BoostER: A Performance Boosting Module for Biomedical Entity Recognition

Rahul Pandey, Md Shamsuzzaman, Sadid A. Hasan, Mohammad S Sorower,
Md Abdullah Al Hafiz Khan, Joey Liu, Vivek Datla, Mladen Milosevic,
Gabe Mankovich, Rob van Ommering, Nevenka Dimitrova
Philips Research North America, Cambridge, MA, USA
{firstname.lastname}@philips.com

***Abstract*—Biomedical Entity Recognition tasks have gained significant importance in the clinical research domain. There has been a lot of prior work on improving entity recognition of biomedical concepts from using rule-based to more context-dependent deep learning-based approaches. However, due to its high domain dependency and distinctive vocabularies, appropriate utilization of contextual knowledge becomes a challenge even for context dependent deep models. To this end, we propose a novel performance-boosting improvement module "BoostER" that can be applied to any existing entity recognition system to boost its performance in terms of precision and F1 scores. The proposed module has been developed on top of a pre-trained BERT model and fine-tuned to give more weights to contextual learning compared to word-specific information. We tested our system with Chemical and Disease Entity Recognition tasks using the BioCreative CDR dataset to demonstrate its effectiveness compared to existing state-of-the-art models.**

***Index Terms*—Disease Named Entity Recognition, Deep Learning, Contextual Embedding**

I. Introduction

Biomedical Named Entity Recognition can be defined as a process for finding references to biomedical entities from a text document including their concept type and location. There exists a plethora of medical documents available in the electronic format, which is continuing to increase over the years. These documents are mainly unstructured, which makes it difficult to manually comprehend useful information. Identifying entities from unstructured documents can provide basic information on what each document is about and can be a first step of several downstream natural language processing (NLP) tasks such as biomedical information retrieval, knowledge discovery, patient profiling, and clinical trial matching [1]. In some of these tasks, achieving high precision is important and in some cases high recall is important. For an automated clinical trial matching system, the precision of finding concepts from clinical trials and patient profiles is very important to minimize the risk of recommending inappropriate clinical trials to patients. It might be more acceptable to not find all patients that fit a specific clinical trial than to match an ineligible patient to a clinical trial - hence, precise biomedical entity recognition is desirable. Such a system should not only recognize the specific clinical concept types, but also needs to disambiguate the same mention into different clinical concepts based on different context. For example, the term *DMD* can refer to both disease and gene in a clinical trial text document. Therefore, the disease entity recognizer must refrain from detecting *DMD* as a disease when it is mentioned as a gene in the document.

We conjecture that an added improvement in precision for entity recognition tasks would have a positive impact on downstream critical biomedical tasks. In contrast to prior works that mostly focus on finding appropriate spans for the entities in a document, we put more emphasis on the context in which an entity is present to determine its type. Existing named entity recognition algorithms incorporate both contextual knowledge and unique word representation into the model for identifying the specific entity type and span. Such design may not be optimum for understanding the context of an entity mention exclusively or, conversely, recognizing a mention without any context e.g., entities mentioned in a table or in a list. Furthermore, appropriate understanding of the context is essential to recognize misspelled or uncommon or newly-invented entities, and to distinguish whether a particular text span denotes a disease or a medication or an irrelevant concept. In addition, disambiguation of the same mentions (mostly acronyms) into different entity types is crucial for a highly accurate biomedical entity recognition system. Therefore, concept recognition model needs to understand the context for specific mentions. In this work, we propose a novel entity recognition system to identify a concept accurately by leveraging the underlying contextual information of the given text while keeping less importance to distinctive word-specific information.

