A Comparison between Named Entity Recognition Models in the Biomedical Domain

Maria Carmela Cariello[1][0000−0001−5001−0360], Alessandro Lenci[2][0000−0001−5790−4308], and Ruslan Mitkov[3][0000−0002−6074−2749]

[1] University of Pisa, Pisa 56126, Italy
m.cariello1@studenti.unipi.it
[2] University of Pisa, Pisa 56126, Italy
alessandro.lenci@unipi.it
[3] University of Wolverhampton, Wolverhampton, WV1 1LY, UK
r.mitkov@wlv.ac.uk

Abstract. The domain-specialised application of Named Entity Recognition (NER) is known as Biomedical NER (BioNER), which aims to identify and classify biomedical concepts that are of interest to researchers, such as *genes*, *proteins*, *chemical compounds*, *drugs*, *mutations*, *diseases*, and so on. The BioNER task is very similar to general NER but recognising Biomedical Named Entities (BNEs) is more challenging than recognising proper names from newspapers due to the characteristics of biomedical nomenclature. In order to address the challenges posed by BioNER, seven machine learning models were implemented comparing a transfer learning approach based on fine-tuned BERT with Bi-LSTM based neural models and a CRF model used as baseline. Precision, Recall and F1-score were used as performance scores evaluating the models on two well-known biomedical corpora: JNLPBA and BIOCREATIVE IV (BC-IV). Strict and partial matching were considered as evaluation criteria. The reported results show that a transfer learning approach based on fine-tuned BERT outperforms all others methods achieving the highest scores for all metrics on both corpora.

Keywords: Biomedical NER · Deep Learning · Transfer Learning.

1 Introduction

Named Entity Recognition (NER) is a task that aims to recognise and classify mentions of named entities in unstructured text into pre-defined semantic categories such as *person*, *organisation*, *location*, *time expression*, *monetary value*, and so on. In Natural Language Processing (NLP), NER not only acts as a tool for information extraction (IE), but plays an essential role in a variety of downstream applications such as information retrieval [7], text summarisation [13], machine translation [2], question answering [14], and many other NLP tasks.

The interest in NER is not a novelty, but it has been increasing in recent years due to the exponential growth of digital information that stimulated domain-specific applications of NER in order to extract entity mentions not only from

Translation and Interpreting Technology Online, pages 76–84,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_009

general texts, such as newspaper articles, but also from specialised texts. In many applied research domains, NER is directly used to alleviate the problem of the search and discovery of information, becoming an invaluable tool particularly in those research areas where it is difficult for researchers to keep up with relevant publications [20].

One specialised application of NER is known as Biomedical named entity recognition (BioNER), which is defined as the task of identifying and classifying Biomedical Named Entities (BNEs), technical terms referring to key concepts that are of interest to biomedical researchers, such as *gene*, *protein*, *chemical compound*, *drug*, *mutation*, *disease*, and so on. BioNER has gained increasing attention from the research community. In fact, many works in medicine focus on the analysis of scientific articles to find out hidden relationships between BNEs, such as *gene* and *protein*, in order to drive experimental research [20]. Although a large body of systems are dedicated to extract BNEs in scientific literature, BioNER tools can be applied to find all kinds of entities in any kind of health related text, including radiology reports and clinical notes [19].

Generally, BioNER is considered a more challenging task compared to domain-independent NER due to the characteristics of biomedical nomenclature. The lack of standardised naming conventions, the frequent crossover in vocabulary, the excessive use of abbreviations, synonyms and variations, the morphological complexity due to the use of unusual characters such as Greek letters, digits, punctuation – these are just some of the factors making the recognition of BNEs particularly difficult for BioNER systems. Moreover, biomedical text often contains complex multi-word BNEs and, especially in the area of gene and protein names, multi-word BNEs are rather the rule than the exception. Not only multi-word BNEs are more difficult to identify, but in many cases there is also no agreement on the exact borders of such names, making the evaluation of BioNER tools complex [11]. For example, many BNEs may contain verbs and adjectives that are embedded in names, making a legitimate gene or protein name hard to distinguish from the general language text surrounding it. Lastly, the biomedical domain is an expanding field where new concepts emerge daily and new names are coined on a daily basis. In addition, new variants are always created for already existing concepts since biomedical concepts are studied in different branches of medicine which use different naming conventions.

