2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) | 978-1-6654-9810-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/INISTA55318.2022.9894270

Biomedical Named Entity Recognition Using Transformers with biLSTM + CRF and Graph Convolutional Neural Networks

Gökberk ÇELİKMASAT
Computer Engineering Department
Marmara University
Istanbul, Turkey
gokberkcelikmasat@marun.edu.tr

Muhammed Enes AKTÜRK
Computer Engineering Department
Marmara University
Istanbul, Turkey
enesakturk@marun.edu.tr

Yunus Emre ERTUNÇ
Computer Engineering Department
Marmara University
Istanbul, Turkey
yunusertunc@marun.edu.tr

Abdul Majeed ISSIFU
Computer Engineering Department
Marmara University
Istanbul, Turkey
abdul.majeed@marun.edu.tr

Murat Can GANİZ
Computer Engineering Department
Marmara University
VeriUs Teknoloji
Istanbul, Turkey
murat.ganiz@marmara.edu.tr

***Abstract*—One of the applications of Natural Language Processing (NLP) is to process free text data for extracting information. Information extraction has various forms like Named Entity Recognition (NER) for detecting the named entities in the free text. Biomedical named-entity extraction task is about extracting named entities like drugs, diseases, organs, etc. from texts in medical domain. In our study, we improve commonly used models in this domain, such as biLSTM+CRF model, using transformer based language models like BERT and its domain-specific variant BioBERT in the embedding layer. We conduct several experiments on several different benchmark biomedical datasets using a variety of combination of models and embeddings such as BioBERT+biLSTM+CRF, BERT+biLSTM+CRF, Fasttext+biLSTM+CRF, and Graph Convolutional Networks. Our results show a quite visible, 4% to 13%, improvements when baseline biLSTM+CRF model is initialized with pretrained language models such as BERT and especially with domain specific one like BioBERT on several datasets.**

***Index Terms*—Natural Language Processing, Named Entity Recognition, Biomedical, LSTM, CRF, GCN**

I. Introduction

Medical or biomedical data comes in various forms, especially in an unstructured text form which makes information extraction difficult and tedious. Medical text data processing is a cross-disciplinary study of computer science and medicine. Structured medical data is needed to productively inspect and mine via using existing analytic methods. Information extraction plays a crucial role, especially in the medical field. Extracting valuable knowledge from these records presents a challenging task since text reports, and narratives include both highly domain-specific terminologies, including jargon, acronyms, and special medical terms.

Named Entity Recognition (NER) is an information extraction method and has practical applications in medical or biomedical sector. The primary objective of NER is to classify words or phrases in the text into categories like person (PER), location (LOC), organization (ORG), etc. In the context of the medical domain, this task aims to identify medical named entities such as diseases, species, drugs, chemicals, genes, cell-type, RNA, and proteins. NER is a supervised sequence classification task that requires labeled data. Supervised machine learning (including deep learning) models are used in this domain. In this paper, we experiment with several different machine learning models including Conditional Random Fields (CRF) for sequence classification, CRF on the top of the bi-directional Long Short-Term Memory (biLSTM) type deep learning models, Bidirectional Encoder Representations from Transformers (BERT), and several contextual or static, generic or domain-specific word embedding models for biLSTM+CRF, and lastly Graph Convolutional Neural Networks (GCN). Our results show that deep learning algorithms such as biLSTM+CRF can achieve state-of-the-art results if combined with powerful word representations obtained from pre-trained language models in a specific domain such as BioBERT. The sections in this paper are organized as follows: Section II presents an overview of the current research and state-of-the-art in named entity recognition. In Section III we go through our approach and the followed methodology for solving the problem. Section IV presents the results of the experiments and highlights the best and worst-performing models and the differences between them. Finally, we conclude the findings of this study in Section V with a quick summary of the potential directions and improvements that could be done in the future.

II. Related Work

Conditional random field (CRF) is a sequence labeling type machine learning algorithm traditionally used for Named

978-1-6654-9810-4/22/31.002022$IEEE$

Entity Recognition (NER) [1] which used to be the state-of-the-art (SOTA) results in many benchmark datasets before the wide applicability of the deep learning models. Even today it is relevant and is used with deep learning models such as LSTM.

