NLP Term Project Autumn 2019

Named Entity Recognition in Biomedical Corpus

Team Name: Vindicators

Team Members:

1. Aaditya Singh(16IE10033)
2. Archit Rungta (18MA20008)
3. Nishant (16BT30015)
4. Vishal Garimella (16CS10061)

PROBLEM STATEMENT

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 in 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 either pre-trained on Bio-medical corpus or general corpus.
Corpora were taken from NCBI Disease, BC5CDR and ChemProt for NER. We studied and tabulated the parameters namely F1, precision and recall for Bi-LSTM-CRF model architecture using these embeddings over these datasets.

MOTIVATION

Biomedical Named Entity Recognition (NER) is a fundamental step in several downstream biomedical text mining and information extraction tasks like relation classification, coreference resolution etc. Traditional Context independent word embeddings and pre-trained embeddings have often been used to model complex syntactic and semantic characteristics of words. But these complementary embedding models fail to capture different word uses across different linguistic contexts (i.e, polysemy). This problem is compounded in biomedical text due to ambiguous usage of words from general text (ex: column in general English means an upright pillar while in medical context can be taken to mean the spine). Another reason is that context independent Word representations obtained from training on biomedical corpora is present in both forms (general English and biomedical)are generally present in the training text. Another issue specific to biomedical domain is the generous usage of abbreviations (ex: gene/protein names like ALA, MEN 1) without explicit mention of their full forms. Therefore Neither character embeddings nor context independent word embeddings are effective in solving this issue.
BERT resolves this issue to some extent using Byte-Pair Encoding BPE which decomposes the unknown word into lexical sub-units.

Hence the survey of context independent word embedding is necessary for named entity recognition in Biomedical corpus

MODEL ARCHITECTURE

We try Bi-LSTM-CRF model as an additional layer over the contextual word-embedding layer obtained over ElMO, BERT, and Flair. It is trained to minimize Viterbi loss if we view this problem as a sequence tagging problem.

RESULT

Elmo Embeddings

	NCBI disease	BC5SDR chem	BC5SDR disease
F1 score	**0.74**	**0.78**	**0.70**
Precision	0.74	0.83	0.76
Recall	0.73	0.74	0.65

Bio-Elmo Embeddings

	NCBI disease	BC5SDR chem	BC5SDR disease
F1 score	**0.87**	**0.93**	**0.84**
Precision	0.86	0.92	0.86
Recall	0.88	0.94	0.83

Flair Embeddings

	NCBI disease	BC5SDR chem	BC5SDR disease
F1 score	**0.8453**	**0.8343**	**0.7164**
Precision	0.8543	0.8186	0.8084
Recall	0.8365	0.8507	0.6432

Bio-Flair Embeddings

	NCBI disease	BC5SDR chem	BC5SDR disease
F1 score	**0.8650**	**0.8997**	**0.8180**

Precision	0.8494	0.8907	0.8416
Recall	0.8812	0.9089	0.7956

BERT Embeddings

	NCBI disease	BC5SDR chem	BC5SDR disease
F1 score	**0.72**	**0.77**	**0.84**
Precision	0.70	0.82	0.85
Recall	0.75	0.73	0.84

Bio-BERT Embeddings

	NCBI disease	BC5SDR chem	BC5SDR disease
F1 score	**0.87**	**0.92**	**0.84**
Precision	0.85	0.91	0.85
Recall	0.90	0.93	0.84

DISCUSSION

We compare our results with other neural network methods known to perform well on these datasets. As a baseline, we implement the Bi-LSTM-CRF tagger described in Lample et al [1], which makes use of domain specific pre-trained embeddings and incorporates subword features using another Bi-LSTM. We implement another baseline that instead uses CNN for calculating character embeddings Ma et al[2]. Finally, we compare our method with a deep multi-task model, which provides state of the art performance on these datasets Wang et al[3]. For the baselines, we performed hyper-parameter tuning with LSTM cell size and dimension of pre-trained word vectors.

