Original Research Paper

Volume - 13 | Issue - 03 | March - 2023 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Engineering

CONTEXTUAL LEARNING APPROACH FOR SYNDROME DISEASE NAMED ENTITY RECOGNITION

Dr. E. Uma*	Assistant Professor Sl. Gr, College of Engineering Guindy, Anna University. *Corresponding Author
Kamatchi K	College of Engineering Guindy, Anna University Research Scholar
Mehala Elangovan	Annamalai University Assistant Professor.

ABSTRACT The groundwork for extracting a significant amount of biomedical information from unstructured texts into structured formats is the difficult research area of biological entity recognition from medical documents. The existing work implemented the named entity recognition for diseases using the sequence labelling framework. The performance of this strategy, however, is not always adequate, and it frequently cannot fully exploit the semantic information in the dataset. The Syndrome Diseases Named Entity problem is presented in this work as a sequence labelling with multi-context learning. By using well-designed text/queries, this formulation may incorporate more previous information and to decode it using decoding techniques such conditional random fields (CRF). We performed experiments on three biomedical datasets, and the outcomes show how effective our methodology is on the BC5CDR-Disease, JNLPBA and NCBI-Disease, compared with other techniques our methodology performs with accuracy levels of 96.70%,98.65 and 96.72% respectively.

KEYWORDS : sequence labelling, context learning, named entity recognition, Gated recurrent unit, Conditional random field.

INTRODUCTION

The goal of biomedical named entity recognition (Bio-NER) is to detect biological entities automatically in provided texts. The requirement for extracting biomedical knowledge stored in unstructured texts and converting them into structured formats is the ability to recognise biological items effectively. The Bio-NER objective thus has substantial scientific value. Traditionally, Bio-NER techniques have relied on carefully engineered features, or the construction of features using a variety of natural language processing (NLP) technologies and subject-matter expertise. DNorm [1], TaggerOne [3] and others are typical examples of such models utilised in the biomedical field. However, the creation of features requires a lot of domain expertise and entity features.

Neural networks with autonomous feature learning capabilities have recently gained popularity in NER tasks [5,6]. Numerous neural network techniques [7–12] have been proposed to recognise biomedical things in the biomedical domain. These techniques often use conditional random fields (CRF) [14] as the input after learning vector representations of each word or token in a phrase using bidirectional long short-term memory (Bi-LSTM) [13]. Language models (such as ELMo [15] and BERT [16]) have just lately achieved state-of-the-art (SOTA) performance on a variety of NLP tasks. The Lee et al method [17] used the SoftMax function and Bio-BERT to achieve SOTA outcomes on various biomedical datasets in order to recognise biological items in the biomedical domain.

Neural network approaches can perform more effectively than feature engineering techniques because they can automatically learn features. By developing a model for sequence labelling to give each token in a given sequence a label, the present techniques typically formalise the Bio-NER issue as a sequence labelling problem. But neither of the models outlined above BERT nor Bio-BERT-Soft-max can successfully learn the semantic information contained in the framework for sequence tagging. When compared to language models, BERT's performance is inadequate [17]. It is challenging for Bio-BERT to employ the semantic knowledge acquired by the system's final layer in the framework for sequence labelling [18]. The present inclination of formalising NLP tasks into machine reading comprehension tasks served as inspiration [19–23]. The recognition of illness named entities has been studied in the literature using a variety of methods, including rule-based, conventional, and deep learning techniques [1, 2]. The classic machine learning techniques and rule-based systems both largely rely on domain knowledge and hand-crafted rules or features. Since they rely heavily on manual involvement, such systems are typically not scalable. Because deep learning approaches can automatically extract features from clinical text to generate meaningful representations, they have recently become more popular.