NER is considered one of downstream tasks of NLP [2]. Medical text contains vital information which comes in the form of clinical notes, patient discharge reports, medical reports, etc. NER models provide the underlying semantics in these clinical texts, which are then used as input to clinical entity resolving [3] and de-identification of the sensitive or personal data [4] as described by Protected Health Information (PHI) standards. NER and deep learning at large, in the medical field goes beyond research. It now helps to improve patient-based service, detection of early spreading of diseases, creating new perspectives on disease mechanisms to provide better treatment methods [5], [6], [7]. Recent methods for NER can be divided into two; traditional Machine Learning (ML) based and Deep Learning (DL) based. Traditional ML methods in this domain includes CRF [8], Hidden Markov Models (HMM), and Maximum entropy Markov models (MEMMs) [9], and feature-based supervised learning approaches [10] [11].

As part of the development of effective NLP tools to assist the arrangement, curating, and retrieval of information in the biomedical domain, B. Settles used CRF for Biomedical NER [1]. Though these methods are commonly used, the emergence of DL methods which use several neural networks provides more robust and better results in many NLP tasks including NER. One of the main difference of DL models is the ability to learn and discovers features automatically, in other words learning the task-specific efficient representations in hidden layers [12]. Based on this we select biLSTM+CRF as a baseline model for our work. In addition, LSTM-based networks are proven to be effective in sequence labeling problems. Unlike LSTM, Bi-directional LSTMs [13] does capture both previous and next context information making it a more robust model to label a token.

In 2018, Google developed a language model BERT [14] and open-sourced the pre-trained and fine-tuned versions of their work, which has proved to be SOTA in many NLP downstream tasks. By using transfer learning, in the medical domain BERT can be fined tuned on domain specific data to form variety of models [15]. Issifu et el. [16] in their work uses domain-specific language model bioBERT [15] with text data augmentation methods to increase the performance of NER Models.

Researchers combine deep learning models such as biLSTM [17] or a more recent transformer-based language model BERT [14] with CRF to achieve better results [18]. Huang et el. [17] use a 50-dimensional vector senna[1] word embedding with a combination of biLSTM+CRF.

Just like transformer models which are built on general attention models, Graph Convolutional Networks (GCNs) have gained popularity recently in the NLP research arena, especially in text classification tasks. The work of Alberto Cetoli et al. [19], inspired by the work of Diego Marcheggiani et el. [20], shows that GCN can be combined with biLSTM+CRF models to get promising results in NER tasks. Even though their work falls short of the SOTA results, the study suggests that the GCN model can improve over the biLSTM+CRF baseline.

In [21], authors suggest the use of multi-lingual pre-trained BERT models with biLSTM+CRF so that they can use English and Spanish biomedical datasets. They evaluate and compare the performance of the biLSTM+CRF model and a multi-lingual BERT model in a general domain and biomedical for both Spanish and English datasets.

It is possible to fine-tune a transformer models such as BERT [14] to a particular application or domain. This usually increase the performance of the downstream application. This transfer learning is one of the most powerful aspects of these models. One specific example is the BioBERT [15]. BioBERT initializes with the weights of a pre-trained BERT model that is trained using a general corpus, then uses a large biomedical corpora to fine-tune the weights. This domain-specific model outperforms the BERT model in several downstream tasks including text classification and NER [15]. In biomedical benchmark datasets, BioBERT outperforms BERT with an F1 score of 0.62% improvement

Another interesting development in the NLP domain is the development of multilingual embedding models which represents the words of many different languages in the same embedding space. The popular examples of such models are multilingual fastText [22] and Multilingual Bert [23]. These are important because they can allow us to train a machine learning model by using a training set in one language and using that model to classify instances in another language. This may prove important for low-resource languages such as Turkish in specific domains such as the medical domain.

III. Approach

In our approach we used different word representations such as fastText embeddings, contextual embeddings obtained from pre-trained BERT model, and medical BioBERT model with the biLSTM+CRF architecture to increase the performance of the baseline model on biomedical benchmark dataset.