II. Related Work

Existing approaches in the biomedical entity recognition domain use dictionary or rule-based sampling to identify biomedical entities [2] (herein referred as clinical concepts). However, they generally fail to provide better results when the input text (i.e, Patients' Electronic Health Record (EHR), biomedical publications, or clinical trials etc.) are not properly structured or contain numerous irregularities. Some of the possible irregularities include presence of acronyms, ambiguous words with other concepts, typographical errors etc. Such

Current contact address of the author, Rahul Pandey is rpandey4@gmu.edu.

irregularities have led researchers to propose new methods for biomedical entity recognition using character-level CNNs [1], [3], sequence-to-sequence models [4], [5] like RNNs & LSTMs along with CRF, as well as including domain knowledge as additional features [6], etc. Recently, researchers focus on utilizing pre-trained language models such as BERT [7] to reduce the training effort with large text and fine-tune these model for the specific entity recognition task. The pre-trained transformer architecture enabled model BERT [8] gains popularity in recent research due to its inherent capability of understanding the contextual information from the given text. BioBERT[9], which pre-trains the BERT model on biomedical datasets like PubMed abstracts & PMC full-text biomedical articles along with generic datasets like Wikipedia & Books corpora has significantly improved the performance of biomedical entity recognition compared to dictionary/rule-based mapping.

However, these approaches still perform worst in recognizing words/concepts that are part of distinctive vocabularies, typos, acronyms, or have polysemous meanings. Many prior research have focused on ambiguity and word sense disambiguation in biomedical text[10], [11], [12], [13]. However, the disambiguation of entities has never been considered together with entity recognition and is highly dependent on dynamic and constantly increasing domain knowledge. Hence, there remains a gap in prior works to improve performance by utilizing additional external guidance and resources. Motivated by this, we address the following research challenges in this work.

1) Can we incorporate an adjudication module to an existing concept identification system to boost its precision?
2) Can our adjudication module be dependent only on the training data (without any external knowledge base) and trained in a way where it gives less importance to unique clinical concept terms and focuses more on its surrounding context?
3) Can we supplement generic knowledge to simplify the context that helps the model learn inherent characteristics efficiently?

To address these research challenges, we propose a novel biomedical entity recognition improvement module, which can be applied to any existing base named entity recognition (NER) model to improve the performance in terms of precision and without impacting the recall, and thus, to improve the overall F1 score. Our module improves the performance of the biomedical entity recognition system by focusing on contextual information learned from the text and assigns less importance to word-specific information.

III. Methodology

In this section, we first describe the generic architecture of our proposed approach and then introduce two variations of our implementation. The overall architecture of our proposed 'BoostER' model is shown in figure 1. Our system comprises of four major components– *i)* 'BoostER' Train Generator, *ii)* 'BoostER' Test Generator, *iii)* Base NER, and *iv)* Pre-Trained BERT. We discuss the details as follows.

- **BoostER Train Generator**: This component prepossesses the training and validation text data and generates training documents required for fine-tuning the pre-trained BERT model.
- **BoostER Test Generator**: This component takes the test data & Base NER to get the base annotation spans and pre-process it to generate the test document to predict the correct concepts given its generalized context.
- **Base NER**: This component represents the existing base Named Entity which is used to get initial annotation of entity type and span. We also compute which spans are correctly identified and which ones are ambiguous or uncertain and pass that information to the 'BoostER' Test Generator.
- **Pre-Trained BERT** We have used state-of-the-art pre-trained language model Bidirectional Encoder Representations from Transformers (BERT) [7] to get the contextual information. We have used the pre-trained model of BERT, *bert-large-cased*, which has 24-layer, 1024-hidden, 16-heads, 340M parameters. The model is pre-trained on Book corpus(800M words)[14] and WIkipedia dump data(2,500M words).

In summary, we first take the training and validation set and preprocess it with the BoostER train generator to generate the pre-processed training documents. Using these training documents, we fine-tune the pre-trained BERT model in two different ways – *i) "Masked LM based BoostER"* and, *ii) "Appropriateness based BoostER"*. In the testing phase, first we use Base NER to get the input spans along with their ambiguous or uncertain nature of the predicted concepts in test data. We then pre-process the test data that is required with BoostER Test Preprocessor. Finally, we predict the pre-processed test document to get the correct prediction and then we report the result. We represent the details of *"Masked LM based Boost ER"*, and *"Appropriateness based BoostER"* as follows.

