Probing Biomedical Embeddings from Language Models

Anonymous NAACL submission

Abstract

Contextualized word embeddings derived from pre-trained language models, such as ELMo (Peters et al., 2018), show significant improvements on several NLP tasks. In this paper, we train a domain-specific version of ELMo on 10M PubMed abstracts, called BioELMo. These embeddings lead to state-of-the-art performance on biomedical named entity recognition and textual entailment tasks, doing significantly better than their general domain ELMo counterparts. We also conduct probing experiments to determine what additional information is carried *intrinsically* in BioELMo compared to ELMo. For this we remove additional sequence modeling layers from the downstream models, which restricts them to only utilize the information carried in BioELMo. Our results suggest that BioELMo encodes fine-grained information about in-domain entity-types, as well as relations between their mentions in context.

1 Introduction

NLP has seen an upheaval in the last year, with contextual word representations, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), setting state-of-the-art performance on many tasks. These empirical successes suggest that unsupervised pre-training from large corpora could be a vital part of NLP models. In specific domains like biomedicine, NLP datasets are much smaller than their general-domain counterparts[1], which leads to a lot of ad-hoc models: some infer through knowledge bases (Chandu et al., 2017), while others leverage existing large-scale general domain datasets for domain adaptation (Wiese et al., 2017). However, unlabeled biomedical texts are abundant, and their full potential has perhaps not yet been fully realized.

In this paper, we present BioELMo, which is a domain-specific version of ELMo trained on 10M PubMed abstracts[2]. When used as input to downstream tasks, it leads to state-of-the-art performance on biomedical named entity recognition (NER), on the BC2GM dataset (Smith et al., 2008), and natural language inference (NLI), on the MedNLI dataset (Romanov and Shivade, 2018). Unsurprisingly, BioELMo performs significantly better than the general-domain ELMo.

Clearly, there is utility in training in-domain contextual word representations, but we would also like to know exactly what extra information is carried *intrinsically* in these embeddings? To answer this question, we design two *probing tasks*, one for NER and one for NLI, where no sequence modeling is allowed on top of the embedding layers. This setting prohibits the model from capturing task-specific contextual patterns, and instead only utilize the information already present in the representations. Our results and analyses from probing tasks show that – (1) BioELMo encodes fine-grained biomedical entity types, such as genes; and (2) certain biomedical relationships, such as disease-symptom pairs, are preserved as linear transformations in the space.

2 Related Work

Embeddings from Language Models: Our work is a direct extension of ELMo (Peters et al., 2018), which trains a deep bidirectional LSTM (biLSTM) language model. ELMo word embeddings are computed by taking a weighted sum of the hidden states from each layer of the LSTM. Recently, Devlin et al. (2018) showed that pre-training transformer networks on a masked language modeling

[1]For example, MedNLI (Romanov and Shivade, 2018) only has about 11k training instances while the general domain NLI dataset SNLI (Bowman et al., 2015) has 550k.

[2]The code and pre-trained model will be released.

objective leads to even better performance. While we focus on ELMo in this paper, our probing tasks can also be conducted on BERT representations.

Biomedical Word Embeddings: Context-independent word embeddings, such as word2vec (w2v) (Mikolov et al., 2013) and Glove (Pennington et al., 2014) trained on biomedical corpora, are widely used in biomedical NLP models. Recently, some works reported better NER performance with in-domain trained ELMo than general ELMo (Zhu et al., 2018; Sheikhshab et al., 2018). We reaffirm these results on biomedical NER and NLI datasets with BioELMo, and further explore *why* these embeddings are superior.

3 Methods

3.1 BioELMo

BioELMo is trained on the **PubMed** corpus. PubMed provides access to MEDLINE, a database containing more than 24M biomedical citations[3]. We used 10M abstracts from PubMed to train BioELMo, containing a total of 2.46B tokens. The statistics of this corpus are very different from more general domains. For example, the token **patients** ranks 22 by frequency in the PubMed corpus while it ranks 824 in the 1B Word Benchmark dataset (Chelba et al., 2013).

We use the Tensorflow implementation[4] of ELMo to train BioELMo. We use vocabulary size of 1M and keep the default hyperparameters. It took about 1.7K GPU hours to train 8 epochs. BioELMo achieves an averaged forward and backward perplexity of 31.37 on a held-out test set.

3.2 Downstream Models

Downstream models are built on top of pre-trained embeddings and include contextual modeling.