In supervised learning, data inputs are composed of both the examples and their respective labels. NER is one of the supervised learning tasks that need labeled data to train a model. In this study, we trained several machine learning / deep learning based supervised NER models on four biomedical benchmark datasets which are available publicly. The datasets are NCBI-disease corpus [24], BC5CDR (BioCreative V CDR corpus) [25], JNLPBA [26], Species-800 [27], and BC2GM [28] datasets. We use the pre-processed version of all datasets that are described in Lee et al [15].

The current studies use transformer models directly for NER tasks. We hypothesize that a combined model of deep learning and traditional machine learning, biLSTM+CRF can

[1] Downloaded from https://ronan.collobert.com/senna/

perform comparably or better. So in this study, we use biLSTM+CRF models that are initialized using different type of embeddings. We use two different contextual embeddings that are produced by pre-trained general domain and a fine-tuned domain-specific transformer model (i.e. BERT and BioBERT). We also use static embedding method fastText that is pre-trained using a general corpus. These models are named as BERT+biLSTM+CRF, BioBERT+biLSTM+CRF and biLSTM+CRF, respectively. Our architecture for the transformer model + biLSTM + CRF is depicted in figure 4.

The biLSTM model consists of two LSTM layers, which processes the sequence in the writing direction and the opposite direction. The ability to work with sequential input and the capacity to recognize long-term dependencies because of a memory cell are the fundamental characteristics of the LSTM. The hidden states vectors $H = (h_1, h_2, \ldots, h_n)$ are the output of the LSTM, which receives a series of vectors $X = (x_1, x_2, \ldots, x_n)$ as input.

In biLSTM, x_t, h_t, and c_t are input at time t, hidden state at time t, and cell state at time t, respectively. Furthermore, σ represents the sigmoid function.

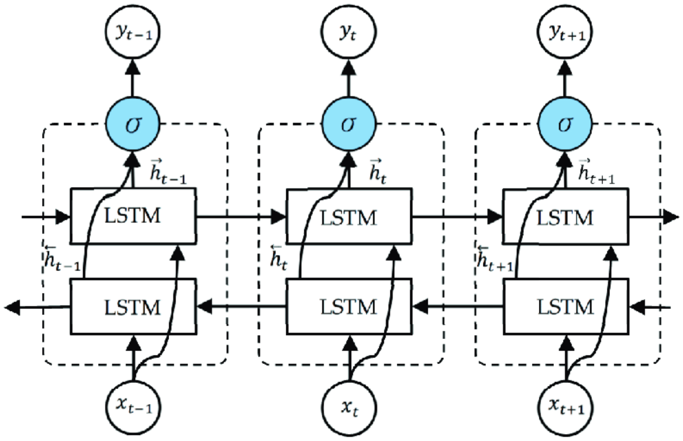

Fig. 1. A General Architecture for biLSTM algorithm (Adapted from [29]).

BiLSTM+CRF model is one of the most popular models used for named entity recognition problems [30]. A CRF model makes a prediction as a graphical model in order to account for the influence of nearby data. A linear chain CRF can make predictions utilizing this improved context after receiving the biLSTM's output. This CRF and biLSTM combination is referred to as a biLSTM+CRF model [31], and its architecture is shown in Figure 2.

A Graph Convolutional Network (GCN) [33] is a type of convolutional neural network (CNN) that efficiently make use of the connections or relations between instances that can be represented as a graphe. GCN creates embedding vectors of the nodes of the input graph based on connectivity of the nodes. Hence, the GCN model requires data to be represented as a graph. In our case a graph is created by representing words or terms as nodes and their pair-wise similarities as edge weights. We use Point-wise Mutual Information (PMI) for measuring similarities. A graph is usually stored and processed as an Adjacency Matrix (A).