A. Masked LM based BoostER

In this approach, we use Masked Language Modeling (Masked LM) technique to fine-tune the BERT model. Masked LM technique predict the words that has been replaced with the token *[MASK]* given the surrounding context. The context for the masked token can be neighboring few words, a sentence, a paragraph or the whole document. In our experiment, we use sentence based context. For each *[MASK]* token, the result of the masked LM approach is the probability for each word in the vocabulary that can replace that mask token. The components of this proposed architecture are:

- **BoostER Train Generator**: Given the training document with n concepts (for example, disease, gene or chemical concepts) and their corresponding spans, we generate more generic pre-processed documents that are not dependent on unique vocabularies of clinical data. For this, we generate $n+1$ sub documents $D_{i=1}^{n+1}$ that are pre-processed such that for each sub document $D_i \ \epsilon \ D_{i=1}^{n}$, all the word annotated with the i_{th} concept is replaced by a unique

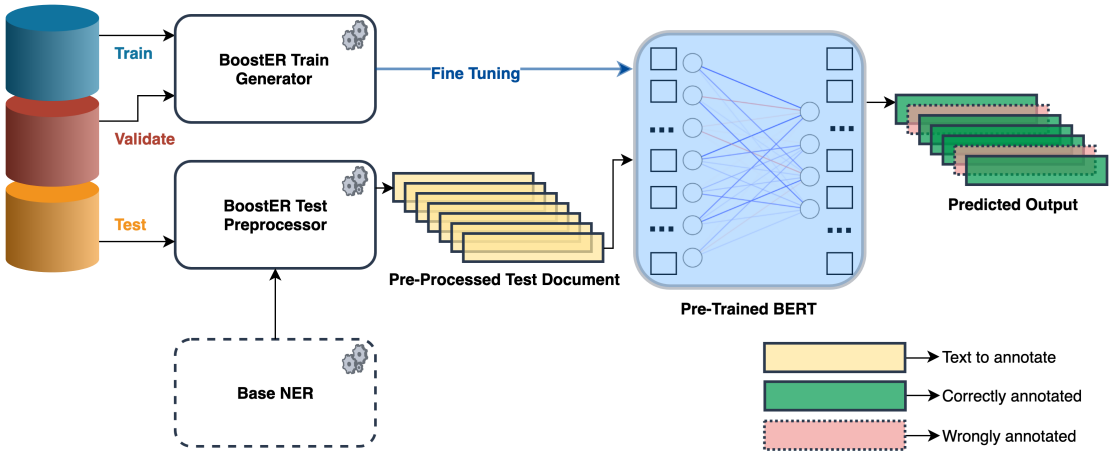

Fig. 1: Proposed System architecture of 'BoostER'

specific i_{th} concept term (i.e, replace all disease terms like pneumonia, arthritis, fever, cough, etc. with a unique term "disease", and replace all drugs/chemical names to "chemical"). For the sub document D_{n+1}, we replace all the words annotated with all different concepts with their corresponding generic concept term. These set of documents are used as input to the pre-trained BERT model for fine-tuning.

- **BoostER Test Preprocessor**: For testing the BoostER model, we first need to choose the Base NER for which we are trying to improve the performance. Given a Base NER and annotations from external sources, *B* we preprocess the test document as follows.
  1) Predict annotation of the tokens using the base annotator, *B*.
  2) To improve the performance of our system, we need to identify inaccurately annotated tokens. Several techniques can be applied to identify these incorrect annotations among them, in this experiment, we introduce the following two distinct techniques to recognize those erroneous annotations.
     a) *Ambiguity Filter:* If the entity is annotated by more than one base entity recognizer as possible concepts, then ambiguity filter suggests that word as ambiguous and needs additional feedback from our BoostER pipeline.
     b) *Uncertainty Filter:* The base entity recognizer predicts concepts for the given entities and generates scores for the corresponding concepts. These scores are then evaluated against the predefined threshold value (determined empirically). These concepts need to re-verify which have lower scores compared to the threshold value.
  3) These filtered entities are then replaced by a generic token *[MASK]*, that is going to predict in the next step.
  4) We choose to select tokens that have higher scores compared to the threshold value and replace these tokens with their corresponding generic concept annotations. This replacement step helps minimize the extra domain-specific word-level information and thus improves the performance of our systems. These pre-processed documents are then forwarded to the pre-trained BERT for predictions.
- **Pre-trained BERT**: This module takes the pre-processed documents as input and predicts the annotation. We finetune this pre-trained BERT model to incorporate domain knowledge from the given documents and finally evaluate the performance.

