CS60075 : Natural Language Processing

TERM PROJECT REPORT

Named Entity Recognition in Biomedical Corpus

Team Name: NLP Term Project Team

Himanshu Mundhra	Aurghya Maiti	Swastika Dutta	Omar Eqbal
16CS10057	16CS10059	16CS10060	16CS30043

Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

Autumn Semester, 2019-20

In recent years, deep contextual embeddings such as (BERT, FLAIR, ELMO) have proved to be competent enough to detect the exact spans of named entities (both in general English Domain and Biomedical Domain).
In this task, we aim to address the gaps of detecting named entities using both context-independent word embeddings and context-dependent word embeddings on the NCBI Disease, BC5CDR and ChemProt Corpus.

Introduction

Textual Data, be it a structured database or an unstructured document, is always rich with information. With the advent of Deep Learning in Natural Language Processing, the task of extraction of important information from relevant text corpora has become increasingly popular, but is yet challenging.

One of the many Natural Language Processing tasks is **Named-Entity Recognition (NER)**, where we are supposed to identify and locate entities in a textual corpus and categorize them into predefined categories. This bucketing can then be used to learn important information relevant to our motive.

A variant of the simple **NER** is **Bio-NER** or **BioMedical Named Entity Recognition** which is one of the most fundamental tasks in biomedical text mining. Similar to NER, the task here involves identifying entities and classifying them in of the predefined categories (like genes, proteins, chemicals and diseases). Biomedical datasets are very few and far between and getting hold of these datasets is expensive, which is why it is essential to develop a model specifically catering to this task and **Bio-NER** achieves this task.

Motivation

NER is an essential component for tasks related to **Natural Language Processing**(NLP) or **Information Retrieval**(IR). It aids to filter essential features from a text corpora and extract meaningful relations or cluster documents into meaningful subgroups. Some of the popular use cases of NERs are to classify contents for news providers, as one of the preprocessing steps for training of ad hoc learning models, to develop content recommendations for a media industry client, to develop efficient and optimised search algorithms, etc. Hence, it is implicit that NER approaches have received much attention in recent years and various approaches for NER has been explored. Specifically, we shall focus on biomedical NER in our task.

There are several reasons for implementation of NER specifically for biomedical tasks. This is because **Bio-NER** has other challenges in addition to those faced by other NER tasks and hence, we need to form NER tools which are specifically and exclusively suited for biomedical datasets.

Approaches for BioNER varies from **dictionary-based**, **rule-based**, **Machine Learning** (ML) to hybrid approaches. In initial stages, Bio-NER was accomplished by the use of explicit set of rules. The widely used ML approaches use annotated data to train a learning model which is then used to classify the unseen BioNEs. Combining the output of different classifiers using ensemble approach is an efficient technique used in BioNER.

Of late, deep learning algorithms based on Artificial Neural Networks (ANNs) are being used to a larger extent to train the learning model for various applications. There are predominantly two approaches in ANNs, i.e., contextual and non-contextual word embeddings.

Context is so vital when working on NLP tasks. Learning to predict the next character based on previous characters forms the basis of sequence modeling. Contextual String Embeddings leverage the internal states of a trained character language model to produce a novel type of word embedding. In simple terms, it uses certain internal principles of a trained character model, such that words can have different meanings in different sentences.

In our project we aim to compare and contrast the different approaches of Bio-NER, in order to evaluate the relative efficiencies of these models.