In order to accomplish this, we propose a syndrome disease NER model for medical trial texts that combines intense contextual embeddings with pertinent domain-specific features, word embeddings, and character embeddings in a framework called Gated Recurrent Neural Network-Conditional Random Field (GRU-CRF). The primary contributions are listed below: By incorporating pertinent domain-specific characteristics, word embeddings, and character embeddings into a Gate recurrent neural network-conditional random field architecture, a disease NER model for clinical trial texts is created, as well as sentence-level contextual information via deep context embeddings learned with language models [4]. Using data from clinical trials, compare the model's performance to other disease NER models that are already in use. To our knowledge, this is the first deep learning-based model that has been suggested for specifically extracting clinical concepts from clinical trial text. Experimental findings reveal that our model outperforms current state-of-the-art methodologies, and additional qualitative research demonstrates the model's efficacy for clinical trial texts.

RELATED WORK:

Deep learning models have become more well-liked recently and have been successfully used in the biomedical NLP sector. The disease NER challenge has been performed using a variety of deep learning models, including convolutional neural networks (CNN) and recurrent neural networks (RNN) [2]. Character embeddings produced by stacking the convolution layers of a character-based CNN model were used in [9] proposed disease-named entity recognition model [8]. To enhance the performance of recognition in our suggested model, we apply external domain knowledge embeddings. The capacity of the Bi-LSTM to consider both the forward and backward contexts with respect to the specific references for identifying clinical events such as diseases and therapies has led to its widespread use in the literature [9,10].

Additionally, CRF maximise the chance of witnessing a sentence given a particular set of mentions, which can lead to improved accuracy for the entity recognition test. CRF consider the entire sentence rather than individual word locations. In order to infer named entities from input text, researchers mix Bi-LSTM and CRF networks and take into account both pertinent input features and sentence-level annotation data [11]. A disease NER model that cascades a CNN model with an RNN to provide character embeddings was proposed by [1]. In contrast, we combine character-based CNN and LSTM models with an integrated embedding strategy. For the purpose of extracting biomedical concepts, deep learning techniques have also been demonstrated to be efficient when combined with contextual information and domain expertise [12,13, 14].

A novel language model-based technique called ELMo [4] can be used to produce word embeddings that accurately reflect the context of the words in a phrase. ELMo was studied by Crichton et al. [3], however they did not use domain knowledge embeddings or test the model on clinical trial texts. Although [1] model design is comparable to ours, they did not take contextualised embeddings from the provided corpus

into account. In contrast, to discover better representations of the clinical trial text, we consider both contextual word embeddings and domain knowledge embeddings.

Word embeddings might retrieve the latent syntactic and semantic information of tokens and map these words/tokens into dense low-dimensional vectors using a vast part of unlabeled data. Several word-embedding techniques have been presented over the past ten years, with Word2Vec [26] and GloVe serving as notable examples. Word2Vec either uses the Continuous Bag-Of-Words (CBOW) model to model the current word/token based on the surrounding context, or it uses the Skip-Gram model to predict surrounding words based on the actual word.

GloVe [17] makes excellent use of statistics and both local and global aspects of the corpus by employing a special weighted least squares model that trains on global word-word co-occurrence counts. The word or token taught using these techniques is mapped to a specific vector, though. As a result, word embeddings developed using these techniques can only simulate context-free representations. Language models from the present day, including ELMo [15] and BERT [16], do improve performance on NLP tasks. Contrary to conventional word embeddings like Word2Vec and GloVe, the embedding that the language model assigns to a word or token relies on the context, which means that the same word or token may have a different representation in other scenarios.

To model the contextual development of the input sequence, ELMo [15] combines separately trained left-to-right and right-to-left LSTM. Transformer is used by Bi-LSTM-CRF to jointly condition on both left and right context in all layers in order to pre-train representations. A vast corpus of data is used to pre-train the model, and it is then tuned using the target dataset as a result of its tremendous success.

In order to better portray expert and knowledge-intensive biomedical literature, Chen [30] reformulated the biomedical named entity recognition challenge as reading comprehension that combines realm specific knowledge from UMLS. By using a multi-way screening reader method, specifically they incorporate three different forms of information, including CUI, semantic type, and evidence snippets, to adaptively calculate contextual representations for the sequence, question, and evidence snippets. On most Bio-NER datasets, experimental outcomes are outperformed the state-of-the-art baseline courtesy to domain-specific expertise.

