to handle the linguistic challenges within biomedical literature. Proposed text representation is then fed to attention-based Bi-directional Long Short Term Memory (BiLSTM) with Conditional random field (CRF) for the BioNER task. The experimental analysis shows that our proposed end-to-end methodology outperforms existing state-of-the-art methods for the BioNER task.**

***Index Terms*—NER, Name Entity, BioBert, Deep Contextual Embedding, Twitter. Disorders.**

I. Introduction

The growing volume of biomedical information in textual form enables advances in the development of pre-training language representations that can be applied to a range of tasks in the biomedical domain, such as pre-trained word-embeddings, sentence-embeddings, and contextual representations [38].

NER (Named-entity Recognition) is the computerized procedure for recognizing and labeling entities, persons, organisations, nationalities, locations, events, date-time, numeric quantities, ordinal and cardinal numerals in texts. In the biomedical domain, typical entities include diseases, chemicals, genes and proteins. BioNER is an essential building block of many downstream text mining applications such as extracting drug-to-drug interactions [19] and disease-treatment relationships [30]. BioNER is also used when building a sophisticated biomedical entity search tool [14] that enables users to pose complex queries to search biomedical texts on the basis of bio-entities.

The approaches on which NER in the mining of biomedical text is primarily based on are the rule, dictionary and machine learning. A drawback of dictionary based systems which causes them to perform poorly, despite their structure being easy to understand, is their inability to handle unknown items, or words with several meanings [32]. This is besides the fact that creating a continuously relevant dictionary needs a lot of manual input [35]. It is easier in rule based approach, for the size or scale of the model to be changed, but then it has to be 'fitted' by hand to a dataset. Despite their potential to obtain high accuracy results [14], these dictionary and rule-based approaches can give wrong suggestions when a word not included in the training data shows up: The out-of-vocabulary problem (OOV). This problem happens quite a lot in the biomedical realm as a result of new terms starting to appear in texts; examples being a novel drug or perhaps a new virus.

Developments in deep learning techniques employed in NLP has made the recent advancement of the models used in biomedical text mining possible. An example is the LSTM (Long Short-Term Memory) which has over the last few years appreciably enhanced BioNER [44]. Another example is the Conditional Random Field (CRF). There are drawbacks to applying these cutting edge NLP methodologies and graph based techniques [33], [34] to biomedical text mining though. Given that word representation models like Word2Vec [22], ELMo [28], BERT [6] and ALBERT [12] are trained primarily on datasets which have general non-specific domain texts (i.e, Wikipedia), it becomes hard to gauge their performance on biomedical text-containing datasets. Apart from that, the way words are distributed in biomedical and general bodies of text are not the same, posing a challenge for BioNER models in mining biomedical text. Biological text mining models thus depend on optimised versions of word descriptions because of these factors [29].

NER yet presents as difficult in biomedicals even though these models exhibit some encouraging results. This is because: (i) BioNER tasks have limited training data to work with, and are plagued with the absence of high quality annotated corpora. (ii) Depending on the context, different entity types could be assumed by one entity. An example is mistakenly labelling gene name "BRCA1" as a disease entity by BiLSTM-CRF based models for disease entities since there are disease terms in training data like "BRCA1 abnormalities" or "Brca1-deficient". Also, since "VHL" (Von Hippel-Lindau disease which is annotated as a disease entity by a training

data set, is also used as name of a gene, it confuses models.

Previously, Word2Vec, a well-established word representation model that works independent of context, was adjudged on biomedical corpora which possessed terms and patterns of words not often included in a corpus from general domain [29]. Though ELMo and ALBERT have shown the usefulness of contextualized representation of words, they do not obtain great results on biomedical corpora as a result of their being prior trained on corpora of general domain. Promising results are obtained with ALBERT on different NLP tasks whilst only changing the structure very minimally across the tasks, optimising ALBERT, as well as other bio versions of word representations like BioELMO and BioWord2Vec for the biomedical domain, such as NER could greatly enhance biomedical NLP research. We propose an end-to-end methodology to handle the above mentioned shortcomings for BioNER. We presented hierarchically text representations by utilising state-of-the-art transformer based word representation trained on biomedical corpora and fused it with deep contextual word-embedding. We evaluate our model on several benchmark datasets. The performance for the BioNER, when using our proposed method, is higher than the state-of-the-art methods. Our contributions can be summarized as follows:

- a novel hierarchical contextual-level words representation by hierarchically fusing representation from transformer (BioBERT) and deep contextual-lvel word representation, devised to comprehensively capture the polysemy in context, semantics, syntax, noise and OOV words.
- an end-to-end methodology to capture deeper contextual word-relationships using attention based BiLTM-CRF layer for BioNER;
- extensive experiments are conducted on several real-world datasets to evaluate the above design. All the results prove that our model constantly outperforms other state-of-the-art methods.