Model Architecture

We have used different contextual and non-textual embeddings to represent words in vector space. These embeddings are then fed to a BiLSTM-CRF model for Named Entity Recognition and exact match F1-scores are calculated and compared. In order to generate word embeddings, we mainly use the following models:

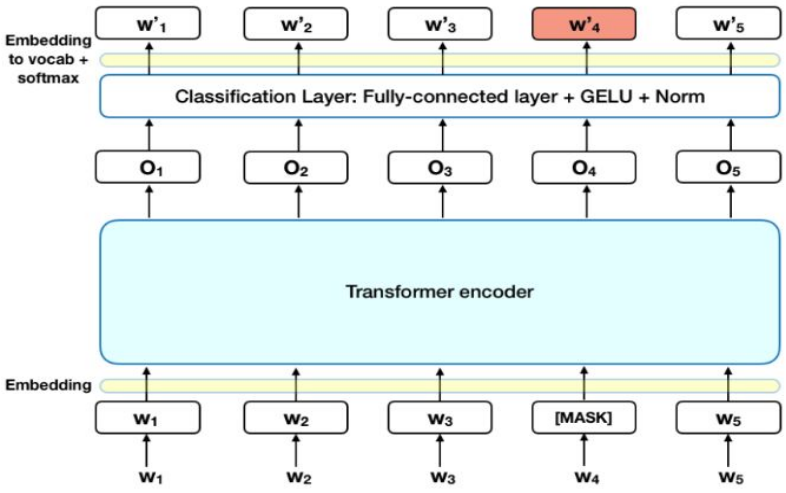
1. BERT

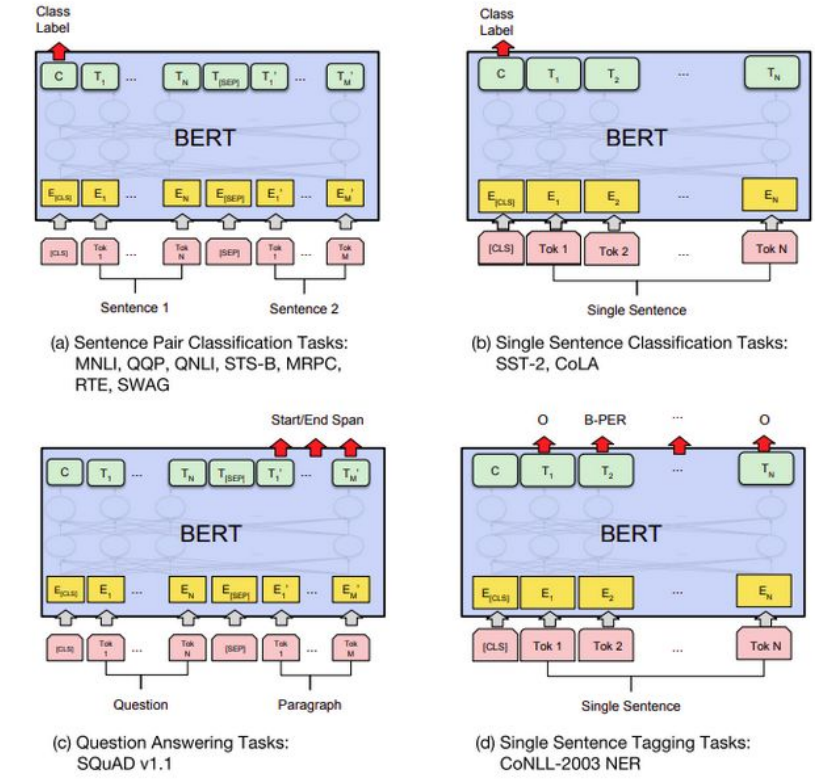
BERT is a deep learning model that has given state-of-the-art results on a wide variety of natural language processing tasks. It has been pre-trained with Masked Language Modeling and Next Sentence Prediction Tasks on Wikipedia and BooksCorpus and requires task-specific fine-tuning.

There are two models of BERT:

- BERT base – 12 layers, 12 attention heads, and 110 million parameters.
- BERT Large – 24 layers, 16 attention heads and, 340 million parameters.

A word starts with its embedding representation from the embedding layer. Every layer does some multi-headed attention computation on the word representation of the previous layer to create a new intermediate representation. All these intermediate representations are of the same size.





