Multi-task Learning Applied to Biomedical Named Entity Recognition Task

Tahir Mehmood[1,2], Alfonso Gerevini[2], Alberto Lavelli[1], and Ivan Serina[2]

[1]Fondazione Bruno Kessler, Via Sommarive, 18 - 38123 Trento, Italy
{t.mehmood,lavelli}@fbk.eu
[2]Department of Information Engineering, University of Brescia, Italy
{t.mehmood,alfonso.gerevini,ivan.serina}@unibs.it

Abstract

Recent deep learning techniques have shown significant improvements in biomedical named entity recognition task. However, such techniques are still facing challenges; one of them is related to the limited availability of annotated text data. In this perspective, with a multi-task approach, simultaneously training different related tasks enables multi-task models to learn common features among different tasks where they share some layers with each other. It is desirable to used stacked long-short term memories (LSTMs) in such models to deal with a large amount of training data and to learn the underlying hidden structure in the data. However, the stacked LSTMs approach also leads to the vanishing gradient problem. To alleviate this limitation, we propose a multi-task model based on convolution neural networks, stacked LSTMs, and conditional random fields and use embedding information at different layers. The model proposed shows results comparable to state-of-the-art approaches. Moreover, we performed an empirical analysis of the proposed model with different variations to see their impact on our model.

1 Introduction

Named entity recognition (NER) consists in recognizing chunks of text and labelling them with predefined categories (e.g., person name, organization, location, etc). NER is an information extraction task and has many applications for instance in co-reference resolution, question answering systems, machine translation, information retrieval etc (Chieu and Ng, 2002). NER is also performed on biomedical data where it involves recognizing biomedical concepts (e.g., cell, chemical, drug, disease, etc) and classifying them into predetermined categories. This is referred as biomedical named entity recognition (BioNER). Large amounts of medical data are available as free, unstructured text and the quantity of annually generated biomedical data like books, scientific papers, and other publications makes it challenging for physicians to stay up to date.

Moreover, biomedical documents are more complex than normal texts and the names of the entities show peculiar characteristics. Long multi-word expressions *(10-ethyl-5-methyl-5,10-dideazaaminopterin)*, ambiguous words (*TNF alpha* can be used for both DNA and Protein) (Gridach, 2017), spelling alternations (e.g., *10-Ethyl-5-methyl-5,10-dideazaaminopterin* vs. *10-EMDDA*) make the BioNER task even more challenging (Giorgi and Bader, 2018). BioNER is also an important preliminary task for other tasks like the extraction of relations between entities (e.g., chemical induced disease relation, drug-drug interaction, . . .).

Recent applications of deep learning in BioNER minimize manual feature engineering process and at the same time produce promising results. Deep learning is now the state-of-the-art technique but, due to the complex structure of biomedical text data, deep learning models have difficulties in performing efficiently. Moreover, these systems require large amounts of input data while the available annotated biomedical data are not enough to train these systems effectively. Manually generating annotated biomedical text data is an expensive and time-consuming job. In order to address this limitation, one solution is to take advantage of a multi-task learning approach. Multi-task learning (MTL) involves training simultaneously different but related tasks together. Such an approach has shown significant improvements in different fields.

In this paper, we propose a multi-task model (MTM-CW) using convolutional neural networks (CNN) (dos Santos and Guimarães, 2015), stacked layers of Bidirectional long-short term memories (BiLSTM), and conditional random fields (CRFs). Furthermore, we have conducted an empirical analysis of the impact of different word input representation to our model.

The rest of the paper is organized as follows; Section 2 gives a brief background of the multi-task learning followed by Section 3 where our multi-task model (MTM-CW) is discussed. Experimental setup is presented in Section 4 which is followed by the results and discussion (Section 5). Section 6 concludes and presents possible future research directions.

2 Multi-task Learning

In general, deep learning model performance highly depends on the amount of annotated data available. It performs better when large amount of data is available. Unfortunately, in different biomedical tasks only a limited quantity of annotated text data is available and in this case deep learning models have difficulties to generalize well. Moreover, manually annotating new data is a time consuming job and this issue can be reduced by using two methods: transfer learning and multi-task learning.