The remainder of this paper is structured as follows: Section II presents related work in the field of BioNER. Section III explains proposed methodology, the results are presented in Section IV, and finally in Section V we present conclusions.

II. Related Work

The rules for naming biomedical entities such as proteins, chemicals, genes, drugs, diseases etc. are not very clear and as such it becomes difficult to resolve a great many abbreviations, phrases and the like. This obstacle makes it more of a challenge to identify biomedical entities compared to non-specific texts. The related work shall be divided into the two ensuing categories:

*1) **Text Representation**:* In data mining research and biomedical text, the prevalent practice of using shared language representations to determine text semantics has long since been customary. The technique called transfer learning is used in such research. This technique involves pre-training language representations on vast corpora, and fine-tuned via NER, relation extraction or other downstream tasks [42]. A well-used trend is using high-dimensional vectors to get semantic meaning to be represented through a form of word-embeddings [48]. To improve word sequences embeddings, similar methods have been developed which involve sentence embeddings [5]. These methods have the constant requirement of the effective use of complex neural networks.

Context-dependent representation is another prevalent trend that has recently been adopted. This permits the possible change of the word meaning as the context changes, thus it is quite different from word-embeddings [37]. Beltagy et al. [3] released SciBERT which is set specific to scientific texts. On different tasks, BERT has typically been observed to be better than ELMo and far better to non-contextual embeddings. The usefulness of contextual models in clinical and biomedical domains have been investigated by a lot of other works. BioELMo, a biomedical version of ELMo, was trained on 10M abstracts sourced from PubMed [8]. It greatly outperforms general ELMo on many biomedical NLP tasks. An ELMo model was trained over a corpus of clinical radiology notes, mixed clinical discharge summaries and Wikipedia articles related to the medical field by Zhu et al. [49] They exhibited improved performance on the i2b2 2010 task[1]. A pre-trained ELMo model was released along with their work, allowing more clinical NLP research with these cogent contextual embeddings. Si et al. [36], trained a clinical note corpus BERT language model, and enhancements over both traditional and ELMo embeddings were obtained on the 12b2 2010 and 2020 tasks by utilising complex task-specific models, thus giving rise to new cutting edge results. Different hybrid models such as DICE [24], DICE+ [23], hybrid contextual word representation [25] and transformer based Deep Intelligent Contextual Embedding [26] were also presented to capture complex language ambiguities in previous studies.

BERT model was trained on corpus of biomedical articles from abstracts[2] as well as full texts sourced from PubMed[3] using BioBERT [13]. This gave rise to enhanced performance on a great many biomedical NLP tasks. The pre-trained Bert model was made available by the authors. Peng et al. [27] recently came out with BLUE, an assembly of resources for determining and evaluating representation models of biomedical natural language. They go ahead to analyse a lot of baselines in their experiments, basing it on BERT and ELMo. Their discovery was that PubMed abstract, and MIMIC-III clinical note-trained BERT model obtained the best outcomes.

*2) **Classifier: Attention-based BiLSTM-CRF**:* For the learning task, previous work on BioNER usually employs conventional machine learning methods. A framework of feature coupling generalization was proposed by Li et al [17] to produce lower dimensional features based on the rate of occurrence of terms and co-occurrence information in unlabelled data. This was done in combination with CRF for BioNER task. Torii et al. [40] developed a BioNER system