To address these challenges, seven Machine Learning (ML) models were implemented following a Sequence Tagging (ST) approach[4]. A transfer learning approach based on fine-tuned BERT is compared to Bi-LSTM-based neural models and a CRF model used as baseline. The impact of pre-trained word embedding models on the performances of neural models is also investigated. The comparison between models is carried out by evaluating the performances on two well-known BioNER corpora.

The rest of the paper is structured as follows. Section 2 presents the data used in this study. Section 3 outlines the models employed in our experiments,

[4] The Colab notebooks used for running the experiments are available here: https://github.com/cariello1/BioNER.

which are described in Section 4. Section 5 discusses the results obtained and finally section 6 summarises the conclusions of this study.

2 Data

BioNER models were evaluated using two benchmark corpora released during well-known and popular shared competitions. The first one is the corpus of the JNLPBA 2004 shared task, which is derived from the popular GENIA corpus. The second one is the BIOCREATIVE (BC-IV) corpus used for the Track 2 of BioCreative IV shared task. Both corpora were made publicly available in the IOB2 annotation format.[5] According to this schema, tokens are labelled with a *B-class* tag at the beginning of every sequence that represents an entity, with an *I-class* tag if the tokens are inside a sequence and with an *O* tag if the tokens are outside of a sequence that represents an entity.

2.1 JNLPBA 2004 Shared Task Corpus

Derived from the GENIA corpus, JNLPBA [8] is a manually annotated collection of articles extracted from the MEDLINE database. Compared to the 36 classes of the original corpus, JNLPBA has 5 classes: *protein*, *DNA*, *RNA*, *cell line* and *cell type*, and does not contain any nested or discontinuous entities. The training set includes entirely the GENIA corpus, while the test set consists of 404 newly annotated MEDLINE abstracts from the GENIA project. The training set contains 18,546 sentences for a total of 472,006 words, while the test set contains 3,856 sentences for a total of 96,780 words.

2.2 BioCreative IV CHEMDNER Corpus

BioCreative IV CHEMDNER (BC-IV) [10] is a collection of PubMed abstracts which contains chemical entity mentions labelled manually by experts in the field, following annotation guidelines specifically defined as part of the BioCreative IV competition. No nested annotations or overlapping entity mentions are included. The original fine-grained annotation schema including seven classes was collapsed into one generic class, *CHEMICAL*. The training set contains 30,682 sentences for a total of 891,948 words, while the test set contains 26,364 sentences for a total of 766,033 words.

3 Models

Considering BioNER as a Sequence Tagging (ST) task, seven models were implemented in order to solve BioNER and compare performances of a traditional ML algorithm used as a baseline with the latest advanced neural models.

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

3.1 Conditional Random Field

Conditional Random Fields (CRF) is a probabilistic graphical model, which provides a framework for modelling global probabilities based on some observations of the local functions and representing a distribution over labels [3]. CRF has the advantage over other ML algorithms to efficiently model dependencies between observations and labels, taking context into account. Among traditional ML algorithms, CRF is known as the most popular solution for solving ST tasks such as BioNER [9].

Given the morphological complexity behind BNEs, which are rich of unusual characters, applying features that have been used for traditionally named entities to identify biomedical instances could be insufficient. A specific set of features that exploit biomedical nomenclature characteristics needs to be engineered, in order to allow the algorithm to efficiently recognise BNEs [1].