NER: Following Lample et al. (2016), we use pre-trained embeddings and a character-based CNN for word representations, which are fed to a biLSTM, followed by a conditional random field (CRF) (Lafferty et al., 2001) layer for tagging.

NLI: We use the ESIM model (Chen et al., 2016), which encodes the premise and hypothesis using a biLSTM. The encodings are fed to a local inference layer with attention, another biLSTM layer and a pooling layer followed by softmax for classification.

[3] https://www.ncbi.nlm.nih.gov/pubmed/

[4] https://github.com/allenai/bilm-tf

Premise: He returned to the clinic three weeks later and was prescribed with **antibiotics**.
Hypothesis: The patient has an **infection**.
Label: Entailment
to treat

Figure 1: Relation information in a MedNLI instance.

3.3 Probing Tasks

We design two probing tasks where no contextual layers are allowed above the embedding layer, to further investigate BioELMo. One task is on NER to probe for entity-type information, and the other is on NLI to probe for relation information.

NER Probing Task: As shown in Figure 2 (left), we embed the input tokens to $\mathbf{R} = [\mathbf{E_1}; \mathbf{E_2}; ...; \mathbf{E_L}] \in \mathbb{R}^{L \times D_e}$, where L is the sequence length and D_e is embedding size. The embeddings are fed to several feed-forward layers:

$$\widetilde{\mathbf{E_i}} = \text{FFN}(\mathbf{E_i}) \in \mathbb{R}^T$$

where T is the number of tags. $[\widetilde{\mathbf{E_1}}; \widetilde{\mathbf{E_2}}; ...; \widetilde{\mathbf{E_L}}]$ is then fed to a CRF output layer. CRF doesn't model the context but ensures the global consistency across the assigned labels, so it's compatible with our probing task setting.

NLI Probing Task: Relational information between tokens of premises and hypotheses is vital to solve MedNLI task: as shown in Figure 1, the hypothesis is an entailment because **antibiotics** are used to treat an **infection**, which is a drug-disease relation. We design the task shown in Figure 2 (right) to probe the relational information: We embed the premise and hypothesis to $\mathbf{P} \in \mathbb{R}^{L_1 \times D_e}$ and $\mathbf{H} \in \mathbb{R}^{L_2 \times D_e}$, where L_1, L_2 are sequence lengths. Then we use bilinear layers to get $\mathbf{S} = [\mathbf{S_1}; \mathbf{S_2}; ...; \mathbf{S_R}] \in \mathbb{R}^{R \times L_1 \times L_2}$ where,

$$\mathbf{S_r} = \mathbf{P}\mathbf{W_r}\mathbf{H^T} \in \mathbb{R}^{L_1 \times L_2},$$

where $\mathbf{W_r} \in \mathbb{R}^{D_e \times D_e}$ is the weight matrix of a bilinear layer. Note that each element of $\mathbf{S_r}$ encodes the interaction between a token from the premise and a token from the hypothesis. We denote

$$\mathbf{h_{ij}} = \left[\mathbf{S_1}[i,j] \quad ... \quad \mathbf{S_R}[i,j]\right]^T \in \mathbb{R}^R, \qquad (1)$$

as the **distributed relation representation** between token i in premise and token j in hypothesis, and R is the tunable dimension of it. We then apply an element-wise maximum pooling layer:

$$\widetilde{\mathbf{h}} = \max_{i,j} \mathbf{h_{ij}} \in \mathbb{R}^R.$$

We use a linear layer to compute the softmax logits of three possible labels, e.g. $p(\text{entailment}) \propto \exp(\widetilde{\mathbf{h}}^T \mathbf{w_{ent}})$, where entailment vector $\mathbf{w_{ent}}$ is the linear weight vector corresponding to the entailment label.