[1]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168320/

[2]https://www.ncbi.nlm.nih.gov/pmc/

[3]https://www.ncbi.nlm.nih.gov/pubmed/

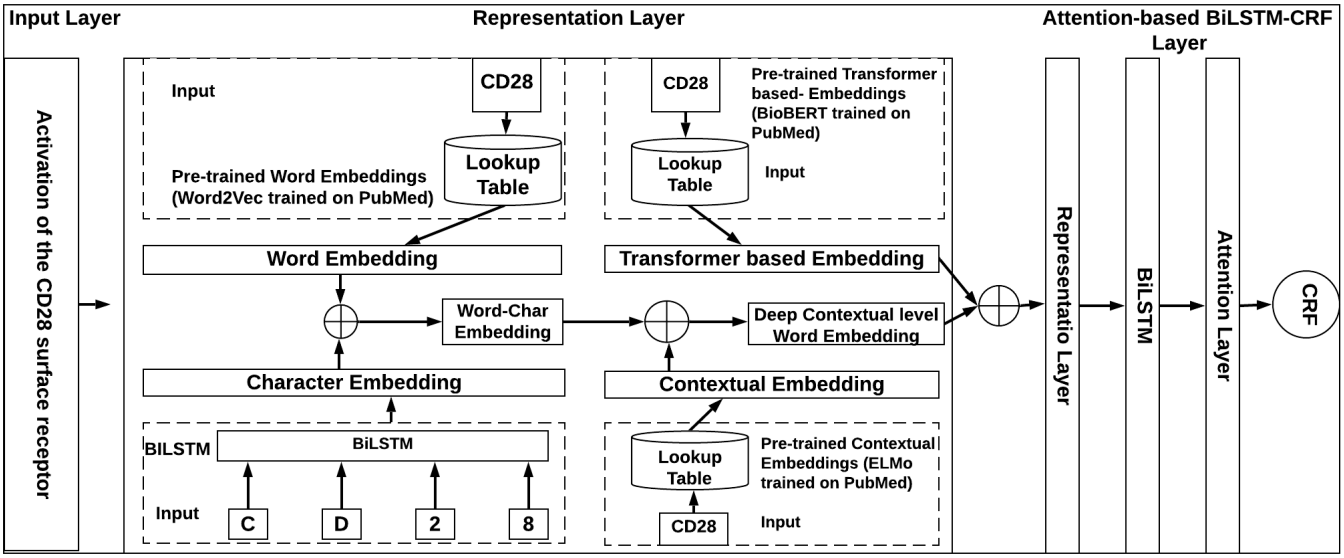

Fig. 1. Overall architecture of proposed method

named BioTagger-GM. It integrates four machine learning models, and is applicable in multiple corpora, according to the authors. Liao and Wu [18] developed the Skip-Chain CRF model for BioNER, wherein the information in a particular context is optimally evaluated as regards its dependence on long sentences. Tang et al. [39] analysed clustering-based representation, word-embeddings and distributional representation; three separate forms of word representation features. These features were sent into CRF. Since it shows good results on various forms of sequence-labelling tasks, CRF became the best choice for BioNER. Still, feature engineering, where complicate features are manually designed, must be relied on in this study. It requires linguistic insight, is time-consuming and relatively done on a need-to-apply basis.

Deep learning has been vastly useful in many fields [4]. As a result of this, and because it omits expensive feature engineering, a lot of studies in NER use neural networks for learning feature representations from text. In BioNER, Yao et al. [45] initially created embeddings of words on unlabelled texts of biological topics using neural networks, going on to establish a multi-layer neural network to obtain cutting edge output. Li et al. [16] mixed sentence vectors and twin word embeddings and utilized the BiLSTM on biomedical texts to identify domain-relevant entities. To identify drug entities, Zeng et al. [47] developed their model, BiLSTM-CRF. A CNN was utilized to get the representation of features on a character level. This was done with representations on a word level as well, to use as data to be fed to the BiLSTM-CRF for BioNER.

In biomedical literature there are a lot of words which can cause information redundancy whilst neural network models are being trained for feature capture, preventing important information being obtained. This may cause the important areas not getting focused on by the BioNER models and loss of information could occur. As such, it is a salient focus to make models of neural networks attentive to areas of importance. In machine translation, Bahdanau et al. [2] suggest the attention focusing mechanism. Taking model of decoder into account, focus can be made on important bits of the initial text as it is decoded, reducing information loss. An attention-based BiLSTM-CRF model is used by Luo et al. [21] for BioNER on a document level. They optimise the tagging inconsistency problem by using, between various sentences, mechanisms that are attention-focused. The best results are obtained on CHEMDNER and CDR corpora using this approach.

III. Proposed Model

In this section we present our proposed model, based on (a) a representation layer and (b) attention-based Bidirectional Long Short Term Memory (BiLSTM) with CRF layer for BioNER. Our model's framework includes five components: Input (Corpus), Representation layer, BiLSTM, Attention and CRF. The overall architecture of our model is given in Fig. 1.

A. Representation Layer

NER involves annotating words in a sentence as named-entities. More formally, given an input sentence $S = (x_1, x_2, ..., x_z)$, we predict corresponding labels $Y = (y_1, y_2, ..., y_z)$. Representations from transformers and deep contextual word-level embeddings, obtained from contextual embedding, and word-Char level embeddings are fused hierarchically at representation level. Detail is given below:

*1) **BioALBERT**:* Lan et al. [12] propose that the ALBERT model is modified based on the architecture of Bidirectional Encoder Representations from Transformers (BERT) proposed by Devlin et al. [6]. In this paper, we pre-trained our own model ALBERT (BioALBERT) on PubMed abstracts and clinical notes (MIMIC-III) to capture word-level representations

which proved to be more better as compared to pre-trained ALBERT model trained on general text.