(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

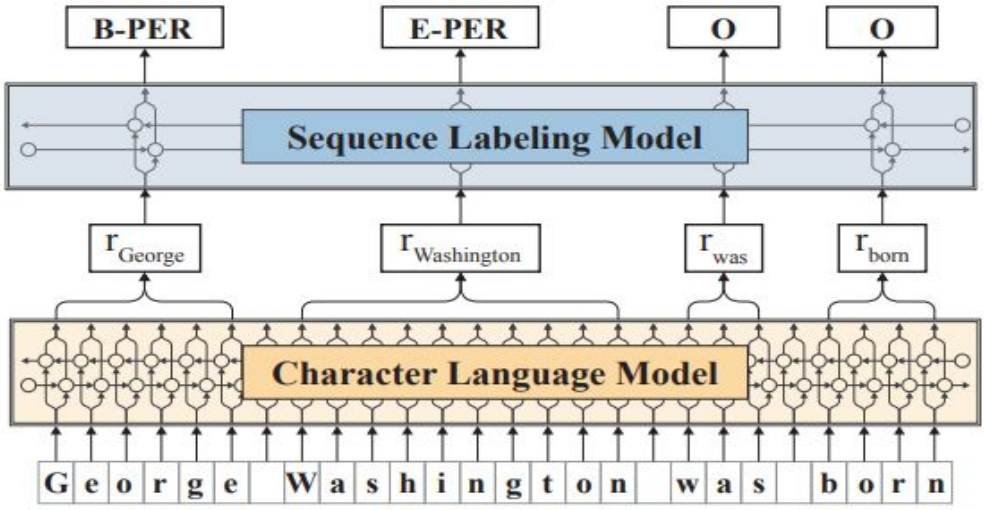
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

2. FLAIR

It leverages the internal states of a trained character language model to contextual string embeddings. FLAIR embeddings have the distinct properties since they fundamentally model words as sequences of characters, and are contextualized by their surrounding text. It utilizes trained Language Models to generate embeddings, but consider Language Models(LMs) at the character level.

A sentence is input as a character sequence into a pre-trained bidirectional character language model. From the learnt Language Model, contextual embeddings corresponding to each word is retrieved and passed into a vanilla

BiLSTM-CRF sequence labeller, achieving robust state-of-the-art results on downstream tasks.



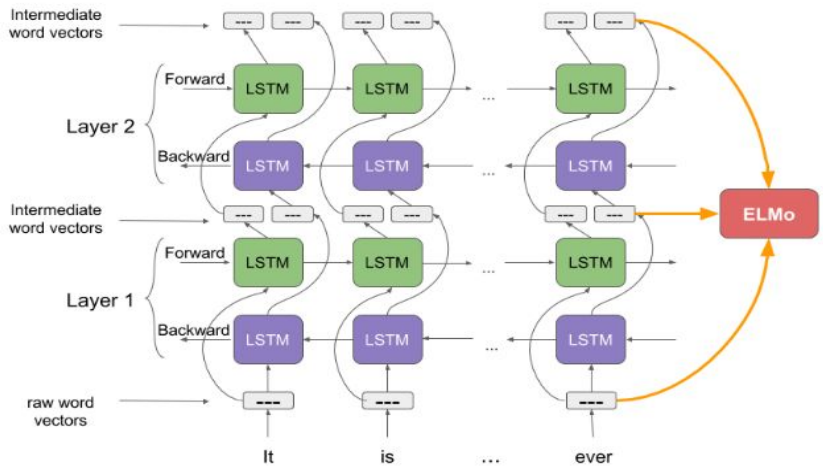
3. ELMO

ELMO embeddings are computed on top of a two-layer bidirectional language model (biLM). This biLM model has two layers stacked together. Each layer has 2 passes — forward pass and backward pass.

The architecture uses a character-level convolutional neural network (CNN) to represent words of a text string into raw word vectors. These raw word vectors are then input to the first layer of biLM.