3.2 Bi-LSTM Based Neural Networks

Bi-LSTM is a type of Recurrent Neural Network (RNN) that is widely used as context-encoder for ST tasks such as BioNER. RNNs are able to model context dependencies storing information during the sequential processing implementing units with self-connections [15]. However, standard RNNs suffer from the exploding gradient problem, which is responsible for the reduction in the ability to learn long-distance relationships, so that they have a limited application to real-world ST. LSTMs extend RNNs with a memory cell unit consisting of several gates to store and access information over long periods of time, efficiently modelling dependencies between far apart sequence elements as well as consecutive elements. Since LSTMs can access context only in one direction, Bidirectional LSTMs (Bi-LSTMs) are used instead, in order to scan the data in both directions and provide access to all surrounding context. Bi-LSTM combines the benefits of long-range memory and bidirectional processing, which make this model perfectly suitable for ST [6].

3.3 Fine-tuned BERT

BERT [5] is a pre-trained system based on Transformer that can be fine-tuned to solve specific language tasks. The Transformer [18] is a neural architecture which dispenses with recurrence entirely relying only on the attention mechanism to draw global dependencies between input and output. Since Transformers do not rely on sequential processing, they can process an input sequence of words all at once, allowing for much more parallelisation and requiring significantly less time to train compared to Bi-LSTM-based models [17]. Since Transformers allowed for a more efficient training on larger datasets than it was possible before they were introduced, they drastically improved the prospects of using Transfer Learning

for Natural Language Processing (NLP). Indeed, in the last years a shift has occurred from the use of pre-trained word vectors for feature extraction to the use of pre-trained systems such as BERT, that has been trained on a huge general language dataset and can be fine-tuned to solve a wide variety of NLP tasks [12].

4 Experiments

The first experiment is aimed at training a CRF model, a traditional ML method widely used for solving BioNER that is easy to implement, provides reasonable results, and does not require much expertise and time to build. The CRF performance is considered as a baseline for evaluating the performance of the other models. For the CRF model, a specific set of features is used to allow the algorithm to recognise BNEs. Linguistic features are selected exclusively to exploit the characteristics of biomedical BNEs such as the morphological complexity. To enable the model to capture contextual information, context features are also provided in a 5-word window.

The first neural model implemented is a Bi-LSTM-based architecture that uses Softmax layer as the decoding layer. The embedding layer is initialised with random weights and computes word vectors during the learning process. Learned representations of data are fed into a Bi-LSTM layer which extracts contextual information. The output is then passed to another Bi-LSTM layer so that the model learns even deeper, more abstract representations from data. The decoding layer uses a Softmax function to transform scores into a probability distribution over classes. Labels for each word are independently predicted without taking into account dependencies between labels. In a second model, Softmax was replaced with CRF to make the model capable of capturing relationships between entity labels. It has been shown indeed that for ST tasks it is more beneficial to jointly decode label sequences using CRF than decoding each label independently [4].

In the next models, the Bi-LSTM+CRF architecture was enhanced replacing randomly initialised word vectors with different pre-trained distributed representation models. First, Bi-LSTM+CRF was combined with pre-trained vectors from FastText.[6] Next, these vectors were replaced with a concatenation of word-level and character-level representations using pre-trained word embedding from GloVe[7] and character embedding learned using an LSTM model. Character representations are able to capture sub-word level information such as prefix, suffix and orthographic characteristics enabling the model to handle the Out-Of-Vocabulary (OOV) problem, which causes GloVe to return many zero values. The last Bi-LSTM based neural model uses contextual embedding incorporating into the embedding layer a pre-trained ELMo [16] model. This model does not consider an additional character-level embedding, unlike the model with GloVe, since ELMo already provides context-dependent character-level representations.

[6] https://fasttext.cc/docs/en/crawl-vectors.html
[7] https://nlp.stanford.edu/projects/glove/

Finally, a fine-tuned BERT-Large model is employed. In the pre-processing step, a WordPiece tokenizer is used in order to allow the model to process words that it has never seen before by decomposing them into known sub-words. For restoring the original tokenisation, a post-processing step is needed in order to compare BERT outputs with those of the other models. The hyper-parameter settings and all the fine-tuning procedure rely on the indication provided on the original paper by Devlin et al. [5]. Due to the high number of parameters, the model is trained on an NVIDIA Tesla K80 16GB GPU.