  Fine-tuning
  1) We incorporate the same language modeling-based approach, that is used during the pre-training of the BERT model for fine-tuning the model. We follow a similar approach that was reported in BERT [7] to finetune our model. We mask 15% of the words randomly out of which 80% will be replaced by token $[MASK]$, 10% by random vocabulary word, and remaining 10% with the actual word itself. In this task, we predict the masked word considering the given context.
  2) In this step, we determine the sentence pairs considering the most probable second sentence in the given sequence of sentences. Note, we use these steps as fine-tuning steps rather than pre-training from scratch. We hypothesize that this fine-tuning the model with these pre-processed training documents helps learn the dependencies of the words including all the generic concept terms that are present in all training documents. In this approach, our model emphasizes contextual dependencies rather than domain-specific actual words due to the replacement of generic concept terms.

  Testing
  1) Once the model is fine-tuned on the data generated by the BoostER train generator, we use this model to predict the masked words of the test documents that

are pre-processed by our BoostER Test Preprocessor.
2) For each masked word, we get a probability (relevancy score) for the presence of all vocabulary words of the training data in the masked position. Now,
   a) We annotate the masked word with i_{th} concept only when:
      i) the generic concept term of i_{th} concept is present in the top k vocabulary words (k being the hyperparameter).
      ii) no other concept is present before the i_{th} concept in the top k vocabulary word.
   b) For all other cases, the masked word is identified as O other concept.
3) We compare the predicted results with the base NER and report the result.

B. Appropriateness based BoostER

In this section, we define a sequential classification problem of *Appropriateness* for predicting the erroneous concept. Table I describes one of the example, how we define *appropriate* vs *not-appropriate* data. As seen in the example, our definition of a *appropriate* sentence is the one that has all recognized entities replaced with their correct generic concept terms. We hypothesize that the sentence with all recognized entities replaced correctly is more appropriate than the sentence with incorrect replacement. In other words, if an entity is identified incorrectly by the base NER system, the sentence replaced with the corresponding generic concept has less appropriateness score and vice versa.

We implement this hypothesis in our proposed architecture. We discuss the details of our proposed architecture for this implementation as follows.

- **BoostER Train Generator**: Similar to the MaskedLM based approach, given the training and validation document with n annotated concepts, create $n+1$ sub-documents in which every i_{th} sub-document will replace the concept words with a generic concept term for the i_th concept. The $n+1_{th}$ document will have all concept replaced with their appropriate generic concept terms, respectively. Each sentence in the replaced document is annotated as *appropriate* document.
  We generate $k-$negative sampling for each sentence, in which each negatively sampled sentence has entities replaced with any other concept term except their corresponding true generic concept terms as well as with the term "other". We assign *non-appropriate* label to these sentences. We have now used this sentence-label pairs to fine-tune the Pre-trained BERT model.
- **BoostER Test Preprocessor**: The first two steps of preprocessing the test data are the same as Masked LM based BoostER. Hence, after filtering out the erroneous entities, the following steps are performed.
  3) All the entities that are not selected in the filtering process, are replaced with the correct generic concept terms predicted by the base NER.
  4) Then, for a sentence S with k possible incorrect annotation and n generic concepts under consideration, $(n+1)^k$ target sentences are created. Each $s_i \epsilon S_{i=0}^{(n+1)^k}$ represents one of the possible combinations where each k_{th} entities are replaced by any of the $n+1$ concepts. Here extra 1 concept denotes the "other" (O) concept, which signifies that it can be anything other than the n concepts at that position.
  5) Finally, all the $S_{i=0}^{(n+1)^k}$ combination sentences are passed to the pre-trained BERT model for its prediction of appropriateness.
- **Pre-trained BERT**: Similar to Masked LM based BoostER, we use two steps for using the pre-trained BERT in *Appropriateness* based BoostER approach– *Fine-tuning* and *Testing*

  Fine-tuning
  1) We use BERT for sequence classification tasks and fine-tune the BERT model to recognize the appropriateness of the sentences.
  2) We take the input sentence-label pairs from the BoostER train generator and use for fine-tuning the Pre-trained BERT model except for 5-10% of the data which is used as a validation set for evaluating the fine-tuning step.