METHODOLOGY

NER can be cast as a sequence-labelling task. In this work, we propose a hybrid Bi-GRU-CRF model to identify disease name mentions. This Gate Recurrent neural network enables considering both forward and backward features in a given sentence and sequential CRF annotates the tags taking the soft-max output of Bi-GRU layer as input. Our model comprises of the following: (i) a feature representation layer, and (ii) a bidirectional GRU-CRF network. Figure 1 depicts the overall architectural layout. We discuss the details of the architecture below.

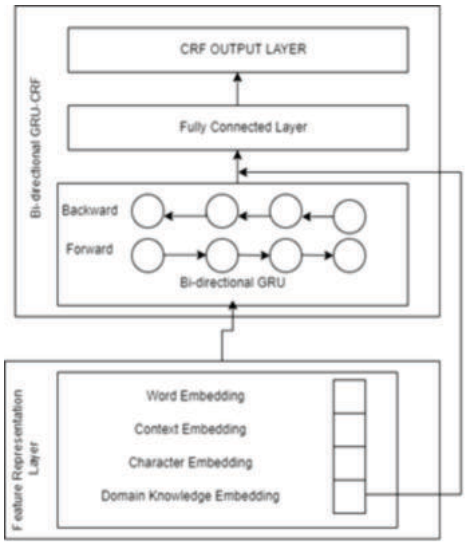

Figure 1: overall architectural layout

Feature Representation Layer

The feature representation layer takes a sequence of input (a_1, a_2, ..., an) containing n words and generates a d dimensional feature vector for each word. Consider several types of embeddings to capture the inherent features of the sentences. Concatenating four different representations word embedding, context embedding, domain knowledge embedding, and character embedding leads to the creation of the d-dimensional feature vector for each phrase. The detailed feature representation layer is shown in figure 2

Domain Knowledge Embedding

Inspired by [1], the domain knowledge dk embedding is obtained from two sources: clinical vocabulary and a hybrid clinical NLP engine [15]. The domain knowledge embedding is denoted as Fdk = [Fclin; Ftag]. We combine numerous dictionaries of diseases, including MEDIC, UMLS, and others to obtain a rich clinical vocabulary to generate the lexicon embeddings. We build a trie like data structure for efficient access of the vocabulary. Generally, trie like data structures are preferable to store words of a dictionary such that adding, modifying, and querying the words become efficient. Such structure also stores tags associated with each word in the vocabulary. Therefore, a given sentence can be easily searched in the trie dictionary tree and corresponding BIO sequence tags can be annotated automatically. Transform the BIO tags to generate lexicon embeddings Fclin. The clinical NLP engine uses a syntactic parser and clinical ontologies such as SNOMED-CT to provide tags. Then, we generate the external tagging embeddings Ftag based on the sequence of tags provided by the NLP engine.

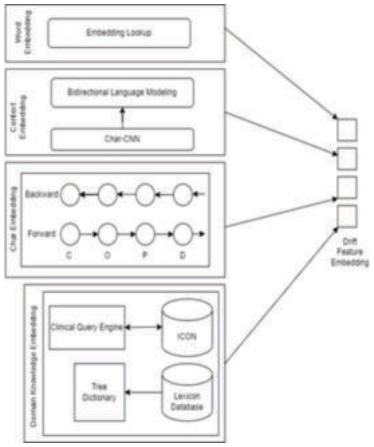

Figure 2: elaborate feature representation layout

Character Embedding

Following [1], Generate a character embedding vector, F_{char}, which is formed by embedding vectors charGRU (V_{GRU}). The model takes words as input, then first looks up in the character embedding matrix $P_d \times |C|$ (where the embedding vector dimension is d) and forms the embedding matrix, Ch_k. The matrix, Ch_k is then convoluted with multiple kernel matrices. Thereafter, we apply a pooling operation to get the final fixed-dimensional embedding vector, F_{cnn}. The charGRU architecture comprises a bi-directional GRU layer and takes the sequence of characters in a word to generate the character embedding vector F_{GRU}, which consists of both direction hidden states [$h_{forward}$; $h_{backward}$].