*2) **Deep contextual level word-embedding**:* In this section, we present our deep contextual level word-embedding where we concatenated representations from (i) BioELMo and (ii) Word-Char Embeddings. Below we explain each of these:

1) **BioELMo**: ELMo [28] are embeddings proposed from a language model which considers different aspects of words according to their usage in context. The use of context-based word representation helps with polysemy and also helps to capture words with different concepts, and are therefore better when dealing with BioNER. These embeddings are based on the representation learned from Bi-language model (BiLM). Unlike traditional word-embeddings, ELMo considers different aspects of words according to their usage in context.
Log-likelihood of sentences in both forward and backward language models is involved in training process of BiLMs and final vector is computed after the concatenation of hidden representations from forward language model $\overrightarrow{h}^{\mathrm{LM}}_{n,j}$ and backward language model $\overleftarrow{h}^{\mathrm{LM}}_{n,j}$, where $j = 1, ..., L$ and is given by eqn. (1).

$$\mathrm{BiLM} = \sum_{n=1}^{k} (\log p(t_n|t_1, ..., t_{n-1}; \Theta_x, \overrightarrow{\Theta}_{\mathrm{LSTM}}, \Theta_s) + \log p(t_n|t_{n+1}, ..., t_n; \Theta_x, \overleftarrow{\Theta}_{\mathrm{LSTM}}, \Theta_s) \quad (1)$$

where θ_x and θ_s are the token representation parameters and softmax parameters respectively that are shared between forward and backward directions. And $\overrightarrow{\Theta}_{\mathrm{LSTM}}$ and $\overleftarrow{\Theta}_{\mathrm{LSTM}}$ are the forward back backward LSTM parameters respectively. ELMo abstracts the representations learned from intermediate layer from BiLM and execute linear combination for each token in a downstream task. BiLM contains *2L+1* set representations as given below.

$$R_n = (X^{\mathrm{LM}}_x, \overrightarrow{h}^{\mathrm{LM}}_{n,j}, \overleftarrow{h}^{LM}_{n,j} | \quad j = 1, ..., L) = (h^{\mathrm{LM}}_{n,j} | \quad j = 0, ..., L)$$

where $h^{\mathrm{LM}}_{n,0} = x^{\mathrm{LM}}_n$ is the layer of token and $h^{\mathrm{LM}}_{n,j} = [\overrightarrow{h}^{\mathrm{LM}}_{n,j}, \overleftarrow{h}^{\mathrm{LM}}_{n,j}]$ for each bi directional LSTM layer. ELMo is a task specific combination of these features where all layers in *M* are flattened to single vector and is given by eqn. (2).

$$ELMo^{\mathrm{task}}_n = E(M_n; \Theta^{\mathrm{task}}) = \gamma^{\mathrm{task}} \sum_{j=0}^{L} s^{\mathrm{task}}_j h^{LM}_{h,j} \quad (2)$$

where s^{task} are weights which are softmax normalized for the combination of different layers representations and γ^{task} is a hyper parameter for optimization and scaling of representation. In this work, we used BioELMo[4] trained on 10M PubMed abstracts by Jin et al. [8].

2) **Word-Char level Representation**: In the Word-Char level representation, we concatenate word-embedding trained on biomedical corpora and character level embeddings. As word-embeddings capture semantic and syntactic meanings of words, they have been widely used in various NLP tasks including NER. In our work, we used word-embeddings[5] provided by Pyysalo et al [29] for BioNER trained on biomedical corpora.
To establish closer representation among words of the same category, the prefix and suffix information of any word provides character-level features. This helps to deal with the challenge of OOV, mitigating issues like unseen words. In our experiment, we have performed character-level representations using Bi-LSTMs in order to produce a character-enhanced embedding for each unique word [11]. We considered the maximum character length is 25 and set forward and backward LSTMs parameters to 25, which results in a 50-dimensional embedding vector. Finally, we concatenate the above vectors to achieve one representation, we called a Word-Char representation.
Finally, we concatenate the final two representations from BioBERT and Deep contextual level word embeddings and fed to an attention-based BiLSTM-CRF layer.