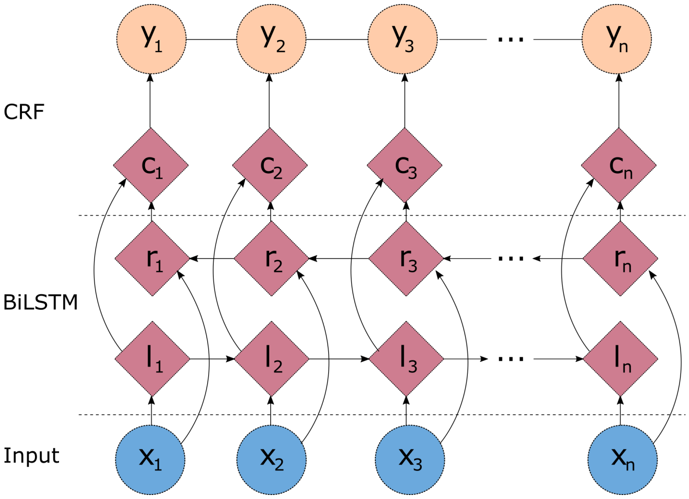

Fig. 2. BiLSTM-CRF structure (Adapted from [32]).

The normalized version of the adjacency matrix A of the graph (A*) will be added as an input to the forward pass in Eq. (1). In this way, the model can learn feature representations based on the connectivity of instances. Bias b is omitted for clarity's sake. We can think about the GCN as a first-order approximation of Spectral Graph Convolution in the form of a message-passing network, in which the information is propagated along the graph's neighboring nodes as explained in more detail in [34].

$$H^{[i+1]} = \sigma(W^{[i]} H^{[i]} A^*) \tag{1}$$

After normalizing the matrix by placing the parameters we have in the $H^{[i+1]}$ (in Fig. (1)) the layer of the neural network, the 2-layer neural network is run with the forward pass method and the results are calculated. The architecture of the two layer GCN model represented in Figure 3.

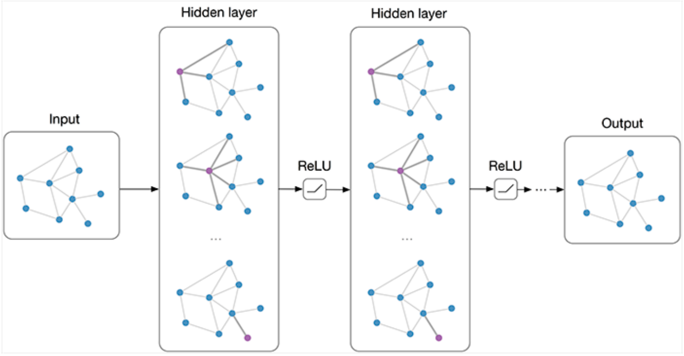

Fig. 3. Two layer Graph Convolutional Network architecture.

The architecture of the biLSTM+CRF model with BERT embedding is shown in Figure 4. In this architecture, the second layer consist of a Bidirectional Long Short-Term

Memory (biLSTM) model as a black-box, and it make use of the natural sequence structure of the words in sentences. On the the top of that, Conditional Random Field (CRF) layer is used for sequence labeling, that is assigning one of the predetermined named entity class label to a particular word in the sequence.