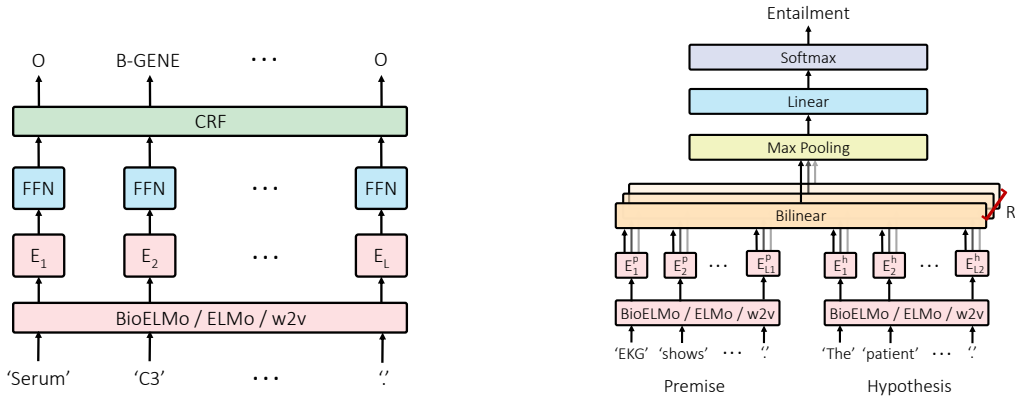

Figure 2: **Left:** NER probing task. The contextual word representations are directly used to predict the NER labels, followed by a CRF layer to ensure label consistency. **Right:** NLI probing task. Bilinear operators map pairs of word representations to relation representations which are used to predict the entailment label.

4 Experiments

4.1 Experimental Setup

Data: For the NER task, we use the BC2GM dataset. BC2GM stands for BioCreative II gene mention dataset (Smith et al., 2008). The task is to detect gene names in sentences. It contains 15k training and 5k test sentences. We also test on the general-domain CoNLL 2003 NER dataset (Tjong Kim Sang and De Meulder, 2003), where the task is to detect entities such as person and location.

For the NLI task, we use the MedNLI dataset (Romanov and Shivade, 2018), where the task is, given a pair of sentences (premise and hypothesis), to predict whether the relation of entailment, contradiction, or neutral (no relation) holds between them. The premises are sampled from doctors' notes in the clinical dataset MIMIC-III (Johnson et al., 2016). The hypotheses and annotations are generated by clinicians. It contains 11,232 training, 1,395 development and 1,422 test instances. We also test on the general-domain SNLI dataset (Bowman et al., 2015), where the premises and hypotheses are drawn from image captions.

Compared Settings: For each dataset, the **Whole** setting describes the state-of-the-art model we used, including contextual modeling layers, and the **Probing** and **Control** settings describe the probing task model introduced in §3.3. The control setting tests the representations on a general-domain dataset/task, to check whether we lose any information in BioELMo compared to ELMo.

Compared Embeddings: We compare three embedding schemes: BioELMo, biomedical w2v trained on a biomedical corpus of 5.5B tokens (Moen and Ananiadou, 2013) and ELMo trained on a general-domain corpus of 5.5B tokens[5][6].

4.2 Main Results

Method	F1 (%)		
	Whole	**Probe**	**Ctrl.**
Ando (2007)	87.2	–	–
Rei et al. (2016)	88.0	–	–
Sheikhshab et al. (2018)	89.7	–	–
General ELMo	87.0	82.9	**84.0**
Biomed w2v	84.9	78.5	67.5
BioELMo	**90.3**	**88.4**	80.9

Table 1: NER test results. **Whole**: whole model performance on BC2GM; **Probe**: Probing task performance on BC2GM; **Ctrl.**: Probing task performance on CoNLL 2003 NER.

Results in Table 1 show that BioELMo in the Whole setting performs much better than the general ELMo and biomedical w2v baselines, setting state-of-the-art performance for this dataset. BioELMo remains competitive in the Probing setting, doing better than most baselines in the Whole setting (including general ELMo). This shows that with the right pre-training, the downstream model can be considerably simplified. Unsurprisingly, in the Control setting BioELMo does worse than the general ELMo embeddings, indicating that the gains come at the cost of losing some general-domain information. However, the gap between general ELMo and BioELMo is larger in the biomedical domain than it is in the general domain. We believe this is because the PubMed corpus contains many mentions of general-domain entities whereas the reverse is not true.

[5] https://allennlp.org/elmo

[6] Note that biomed w2v and ELMo are trained on much larger corpora than BioELMo. Despite this we see significant improvements on in-domain tasks using BioELMo.

Method	Accuracy (%)		
	Whole	Probe	Ctrl.
Romanov and Shivade (2018)	76.6	–	–
General ELMo	75.8	69.6	**78.7**
Biomed w2v	74.2	71.1	73.8
BioELMo	**78.2**	**75.5**	75.7

Table 2: NLI test results. **Whole**: whole model performance on MedNLI; **Probe**: Probing task performance on MedNLI; **Ctrl.**: Probing task performance on SNLI.