B. Attention-based BiLSTM-CRF layer

1) BiLSTM with Attention layer: We placed the BiLSTM layer on top of our representation layer with the attention layer to capture information from both directions. A BiLSTM takes an input of a vector with a sequence of x_z tokens and produces hidden representation h_i at a given time i by concatenating the hidden representations from both forward $\overrightarrow{h_i}$ and backward $\overleftarrow{h_i}$ LSTM, given by eqn. (3).

$$h_i = [\overrightarrow{h_i} \parallel \overleftarrow{h_i}] \quad (3)$$

where $\parallel$ in eqn. (3), denotes the concatenation of outputs from both forward and backward LSTM.

Different words play different roles in developing understanding, thus not all words contribute equally in understanding a sentence. In this work, we have further used the attention mechanism to enforce the contribution of important words in understanding the sentence. We have assigned weight a_i to each token through a softmax function and finally, representation **R** which is a weighted-sum of all tokens, computed as shown in eqn. (4).

$$R = \sum_{i=1}^{z} a_i h_i, \quad (4)$$

where,

$$a_i = \frac{exp(e_i)}{\sum_{t=1}^{z} exp(e_t)}, \quad \sum_{i=1}^{z} a_i = 1$$

$$e_i = tanh(W_h h_i + b_h)$$

[4]https://github.com/Andy-jqa/bioelmo

[5]http://bio.nlplab.org/

TABLE I
STATISTICS OF DATASETS

Corpus	Entity Type	# Annotations
BC5CDR- Disease	Disease	12,694
NCBI Disease	Disease	6,881
BC5CDR- chemical	Chemical	15,411
BC4CHEMD	Chemical	79,842
JNLPBA	Gene/protein	35,460
BC2GM	Gene/protein	20,703

and W_h and b_h are learned parameters, h_i is the concatenation of the representations of the forward and backward LSTM, introduced in eqn. (3).

2) Output layer: At the output layer, we applied a Conditional Random Field (CRF) to the representation **R** generated from an attention layer.

IV. EXPERIMENTAL ANALYSIS

In this section, first, we present experimental settings, experimental evaluation and results to show the effectiveness of our proposed model.

A. Experimental settings

In this section, we present the experimental settings used in our experiments.

*1) **Datasets**:* We used datasets from three different entities related to BioNER. Table I shows the statistics of the datasets used. Below we briefly explained each dataset:

- **BC5CDR**: The Biocreative community challenge for the chemical-disease relation extraction task (BC5CDR) corpus was made available in a Biocreative workshop [15]. The two sub-tasks of BC5CDR are identifying; (i) **chemical** and (ii) **disease** entities from Medline abstracts. The corpus has 1,500 abstracts from Pubmed and chemical entities are hand-annotated by a team of indexers from Medical Subject Headings (MeSH).
- **BC4CHEMD**: This dataset is provided by BioCreative community challenge IV for the development and evaluation of tools for Chemical NER [10]. BC4CHEMDNER was used for the recognition of chemical compounds and drugs from Pubmed abstracts. We have downloaded training, validation and test-sets in the IOBES tagging scheme from Github[6].
- **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. [7].
- **JNLPBA**: We also used the JNLPBA corpus as an experimental dataset, provided by Kim et al. [9]. This corpus contains five entity types including DNA, RNA, Cell Type, Cell Line and Protein.
- **BC2GM** Our last dataset is provided by Ando [1], 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.

[6]https://github.com/cambridgeltl/MTL-Bioinformatics-2016/tree/master/data/BC4CHEMD

The Grid search optimization technique was used to find the hyper-parameters used in our experiments. Further, F1-Score is used for the evaluation.

B. Experimental Evaluation

This section presents the baselines models used in comparison to our proposed model and discusses the results.

*1) **Baselines**:* Several cutting-edge methodologies and their published results are extensively compared to assess the performance of the proposed method; published results being obtained from each of their original publications (each are indicated in their respective tables). The following methods are compared with our model. CollaboNet, introduced by Yoon et al. [46], comprised of multiple BiLSTM-CRF models, for BioNER. Multiple datasets can be leveraged, and also it obtains the best F1 scores unlike existing models which were capable of only datasets with one entity type. CollaboNet is constructed with multiple single-task NER models (STMs) that give more accurate predictions by exchanging information with each other. This is an improvement from other recently suggested multi-task models. It also creates distinctions among polysemous entities of biomedical nature, or those whose orthography are alike. BC5CDR-both was used by Wang et al. [41] in their studies, while Yoon et al. [46] re-ran their models on BC5CDR-chem and BC5CDR-disease and obtained outputs for a fair comparison with other models. For joint disease entity-recognition and normalization, Lou et al. [20] suggested a transition-based model: structured prediction framework using structured perceptron with beam-search decoding and early-update training.