In transfer learning, the model is partially trained on an auxiliary task and is then reused on the main task. This enables the model to fine tune the weights of the layers which are learned during the training on the auxiliary task. This helps the model to generalize well on the main task, which implies learning generalized features between the auxiliary and the main task. This method learns and transfers shallow features from one domain to another domain (Luong et al., 2016).

On the other hand, multi-task learning (MTL) is an approach where different related tasks are trained simultaneously. Unlike transfer learning, multi-task learning optimizes the model under construction concurrently. In MTL approach, some of the layers in the model are shared among different tasks while keeping some layers task-specific. Training jointly on related tasks helps the multi-task model to learn common features among different tasks by using shared layers (Bansal et al., 2016). The task-specific layers, usually the lower layers, learn features that are more related to the current task. MTL lowers the chances of overfitting as the model has to learn the common representation among all tasks. MTL has been widely adopted in many different domains (Luong et al., 2016).

Crichton et al. (2017) proposed a multi-task model (MTM) based on CNN to perform BioNER. However, they only focused on the word level features ignoring the character level ones. Although word level features give much information about the entities, character level features help to extract common sub-word structures among the same entities. Moreover, depending solely on the word level features can lead to out-of-vocabulary problems when a specific word is not found in the pre-trained word embedding. Wang et al. (2019) also performed BioNER using different multi-task models. They found that the MTM with the word level features and extraction of the character level features using BiLSTM enhances performance of the model. They concluded that the character level feature should be considered for the BioNER task. A similar model is proposed by Mehmood et al. (2019) where, apart from single shared BiLSTM, they introduce the task-specific BiLSTM as well to learn the features that are more specific to the task. Introduction of task-specific BiLSTM and use of CNN instead of BiLSTM at character level showed performance improvement.

3 Our Proposal

Neural networks work on a concept of hierarchical feature learning (Xiao et al., 2018). Hierarchical feature learning is done as sequences propagates through the network (LeCun et al., 2015). Deep learning can learn the complex hierarchical structure of the sequence with multiple layers. Moreover, it is always desirable to stack LSTMs when a large amounts of training data is available (Li et al., 2018). Such intuition can be noticed in the model proposed by Mehmood et al. (2019) where increasing the layer of BiLSTM leads to performance enhancement. However, moving towards deep LSTMs network can causes gradient vanishing problem as well (Li et al., 2018).

To tackle this issue we are proposing a model which induces the input information at different layers. Our proposed multi-task model with character and word input representations (MTM-CW) propagates input embedding information along different shared layers as shown in Figure 1. This not only helps lower layers to learn the complex structure from encoded representation of the previous layer but also considers inputs embeddings as well to overcome the gradient vanishing problem in stacked LSTMs.

Furthermore, using stacked BiLSTMs will help hidden states of BiLSTM to learn hidden structure of the data presented at different level. This will help BiLSTM to learn features at a more abstract level. Apart from the shared stacked BiLSTMs, our model also uses task-specific BiLSTM as well to extract task-specific features. Furthermore, we use CNN to extract features at character level. Many of the previous approaches have used CNN at character level (dos Santos et al., 2015; Collobert et al., 2011) due to its finer ability of features extraction. CNN learns global level features from local level features. This enables CNN to extract more hidden features. More specifically, lower layers in our proposed MTM-CW model are task-specific. So for the specific task, both shared layers and layers belonging to that specific task are activated.