Table 2 shows that BioELMo in the Whole setting performs better than the general ELMo and biomedical w2v baselines for NLI, setting state-of-the-art performance for this dataset as well. Once again, we see the same trends for the Probing and Control settings, as we did for the NER tasks. Note that the Probing task *only* models relationships between tokens, but we still see competitive accuracy in that setting (75.5% vs 76.6% previous best). This suggests that, (i) many instances in MedNLI can be solved by identifying token-level relationships between the premise and the hypothesis, and (ii) that BioELMo already captures this kind of information in its embeddings. We explore this in more detail in the next section.

4.3 Analysis

Entity-type Information in BioELMo. In the biomedical literature, the acronym *ER* has multiple meanings: out of the 124 mentions we found in 20K recent PubMed abstracts, 47 refer to the gene "estrogen receptor", 70 refer to the organelle "endoplasmic reticulum" and 4 refer to the "emergency room" in hospital. We use t-SNE (Maaten and Hinton, 2008) to visualize the contextualized embeddings of these mentions in Figure 3.

For general ELMo, by far the strongest signal separating the mentions is whether they appear inside or outside parentheses. This is not surprising given the language modeling training objective for learning these embeddings. BioELMo does a better job of grouping mentions of the same entity together, which is clearly helpful for the NER task.

Relational Information in BioELMo. We manually went through all test instances with the "entailment" label in MedNLI, and found 81 token pairs across the premises and hypotheses which strongly suggest entailment. Among them, 22 are disease-symptom pairs, 13 are disease-drug pairs, 19 are numbers and their indications (e.g.: 150/93 and hypertension) and 24 are synonyms

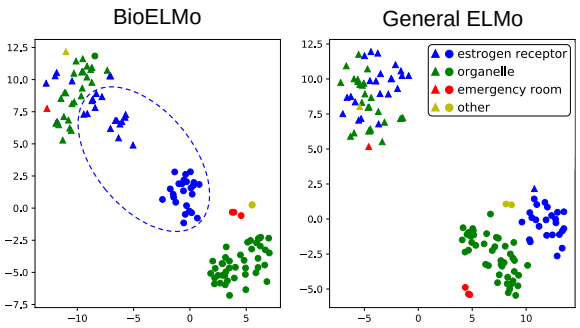

Figure 3: t-SNE visualizations of the token **ER** embeddings in different contexts by BioELMo (left) and general ELMo (right). ● and ▲ represent mentions within and outside parentheses, respectively. Colors refer to different actual meanings of the **ER** mention.

or closely related concepts (e.g.: Lasix® and diuretic). Figure 1 shows an example. We calculated the cosine similarities between the **distributed relation representation** of these token pairs (Eq. 1) and linear weight vectors for each of the 3 labels in the dataset. Table 3 shows the proportion of token pairs, for each type, which are nearest to the entailment vector.

For all relations, more pair representations from BioELMo are closer to the entailment vector than those from other embeddings. Hence, at the very least, we can conclude that these complex semantic relationships can be modeled linearly in the space of BioELMo representations.

Relation Type	NN w/ Entailment Vector (%)		
	BioELMo	ELMo	Biomed w2v
disease-symptom	**63.6**	40.9	54.5
disease-drug	**69.2**	**69.2**	61.5
number-indication	**73.7**†‡	31.6	31.6
synonyms	**95.8**†‡	70.8	66.7
All	**76.9**†‡	52.6	53.8

Table 3: Nearest neighbor (NN) pairs to the entailment weight vector. † and ‡ show significant ($p < 0.05$ in two-proportion z test) difference with ELMo and biomedical w2v, respectively.

5 Conclusion

We have shown that BioELMo representations are highly effective on biomedical NER and NLI, even without complicated downstream models. This effectiveness comes from their ability to encode entity types and their relations, and hence we believe they should benefit any task which requires such information. Our probing tasks can also be used to test other representation schemes, such as BERT, in a similar manner.

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.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Khyathi Chandu, Aakanksha Naik, Aditya Chandrasekar, Zi Yang, Niloy Gupta, and Eric Nyberg. 2017. Tackling biomedical text summarization: Oaqa at bioasq 5b. *BioNLP 2017*, pages 58–66.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

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

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*, pages 39–43.

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 E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

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

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Golnar Sheikhshab, Inanc Birol, and Anoop Sarkar. 2018. In-domain context-aware token embeddings improve biomedical named entity recognition. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 160–164.

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.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural domain adaptation for biomedical question answering. *arXiv preprint arXiv:1706.03610*.

Henghui Zhu, Ioannis Ch Paschalidis, and Amir Tahmasebi. 2018. Clinical concept extraction with contextual word embedding. *arXiv preprint arXiv:1810.10566*.