Results of our hybrid model, with normalisation reasonably compared to the pipeline baseline, indicated a vast improvement in disease entity-recognition. Lou et al. [21] for chemical NER on document level, proposed a neural network approach; an Att-BiLSTM-CRF which means an "attention-based bidirectional Long Short-Term Memory with a conditional random field layer". To ensure consistency in tagging a document across multiple occurrence of the same token, document-level global information acquired by attention mechanism is leveraged in this approach. Better outcomes with minimal need for engineering of features are achieved. A novel document-level attention and dictionary-based mechanism named as DABLC was suggested by Xu et al. [43]

DABLC operates with a deep neural network NER method, tagging the consistency of multiple instances of entities at the document level and incorporates an external disease dictionary with a detailed collection, being served by five disease resources. The exact string matching method was adopted for dictionary matching; correctly and efficiently matching disease names. BioBERT; a prior trained model of language representation for mining of text concerned with the biomedical field, was introduced by Lee et al. [13] It was shown that a prior training of BERT on biomedical bodies of text is salient for adapting it to this domain. BioBERT outperforms previous

TABLE II
COMPARISON OF "DISEASE" TYPE BIONER

Model\Dataset		BC5CDR-Disease	NCBI Disease
Wang et al. [42]		83.33*	86.14
Yoon et al. [46]		84.08	86.36
Lou et al. [20]		86.23	—
Xu et al. [43]		—	88.60
Lee et al. [13][7]	BERT (wiki+book)	82.41	85.63
	BioBERT (PubMEd)	86.20	87.38
	BioBERT (PMC)	85.27	87.79
	BioBERT (PubMED+PMC)	86.56	89.36
Peng et al. [27]	ELMO	83.90	—
	BioBERT	85.90	—
	Peng BERT-base (PubMed)	86.60	—
	Peng BERT-base (PubMed+MMIC)	85.40	—
Proposed		**88.34**	**91.23**

models such as QA, NER, and RE, and requires very little task-specific architectural modification

BLUE, an assembly of resources for determining and evaluating representation models of biomedical natural language, was introduced by Peng et al. [27] It was determined that most state-of-the-art models are not as good as the BERT models pre-trained on PubMed abstracts and clinical notes.

The published results were acquired from the respective original publications (the reference publication is indicated in the respective tables). We selected those methods because they are the state-of-the-art, and based on the conducted meta-analysis exhibit the highest performance among the techniques so far developed.

TABLE III
COMPARISON OF "DRUG/CHEM" TYPE BIONER

Model\Dataset		BC5CDR-chem	BC4CHEMD
Wang et al. [42]		91.29	89.37
Yoon et al. [46]		93.21	88.85
Lou et al. [20]		86.23	—
Lou et al. [21]		—	91.14
Lee et al. [13]	BERT (wiki+book)	91.16	90.04
	BioBERT (PubMEd)	92.64	91.26
	BioBERT (PMC)	92.54	90.97
	BioBERT (PubMED+PMC)	93.44	91.41
Peng et al. [27]	ELMO	91.50	—
	BioBERT	93.00	—
	Peng BERT-base (PubMed)	93.50	—
	BERT-base (PubMed+MMIC)	85.40	—
Proposed		**94.24**	**92.28**

C. Results

Results are summarized in the Tables II–IV. Table IV presents the performance results of our proposed method

TABLE IV
COMPARISON OF "GENE/PROTEIN" TYPE BIONER

Model\Dataset		JNLPBA	BC2GM
Wang et al. [42]		73.52	80.74
Yoon et al. [46]		78.58	79.73
Sachan et al. [31]		75.03	81.69
Lee et al. [13]	BERT(wiki+book)	74.94	81.79
	BioBERT(PubMEd)	76.65	82.54
	BioBERT(PMC)	76.53	83.53
	BioBERT(PubMED+PMC)	77.59	84.40
Proposed		**79.97**	**86.05**

(**Proposed**) and contrasts them to other state-of-the-art baseline methodologies, along with published results using the same dataset. As can be observed, the proposed model outperforms all other approaches, as well as all methods with published results for the respective BioNER tasks. Our model achieved better performance because: (i) it handles language ambiguities by capturing deeper relationships within the text. Our proposed model learns high-quality representations by adding polysemy, OOV words, semantics and syntactical information of words and learns high-level representations from transformer (ALBERT) trained on biomedical corpora which helps to achieve better text representation and; (ii) we used Attention-based BiLSTM-CRF which is currently the most widely used in-sequence labeling tasker. For, (i) Disease-related datasets, our model improved the performance ratio (Δ) by 2.00% and 2.10% for BC5CDR-Disease and NCBI-Disease datasets respectively; (ii) Drug/Chem related datasets increase in performance ratio (Δ) by 0.70% and 0.95% for BC5CDR-Chem and BC4CHEMD datasets respectively and; (iii) Gene/Protein related datasets increase in performance ratio (Δ) by 1.76% and 1.96% for JNLPBA and BC2GM datasets respectively, when compared to previous best results. As our model offers consistent improvement over all other methods for all tested datasets we can conclude that it is a robust solution for BioNER.