Context Embedding

GloVe-based word embedding mainly relies on word-level co-occurrence statistics. In order to encode context-level information, obtain context-aware word representations using the language model-based method, ELMo [4]. It comprises of a character-based Convolutional Neural Network (char-CNN) and two layer bidirectional-Language Model (bi-LM) to embed contextual information through a highway connection and a low-dimensional projection layer, which are introduced after stacking the char-CNN and bi-GRU layers. Unlike traditional word embedding that represents a stable embedding vector for downstream tasks, ELMo captures contextual information dynamically for each word as each word is represented as a function of the given sentence.

Word Embedding

Obtain word embeddings using the publicly available GloVe1 representation. It generates word embeddings by considering both local context window and the global matrix factorization. Use 400-

dimensional vector representation for each word and denote this embedding as $F_{word\text{-}embedding}$.

In summary, the feature representation layer concatenates the above four embeddings to represent a given sentence. The final feature embedding as $F_{feature} = [F_{word\text{-}embedding}; F_{elmo}; F_{char}; F_{dk}]$. These features are then fed into the bidirectional GRU-CRF layer for tagging the sequence of the clinical trial text.

BIDIRECTIONAL GRU-CRF LAYER

Bidirectional GRU-CRF architecture to predict the corresponding tags: O (O=outside), B-disease (B=beginning), and I-disease (I=intermediate) from a given clinical trial text. In particular, Bidirectional GRU takes a sequence of embedded feature, F_{seq} as input and generates the sequence $y = (m_1, m_2, ..., m_n)$ that represents feature encodings from the embedded features. We denote this encoded feature as $F_{GRU} = [h ; h_{backward}]$. This feature vector, F_{GRU} concatenated with domain knowledge embedding, F_{dk} to generate the input feature vector, $[F_{GRU,} F_{dk}]$ for the fully connected layer. We consider the output of the fully connected layer that is multiplied with corresponding weights, W_f and bias, bf, as input for the CRF layer and finally, the model envisions the most excepted tag sequence.

RESULT

On datasets from the BC5CDR [25], NCBI- Disease [26], and JNLPBA [27], all of which have been pre-processed and tested using our technique. The BC5CDR dataset comprises one of these datasets, and it is used to assess chemical and disease entities, respectively. Because most of the existing methods were evaluated on BC5CDR-Disease respectively, we did the same. Table 1 lists the statistics of these datasets.

The initial training and development set were combined to create a new training set for the trials. Then 10% of the new training set was sampled as the validation set to tune hyper-parameters. The test set was only used to evaluate the model. Most existing works [3,9,12,17] split data in this way, and we also followed this way. Because the limitation of computational complexity, most of the existing works are based on BERT base model [16]. To facilitate comparison with these works, all BERT models in this work are based on the BERT base model.

Table 1 Statistics of BioNER datasets

Dataset	Entity Type	No. of annotations	No of sentences
BC5CDR-Disease	Disease	12,694	14,228
NCBI-Disease	Disease	6,881	7,639
JNLPBA	Disease	35,460	22,562

The performance is measured with the accuracy, whose attributes equal importance to true predicted and false predicted. In this work, each experiment is repeated five times, and we report the maximum accuracy. Moreover, we also exploit T-test to perform statistical significance tests and report the confidence interval. In our experiments, Bi-LSTM CRF reaches its highest performance at 1 or 3 epochs on the BC5CDR-Disease, NCBI and JNLPBA datasets. The reasons are twofold: 1) these data- sets, especially BC5CDR, are large in scale; and 2) Bi-LSTM CRF has powerful feature learning capabilities.

Performance comparison for different models

We examined outstanding BERT models in the biomedical area to examine the impact of model. These models each achieved SOTA performance in their individual works.

