In-domain Context-aware Token Embeddings Improve Biomedical Named Entity Recognition

Golnar Sheikhshab
Simon Fraser University
gsheikhs@sfu.ca

Inanc Birol
British Columbia Cancer Agency
ibirol@bcgsc.ca

Anoop Sarkar
Simon Fraser University
gsheikhs@sfu.ca

Abstract

Rapidly expanding volume of publications in the biomedical domain makes it increasingly difficult for a timely evaluation of the latest literature. That, along with a push for automated evaluation of clinical reports, present opportunities for effective natural language processing methods. In this study we target the problem of named entity recognition, where texts are processed to annotate terms that are relevant for biomedical studies. Terms of interest in the domain include gene and protein names, and cell lines and types. Here we report on a pipeline built on Embeddings from Language Models (ELMo) and a deep learning package for natural language processing (AllenNLP). We trained context-aware token embeddings on a dataset of biomedical papers using ELMo, and incorporated these embeddings in the LSTM-CRF model used by AllenNLP for named entity recognition. We show these representations improve named entity recognition for different types of biomedical named entities. We also achieve a new state of the art in gene mention detection on the BioCreative II gene mention shared task.

1 Introduction

Last decade witnessed substantial improvements in machine learning methods and their application to natural language processing tasks. Recently, Peters et al. (2018) introduced ELMo (Embeddings from Language Models), a system for deep contextualized word representation, and showed how it can be used in existing task-specific deep neural networks. The method improves the state of the art over a variety of NLP tasks such as question answering, word sense disambiguation, sentiment analysis, and named entity recognition. The developers of the tool also provide an ELMo model pre-trained on the Billion-word Language Model (LM) dataset (Chelba et al., 2014) as an off-the-shelf tool for use in a wide variety of NLP tasks and domains.

This begs the question of how the performance of downstream analysis would improve if the model were to be adapted to work with domain-specific texts. In this paper, we investigate the effect of an in-domain training set for ELMo in Named Entity Recognition (NER) applications. Our contributions are as follows:

1. Off-the-shelf ELMo has room for improvement in domain-specific applications

2. ELMo consistently improves biomedical named entity recognition when trained on in-domain data

3. Such improvement can be achieved even when the in-domain training dataset is smaller than the Billion-word LM data.

4. The resulting model achieves the highest precision/recall/F1 scores so far on BioCreative II Gene mention detection shared task (BC2GM).

We explain ELMo and AllenNER, the named entity recognizer we used, in sections 2 and section 3. Then, we describe our datasets in section 4, and we move on to report the results in section 5.

2 ELMo

ELMo (Peters et al., 2018) is a system that produces context-aware embeddings for word tokens. Similar to traditional context-independent word embeddings such as GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013), ELMo representations can be used as input to a neural network for downstream tasks. Though, ELMo is different from the traditional word embeddings in that it gives the representation of the

Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pages 160–164
Brussels, Belgium, October 31, 2018.