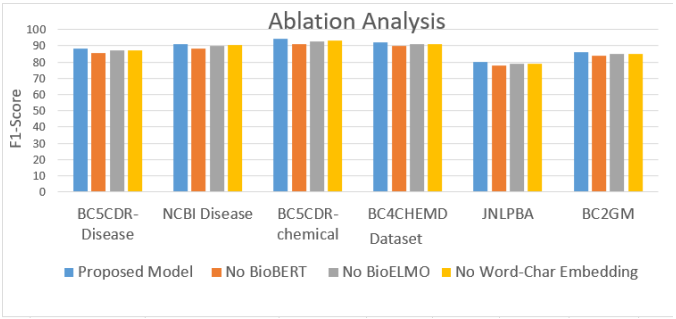

Fig. 2. Ablation Analysis of proposed model.

D. Ablation Analysis

It is evident from Fig. 2 that all layers in our proposed model improve overall performance. A noticeable drop is observed in the performance when we remove BioBERT from our model for all datasets. Further, experimental analysis

also indicates that performance drops in both cases when we remove BioELMO or word-Char Embedding from our deep contextual embedding. Hence, we can conclude that one of the strengths of our model lies in the combination of different components that builds the diversity which contributes to increased performance of BioNER.

V. Conclusion

In this study we presented an end-to-end methodology for BioNER. We proposed a novel representation layer which is then fed to attention based BiLSTM-CRF layer. Our representation layer consists of hierarchically fusing representations from BioBERT and deep contextual-level word representation. Our representation layer is trained on biomedical corpora and well-equipped to handle linguistic challenges such as polysemy, semantic, syntax, OOV and noise within the biomedical literature. The experiment shows that our proposed methodology outperforms different baselines and achieves state-of-the-art performance for BioNER. In future, we plan to explore different ways to incorporate more data characteristics, handle other language complexities, and apply our model on different domains.

References

[1] Rie Kubota Ando. Biocreative ii gene mention tagging system at ibm watson. 2007.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text, 2019.

[4] Min Chen, Yiming Miao, Xin Jian, Xiaofei Wang, and Iztok Humar. Cognitive-lpwan: Towards intelligent wireless services in hybrid low power wide area networks. *IEEE Transactions on Green Communications and Networking*, 3(2):409–417, Jun 2019.

[5] Qingyu Chen, Yifan Peng, and Zhiyong Lu. Biosentvec: creating sentence embeddings for biomedical texts. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5, 2018.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Rezarta Islamaj Doundefinedan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus. *J. of Biomedical Informatics*, 47(C):1–10, February 2014.

[8] Qiao Jin, Bhuwan Dhingra, William W. Cohen, and Xinghua Lu. Probing biomedical embeddings from language models, 2019.

[9] Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, JNLPBA '04, page 70–75, USA, 2004. Association for Computational Linguistics.

[10] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel Lowe, Roger Sayle, Riza Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, and Alfonso Valencia. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:S2, 03 2015.

[11] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360, 2016.

[12] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2019.

[13] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining, 2019.

[14] Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, DongHee Choi, Sunkyu Kim, Aik-Choon Tan, and Jaewoo Kang. Best: Next-generation biomedical entity search tool for knowledge discovery from biomedical literature. In *PloS one*, 2016.

[15] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database : the journal of biological databases and curation*, 2016, 2016.

[16] Lishuang Li, Liuke Jin, Yuxin Jiang, and Degen Huang. Recognizing biomedical named entities based on the sentence vector/twin word embeddings conditioned bidirectional lstm. In *CCL*, 2016.

[17] Yanpeng Li, Hongfei Lin, and Zhihao Yang. Incorporating rich background knowledge for gene named entity classification and recognition. *BMC Bioinformatics*, 10:223 – 223, 2008.

[18] Zhihua Liao and Hongguang Wu. Biomedical named entity recognition based on skip-chain crfs. *2012 International Conference on Industrial Control and Electronics Engineering*, pages 1495–1498, 2012.

