2020

LM-Based Word Embeddings Improve Biomedical Named Entity Recognition: A Detailed Analysis

Liliya Akhtyamova

John Cardiff

LM-Based Word Embeddings Improve Biomedical Named Entity Recognition: A Detailed Analysis

Liliya Akhtyamova and John Cardiff(✉)

Social Media Research Group, Technological University Dublin, Dublin, Ireland
akhtyamova@phystech.edu, john.cardiff@tudublin.ie

Abstract. Recent studies have shown that contextualized word embeddings outperform other types of embeddings on a variety of tasks. However, there is little research done to evaluate their effectiveness in the biomedical domain under multi-task settings.

We derive the contextualized word embeddings from the Flair framework and apply them to the task of biomedical NER on 5 benchmark datasets, yielding major improvements over the baseline and achieving competitive results over the current best systems. We analyze the sources of these improvements, reporting model performances over different combinations of word embeddings, and fine-tuning and casing modes.

Keywords: Deep learning · Biomedical named entity recognition · Contextualized word embeddings

1 Introduction

Named entity recognition (NER) is a fundamental basis for many applications such as speech recognition [6], question answering [13], knowledge base population [14], query. One of the areas in which NER and its applications are most useful is the biomedical domain.

However, as the labeling of a corpus for biomedical NER requires domain knowledge, the preparation of high-quality training corpora is usually quite expensive and time-consuming. *Transfer learning* introduced in NLP through the concept of pretrained word embeddings allows us to leverage knowledge about the language semantics more accurately. One of the recent advances of it is *contextualized language modeling* based concept representations.

The release of contextualized word embeddings [3,7,21] has substantially advanced the state-of-the-art in many NLP tasks. It has become possible by learning the *contextual* representations of terms and training of models based on fragments of *contiguous* text that typically span multiple sentences thus capturing long distance relationships within the text fragments better.

Models based on contextualized word embeddings due to the more complex structure of latter in comparison to the standard word embeddings such as

I. Rojas et al. (Eds.): IWBBIO 2020, LNBI 12108, pp. 624–635, 2020.
https://doi.org/10.1007/978-3-030-45385-5_56

Word2vec [19], Glove [20], FastText [4] are better at capturing information from domain-restricted corpora or even the unrelated or general nature corpora.

These advantages of contextualized word embeddings motivate us to apply them to biomedical NER tasks. The identification of biomedical instances in texts can lead to an improvement in the structuring of biomedical and medical knowledge (e.g., biomedical knowledge bases' population) and revealing hidden or unknown previously phenomena from biomedical texts to help clinicians and medical professionals in their routine (e.g., medical database query, decision support systems).

While the most popular version of language representation – BERT [7] is well investigated and has yielded state-of-the-art results on many biomedical benchmark datasets [16], the capabilities of Flair language model (LM) [3] have not yet been researched comprehensively for biomedical NER. Although, Sharma and Daniel [24] in their paper present the BioFlair system learnt over the part of benchmark datasets from Lee et al. [16], they do not learn the model extensively over a variety of combinations of word embeddings and different model architectures.

In this work, we aim to close this gap by (1) incorporating pre-trained contextualized embeddings in a state-of-the-art NER multi-task system [27], obtaining major performance improvements over previous state-of-the-art and competitive to other systems result; (2) for comparability of single-task models, we also experiment with contextualized embeddings integrated into the of-the-shelf Flair NER system[1]; (3) we test model performances over different combinations of the standard, character and contextualized word embeddings as well as parameter settings (casing, fine-tuning).

2 Materials and Methods

The following sections present the technical details of the NER architectures used in this study [1,27]. We first briefly give the problem definition, then describe single-task and multi-task learning systems. We also describe word embeddings and datasets used in the experiments. And finally, we give details on the evaluation metric.

2.1 Problem Definition

The problem of biomedical NER is a sequence labeling task where the goal is to detect the correct spans of entities and assign them the right labels.

To be in line with the results of the original work, to classify entities in Wang's model [27], we used a BIO schema. These classify entities in a document as [B]eginning, [I]nside, [O]utside.

For the Flair NER system, we used the default best settings which include BIOES tagging schema, where B stands for Beginning, I for Inside, O for Out, E for End, and S for Single entity.

[1] https://github.com/zalandoresearch/flair.