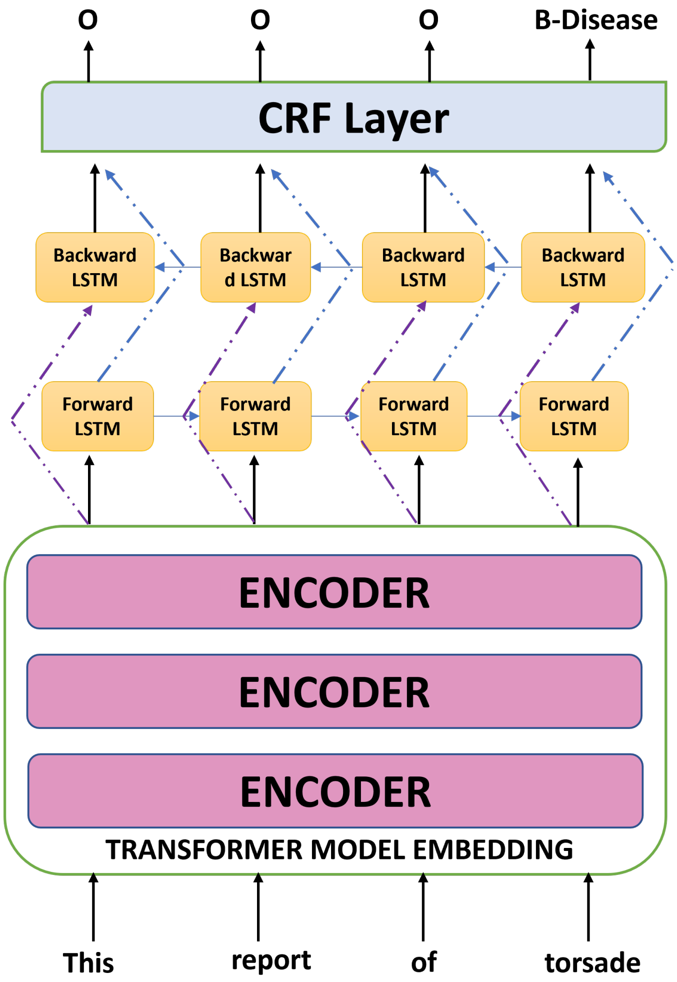

Fig. 4. BiLSTM+CRF model with Transformer embedding architecture (Adapted from [35]).

As mentioned, the very first layer of the biLSTM+CRF is the input layer which can take the vector representations of individual terms. A popular approach is to use short and dense vector representations of terms using pre-trained embedding models. Transformer based large language models such as BERT [14] can be used to obtain contextual embeddings of words. As a result, we use different BERT models in our experiments. The BERT model uses bidirectional self-attention, and the model we use has 12 transformer blocks. Each transformer block has 768 layers with 12 attention heads. The vocabulary size parameter is set to 110,000. The BERT algorithm can be used for pre-training from the scratch or fine-tuning an existing pre-trained model for a specific domain or application. The model is trained using usually a very large amount of unlabeled data in pre-training. On the other hand, a pre-trained model can be fine-tuned using labeled data for a specific downstream task or domain specific unlabeled data. A downstream task training for a model employing BERT will have its own fine-tuned model at the end. BioBERT is a domain specific variation of BERT. It is pre-trained in a bidirectional way using a large-scale biomedical corpus. We use this models embeddings to the biomedical NER task.

We trained our GCN model from scratch using the medical datasets mentioned at the beginning of this section. The GCN provides message passing between nodes. In our case we use a two-layer GCN which means the message passing can be between nodes that are at most two steps away. This is interesting because although there are no direct document-document connections or edges in the graph, due to its two-layer structure it may facilitate information exchange between pairs of documents.

For the training set, development set and test set split we use the settings provided by the datasets. We use several evaluation metrics in our experiments. For precision, recall, and F1 we use macro-averaging.

IV. Experiment Results and Discussion

To establish a baseline, we start with the simplest model, which is the traditional machine learning method of Conditional Random Fields (CRF) for Named Entity Recognition (NER). After building biLSTM+CRF model, we use biLSTM+CRF that is initialized using embeddings from a pre-trained BERT model from general domain and BioBERT model from biomedical domain. We also use pre-trained fastText embeddings with biLSTM+CRF. The above mentioned models have following parameters: Loss Function is ”neg-log-likelihood-loss”, optimizer is ”Adam”, and learning rate is ”1e-4”. We got optimal experiment results by using 10 epochs with a batch size of 32.

We also experiment with GCN models. Like biLSTM+CRF, we apply GCN directly to sequence labeling applications in particular NER. Please note that we don't use a pre-trained embedding model with GCN.

TABLE I
Performance Comparisons of Models on JNLPBA

Method	Accuracy	Precision	Recall	F1
GCN	83.87	61.06	55.28	57.16
CRF	85.07	65.91	57.03	60.47
biLSTM+CRF	89.90	74.10	81.30	77.27
FastText+biLSTM+CRF	81.56	33.31	34.15	32.14
BERT+biLSTM+CRF	93.32	82.04	**85.14**	**83.52**
BioBERT+biLSTM+CRF	**93.40**	**86.41**	78.53	82.02

From the result tables, we can clearly see our biLSTM+CRF model with BERT embedding generally outperforms the other models in comparison. Other models use attention based (Att + biLSTM + CRF) approach [30] or Glove word embedding [36] or other word embeddings, but in the medical domain, like in this research, BERT and BioBERT word embeddings are the best choice since we use medical corpora.