  Testing
  1) For each pre-processed test sentence, there are $(n+1)^k$ derived sentences and we predict the appropriateness score for each of these derived sentences.
  2) We choose a derived sentence with a certain combination of concepts that has scored the maximum appropriateness score among all the derives sentences for an input test sentence.
  3) We compare the predicted concept combination against the ones provided by Base NER and replace all the erroneous concepts with its correct concepts from the predictions.

We explain the Base NER component in our experiment section.

IV. Experiment

We use two clinical NER tasks in our experiment: *Disease Concept Identification* and *Chemical Concept Identification*. In this section, we discuss the details of the datasets used in this experiment.

a) Dataset: We have used BioCreative CDR data[15] for our tasks. For the sentence tokenized document, we have taken BC5CDR data fromt the MTL-Bioinformatics[16] repo. Table II gives the statistics of number of training, development, and test sentences in the dataset. # Sentences represent total number of sentences in the whole document. While # Concept Enity is the total number of whole concepts (disease/chemical/... etc.) present, # Concept Tokens are the sum of all individual token word, which are a part of concepts present in the whole

https://github.com/cambridgeltl/MTL-Bioinformatics-2016

TABLE I: Example of appropriate vs non-appropriate sentences.

Text	Label
I am suffering from disease and that's why I have taken medicine that contains chemical.	*appropriate*
I am suffering from chemical and that's why I have taken medicine that contains disease.	*not-appropriate*

document. We have used both training and development data for fine-tuning and reported our results on the test data.

b) Experimental setup: We have performed our experiment on a system with Intel Xeon CPU E5-2623 having 8 $\times$ 32 GB DDR4 RAM, and 4 $\times$ Tesla K80 GPU. We have fine-tuned on all 4 GPUs while during testing we have only used 1 GPU that suggests that our module can run and give correct predictions using only 1 GPU.

For implementing BERT, we have used pytorch-transformers (previously pytorch-pretrained-bert) developed by HuggingFace for the experiments. We have taken $threshold$ for uncertainty filter as 0.99 throughout the experiment. Similarly for Masked LM based BoostER, we have taken top k vocabulary word as 5 throughout the experiment for generalization. For the $k-$negative sampling in Appropriateness based BoostER, we have generated at max 5 negative samples per sentence. After BoostER train generator in Masked LM based BoostER, 27,420 sentences were generated in 2,998 paragraphs that is used for fine-tuning. Meanwhile, for Appropriateness based BoostER; with negative sampling, 130,000 sentences were used for fine-tuning and 9,579 for validation. The distribution of *appropriate* vs *not-appropriate* in the training set were 25,585 and 104,415 respectively. Both the fine-tuning process took less than 10 hours to train on the above specified tech specs.

c) Base NER: For both variant of our proposed BoostER system, we have used the state-of-the-art BioBERT annotator trained on biomedical corpora. BioBERT uses powerful BERT architecture that has been very significant in many NLP tasks as it can intelligently capture the contextual information along with positional and word-specific information. Also, since BioBERT has been trained on large amount of biomedical texts, it incorporates biomedical specific vocabulary to train their dependencies. Hence, it has outperformed BERT on every tasks and became state-of-the-art in 6 out of 9 tasks. Our major goal in improving the performance of the base NER was to increase the precision without much forgetting their initial results i.e. recall. Given the spans of the predicted *Chemical* or *Disease* concepts, we want to pass those spans to our BoostER system to get additional feedback if the spans may contain the concepts or not given its context.

BioBERT for entity recognition gives the probability of each sub-word (part of a whole word) if it is among *B* beginning of the concept, *I* intermediate sub-word(s) of the concept, *O* all other words that do not belong to the concept, *X*: part of the previous sub-word, *[CLS]* the beginning of the document, *[SEP]* separator between sentences. Now for each sub-word, we get a logit score for each of the target classes (*B, I, O, X, [CLS], [SEP]*). To get the score for *Uncertainty Filter*, we compute the softmax value of each logit score for a sub-word. The score for each sub-word would be the max of the softmax value and class would be the one with max softmax value. We compute the score for each whole word by taking the minimum of the score of its sub-word. We will compare the score with the $threshold$ to get the uncertain erroneous spans.

d) Metric: Since, our initial input spans are coming from the base NER itself, our recall cannot be increased. But we can increase the precision by learning what was incorrectly classified as the concepts. Hence to compare our proposed system, we mainly focus on precision score and also show how much recall has been decreased or how much percent of data is actually forgotten by our system in order to improve the precision. We have tested both our implementation with two different concept identification tasks. Our F1 score signifies if we were successful in improving the performance of base NER or not.