Table 2 illustrates the effect of different model performance. Overall, the performance of Bi-GRU-CRF is better than BERT. Compared with BERT is sensitive to uppercase and lowercase characters. This experimental result shows that the uppercase and lowercase character information are useful for SDNER tasks on most datasets. Moreover, we also noticed that the performance of Bi-GRU-CRF shown in Figure 3 (98% in accuracy) is superior to BERT (87% in the accuracy) shown in Figure 4.

Table 2 Performance comparison for different datasets with models.

Dataset	Model	Accuracy
BC5CDR-Disease	BERT	87.56
	Bi-LSTM-CRF	96.70
NCBI-Disease	BERT	85.11
	Bi-LSTM-CRF	96.72
JNLPBA	BERT	78.45
	Bi-LSTM-CRF	98.65

Summary of results are shown in Table 3. In the first illustration, false negatives (FN) were identified by BERT, but they were changed to true positives using Bi-GRU-CRF (TPs). This example shows that SDNER has certain advantages over BERT in terms of learning syntactic information

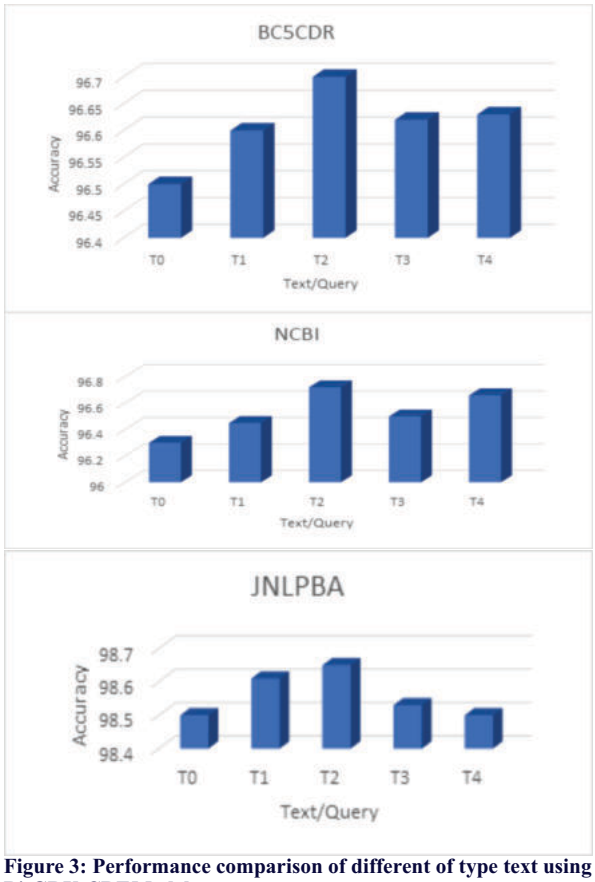

Figure 3: Performance comparison of different of type text using Bi-GRU-CRF Model