TABLE II
PERFORMANCE COMPARISONS OF MODELS ON S800

Method	Accuracy	Precision	Recall	F1
GCN	96.32	48.82	37.36	39.34
CRF	95.68	67.75	37.81	40.70
biLSTM+CRF	96.54	73.25	61.03	65.94
FastText+biLSTM+CRF	94.28	33.41	33.23	33.25
BERT+biLSTM+CRF	96.94	74.98	74.01	74.42
BioBERT+biLSTM+CRF	**97.37**	**78.46**	**77.54**	**77.99**

TABLE III
PERFORMANCE COMPARISONS OF MODELS ON NCBI-DISEASE

Method	Accuracy	Precision	Recall	F1
GCN	92.53	70.51	42.65	47.88
CRF	91.91	67.27	43.00	47.62
biLSTM+CRF	**96.53**	72.32	65.46	68.38
FastText+biLSTM+CRF	89.28	39.72	35.72	37.01
BERT+biLSTM+CRF	95.99	81.70	**82.15**	81.93
BioBERT+biLSTM+CRF	96.25	**85.80**	78.86	**82.06**

TABLE IV
PERFORMANCE COMPARISONS OF MODELS ON BC5CDR-DISEASE

Method	Accuracy	Precision	Recall	F1
GCN	94.55	52.36	37.28	38.79
CRF	94.15	54.01	35.44	36.38
biLSTM + CRF	94.62	63.92	65.42	64.52
FastText+biLSTM+CRF	93.92	38.47	33.52	32.89
BERT + biLSTM + CRF	96.32	**76.98**	72.49	74.55
BioBERT+biLSTM+CRF	**96.94**	76.56	**78.26**	**77.39**

TABLE V
PERFORMANCE COMPARISONS OF MODELS ON BC2GM

Method	Accuracy	Precision	Recall	F1
GCN	90.37	61.10	48.72	51.63
CRF	89.90	66.08	43.99	48.56
biLSTM + CRF	92.41	74.39	65.53	69.34
FastText+biLSTM+CRF	89.37	30.97	33.32	31.51
BERT + biLSTM + CRF	94.50	81.63	77.12	79.22
BioBERT+biLSTM+CRF	**95.10**	**83.46**	**80.32**	**81.81**

The GCN model was trained for 500 epochs for s800, NCBI-disease, and JNLPBA datasets. The results on the JNLPBA dataset show 57% F1 which is a significant increase in the performance.

By optimizing the model and increasing the epochs in training, one may achieve better results. We reserve the GCN model as a viable option for NER.

Our experiment results shows that BioBERT+biLSTM+CRF model generally outperforms other models for the NER task in almost all of the datasets. The results of these datasets can be seen in Table II for s800 dataset, Table III for NCBI-Disease dataset, Table IV for BC5CDR-Disease dataset and Table V for BC2GM dataset. There is an exception for BERT+biLSTM+CRF results that can be seen in Table I for the JNLPBA dataset, in this case, BERT pre-trained model from a general corpus outperforms other models for Recall and F1 metrics. These are interesting results that require further investigation.

As can be seen in all tables, and as expected our simplest model, CRF has lower performance compared to the more complicated models of biLSTM+CRF with different embeddings (BERT, and BioBERT). One exception to this is FastText+biLSTM+CRF. A generic FastText pre-trained model decreases the performance of biLSTM+CRF considerably, showing the lowest results for all our datasets. This also requires further investigation.

It is important to underline that we reach the state-of-the-art (SotA) result in the JNLPBA dataset, our BERT+biLSTM+CRF model achieves an 83.52% F1 score and the BioBERT+biLSTM+CRF model achieves 82.02% F1 score can be seen in Table I. Both of the models outperform the current SotA KeBioLM [37] of 82.0% F1 score and are also better than the original BioBERT of 77.59% F1 score.

Also in the s800 dataset, we got the SotA using our BioBERT+biLSTM+CRF model with an F1 score of 77.99% can be seen in Table II, outperforming the current SotA of SciFive-Base [38] 76.55% F1 score.

We have mixed results for GCN. GCN usually performs poorly and outputs lower results than our baseline model of CRF. Only on NCBI-disease dataset GCN can catch CRF as can be seen in Table III. However, we use a vanilla GCN and believe that it can be modified to have a better performance for this task.