V. Results and Discussion

We have implemented the BioBERT from the dmis-lab repo and taken the pre-trained model from the naver repo. We have taken the state-of-the-art BioBERT v1.0 (+ PubMed 200K + PMC 270K) pre-trained model to fine-tune the BC5CDR tasks of both disease and chemical entity recognition. After fine-tuning, we got the results of the test data along with their logit scores to compute the scores for *Uncertainty Filtering* during Test pre-processing (described in Base NER sub-section). Table III represents all the results of our proposed method against the base NER. we observe that the results of BioBERT is a little different from what they have mentioned in the paper. But since our approach requires the predicted input and their scores for processing, we are reporting only what we have observed after running the fine-tuning of their pre-trained model on our system. We notice that both of our approach have performed better than the base NER model on both the tasks. Also, our proposed approach obtains a higher F1 score compared to the base NER model although there was a slight decrease in recall compared to a considerable increase in precision.

We could have taken BioBERT pre-trained model as our pre-trained model in BoostER. However, we chose to use the *bert-large-cased* model mainly because the BioBERT has been pre-trained with the *bert-base-cased* model, which has a smaller architecture. Also, since we were already pre-processing the document and replacing all distinctive vocabularies word to their generic concept term, we did not feel the need of using domain-specific pre-trained model. Also, even the F1 increase is not very high of our proposed approach (0.23-0.33), our precision is higher compared to recall value when compared to

https://github.com/huggingface/pytorch-transformers

https://github.com/dmis-lab/biobert

https://github.com/naver/biobert-pretrained/

TABLE II: BC5CDR dataset statistics

Dataset	Entity type	Doc type	# Sentences	# Concept Ety	# Concept Tkns
BC5CDR	Disease	Train	4559	4182	7100
BC5CDR	Disease	Development	4580	4246	6969
BC5CDR	Disease	Test	4796	4424	7161
BC5CDR	Chemical	Train	4559	5203	7103
BC5CDR	Chemical	Development	4580	5347	7095
BC5CDR	Chemical	Test	4796	5385	7013

TABLE III: Results of our proposed method on BC5CDR data

Method	Entity type	Precision(↑)	Recall(↓)	F1(↑)
BioBERT	Disease	84.49%(–)	86.66%(–)	85.56%(–)
(ours) Masked LM	Disease	85.13%(0.64↑)	86.32%(**0.20**↓)	85.79%(0.23↑)
(ours) Appropriateness	Disease	**85.44%(0.95↑)**	86.46%(0.34↓)	**85.88%(0.32↑)**
BioBERT	Chemical	92.05%(–)	92.40%(–)	92.23%(–)
(ours) Masked LM	Chemical	92.75%(0.70↑)	92.22%(**0.18**↓)	**92.49%(0.26↑)**
(ours) Appropriateness	Chemical	**93.04%(0.99↑)**	91.88%(0.52↓)	92.46%(0.23↑)

Heart mitochondria isolated from doxorubicin-treated rats exhibited depressed rates for state 3 respiration (336 +/- 26 versus 425 +/- 53 natom O/min/mg protein) and a lower respiratory control ratio (RCR) (4.3 +/- 0.6 versus 5.8 +/- 0.4) compared with cardiac mitochondria isolated from saline-treated rats.

Surprisingly, by 13 weeks following the last DOX treatment (late stage), MHC-CB7 exhibited a progressive decrease in cardiac function and higher rates of cardiomyocyte apoptosis when compared with NON-TXG mice.