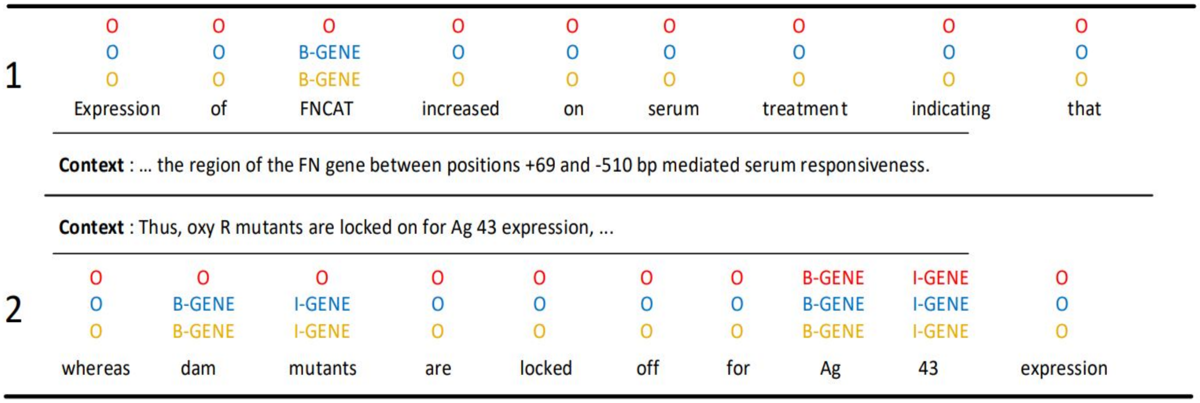
Baseline	NCBI disease	BC5SDR (chem+disease)
Lample et al[1]	0.85	0.86
Ma et al [2]	0.83	0.85
Wang et al [3]	0.86	0.88

ABLATION ANALYSIS

The sources of error and ambiguity in BioNER's stems from polysomy, abbreviation, non-conforming synonyms(CASP3, caspase-3, and CPP32), multiword Bio named entities and finally lack of standard nomenclature

To look into how contextualized representations help(importance) in Biomedical NER, we select some examples from the test set. In the following Figure, two recurring cases where using contextualized word embeddings have helped are shown. Example 1 exhibits the behavior of the baseline Lample et al (which uses pretrained context independent embeddings) and our model(which uses ELMO + Bi-LSTM CRF) in case of an acronym, the usage of which is very common in scientific texts.

Looking at the figure, we can understand how using contextualized embeddings help to deal with polysemy. The token dam, which is more commonly associated with a reservoir structure is incorrectly labeled by the context independent embeddings model. The baseline system correctly labeled oxy R mutants as a gene/protein entity but did not recognize dam mutants(synonyms). Looking at the context suggests that like oxy R mutants, dam mutants should also be a protein. Again in such cases, looking at the larger context might have helped. Finally, on analysis, we find that a fully contextualized embeddings also does much better on longer entities.



1	O	O	O	O	O	O	O	O	O
	O	O	B-GENE	O	O	O	O	O	O
	O	O	B-GENE	O	O	O	O	O	O
	Expression	of	FNCAT	increased	on	serum	treatment	indicating	that

Context : ... the region of the FN gene between positions +69 and -510 bp mediated serum responsiveness.

Context : Thus, oxy R mutants are locked on for Ag 43 expression, ...

2	O	O	O	O	O	O	O	B-GENE	I-GENE	O
	O	B-GENE	I-GENE	O	O	O	O	B-GENE	I-GENE	O
	O	B-GENE	I-GENE	O	O	O	O	B-GENE	I-GENE	O
	whereas	dam	mutants	are	locked	off	for	Ag	43	expression

Fig : Example outputs of our model in cases where context helps determine the tags. Example 1 shows a case where the token is an acronym and Example 2 is a case of polysemy (for token dam). Entity tags highlighted with gold are the gold-standard tags, with red are the ones from baseline, and with blue, are from our model. Related context is also shown

CONCLUSION

In this report, we show that a fully contextualized embedding like BERT, ELMO & FLAIR which makes the use of local context more effectively, therefore giving higher exact-span match F1-score for Named Entity Recognition tasks. We compare our results with some of the baselines which use context-independent word embeddings, embeddings trained on a general domain and only simple multi-task Dense layers at each step for prediction.