5 Results

The results obtained on the test set for both corpora are shown in Tables 1 and 2. Precision, Recall and F1-score are reported according to the strict matching criterion, and the overall scores for JNLPBA are computed using micro-average. The F1-score computed according to the partial matching criterion is also reported. For what concerns the neural models, each model is run five times and the final reported result is the average among the runs.

JNLPBA				
Model	**Precision**	**Recall**	**F1**	**F1 (partial)**
CRF	0.68	0.69	0.69	0.78
Bi-LSTM+Softmax	0.68	0.69	0.69	0.78
Bi-LSTM+CRF	0.68	0.70	0.69	0.77
FastText+Bi-LSTM+CRF	0.67	**0.74**	**0.70**	0.77
GloVe+Char+Bi-LSTM+CRF	**0.68**	**0.75**	**0.71**	**0.79**
ELMO+Bi-LSTM+CRF	0.63	**0.77**	0.69	0.78
Fine-tuned BERT	**0.68**	**0.77**	**0.72**	**0.79**

Table 1. Overall performance of the models on JNLPBA.

BIOCREATIVE IV				
Model	**Precision**	**Recall**	**F1**	**F1 (partial)**
CRF	0.86	0.73	0.79	0.83
Bi-LSTM+Softmax	0.85	0.74	0.79	0.83
Bi-LSTM+CRF	0.77	0.83	0.80	0.87
FastText+Bi-LSTM+CRF	0.82	0.77	0.80	0.88
GloVe+Char+Bi-LSTM+CRF	0.83	**0.82**	**0.83**	0.88
ELMO+Bi-LSTM+CRF	0.77	**0.87**	**0.82**	0.86
Fine-tuned BERT	**0.89**	**0.87**	**0.88**	**0.93**

Table 2. Overall performance of the models on BIOCREATIVE IV.

The overall results show that BERT outperforms the other models, since it achieves the highest scores on both corpora. BERT proves to be able to effectively recognise and classify BNEs, despite being a model trained on text different from

the target domain. For JNLPBA the scores do not differ in terms of precision compared to the other models but recall shows an improvement of 8% over the baseline model. A slight increase is also recorded for the F1-score compared to the baseline model, achieving 72% (against 69%) and 79% (against 78%) according to, respectively, the strict and partial matching evaluation criteria. The second and the third best performing models are respectively the Bi-LSTM+CRF that incorporates the GloVe+Character embedding and the model that incorporates the FastText embedding. A significant increase for the recall and a slight increase for the F1-score are recorded for both strict and partial matching over the baseline model. For BC-IV, instead, BERT stands out significantly over the other models, achieving outstanding scores on all metrics. Specifically, BERT outperforms the baseline model by 14% on recall and achieves an F1-score of 88% (against 79%) and 93% (against 83%) according to, respectively, the strict and partial matching evaluation criteria. For what concerns the precision the increase on the baseline is instead less remarkable. The BERT model outperforms also the second and the third best performing models that in this case are, respectively, the Bi-LSTM+CRF that incorporates the GloVe+Character embedding and the model that incorporates the ELMo embedding. A significant increase is recorded for all the metrics with the exception of the recall, where the ELMo model achieves a score comparable to BERT.

Using the GPU, the training of BERT on the BC-IV corpus requires only 20 minutes, while the Bi-LSTM models require more then 30 minutes. Therefore, even if the use of a GPU is required to fine-tune BERT, this model clearly outperforms the other approaches for the recognition of BNEs on both the biomedical test corpora.