The forward pass contains information about a certain word and the context (other words) before that word while the backward pass contains information about the word and the context after it.

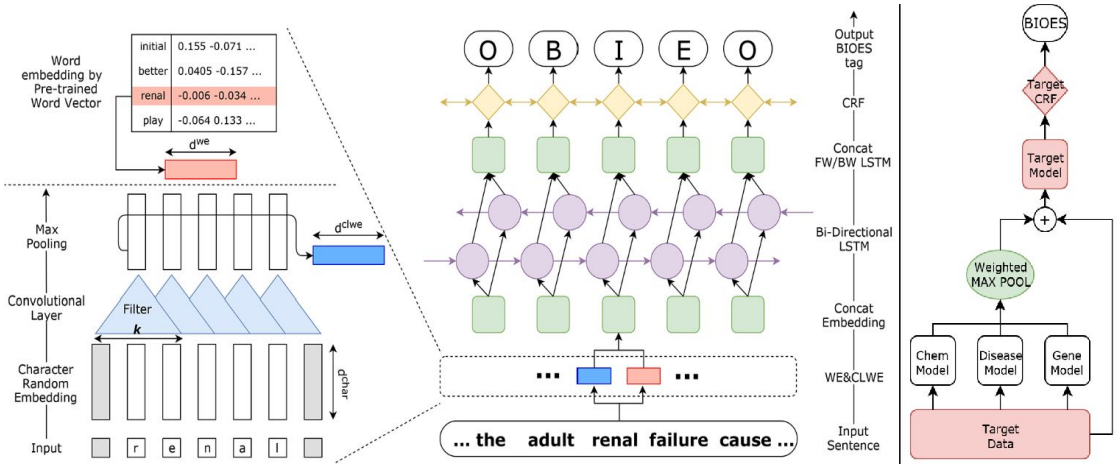
This pair of information, from the forward and backward pass, forms the intermediate word vectors and are fed into the next layer of biLM. The final representation of corresponding text is then calculated using the weighted sum of the raw word vectors and the 2 intermediate word vectors.



4. CollaboNet

It utilizes a combination of multiple NER models; models trained on different datasets are connected to each other so that a target model obtains information from other collaborator models to reduce false positives. Every model is an expert on their target entity type and take turns serving as a target and a collaborator model during training time. This approach helps to tackle problems associated with lack of data and the entity type misclassification problem in BioNER tasks.

Single-task model structure consists of character level word embedding using CNN and Bidirectional LSTM with Conditional Random Field (BiLSTM-CRF).
Overall structure of CollaboNet is shown in the right of the figure. Arrows show the flow of information when target model M_{Target} is training. The models in CollaboNet take turns in being the target model.



Related Works

State-of-the-art BioNER systems often require handcrafted features (e.g., capitalization, prefix and suffix) to be specifically designed for each entity type (Ando, 2007; Leaman and Lu, 2016; Zhou and Su, 2004; Lu et al., 2015). This feature generation process takes the majority of time and cost in developing a BioNER system (Leser and Hakenberg, 2005), and leads to highly specialized systems that cannot be directly used to recognize new types of entities. The accuracy of the resulting BioNER tools remains a limiting factor in the performance of biomedical text mining pipelines (Huang and Lu, 2015).

Recent NER studies consider neural network models to automatically generate quality features. Habibi et al. adopted the model from Lample et al. and used word embeddings as input into a bidirectional long short-term memory-conditional random field (BiLSTM-CRF) model. These neural network models free experts from manual feature engineering. However, these models have millions of parameters and require very large datasets to reliably estimate the parameters.

However, directly applying state-of-the-art NLP methodologies to biomedical text mining has limitations. First, as recent word representation models such as Word2Vec (Mikolov et al., 2013), ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) are trained and tested mainly on datasets containing general domain texts (e.g.Wikipedia), it is difficult to estimate their performance on datasets containing biomedical texts.