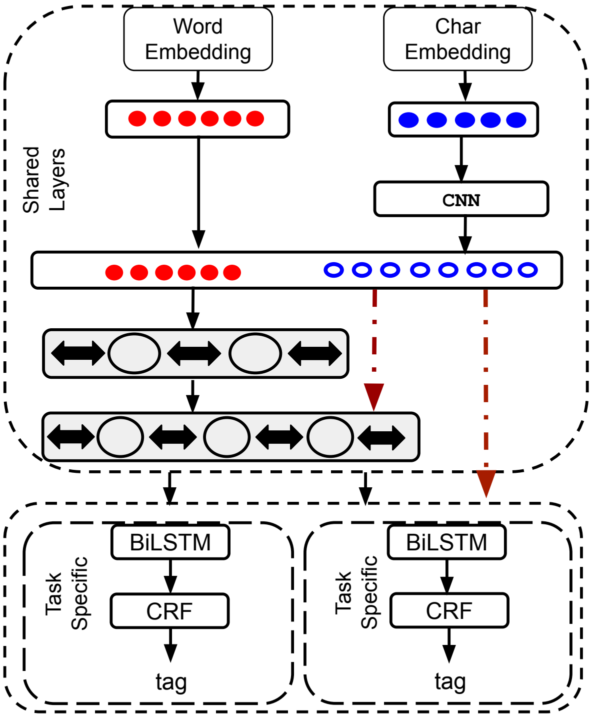

Figure 1: Proposed MTM-CW Model where dashed arrows show skip connections

Finally, we use CRFs for output labeling. CRFs have the ability to tag the current token by considering neighboring tags at sentence level (Huang et al., 2015). Yang et al. (2018) performed experiments comparing CRF and Softmax and found out that CRF produces better results compared to Softmax.

An alternative training approach was adopted for the training phase. Let suppose we have D_1,D_2,..., D_t training sets, related to the T_1, T_2, ..., T_t tasks respectively. During training, a training set D_i is selected randomly and both shared layers and layers specific to the corresponding task T_i are activated. Every task has its own optimizer so during training only the optimizer specific to the task T_i is activated and the loss function related to that optimizer is optimized. It means that the parameters of the shared layers and of the task-specific layers are changed during the training of the specific task. Optimizing parameters of the shared layers for all the tasks helps the model to find the common features among different tasks.

4 Experiments

We performed experiments on the 15 datasets which were also used by Crichton et al. (2017), Wang et al. (2019), and Mehmood et al. (2019). The bio-entities in these datasets are Chemical, Species, Cell, Gene/Protein, Cell Component, and Disease[1]. Descriptions of the datasets can be found in Crichton et al. (2017). Moreover, to represent words, we use domain-specific pre-trained word embeddings since generic word embeddings can cause a high rate of out-of-vocabulary words. In particular, we use WikiPubMed-PMC word embedding which is trained on a large set of the PubMedCentral(PMC) articles and PubMed abstracts as well as on English Wikipedia articles (Giorgi and Bader, 2018). On the other hand, character embedding is initialized randomly while orthographic (case) embedding is represented by the identity matrix where each diagonal 1 represents the presence of a word's orthographic feature. Moreover, we analyse the effect of different input representations (word level, character level, and case level) of a word on the performance of our proposed architecture. Furthermore, this paper reports the average F1-score where each experiment is run for 10 times. We use the Nadam

[1] The datasets can be found at the following link https://github.com/cambridgeltl/MTL-Bioinformatics-2016

optimizer in our model and use CNN with a filter size of 30 while each LSTM in the model consists of 275 units and the experiment is run for 50 epochs and early stop is set to 10 epochs.

5 Results and Discussion

In Table 1 we compare the results produced by our model with state-of-the-art models (Wang et al., 2019; Mehmood et al., 2019). We can see a substantial improvement in the F1-score by MTM-CW compared to these models. However, to observe whether connecting embedding layers to the middle layers has truly contributed to the performance of the model, we made a variation in the model and dropped the skip connections coming from embedding layers (refer to Figure 1). Dropping these skip connections makes our model similar to the model by Mehmood et al. (2019) where we have introduced another layer of shared BiLSTM. The effect of such variation is reported in Table 2 where it can be noted that few datasets show moderate performance increase while for most of them performance degrades. This supports our intuition that passing embedding layer information to the lower layers has positive impact on the model. Moreover, it is interesting that, even after dropping those skip connections, our model is still able to perform better compared to state-of-the-art models. This suggests that, with increasing size of training examples, more layers of LSTM should be considered (Li et al., 2018). For this reason, the proposed model by Mehmood et al. (2019) performed better compared to model proposed by Wang et al. (2019) which used single layer of LSTM.