[19] Sangrak Lim, Kyubum Lee, and Jaewoo Kang. Drug drug interaction extraction from the literature using a recursive neural network. *PLOS ONE*, 13:e0190926, 01 2018.

[20] Yinxia Lou, Yue Zhang, Tao Qian, Fei Li, Shufeng Xiong, and Donghong Ji. A transition-based joint model for disease named entity recognition and normalization. *Bioinformatics (Oxford, England)*, 33, 03 2017.

[21] Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34:1381–1388, 2018.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–9, 2013.

[23] Usman Naseem, Shah Khalid Khan, Imran Razzak, and Ibrahim A Hameed. Hybrid words representation for airlines sentiment analysis. In *Australasian Joint Conference on Artificial Intelligence*, pages 381–392. Springer, 2019.

[24] Usman Naseem and Katarzyna Musial. Dice: Deep intelligent contextual embedding for twitter sentiment analysis. *2019 15th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1–5, 2019.

[25] Usman Naseem, Imran Razzak, and Ibrahim A Hameed. Deep context-aware embedding for abusive and hate speech detection on twitter. *Australian Journal of Intelligent Information Processing Systems*, page 69.

[26] Usman Naseem, Imran Razzak, Katarzyna Musial, and Muhammad Imran. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, 2020.

[27] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets, 2019.

[28] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018.

[29] Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. Distributional semantics resources for biomedical text processing. 2013.

[30] Barbara Rosario and Marti Hearst. Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 430–437, Barcelona, Spain, July 2004.

[31] Devendra Singh Sachan, Pengtao Xie, Mrinmaya Sachan, and Eric P Xing. Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition, 2017.

[32] Zafar Saeed, Rabeeh Ayaz Abbasi, Onaiza Maqbool, Abida Sadaf, Imran Razzak, Ali Daud, Naif Radi Aljohani, and Guandong Xu. What's happening around the world? a survey and framework on event detection techniques on twitter. *Journal of Grid Computing*, pages 1–34, 2019.

[33] Zafar Saeed, Rabeeh Ayaz Abbasi, Imran Razzak, Onaiza Maqbool, Abida Sadaf, and Guandong Xu. Enhanced heartbeat graph for emerging event detection on twitter using time series networks. *Expert Systems with Applications*, 2019.

[34] Zafar Saeed, Rabeeh Ayaz Abbasi, Muhammad Imran Razzak, and Guandong Xu. Event detection in twitter stream using weighted dynamic heartbeat graph approach. *arXiv preprint arXiv:1902.08522*, 2019.

[35] Zafar Saeed, Rabeeh Ayaz Abbasi, Abida Sadaf, Muhammad Imran Razzak, and Guandong Xu. Text stream to temporal network-a dynamic heartbeat graph to detect emerging events on twitter. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 534–545. Springer, 2018.

[36] Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, Jul 2019.

[37] Noah A. Smith. Contextual word representations: A contextual introduction, 2019.

[38] Shane Storks, Qiaozi Gao, and Joyce Y. Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches, 2019.

[39] Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international*, 2014:240403, 03 2014.

[40] Manabu Torii, Zhangzhi Hu, Cathy Wu, and Hongfang Liu. Biotagger-gm: A gene/protein name recognition system. *Journal of the American Medical Informatics Association : JAMIA*, 16:247–55, 03 2009.

[41] Xu Wang, Chen Yang, and Renchu Guan. A comparative study for biomedical named entity recognition. *International Journal of Machine Learning and Cybernetics*, 9(3):373–382, Mar 2018.

[42] Xuan Wang, Yu Lin Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis P. Langlotz, and Jiawei Han. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35 10:1745–1752, 2018.

[43] Kai Xu, Zhenguo Yang, Peipei Kang, Qi Wang, and Wenyin Liu. Document-level attention-based bilstm-crf incorporating disease dictionary for disease named entity recognition. *Computers in biology and medicine*, 108:122–132, 2019.

[44] Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models, 2019.

[45] Lin Yao, Hong Liu, Yi Liu, Xinxin Li, and Muhammad Anwar. Biomedical named entity recognition based on deep neutral network. *International Journal of Hybrid Information Technology*, 8:279–288, 08 2015.

[46] Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinformatics*, 20(S10), May 2019.

[47] Donghuo Zeng, Chengjie Sun, Lei Lin, and Bingquan Liu. Lstm-crf for drug-named entity recognition. *Entropy*, 19(6), 2017.

[48] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. In *Scientific Data*, 2019.

[49] Henghui Zhu, Ioannis C. Paschalidis, and Amir M. Tahmasebi. Clinical concept extraction with contextual word embedding. *NIPS Machine Learning for Health Workshop*, Dec 2018.