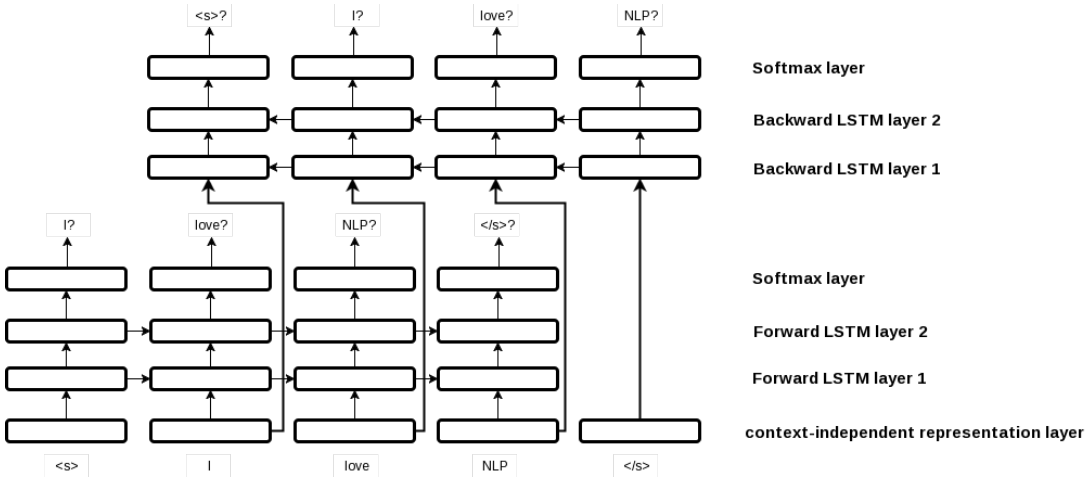

Figure 1: ELMo is a bidirectional LSTM for language modelling where the next or precedent tokens are predicted from the softmax layers over forward and backward LSTMs respectively.

word in the context of the specific given sentence; hence it is a context-aware word token representation as opposed to a word type representation.

It is trained using a language modeling objective function, where the objective is to predict the next word in the sequence; either sequentially left to right or right to left. As such, it can be viewed as learning a token level representation of words for a task that can be trained on unannotated data. These word representations can then be used for a task that is trained on labeled data. In our case, the task is biomedical named-entity recognition.

Figure 1 shows the architecture of ELMo as a recurrent language modelling network. The input to this system is a sequence of words $w_1 w_2 \dots w_i \dots w_n$. First, each word is converted to a context-independent embedding by a convolutional neural network (CNN) over its characters. These character-based representations are then fed into a two-layer bidirectional long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) recurrent neural network. Output of the second layers of the forward and backward LSTMs are fed to a soft-max layer to predict w_{i+1} and w_{i-1}, respectively, at each position i.

Task-specific learned weights can be used later to combine all layers in ELMo model at position i and form the task-specific "ELMo representation of w_i".

Peters et al. (2018) showed that different layers in this deep recurrent model learn different aspects of a given token. The lower layers learn more syntactic features whereas higher layers learn the contextual aspects of the word. They linearly combined the layers using task-dependent weights, and their experiments show that for Named Entity Recognition tasks, the layers are combined with effectively the same weights.

3 Named Entity Recognition with AllenNLP

In our pipeline, we couple ELMo embeddings to AllenNLP (Gardner et al., 2017) for NER tasks.

AllenNLP uses a bidirectional two-layer LSTM-CRF (Lample et al., 2016) to perform NER as a sequence tagging task. Each word is tagged with an output that marks if it is at the beginning (B), in the middle (I), at the end (E or L), or outside (O) of an entity type. One-word entities are also marked (as S or U). For example B-Gene and I-Gene stand for beginning and inside of a Gene, whereas B-DNA and E-DNA stand for beginning and ending of a DNA entity type.

AllenNLP embeds the input words using a Convolutional Neural Network over characters. Rei et al. (2016) showed that word embeddings from character compositions outperform lookup embeddings such as word2vec, when used for named entity recognition.

AllenNLP combines the layers in ELMo Model using learned task-specific weights, concatenates the result for each token to context-independent word embeddings, and feed the concatenation into the LSTM-CRF as illustrated in Figure 2.

4 Datasets

We collected a focused domain-specific subset of PubMed Central (PMC) documents, and used

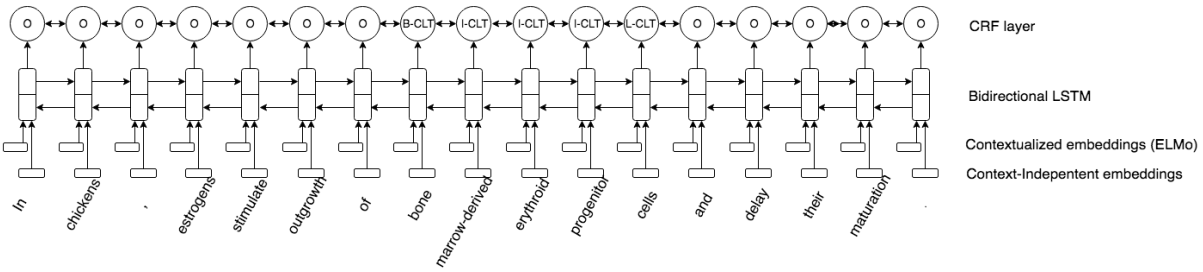

Figure 2: Architecture of LSTM-CRF (Lample et al., 2016) with ELMo. Traditional word embeddings and ELMo representations are concatenated and fed into a bidirectional LSTM. A CRF layer on top of bidirectional LSTM takes local label dependencies into account. At training time the log likelihood of gold label sequences is maximized. At test time, Viterbi (Viterbi, 1967) algorithm is used to decode the complete label sequence. -CLT in the labels of the example indicate cell_type entity.

them for training ELMo. This dataset is described in detail in section 4.1. We report results on two benchmark datasets, which we describe in sections 4.2 and 4.3.