Datasets	Wang et al.	Mehmood et al.	MTM-CW
AnatEM	86.04	86.99	**87.50**
BC2GM	78.86	80.82	**81.57**
BC4CHEMD	88.83	87.39	**89.24**
BC5CDR	88.14	87.85	**88.54**
BioNLP09	88.08	**88.74**	88.52
BioNLP11EPI	83.18	84.75	**85.36**
BioNLP11ID	83.26	**87.65**	87.19
BioNLP13CG	82.48	84.25	**84.94**
BioNLP13GE	79.87	79.82	**80.91**
BioNLP13PC	88.46	88.84	**89.16**
CRAFT	82.89	83.15	**85.23**
Ex-PTM	80.19	80.95	**81.72**
JNLPBA	72.21	**74.05**	72.10
linnaeus	**88.88**	87.79	88.12
NCBI-disease	85.54	**85.66**	85.07

Table 1: Multi-task Models Comparison where CW represents character and word respectively

We then extended our experiments by introducing orthographic-level representation of a word in our model. Dugas and Nichols (2016) Segura-Bedmar et al. (2015) Huang et al. (2015) have shown that orthographic-level information can improve model's performance. In addition, statistical models (e.g. CRF at the output layer) are also highly dependent on hand-crafted features (Limsopatham and Collier, 2016). In this work, the orthographic-level feature includes information on the structure of the word, i.e. either the word is starting with a capital letter followed by small letters or all the letters in the word are capital or contain digits, etc. Table 2 reports the comparison between MTM-CW and its variant with orthographic-level features (we name it case, MTM-CW-Case). We observe that, for some datasets, orthographic-level features moderately improved the results. Thus, we can conclude that orthographic-level features might help the model to implicitly learn hidden features at an orthographic level which could be helpful for some entities. However, for simplicity we are limiting our work to explicitly representing the word-level features; thus we stick to the character-level representation and the word itself. We also replaced CRF with Softmax at the output layer to see the impact of both methods on predicting the output label of the entities. Table 2 also depicts the comparison of our proposed model with softmax (MTM-CW-Softmax) and CRF (proposed MTM-CW) at the output layer and model with CRF produce better results compared to the model with Softmax.

To statistically evaluate the results obtained by different variants of our model we perform the Friedman test (Zimmerman and Zumbo, 1993). We also analyse the pairwise comparison of different models to see which model is statistically better than the other. The graphical representation of the pairwise comparison is shown in Figure 2 as it can be seen in variant of the model proposed with softmax (MTM-CW-Softmax represented as just Softmax) which is statistically worse compared to the others and to other variants of the model. Figure 3 shows the post-hoc Conover Friedman test where it can be seen that the difference between results produced by all the models is significant with different p values.

Datasets	MTM-CW	MTM-CW (w/out skip connections)	MTM-CW Case	MTM-CW Softmax
AnatEM	**87.50**	86.94	87.37	86.36
BC2GM	81.57	81.29	**81.66**	80.04
BC4CHEMD	**89.24**	87.44	89.13	86.88
BC5CDR	88.54	88.11	**88.64**	87.39
BioNLP09	88.52	89.31	88.61	88.18
BioNLP11EPI	**85.36**	85.01	85.04	84.16
BioNLP11ID	87.19	88.16	87.76	87.28
BioNLP13CG	**84.94**	84.61	84.86	84.00
BioNLP13GE	80.91	82.28	80.16	80.49
BioNLP13PC	89.16	89.04	**89.26**	88.37
CRAFT	**85.23**	83.44	85.04	82.86
Ex-PTM	81.72	82.40	81.50	80.64
JNLPBA	72.10	72.02	**72.21**	70.31
linnaeus	88.12	88.69	**88.74**	88.33
NCBI-disease	85.07	85.12	**85.56**	84.36