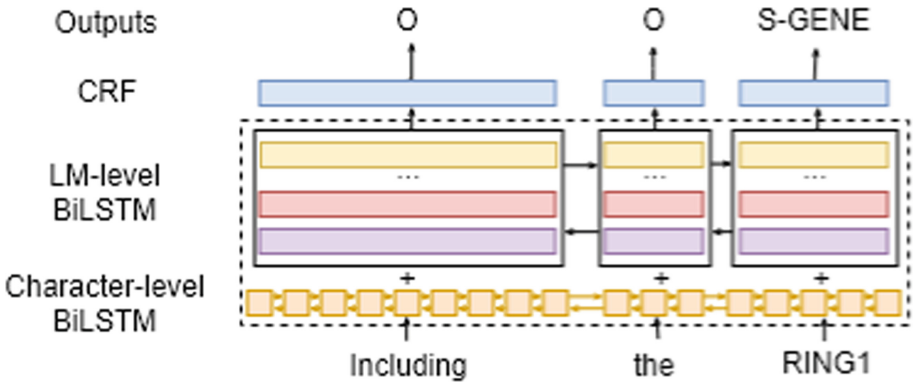

Fig. 1. Architecture of modified Wang et al. deep multi-task learning system. Instead of the original word embeddings, concatenated LM-based word embeddings are used.

2.2 Single-Task Learning

The single-task model (STM) learns on one task at a time. In this work, we experiment with two STMs. The first is the model of Wang et al [27] who in turn adopted their model from Liu et al. [17]. It is the bi-LSTM-CRF model with integrated character-level embeddings to be combined with the word level embedding representations. The character-level embeddings in their models are learned through another bi-LSTM model. As stated by authors, the advantage of their architecture from vanilla bi-LSTM is that its character level word representations allow capturing out-of-vocabulary (OOV) terms and context around words, thus being "contextualized" in some degree. In their architecture, we expanded the word embedding layer with other types of embeddings including contextualized word embeddings. The augmented architecture is presented in Fig. 1.

Another part of experiments was conducted using the Flair NER framework (see Footnote 1) which goes on top of Theano providing a convenient way to experiment with the combination of different types of word embeddings. It consists of a bidirectional Long Short Term Memory (bi-LSTM) network of Huang et al. [12] with the options to select and tune its parameters. We trained it with Conditional Random Fields (CRF) using the default best parameter settings: one LSTM layer, hidden state dimension 256, initial learning rate 0.1 with subsequent halving if the loss does not decrease for 5 epochs, mini-batch size 8, and Adam objective function optimizer.

2.3 Multi-Task Learning

Multi-task learning (MTL) is the task of learning on many tasks in *parallel* by sharing some part of learning model representation between tasks. This approach became popular in NLP [5], computer vision [18], speech recognition [6],

simulation of electrocardiogram signals [23] and other tasks [22] outperforming results gained with STL model architectures.

For our MTL experiments, we utilize again the MTL architecture of Wang et al. [27]. Authors using the single-task based model of Liu et al. [17] built the competitive multi-task model for solving biomedical NER tasks. Using 15 benchmark datasets they showed the substantially better performance of their approach in comparison to STL and other baseline systems.

We took their best-performed MTL architecture - with shared both word and character-level layers - as a foundation for comparison with STL while integrating contextualized, standard and character-level word embeddings.

2.4 Flair Embeddings

In our experiments, we consider the variant of LM-based word embeddings called Flair [3] and ELMo [21]. As in Wang et al. [27] these embeddings are also character-level and trained using the bi-LSTM network. The principal difference of them from Wang et al. [27] is that they are trained in a separate task of LM with next character in a text being the target to predict, while in the model of Wang et al. [27] the character-level embeddings are trained *jointly* as part of NER model with the word label to be the final target to predict.

In comparison to, for example, BERT language representation which allows after pre-training it further fine-tune it for the downstream task, Flair, ELMo and other similar LM architectures do not allow to do it and should be used in a pre-trained form to *embed* sentences of the downstream task's input to form contextualized embeddings.

In this work, we use the *pooled* version of Flair embeddings which keeps the information on each encountered word and reduces the word representation bias when the same word occurs in noisy, under specified context [2]. For example, the biomedical term "hemoglobin" is not ambiguous and its word embedding should not vary heavily in different contents.

We chose Flair and ELMo embeddings both pre-trained on in-domain English PubMed articles[2].