4.1 ELMo Training Set

We downloaded the text files of a subset of PMC documents that are available at ftp://ftp.ncbi.nlm.nih.gov/pub/pmc in May 2018, and picked 3960 full-text documents that had a Medical Subject Heading (Mesh) term 'cancer'. We ran StanfordNLP/CoreNLP toolkit (Manning et al., 2014) on these documents for sentence splitting and tokenization. Tokens of each sentence were joined with space character in between to form the sentences in the training set. This dataset contains about 21 million tokens, and is substantially smaller than the One Billion Word Benchmark (Chelba et al., 2014) that Peters et al. (2018) used for training ELMo but contains in-domain text that is more likely to benefit the biomedical text analysis of interest in this paper.

4.2 BC2GM

BC2GM is the data set for BioCreative II Gene Mention detection shared task (Smith et al., 2008). This dataset contains 15000 training and 5000 test sentences, all from PubMed abstracts. Gold annotations give the gene mentions by providing the sentence ID, the start and end characters of the mention (ignoring all space characters), and the mention itself.

4.3 JNLPBA

JNLPBA (Kim et al., 2004) is the dataset for a shared task on biomedical entity detection. Its training set contains 2000 GENIA (Kim et al., 2003) abstracts, which the authors had collected by searching MEDLINE abstracts for Mesh terms 'human', 'blood cells' and 'transcription factors'. The test set contain 404 abstracts, half of which are from the same domain and the other half are from a super-domain of 'blood cells' and 'transcription factors'. The documents are annotated for protein, DNA, RNA, cell_line, and cell_type entity classes.

5 Results

Table 1 shows the leading results in the literature (top four rows) in comparison with our results (bottom three rows) on BC2GM dataset.

In the year it was held, Ando (2007) had won the challenge with a semi-supervised system equipped with a lexicon and a combination of several classifiers. Gimli (Campos et al., 2013) is a supervised method based on conditional random fields (CRF) (Lafferty et al., 2001) with hand-engineered features that was the state of the art for gene mention detection before GraphNER (Sheikhshab et al., 2018) obtained a higher F-score. GraphNER, obtained the distributions over labels from the CRF and propagated them on a graph of 3-grams similarities constructed over BC2GM.

Rei et al. (2016) set the previous state of the art on BC2GM by applying an LSTM-CRF based system with attention to characters. Our baseline, AllenNER (described in detail in section 3) is similar to their system, except AllenNER uses a convolutional neural network (CNN) over characters instead of using attention mechanism.

Our results, the lower part of Table 1, show that using the off-the-shelf ELMo, that is trained on the one Billion word language model benchmark (Chelba et al., 2014), improves the preci-

Model	Prec. (%)	Rec. (%)	F1 (%)
Ando (2007)	88.48	85.97	87.21
Gimli (2013)	90.22	84.32	87.17
GraphNER (2018)	89.18	85.57	87.34
Rei et al. (2016)	-	-	87.99
AllenNER with no ELMo (Baseline)	88.05	88.72	88.39
AllenNER + off-the-shelf ELMO	89.03	87.95	88.49
AllenNER + ELMO Trained In-Domain	**89.86**	**89.59**	**89.72***

Table 1: Leading results in the literature (up) in comparison with our results (down) on BC2GM dataset

Model	protein	DNA	RNA	cell_line	cell_type
AllenNER with no ELMo (Baseline)	70.47	70.87	63.94	57.17	73.55
AllenNER + off-the-shelf ELMO	69.96	70.56	**65.38**	59.70	73.21
AllenNER + ELMO Trained In-Domain	**75.08***	**73.13**	65.17	**61.15**	**75.87***

Table 2: F1-scores (%) for different entity types in JNLPBA dataset

sion on the expense of recall, modestly improving the F1 score. When ELMo is trained on approximately 21 million in-domain tokens both precision and recall are considerably improved resulting in a more than 1 percentage point improvement in the F1-score. A significance test using sigf (Padó, 2006) showed that this improvement is statistically significant ($p < 10^{-5}$), and the one from off-the-shelf ELMo is not ($p > 0.02$).