6 Conclusion

Seven Machine Learning (ML) models were implemented following a Sequence Tagging (ST) approach for solving BioNER on two well-known corpora. A transfer learning approach based on fine-tuned BERT was compared with Bi-LSTM-based neural models and a CRF model used as baseline. The fine-tuned BERT model achieved the highest scores for all metrics on both corpora. Thus, according to what emerged from these experiments, the use of pre-trained vectors has a significant impact on the performance of the Bi-LSTM models, leading to an OOV error reduction and an increase of the recall. In addition, the inclusion of sub-word level information into the models proved to be particularly beneficial for solving BioNER on both corpora. Based on these results, the use of pre-trained transformer-based neural models such as BERT for solving BioNER looks promising. Specifically, the advantage of using BERT for BioNER lies in the fact that it can be employed as a ready-to-use model that can be easily fine-tuned for solving the task, requiring significantly less time to train and achieving superior performance scores compared to other approaches.

References

References

1. Alshaikhdeeb, B., Ahmad, K.: Biomedical Named Entity Recognition: A Review. Artificial Intelligence Review, 6(6), pp. 889, (2016)
2. Babych, B., Hartley, A.: Improving Machine Translation Quality with Automatic Named Entity Recognition. In: Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003, pp. 1-8. Association for Computing Machinery, Budapest, Hungary (2003)
3. Bengong Y., Zhaodi F.: A comprehensive review of conditional random fields: variants, hybrids and applications. Artificial Intelligence Review, 6(53), pp. 4289–4333, (2020)
4. Cho, H., Lee, H.: Biomedical named entity recognition using deep neural networks with contextual information. BMC Bioinformatics, 20, 735 (2019)
5. Devlin, J.,Chang, M., Lee, K., Toutanova, K. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019)
6. Graves, Alex: Supervised Sequence Labelling with Recurrent Neural Networks. Springer, Berlin (2012)
7. Guo, J., Xu, G., Cheng, X., Li, H.: Named Entity Recognition in Query. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 267–274. Association for Computing Machinery, Boston, MA, USA (2009)
8. Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N.: 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, pp. 70–75. Association for Computational Linguistics, Geneva, Switzerland (2004)
9. Kocaman V., Talby D.: Biomedical Named Entity Recognition at Scale. ICPR Workshops (2020)
10. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D., Sayle, R., Batista-Navarro, R., Rak, R., Huber, T., Rocktäschel, T., Matos, S., Campos, D., Tang, B., Xu, H., Valencia, A.: The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of Cheminformatics, 7(1), S2 (2015)
11. Leser, U. and Hakenberg, J.: What makes a gene name? Named entity recognition in the biomedical literature. Briefings in bioinformatics, 6(4), pp.357-369 (2006)
12. Malte A., Ratadiya P.: Evolution of transfer learning in natural language processing. ArXiv, abs/1910.07370 (2019)
13. McDonald, D. M., Chen, H.: Summary in Context: Searching versus Browsing. ACM Transactions on Information Systems (TOIS) **24**(1), 111–141 (2006)
14. Molla-Aliod, D., Zaanen, M., Smith, D.: Named entity recognition for question answering. In: Proceedings of the Australasian Language Technology Workshop 2006, pp. 51–58. Australasian Language Technology Association, Sancta Sophia College, Sydney, Australia (2006)
15. Nayel, H., Shindo, H., Shashirekha, H., Matsumoto, Y.: Improving Multi-Word Entity Recognition for Biomedical Texts. International Journal of Pure and Applied Mathematics, 118(16), pp. 301-320 (2018)

16. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: 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), pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana, USA (2018)
17. Li, J., Sun, A., Han J., Li, C.: A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge & Data Engineering, 1, pp. 1-1, (2020)
18. Vaswani A., Shazeer N.M., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I.: Attention is All you Need. ArXiv, abs/1706.03762 (2017)
19. Zhang S., Elhadad N.: Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. Journal of Biomedical Informatics, 46(6), pp. 1088-1098 (2013)
20. Zweigenbaum, P., Demner-Fushman, D., Yu, H., Cohen, K. B.: Frontiers of biomedical text mining: current progress. Briefings in bioinformatics, 8(5), pp. 358–375 (2007)