The word distributions of general and biomedical corpora are quite different, which can often be a problem for biomedical text mining models. As a result, recent models in biomedical text mining rely largely on adapted versions of word representations (Habibi et al., 2017; Pyysalo et al, 2013).

Previously, Word2Vec, which is one of the most widely known context independent word representation models, was trained on biomedical corpora which contain terms and expressions that are usually not included in a general domain corpus (Pyysalo et al., 2013).

While ELMo and BERT have proven the effectiveness of contextualized word representations, they cannot obtain high performance on biomedical corpora because they are pre-trained on only general domain corpora. Hence, Bio-BERT(2019) and Bio-Flair was introduced to fill this void. In our project, we probe into the differences between these models in order to analyse their merits and demerits.

Results and Discussions

NCBI disease

	F1 Score	Precision	Recall
Elmo	0.7577	0.7514	0.7643
Bio-Elmo	0.8758	0.8635	0.8816
Flair	0.8253	0.8343	0.8165
Bio-Flair	0.8791	0.8667	0.8845
Bert	0.7373	0.7317	0.7431
Bio-Bert	0.9058	0.8918	0.9127
Bio-Bert (Fine Tuning)	**0.9279**	**0.9123**	**0.9335**
CollaboNet	0.8481	0.8315	0.8653

BC5CDR Chem

	F1 Score	Precision	Recall
Elmo	0.7826	0.8153	0.7524
Bio-Elmo	0.9153	0.9058	0.9276
Flair	0.8283	0.8106	0.8498
Bio-Flair	0.9135	0.9215	0.9023
Bert	0.7795	0.8251	0.7387
Bio-Bert	0.9172	0.9074	0.9254
Bio-Bert (Fine Tuning)	**0.9388**	**0.9285**	**0.9492**
CollaboNet	0.9079	0.9205	0.8913

BC5CDR Disease

	F1 Score	Precision	Recall
Elmo	0.7103	0.7517	0.6732
Bio-Elmo	0.8712	0.8813	0.8416
Flair	0.8291	0.8201	0.8461
Bio-Flair	0.8585	0.8671	0.8501
Bert	0.8263	0.8145	0.8327
Bio-Bert	0.8759	0.8823	0.8656
Bio-Bert (Fine Tuning)	**0.9162**	**0.9071**	**0.9253**
CollaboNet	0.8322	0.8296	0.8349

We notice that CollaboNet is more efficient than Elmo and Bert(contextual embeddings). However, as expected, contextual embeddings trained on biomedical corpus(Bio-BERT, Bio-Elmo, Bio-Flair) surpass the results obtained by non-contextual embeddings. Finally, we observe that BioBERT fine-tuned on respective datasets performs the best with extremely high accuracy.

Ablation Analysis

The possible sources of errors exclusive to biomedical NERs are as follows:

- **Ambiguity due to abbreviation:** For example, EGFR corresponds to epidermal growth factor receptor or estimated glomerular filtration rate. For the first case, the entity is the name of a protein, while the second one is not.

- **Polysemy**, i.e, a word refers to different entities. For example, myc-c refers to the name of a gene or protein..

- **Synonyms identification:** For example, CASP3, caspase-3, and CPP32 denote the same entity.

- **Out of dictionary**: Results due to the frequent insertion of new names into the dictionary. Due to IUPAC nomenclature new proteins can be named, but they are not present in the vocabulary frequent enough to learn a proper embedding.

- **Multi-word/Hyphenated BioNE's**: If the word is hyphenated then should we tokenize into multiple words or learn it.
  " . . . we demonstrate a high frequency of ATM mutations in T - PLL ."