2.5 Additional Embeddings

In numerous papers, it was shown that stacking different types of embeddings mostly improves the quality of NLP models [3,21]. In this work, we integrate the following types of additional embeddings:

[2] http://evexdb.org/pmresources/language-models/.

Table 1. Statistics on biomedical NER datasets

Dataset	Size	Entity types and counts
BC2GM	20,000 sentences	Gene/Protein (24,583)
BC4CHEMD	10,000 abstracts	Chemical (84,310)
BC5CDR	1,500 articles	Chemical (15,935), Disease (12,852)
NCBI-Disease	793 abstracts	Disease (6,881), Gene/Protein (35,336)
JNLPBA	2,404 abstracts	Cell Line (4,330), DNA (10,589), Cell Type (8,649), RNA (1,069)

1. **General-domain word2vec embeddings.**[3] These embeddings are trained over news and Wikipedia data.
2. **Biomedical word2vec embeddings.**[4] These embeddings were trained on 5.5M terms over PubMed, PubMed Central and Wikipedia texts with the window size 200.
3. **Byte-pairwise encoded embeddings (BPE).** They are statistically calculated based on occurrences of sub-word tokens of words [11].
4. **Character-level word embeddings** are trained in the model of Wang et al. [27] using the methodology developed by Lample et al. [15].

2.6 Datasets

For comparative purposes, we test our models on the same datasets as used by [27] (Table 1). Here, NER on BC2GM, BC4CHEMD and NCBI-Disease are *binary classification* problems, and NER on JNLPBA and BC5CDR are *multi-classification* ones.

These datasets cover major biomedical entity types (genes, proteins, chemicals, diseases) and thus were chosen as a standalone set of biomedical datasets by many researchers. All datasets could be downloaded from the GitHub repository of MTL Bioinformatics Lab[5].

In line with Wang et al. [27], below we also briefly mention the origin of datasets and state-of-the-art results on them to the current moment for the systems similar to ours.

BC2GM This dataset was used in the BioCreative II gene mention recognition task. The best result to the moment are holded by to the moment are holded by [26] and Lee et al. [16] (with insignificant difference). Lee et al. [16] utilized LM-based BERT system. They trained a large BERT system on relevant corpora and fine-tuned on a downstream tasks. Wang et al. [26] applied the multi-task learning techniques cross-sharing structure for their neural-network based model.

[3] https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings/CLASSIC_WORD_EMBEDDINGS.md.

[4] http://evexdb.org/pmresources/vec-space-models/wikipedia-pubmed-and-PMC-w2v.bin.

[5] https://github.com/cambridgeltl/MTL-Bioinformatics-2016.

Table 2. Comparative evaluation of proposed STM systems against state-of-the-art systems on five NER tasks.

	NCBI-disease	JNLPBA	BC5CDR	BC4CHEMD	BC2GM
STM_LM	86.47	75.17	88.98	89.34	81.66
STM_Flair	87.13	76.81	90.33	–	82.89
Best published					
STM, Wang et al. [27]	83.92	72.17	86.96	88.75	80.00
Collabonet, Yoon et al. [29]	84.69	–	–	88.19	78.56
BiLM, Sachan et al. [25]	87.34	75.03	89.28	–	81.69
FullyNER, Gupta et al. [10]	88.31	76.20	88.64	–	82.06
BioBERT, Lee et al. [16]	–	–	–	91.41	84.40
BioFlair, Sharma and Daniel [24]	–	–	89.42	–	–
TransferSM, Giorgi and Bader [9]	87.66	–	–	88.98	80.65

BC4CHEMD used in BioCreative IV shared task on chemical entity mention recognition. The state-of-the-art is again belongs to Lee et al. [16] with BioBERT system and Watanabe et al. [28] with multi-task paraphrasing neural network model to utilize paraphrase pairs.

BC5CDR used in BioCreative V challenge on chemical and disease mention recognition. The state-of-the-art for STM is obtained by Sharma and Daniel [24] with their BioFlair system. They combined Flair embeddings with ELMo embeddings trained both over the biomedical corpora using the FLair framework. For MTM, we did not find publicly available recent results on 3 class problem.

NCBI-Disease A collection of 793 fully annotated PubMed abstracts obtained by Doğan et al. [8]. It was widely used by researchers and the current state-of-the-art result belongs again to Wang et al. [26] for SMT and for MTM to Zhao et al. [30] with their jointly performed NER and normalization tasks.