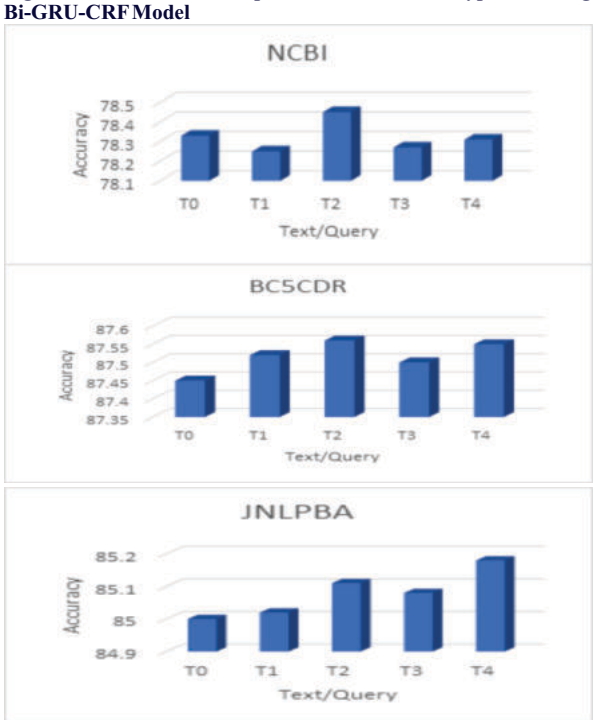

Figure 4: Performance comparison of different of type text using

BERT Model

(e.g., phrases and segments). The second example is a consistency problem. It can be seen that BERT only recognized one "obstructing" in the whole sequence, while Bi-GRU-CRF corrected the error of BERT. This example shows that Bi-LSTM-CRF can alleviate the problem of label inconsistency by learning the semantic information of the entire sequence. In the third and final example, false positives (FPs) were identified by BERT, but real negatives were rectified using Bi-LSTM-CRF (TNs). The fourth and fifth are segmentation problem examples. Compared with BERT, Bi-LSTM-CRF can better distinguish the boundary information of entities. These four examples all demonstrate the effectiveness of Bi-GRU-CRF in the syntactic learning. Through the case Through the case study, we can infer that compared with BERT, Bi-GRU-CRF has better performance in syntactic and semantic learning. Specifically, Bi-GRU-CRF can eliminate the issue of label inconsistency, fix some FNs and FPs, and appropriately identify entity boundaries

TABLE-3 SUMMARY OF RESULT

No	Dataset	Entity	Model	Result
1.	BC5CDR-Disease	Cholecystectomy -syndrome Disease Obstructing - Others	BERT	ERCP versus MRCP is recommended to exclude obstructing mass. The findings could reflect changes of cholecystectomy.
			Bi-LSTM-CRF	ERCP versus MRCP is recommended to exclude obstructing mass. The findings could reflect changes of cholecystectomy.
2.	NCBI-Disease	left atrial-Others intracardiac thrombus - syndrome Disease	BERT	echo smoke is seen, and in fact, an intracardiac thrombus is identified and circumscribed at 1.83 cm in circumference at the base of the left atrial appendage.
			Bi-LSTM-CRF	echo smoke is seen, and in fact, an intracardiac thrombus is identified and circumscribed at 1.83 cm in circumference at the base of the left atrial appendage.
3.	JNLPBA	pituitary adenoma, noncalcified craniopharyngioma, Rathke's cleft cyst - syndrome Disease retrospect - others	BERT	In retrospect sellar enlargement could be seen on the angiogram X-rays. Differential consideration was given to cystic pituitary adenoma, noncalcified craniopharyngioma, or Rathke's cleft cyst with solid component
			Bi-LSTM-CRF	In retrospect sellar enlargement could be seen on the angiogram X-rays. Differential consideration was given to cystic pituitary adenoma, noncalcified craniopharyngioma, or Rathke's cleft cyst with solid component

CONCLUSION AND FUTURE WORK

This paper delineate the syndrome disease NER model for bio-medical trial texts by using realm contextual embeddings with relevant domain-specific features, word embeddings, and character embeddings in a bidirectional gated recurrent neural network - conditional random field framework. Extensive experiments and analyses on a clinical trial dataset and the benchmark datasets dataset show the effectiveness of the proposed model. In the future will experiment with deep bidirectional transformer-based language models to generate deep contextualized embeddings of clinical trial texts.

REFERENCES:

[1] Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2021), Biomedical named entity recognition using BERT in the machine reading comprehension framework. Journal of Biomedical Informatics, 118, 103799.

[2] Ling, Y., Hasan, S. A., Farri, O., Chen, Z., van Ommering, R., Yee, C., & Dimitrova, N. (2019). A domain knowledge-enhanced LSTM-CRF model for disease named entity recognition. AMIA summits on translational science proceedings, 2019, 761.

[3] Crichton, G., Pyysalo, S., Chiu, B., & Korhonen, A. (2017). A neural network multi-task learning approach to biomedical named entity recognition. BMC bioinformatics, 18(1), 1-14.

[4] Sahu, S. K., & Anand, A. (2016). Recurrent neural network models for disease name recognition using domain invariant features. arXiv preprint arXiv:1606.09371.