- **Nested BioNEs**, i.e., a BioNE may occur as part of longer BioNE as a proper string. For example, colorectal adenomas and carcinoma

Since BioBERT provides us with the most promising results, we shall probe into the flaws or strengths of BioBERT against other models. By analysing our data, we notice the following patterns:

- **BioBERT performs well with abbreviations:**

*"In **SCA3** , gaze - evoked nystagmus was often present as was saccade hypometria and smooth pursuit gain was markedly decreased"*

*"Because **IVF** causes cardiac rhythm disturbance , we investigated whether malfunction of ion channels could cause the disorder by studying mutations in the cardiac sodium channel gene SCN5A ."*

This might be explained by the fact that BioBERT resorts to sub-word embeddings to handle OOV words. Hence, it can correctly map the abbreviated words to its corresponding context and provide reasonable embeddings.

- **BERT often ignores some parts of the entity, in short instances or as a part of longer entity.**

*"This mutation may be valuable for developing models of dominantly **inherited neurodegeneration** (predicted O but I), as the early age of onset of symptoms suggests that this mutation may be particularly deleterious to the magnocellular neurons . "*

We have studied a set of 164 patients with multiple colorectal adenomas and / or carcinoma and analyzed codons 1263 - 1377 (exon 15G) of the APC gene for germ - line variants .

This trend is observed commonly with Nested Bio-NEs. Presence of these inherent challenges in biomedical corpus makes it difficult for models to parse them into suitable sentence frame and label correct NERs. This reduces exact match F1 scores for BERT.

- **BioBERT often predicts I instead of O when the entities are followed by OOV words or followed by other entities**
  Mucopolysaccharidosis IVA (MPS IVA) is an autosomal recessive lysosomal storage disorder caused by a genetic(B) ***defect****(actual I)* ***in N - acetylgalactosamine - 6 - sulfate****(the whole thing in bold is predicted I) sulfatase (GALNS) .*

- **Sometimes if the word is commonly known to be a non-entity, then it will be predicted O instead I**
  Immunohistochemical staining of human breast specimens also revealed BRCA1 nuclear foci in benign breast , invasive lobular cancers and low - ***grade*** *(marked as O actually I) ductal carcinomas .*

  As evident from our results, non-textual embeddings such as CollaboNet could not detect instances of disease entity when the disease is abbreviated and out of vocabulary in pubmed word2vec.
  Also some cases of polysemy could not be handled by word2vec but could be resolved in contextual embeddings.

Conclusion

We explored a BiLSTM-CRF based model - non-contextual embedding model CollaboNet and some contextual embedding models like Bio-BERT, Bio-ELMO and Bio-FLAIR for the Bio-NER task. We show that fully contextualized embeddings which make effectively use local context, give higher exact-span match F1-score than CollaboNet which uses pubmed word embedding, i.e., context independent embeddings. Contextual embeddings trained on bio-medical domain perform better than the embeddings trained on a general domain. Fine tuning the Bio-BERT model on task-specific data further increases the performance.

We also notice that CollaboNet outperforms BERT and Elmo. This might be due to the unique nature of biomedical corpuses. As discussed earlier, biomedical texts has several additional challenges such as ambiguity, polysemy etc; and are underrepresented in the training corpus used for Bert or Elmo. Consequently, the word embeddings generated by these models are not very efficient in capturing the semantic or syntactic relations in such corpus. However, CollaboNet is trained on biomedical corpus itself and hence better represents the words in their corresponding word vectors.

Acknowledgement

We wish to express my sincere thanks to Ishani Mondal for providing me with all the necessary facilities for the project. We are also grateful to Prof. Sudeshna Sarkar, for giving us the opportunity to work on this project.

References

1. Improving Multi-Word Entity Recognition for Biomedical Texts. Hamada A. Nayel et al. International Journal of Pure and Applied Mathematics, Volume 118 No. 16, 2018.
2. Flair: State-of-the-Art Natural Language Processing (NLP)
3. BioBERT paper Analysis
4. Demystifying BERT: A Comprehensive Guide to the Groundbreaking NLP Framework
5. ELMo: Deep contextualized word representations - AllenNLP
6. CollaboNet: collaboration of deep neural networks for biomedical named entity recognition

---