The biLSTM+CRF model only, without initializing the embedding layer with a pre-trained model embedding layer such as BERT, BioBERT or FastText, as expected performs relatively poorly.

V. CONCLUSIONS AND FUTURE WORK

In the medical domain, domain-adapted transformer models such as BioBERT [15] has the state-of-the-art results for Named Entity Recognition (NER). In our study, we focus on improving the biLSTM+CRF using a variety of word embedding approaches including the BioBERT. We also attempt to use Graph Convolutional Networks (GCN) in this domain; biomedical named entity recognition. We conduct extensive experiments with several different biomedical datasets and several different models. Our results show that domain specific large language models based on transformers such as BioBERT can also be used to improve the performance of biLSTM+CRF, a deep learning based NER model, when used in the embedding layer. Furthermore, even the general domain BERT shown to be useful and as expected, works better than static embeddings such as fastText. Interestingly, we could be able to achieve and surpass state-of-the-art results of domain specific language models based on transformers using our BioBERT+biLSTM+CRF setting on some of the datasets.

For future work, we shall explore more medical domain embeddings based on transformers such as BioELECTRA [39] , SciBERT [40], and static embeddings such as BioWordVec [41], and cui2vec [42]. With this domain specific embeddings +biLSTM+CRF, we think there will be more interesting results. We also plan to significantly improve the GCN model by modifying it for NER. We will also apply our current solutions for multilingual cases. By using multilingual models we can apply our solution to datasets in different languages.

References

[1] B. Settles, “Biomedical named entity recognition using conditional random fields and rich feature sets,” in *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, pp. 107–110, 2004.

[2] V. Yadav and S. Bethard, “A survey on recent advances in named entity recognition from deep learning models,” *arXiv preprint arXiv:1910.11470*, 2019.

[3] D. Tzitzivacos, “International classification of diseases 10th edition (icd-10),” *CME: Your SA Journal of CPD*, vol. 25, no. 1, pp. 8–10, 2007.

[4] Ö. Uzuner, Y. Luo, and P. Szolovits, “Evaluating the state-of-the-art in automatic de-identification,” *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 550–563, 2007.

[5] K. Y. He, D. Ge, and M. M. He, “Big data analytics for genomic medicine,” *International journal of molecular sciences*, vol. 18, no. 2, p. 412, 2017.

[6] A. Kankanhalli, J. Hahn, S. Tan, and G. Gao, “Big data and analytics in healthcare: Introduction to the special section,” *Information Systems Frontiers*, vol. 18, no. 2, pp. 233–235, 2016.

[7] B. Ristevski and M. Chen, “Big data analytics in medicine and healthcare,” *Journal of integrative bioinformatics*, vol. 15, no. 3, 2018.

[8] J. Lafferty, A. McCallum, and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” 2001.

[9] A. McCallum, D. Freitag, and F. C. Pereira, “Maximum entropy markov models for information extraction and segmentation.,” in *Icml*, vol. 17, pp. 591–598, 2000.

[10] J. Li, A. Sun, J. Han, and C. Li, “A survey on deep learning for named entity recognition,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2020.

[11] M. Taşpınar, M. C. Ganiz, and T. Acarman, “A feature based simple machine learning approach with word embeddings to named entity recognition on tweets,” in *International Conference on Applications of Natural Language to Information Systems*, pp. 254–259, Springer, 2017.

[12] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[13] B. Y. Lin, F. F. Xu, Z. Luo, and K. Zhu, “Multi-channel bilstm-crf model for emerging named entity recognition in social media,” in *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 160–165, 2017.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[15] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[16] A. M. Issifu and M. C. Ganiz, “A simple data augmentation method to improve the performance of named entity recognition models in medical domain,” in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pp. 763–768, IEEE, 2021.

[17] Z. Huang, W. Xu, and K. Yu, “Bidirectional lstm-crf models for sequence tagging,” *arXiv preprint arXiv:1508.01991*, 2015.

[18] A. O. B. Sapci, O. Tastan, and R. Yeniterzi, “Focusing on possible named entities in active named entity label acquisition,” *arXiv preprint arXiv:2111.03837*, 2021.