[5] Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research, 304, 114135.

[6] Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., Bedrick, S., Lazar, A. J., & Pant, S. (2017, November). Overview of the TREC 2017 Precision Medicine Track. In TREC.

[7] Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics, 47, 1-10.

[8] Jagannatha, A. N., & Yu, H. (2016, June). Bidirectional RNN for medical event detection in electronic health records. In Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting (Vol. 2016, p. 473). NIH Public Access.

[9] Zhao, Z., Yang, Z., Luo, L., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2017). Disease named entity recognition from biomedical literature using a novel convolutional neural network. BMC medical genomics, 10, 75-83.

[10] Boag, W., Sergeeva, E., Kulshreshtha, S., Szolovits, P., Rumshisky, A., & Naumann, T. (2018). Cliner 2.0: Accessible and accurate clinical concept extraction. arXiv preprint arXiv:1803.02245.

[11] Chalapathy, R., Borzeshi, E. Z., & Piccardi, M. (2016). Bidirectional LSTM-CRF for clinical concept extraction. arXiv preprint arXiv:1611.08373.

[12] Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

[13] Ling, Y., An, Y., & Hasan, S. A. (2017, April). Improving clinical diagnosis inference through integration of structured and unstructured knowledge. In Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications (pp. 31-36).

[14] Ling, Y., An, Y., Liu, M., Hasan, S. A., Fan, Y., & Hu, X. (2017, May). Integrating extra knowledge into word embedding models for biomedical NLP tasks. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 968-975). IEEE.

[15] Si, Y., Wang, J., Xu, H., & Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. Journal of the American Medical Informatics Association, 26(11), 1297-1304.

[16] Datla, V., Hasan, S. A., Qadir, A., Lee, K., Ling, Y., Liu, J., & Farri, O. (2017, November). Automated clinical diagnosis: The role of content in various sections of a clinical document. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1004-1011). IEEE.

[17] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[18] Zhu, H., Paschalidis, I. C., & Tahmasebi, A. (2018). Clinical concept extraction with contextual word embedding. arXiv preprint arXiv:1810.10566.

[19] Dogan, R. I., & Lu, Z. (2012, June). An improved corpus of disease mentions in PubMed citations. In BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (pp. 91-99).

[20] Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., & Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics, 33(14), i37-i48.

[21] Bhatia, P., Busra Celikkaya, E., & Khalilia, M. (2020). End-to-end joint entity extraction and negation detection for clinical text. Precision Health and Medicine: A Digital Revolution in Healthcare, 139-148.

[22] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[23] Liang, T., Xia, C., Zhao, Z., Jiang, Y., Yin, Y., & Philip, S. Y. (2023). Transferring from Textual Entailment to Biomedical Named Entity Recognition. IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[24] Fabregat, H., Duque, A., Martinez-Romo, J., & Araujo, L. (2023). Negation-based transfer learning for improving biomedical Named Entity Recognition and Relation Extraction. Journal of Biomedical Informatics, 104279.

[25] Jaura, S., & Ramanna, S. (2023). Named Entity Recognition on CORD-19 Bio-Medical Dataset with Tolerance Rough Sets. In Transactions on Rough Sets XXIII (pp. 23-32). Berlin, Heidelberg: Springer Berlin Heidelberg.

[26] Liu, S., Duan, J., Gong, F., Yue, H., & Wang, J. (2023, January). Fusing Label Relations for Chinese EMR Named Entity Recognition with Machine Reading Comprehension. In Bioinformatics Research and Applications: 18th International Symposium, ISBRA 2022, Haifa, Israel, November 14–17, 2022, Proceedings (pp. 41-51). Cham: Springer Nature Switzerland.

[27] Chen, P., Wang, J., Lin, H., Zhang, Y., & Yang, Z. (2023). Knowledge Adaptive Multi-way Matching Network for Biomedical Named Entity Recognition via Machine Reading Comprehension. IEEE/ACM Transactions on Computational Biology and Bioinformatics.