Table 2 shows our F1 scores on JNLPBA. It is evident from the table that using ELMo leads to salient improvements over the baseline if it is trained in-domain. The off-the-shelf ELMo has improved the performance for RNA and cell_line entity types but hurt the performance for protein, DNA, and cell_type. In-domain ELMo always obtains the best performance with the exception of RNA entity type where it is competitive with off-the-shelf ELMo and considerably better than the baseline.

Statistical significance tests using sigf (Padó, 2006) showed that most differences in Table 2 are not statistically significant after Bonferroni correction for multiple testing. The only statistically significant improvements are those of in-domain ELMo for protein and cell_type mention detections over both off-the-shelf ELMo and baseline. This could be due to the fact that proteins and cell_types are more frequent in JNLPBA when compared to other entities. Still, it is interesting to note that in-domain trained ELMo model is consistently performed better than the alternative ELMo models in all but one NER task. Table 3 shows the frequencies of different entity types in training and test sets of JNLPBA.

Our results on JNLPBA are not the state of the art. Habibi et al. (2017) report F1 scores as high as 77.25% for protein and 63.31% for cell_line entity types when they use word embeddings trained on the union of (nearly 23 million) PubMed abstracts, (nearly 700,000) PMC full articles, and (approximately four million) English Wikipedia articles as input to an LSTM-CRF. Nevertheless, our results show the positive effect of using in-domain trained ELMo representations compared to a very strong baseline. We believe new state of the art will be achieved if in-domain ELMo representations are used to augment current state-of-the-art systems.

Entity type	Training	Test
protein	30,269	5,067
DNA	9,533	1,056
RNA	951	118
cell_type	6,718	1,921
cell_line	3,830	500

Table 3: Frequencies of different entity types in training and test sets of JNLPBA

6 Conclusion

We show that token level context-aware embeddings trained on an auxiliary task of language modeling using the ELMo toolkit can be used to consistently improve biomedical named entity recognition tasks, but only when the pre-trained embeddings are trained on in-domain biomedical data. Using this technique we produce a new state of the art result on the BioCreative II dataset for gene mention detection.

Acknowledgments

The authors thank the funding organizations, Genome Canada, British Columbia Cancer Foundation, and Genome British Columbia for their

partial support. The research was also partially supported by the Natural Sciences and Engineering Research Council of Canada grants NSERC RGPIN-2018-06437 and RGPAS-2018-522574 and a Department of National Defence (DND) and NSERC grant DGDND-2018-00025 to the third author.

References

Rie Kubota Ando. 2007. BioCreative II gene mention tagging system at IBM Watson. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, volume 23, pages 101–103. Centro Nacional de Investigaciones Oncologicas (CNIO) Madrid, Spain.

David Campos, Sérgio Matos, and José Luís Oliveira. 2013. Gimli: open source and high-performance biomedical name recognition. *BMC bioinformatics*, 14(1):54.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. *INTERSPEECH*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. 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*, pages 70–75. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. 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*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Sebastian Padó. 2006. *User's guide to `sigf`: Significance testing by approximate randomisation*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Marek Rei, Gamal KO Crichton, and Sampo Pyysalo. 2016. Attending to characters in neural sequence labeling models. *arXiv preprint arXiv:1611.04361*.

Golnar Sheikhshab, Elizabeth Starks, Readman Chiu, Aly Karsan, Anoop Sarkar, and Inanc Birol. 2018. Graphner: Using corpus level similarities and graph propagation for named entity recognition. In *Proceedings of the 2018 IEEE International Parallel and Distributed Processing Symposium Workshops*, pages 229–238.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2):S2.

Andrew Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269.