Table 2: Comparison between the Results of Different Variants of the Model Proposed

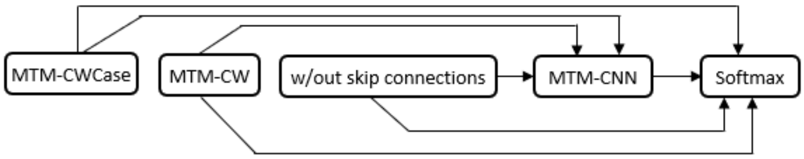

Figure 2: Pairwise Models Comparison w.r.t to Friedman Test

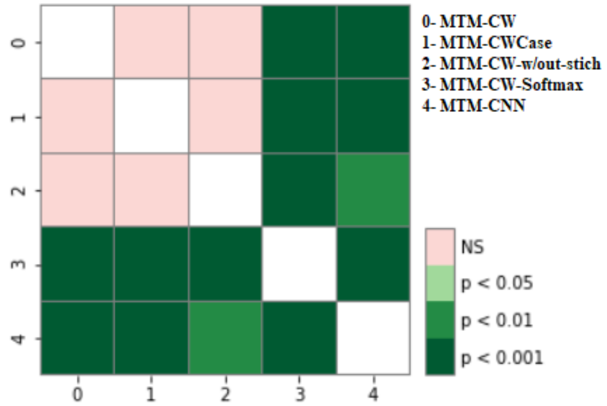

Figure 3: Post-hoc Conover Friedman Test (NS represents not significant)

6 Conclusion and Future Work

In this paper we showed that the BioNER performance can be drastically improved by using a multi-task approach. We showed that using stacked LSTMs in such models are effective to learn hidden structure of the data. Moreover, to overcome the vanishing gradient problem in using stacked LSTMs is addressed by passing embedding information layers to layers. We showed that our model outperforms in F1-score compared to the state-of-the-art models.

For future work, we will extend the multi-task approach for relation extraction task. In such approach, BioNER can be used as an auxiliary task while keeping relation extraction task as the main task in the multi-task approach.

References

Trapit Bansal, David Belanger, and Andrew McCallum. 2016. Ask the gru: Multi-task learning for deep text recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 107–114. ACM.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa.

2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):368.

Cícero Nogueira dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop, NEWS@ACL 2015, Beijing, China, July 31, 2015*, pages 25–33.

Cıcero dos Santos, Victor Guimaraes, RJ Niterói, and Rio de Janeiro. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 25.

Fabrice Dugas and Eric Nichols. 2016. DeepNNNER: Applying BLSTM-CNNs and extended lexicons to named entity recognition in tweets. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 178–187.

John M Giorgi and Gary D Bader. 2018. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, 34(23):4087–4094.

Mourad Gridach. 2017. Character-level neural network for biomedical named entity recognition. *Journal of biomedical informatics*, 70:85–91.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Jinyu Li, Changliang Liu, and Yifan Gong. 2018. Layer trajectory LSTM. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, pages 1768–1772.

Nut Limsopatham and Nigel Collier. 2016. Learning orthographic features in bi-directional LSTM for biomedical named entity recognition. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining, BioTxtM@COLING 2016, Osaka, Japan, December 12, 2016*, pages 10–19.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multitask sequence to sequence learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Tahir Mehmood, Alfonso Gerevini, Alberto Lavelli, and Ivan Serina. 2019. Leveraging multi-task learning for biomedical named entity recognition. In *International Conference of the Italian Association for Artificial Intelligence*. Springer.

Isabel Segura-Bedmar, Víctor Suárez-Paniagua, and Paloma Martínez. 2015. Exploring word embedding for drug name recognition. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, Louhi@EMNLP 2015, Lisbon, Portugal, September 17, 2015*, pages 64–72.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2019. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752.

Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *JAMIA*, 25(10):1419–1428.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3879–3889.

Donald W Zimmerman and Bruno D Zumbo. 1993. Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks. *The Journal of Experimental Education*, 62(1):75–86.