JNLPBA is the 2004 year shared task on biomedical entity recognition of wide range of entities (5 classes). The best STM belongs to Gupta et al. [10] who trained their own version of contextualized word embeddings.

2.7 Evaluation Metric

All datasets are provided with training, development and test data. In our experiments, we merge training and development data, shuffle it and select 10% of it for evaluation of results.

We compared all methods in terms of macro-averaged F-score. It is computed as the harmonic mean of precision and recall. Here, precision is computed as the percentage of the predicted entities that are gold ones, and recall as the percentage of the gold entities that are correctly predicted. The exact entity span match is used for evaluation.

Table 3. Comparative evaluation of proposed MTM systems against state-of-the-art systems on five NER tasks.

	NCBI-disease	JNLPBA	BC5CDR	BC4CHEMD	BC2GM
MTM_LM	86.56	76.01	89.33	89.52	81.82
Best published					
MTM, Wang et al. [27]	86.14	73.52	88.78	89.37	80.74
CollabonetMulti, Yoon et al. [29]	86.36	–	–	88.85	79.73
CompParaph, Watanabe et al. [28]	–	–	–	92.57	–
CrossSharing, Wang et al. [26]	86.50	–	–	–	84.40
TransferMM, Giorgi and Bader [9]	86.89	–	–	88.81	79.60
JointNER, Zhao et al. [30]	87.43	–	–	87.63	–

3 Results and Discussion

In this section, we provide details of the NER results for STMs and MTMs. The section is divided into two broad parts – the first presenting the results of the NER task and comparison with other works, and the second providing comparative, selective results over different variants of word embedding stacking and parameter settings.

3.1 NER Results

The experimental results of the baseline models, models with integrated LM-based word embeddings and current state-of-the-art models are provided in Tables 2 and 3, respectively. Table 2 shows the comparison between the existing state-of-the-art STMs and STM of Wang et al. [27] with and without integrated LM-based word embeddings as well as Flair STM trained using the Flair NER framework [1]. For the MTM, in Table 3 we present the comparison of the MTM system of Wang et al. [27] with and without integrated LM-word embeddings and best published MTM systems.

Note that we do not report results on the BC4CHEMD dataset using the Flair NER framework due to limited computation sources (BC4CHEMD dataset is around four times larger than the next largest dataset used in our analysis). Also, it should be noted that some authors solved only binary entity classification problems, i.e. with one biomedical entity to be predicted. In these cases, we do not report results for these evaluations (missing BC5CDR and JNLPBA datasets' result entries).

Overall, for both STM and MTM the NER performances in all cases using Wang et al. [27] model architecture significantly benefit from incorporating additional embeddings with the maximum gain of 2.4% achieved by MTM on the JNLPBA dataset.

Moreover, for STM in all cases *STM_Flair* results outperform *STM_LM* results. It is probably due to the fact that in Flair NER, the default hyperparam-

Table 4. Results of experiments over different combinations of word embeddings

	Combination	F-score
BC2GM	embeddings pubmed	78.96
	embeddings pubmed+pubmed Flair	81.05
	embeddings pubmed+pubmed Flair+bpe	81.66
	embeddings pubmed+ELMo+bpe	81.40
NCBI	embeddings pubmed+pubmed Flair+bpe	86.40
	embeddings pubmed+pubmed Flair+ELMo+bpe	85.10
	embeddings pubmed+ELMo+bpe	86.42
	ELMo+bpe	86.30

Table 5. Case sensitivity of model

	Combination	F-score (caseless)	F-score (case-sensitive)
NCBI	embeddings pubmed+pubmed Flair+bpe	85.46	86.07
BC2GM	embeddings pubmed+pubmed Flair+bpe	80.28	81.66

eters values were more thoughtfully selected and some additional mechanisms of Flair NER model such as learning rate annealing, gradient clipping, etc.

Overall, the constructed STM and MTM achieve higher F1-score than most other models of similar complexity on all datasets.

With relation to *BioBERT*, we can only compare results on BC2GM dataset (1.8% lower) as for BC4CHEMD dataset we did not calculate *STM_Flair* results however *STM_LM* results stand not far from *BioBERT* ones on the BC4CHEMD dataset. This is good results taking into account the complexity of the BERT model (large BERT consists of 24 layers, 1024 hidden layers, and total of 340 M parameters).