[19] A. Cetoli, S. Bragaglia, A. D. O’Harney, and M. Sloan, “Graph convolutional networks for named entity recognition,” *arXiv preprint arXiv:1709.10053*, 2017.

[20] D. Marcheggiani and I. Titov, “Encoding sentences with graph convolutional networks for semantic role labeling,” *arXiv preprint arXiv:1703.04826*, 2017.

[21] R. M. Rivera-Zavala and P. Martínez, “Analyzing transfer learning impact in biomedical cross-lingual named entity recognition and normalization,” *BMC bioinformatics*, vol. 22, no. 1, pp. 1–23, 2021.

[22] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” *arXiv preprint arXiv:1710.04087*, 2017.

[23] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual bert?,” *arXiv preprint arXiv:1906.01502*, 2019.

[24] R. I. Doğan, R. Leaman, and Z. Lu, “Ncbi disease corpus: a resource for disease name recognition and concept normalization,” *Journal of biomedical informatics*, vol. 47, pp. 1–10, 2014.

[25] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, “Biocreative v cdr task corpus: a resource for chemical disease relation extraction,” *Database*, vol. 2016, 2016.

[26] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. 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*, pp. 70–75, Citeseer, 2004.

[27] E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen, “The species and organisms resources for fast and accurate identification of taxonomic names in text,” *PloS one*, vol. 8, no. 6, p. e65390, 2013.

[28] L. Smith, L. K. Tanabe, C.-J. Kuo, I. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, M. Torii, *et al.*, “Overview of biocreative ii gene mention recognition,” *Genome biology*, vol. 9, no. 2, pp. 1–19, 2008.

[29] Y.-H. Li, L. N. Harfiya, K. Purwandari, and Y.-D. Lin, “Real-time cuffless continuous blood pressure estimation using deep learning model,” *Sensors*, vol. 20, no. 19, p. 5606, 2020.

[30] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, “An attention-based bilstm-crf approach to document-level chemical named entity recognition,” *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, 2018.

[31] “BiLSTM-CRF network bilstm-crf network for ner.” https://blog.dominodatalab.com/named-entity-recognition-ner-challenges-and-model. Accessed: 2022-05-14.

[32] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural architectures for named entity recognition,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 260–270, Association for Computational Linguistics, June 2016.

[33] R. v. d. Berg, T. N. Kipf, and M. Welling, “Graph convolutional matrix completion,” *arXiv preprint arXiv:1706.02263*, 2017.

[34] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.

[35] H. Yang and W. H. Hsu, “Named entity recognition from synthesis procedural text in materials science domain with attention-based approach.,” in *SDU@ AAAI*, 2021.

[36] H. Poostchi, E. Z. Borzeshi, and M. Piccardi, “Bilstm-crf for persian named-entity recognition armanpersonercorpus: the first entity-annotated persian dataset.,” in *LREC*, 2018.

[37] Z. Yuan, Y. Liu, C. Tan, S. Huang, and F. Huang, “Improving biomedical pretrained language models with knowledge,” *arXiv preprint arXiv:2104.10344*, 2021.

[38] L. N. Phan, J. T. Anibal, H. Tran, S. Chanana, E. Bahadroglu, A. Peltekian, and G. Altan-Bonnet, “Scifive: a text-to-text transformer model for biomedical literature,” *arXiv preprint arXiv:2106.03598*, 2021.

[39] K. raj Kanakarajan, B. Kundumani, and M. Sankarasubbu, “Bioelectra: pretrained biomedical text encoder using discriminators,” in *Proceedings of the 20th Workshop on Biomedical Language Processing*, pp. 143–154, 2021.

[40] I. Beltagy, K. Lo, and A. Cohan, “Scibert: A pretrained language model for scientific text,” *arXiv preprint arXiv:1903.10676*, 2019.

[41] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, “Biowordvec, improving biomedical word embeddings with subword information and mesh,” *Scientific data*, vol. 6, no. 1, pp. 1–9, 2019.

[42] A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai, and I. S. Kohane, “Clinical concept embeddings learned from massive sources of multimodal medical data,” in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pp. 295–306, World Scientific, 2019.