Fig. 2: Example of erroneous annotations that our proposed approach has corrected

the base NER. Since precision is more important metric than recall for evaluating downstream critical clinical application, improvement of precision score by our proposed system validates our model performance gain. We analyzed some of the predicted sentences in both disease and chemical entity recognition to see which concepts were correctly identified by our proposed approach when compared to the base NER. Figure 2 shown two such examples for both chemical and disease entity recognition tasks. For the first example, BioBERT annotated the word "depressed" as *disease*. However, as seen in the example, "depressed" word was used as a verb that symbolizes less rates. Our approach successfully classified this word as *O* other category. Similarly, in the second example, "MHC" annotated as chemical by BioBERT. However, the following part of the sentence (highlighted in blue) expressed that "MHC" must be something other than a chemical as chemical could not exhibit decrease in cardiac function. Our proposed approach understood that context and correctly identified this concept as *O* concept. In reality "MHC" was denoting the trans gene of mouse. Hence, we observed that the contextual information played a vital role in determining the concepts.

VI. Conclusion, Limitation, and Future Work

In this paper, we proposed a novel Biomedical Entity Recognition improvement module BoostER that gained more generalized contextual information and improved the performance of the existing base entity recognition system in terms of improved precision and hence improved F1 score. We proposed two different implementation of the BoostER and conducted two entity recognition experiments with state-of-the-art BioBERT as the base NER. We observed that even a fine-tuned BERT model with generalized data was enough to predict whether the spans given by the base NER was correct or not and achieved better precision by correctly identifying the spans that was predicted incorrectly. We also observed that despite the lack of recall improvement, our proposed method still was able to improve the F1 with increasing precision value.

However, our system has some limitation. As our target spans are completely dependent on the base NER itself, and improving recall may not possible. Hence, to improve the overall accuracy, our system has to sustain the same recall while increasing the precision. Also, in some cases, we lose a lot of information when we replace every possible annotations with their generic concept term. This results in bad interpretation and it happens mostly when there are not enough contextual information present. Hence in those cases, our model still underperforms. Our model will also perform poorly for the concept mentions without context – concepts mentioned in a table or in a list. We can think of developing a separate algorithm that can recognize this sort of structure and section in a document and develop a concept recognition algorithm which heavily depends on unique word representation or as simple as matching with the medical concept dictionary.

In future, we will select only the replaceable spans intelligently during generating pre-processed train and test documents. We can also combine different base NER to get more target spans to predict. We can also apply our technique to other tasks like gene identification as well. Since our approach is modular and generic, it can also be applied to other domain dataset and different other tasks with different base NER system.

REFERENCES

[1] 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. In *Machine Learning for Healthcare Conference*, pages 383–402, 2018.

[2] Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. Prominer: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6(1):S14, 2005.

[3] Zhehuan Zhao, Zhihao Yang, Ling Luo, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. Disease named entity recognition from biomedical literature using a novel convolutional neural network. *BMC medical genomics*, 10(5):73, 2017.

[4] Qikang Wei, Tao Chen, Ruifeng Xu, Yulan He, and Lin Gui. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. *Database: The Journal of Biological Databases and Curation*, 2016.

[5] Sunil Kumar Sahu and Ashish Anand. Recurrent neural network models for disease name recognition using domain invariant features. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[6] Yuan Ling, Sadid A Hasan, Oladimeji Farri, Zheng Chen, Rob van Ommering, Charles Yee, and Nevenka Dimitrova. A domain knowledge-enhanced lstm-crf model for disease named entity recognition. *AMIA Summits on Translational Science Proceedings*, 2019:761, 2019.

[7] 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.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[9] 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. *arXiv preprint arXiv:1901.08746*, 2019.

[10] CM Fauquet and CR Pringle. Abbreviations for vertebrate virus species names. *Archives of virology*, 144(9):1865–1880, 1999.

[11] Marc Weeber, James G Mork, and Alan R Aronson. Developing a test collection for biomedical word sense disambiguation. In *Proceedings of the AMIA Symposium*, page 746. American Medical Informatics Association, 2001.

[12] Hongfang Liu, Alan R Aronson, and Carol Friedman. A study of abbreviations in medline abstracts. In *Proceedings of the AMIA Symposium*, page 464. American Medical Informatics Association, 2002.

[13] Mark Stevenson and Yikun Guo. Disambiguation in the biomedical domain: the role of ambiguity type. *Journal of biomedical informatics*, 43(6):972–981, 2010.

[14] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

[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*, 2016, 2016.

[16] Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):368, 2017.