3.2 Discussion

Combination of Word Embeddings. We wanted to compare the results over different combinations of word embeddings to evaluate the performance gain while increasing the complexity of embedding layer. The results over two benchmark datasets using the single-task model of Wang et al. [27] are presented in Table 4.

It should be noted that for experiments where Flair embeddings are used we did not fine tune the model, however for ELMo model we fine-tune the model as from our observations fine-tuning for model with ELMo embeddings works the best.

From the results, presented in Table, it could be seen that overall increasing the complexity of word level representation by adding more different types of

Table 6. Results of fine-tuning the model

	Combination	F-score (no fine-tuning)	F-score (fine-tuning)
NCBI	embeddings pubmed+pubmed Flair+bpe	86.40	86.07
	embeddings pubmed+pubmed ELMo+bpe	86.33	86.47

embeddings improves results. However, two complex similarly constructed word embeddings such as ELMo and Flair coupled together in one model deteriorate results. Overall, Flair embeddings usually give on par or better results.

Case Sensitivity. The results of experiments with lower-casing the words and without lower-casing on two benchmark datasets are presented in Table 5.

Lower-casing always positively influences the performance of model. Indeed, many biological terms are upper-cased or start with upper-cased letter. In addition to the formal nature of benchmark datasets, all these requires leaving the textual data "as it is".

Fine Tuning. The results of experiments with and without fine-tuning the model for NCBI dataset are presented in Table 6.

While fine-tuning process works better for ELMo embeddings, it deteriorates results when using Flair embeddings. The reason for that should be investigated further.

4 Conclusion

In this paper, we focused on the problem of biomedical NER. In particular, we attempted to investigate approaches by which LM-based word embeddings can be applied to improve the automatic NER on textual data containing biomedical entities. Our particular focus was on scientific literature texts. Our results strongly suggest that integrating contextualized embeddings and combining them with other types of embeddings can improve sequence labeling accuracy. As such, there is a strong motivation to explore other ways to integrate contextualized information into the current state-of-the-art NER models to further boost their performance. types of advances in language representation such as transformers, etc.

We explored the incorporation of LM-based embeddings in the strong multi-task learning framework. The incorporation of such embeddings has shown to improve the baseline on all tasks. We suggest investigating further the behavior of LM-based embeddings under multi-task learning settings.

Lastly, we conducted a comparative analysis of different model architectures, text preprocessing techniques, and model parameter settings. Simple off-shelf Flair NER architecture turned out giving better performance rather than the more sophisticated architecture of Wang et al. [27]. Preprocessing in terms of

lower-casing and fine-tuning the model deteriorates results. However, fine-tuning showed to work well on ELMo embeddings for Wang et al. [27] architecture.

In the future, we would like to explore other combinations of word embeddings and different NN architectures. Moreover, as mostly NER tasks in the biomedical domain are unbalanced problems, future research on improving model parameter settings to handle this problem should improve the results of biomedical sequence labeling as well.

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: an easy-to-use framework for state-of-the-art NLP. In: Proceedings of the 2019 Conference of the North, pp. 54–59. Association for Computational Linguistics, Stroudsburg (2019). https://doi.org/10.18653/v1/N19-4010, http://aclweb.org/anthology/N19-4010
2. Akbik, A., Bergmann, T., Vollgraf, R.: Pooled contextualized embeddings for named entity recognition. In: NAACL (2019). https://github.com/zalandoresearch/flair
3. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: COLING (2018). https://github.com/zalandoresearch/flair
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**(2307–387X), 135–146 (2017). http://arxiv.org/abs/1607.04606
5. Collobert, R., Weston, J.: A unified architecture for natural language processing, pp. 160–167. Association for Computing Machinery (ACM) (2008). https://doi.org/10.1145/1390156.1390177
6. Deng, L., Hinton, G., Kingsbury, B.: New types of deep neural network learning for speech recognition and related applications: an overview. In: IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, ICASSP, pp. 8599–8603, October 2013. https://doi.org/10.1109/ICASSP.2013.6639344, ISBN 9781479903566, ISSN 15206149
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, October 2018. http://arxiv.org/abs/1810.04805
8. Doğan, R.I., Leaman, R., Lu, Z.: NCBI disease corpus: a resource for disease name recognition and concept normalization. J. Biomed. Inform. **47**, 1–10 (2014). https://doi.org/10.1016/j.jbi.2013.12.006. ISSN 15320464
9. Giorgi, J.M., Bader, G.D.: Towards reliable named entity recognition in the biomedical domain. Bioinformatics **36**(1), 280–286 (2020). https://doi.org/10.1093/bioinformatics/btz504. https://academic.oup.com/bioinformatics/article/36/1/280/5520946, ISSN 1367–4803
10. Gupta, A., Goyal, P., Sarkar, S., Gattu, M.: Fully contextualized biomedical NER. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019. LNCS, vol. 11438, pp. 117–124. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15719-7_15
11. Heinzerling, B., Strube, M.: BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 18–1473 (2018). https://aclweb.org/anthology/papers/L/L18/L18-1473/

12. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF Models for Sequence Tagging, August 2015. http://arxiv.org/abs/1508.01991
13. Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., Lu, X.: PubMedQA: a dataset for biomedical research question answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 2567–2577, September 2019. http://arxiv.org/abs/1909.06146
14. Kim, D., et al.: A neural named entity recognition and multi-type normalization tool for biomedical text mining. IEEE Access **7**, 73729–73740 (2019). https://doi.org/10.1109/ACCESS.2019.2920708. https://ieeexplore.ieee.org/document/8730332/, ISSN 2169-3536
15. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: 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, pp. 260–270. Association for Computational Linguistics, Stroudsburg (2016). https://doi.org/10.18653/v1/N16-1030, http://aclweb.org/anthology/N16-1030
16. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (btz682) (2019). https://doi.org/10.1093/bioinformatics/xxxxxx, https://github.com/dmis-lab/biobert
17. Liu, L., et al.: Empower sequence labeling with task-aware neural language model, September 2017. http://arxiv.org/abs/1709.04109
18. Liu, S., Johns, E., Davison, A.J.: End-to-End Multi-Task Learning with Attention, March 2018. http://arxiv.org/abs/1803.10704
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119 (2013). https://arxiv.org/pdf/1310.4546.pdf
20. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014). https://nlp.stanford.edu/pubs/glove.pdf
21. Peters, M.E., et al.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365, February 2018. http://arxiv.org/abs/1802.05365
22. Ruder, S.: An overview of multi-task learning in deep neural networks, June 2017. http://arxiv.org/abs/1706.05098
23. Sarkar, P., Ross, K., Ruberto, A.J., Rodenburg, D., Hungler, P., Etemad, A.: Classification of cognitive load and expertise for adaptive simulation using deep multitask learning, July 2019. http://arxiv.org/abs/1908.00385
24. Sharma, S., Daniel, R.: BioFLAIR: pretrained pooled contextualized embeddings for biomedical sequence labeling tasks. arXiv preprint arXiv:1908.05760, August 2019. http://arxiv.org/abs/1908.05760
25. Sachan, D.S., Xie, P., Sachan, M., Xing, E.P.: Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. Technical report (2018). https://arxiv.org/pdf/1711.07908.pdf
26. Wang, X., Lyu, J., Dong, L., Xu, K.: Multitask learning for biomedical named entity recognition with cross-sharing structure. BMC Bioinformatics **20**(1), 427 (2019). https://doi.org/10.1186/s12859-019-3000-5. ISSN 14712105
27. Wang, X., et al.: Cross-type biomedical named entity recognition with deep multi-task learning. Bioinformatics **35**(10), 1745–1752 (2018). https://doi.org/10.1093/bioinformatics/xxxxxx. https://github.com/yuzhimanhua/lm-lstm-crf

28. Watanabe, T., Tamura, A., Ninomiya, T., Makino, T., Iwakura, T.: Multi-task learning for chemical named entity recognition with chemical compound paraphrasing. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 6243–6248 (2019). https://pubchem.ncbi.nlm.nih.gov/
29. Yoon, W., So, C.H., Lee, J., Kang, J.: CollaboNet: collaboration of deep neural networks for biomedical named entity recognition (2019). https://doi.org/10.1186/s12859-019-2813-6
30. Zhao, S., Liu, T., Zhao, S., Wang, F.: A Neural multi-task learning framework to jointly model medical named entity recognition and normalization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 817–824 (2019). https://doi.org/10.1609/aaai.v33i01.3301817, https://aaai.org/ojs/index.php/AAAI/article/view/3861, ISSN 2374–3468