Simple Hierarchical Multi-Task Neural End-To-End Entity Linking for Biomedical Text

Maciej Wiatrak, Juha Iso-Sipilä
BenevolentAI
4-8 Maple St, London
W1T 5HD
{maciej.wiatrak, juha.iso-sipila}@benevolent.ai

Abstract

Recognising and linking entities is a crucial first step to many tasks in biomedical text analysis, such as relation extraction and target identification. Traditionally, biomedical entity linking methods rely heavily on heuristic rules and predefined, often domain-specific features. The features try to capture the properties of entities and complex multi-step architectures to detect, and subsequently link entity mentions. We propose a significant simplification to the biomedical entity linking setup that does not rely on any heuristic methods. The system performs all the steps of the entity linking task jointly in either single or two stages. We explore the use of hierarchical multi-task learning, using mention recognition and entity typing tasks as auxiliary tasks. We show that hierarchical multi-task models consistently outperform single-task models when trained tasks are homogeneous. We evaluate the performance of our models on the biomedical entity linking benchmarks using MedMentions and BC5CDR datasets. We achieve state-of-the-art results on the challenging MedMentions dataset, and comparable results on BC5CDR.

1 Introduction & Related Work

The task of identifying and linking mentions of entities to the corresponding knowledge base is a key component of biomedical natural language processing, strongly influencing the overall performance of such systems. The existing biomedical entity linking systems can usually be broken down into two stages: (1) Mention Recognition (MR) where the goal is to recognise the spans of entity mentions in text and (2) Entity Linking (EL, also referred as Entity Normalisation or Standardisation), which given a potential mention, tries to link it to an appropriate type and entity. Often, the entity linking task includes the Entity Typing (ET) and Entity Disambiguation (ED) as separate steps, with the former task aiming to identify the type of the mention, such as *gene*, *protein* or *disease* before passing it to the entity disambiguation stage, which effectively grounds the mention to an appropriate entity.

Widely studied in the general domain, entity linking is particularly challenging for the biomedical text. This is mostly due to the size of the ontology, (here referred to as the knowledge base), high syntactic and semantic overlap between types and entities, the complexity of terms, as well as the lack of availability of annotated text.

Due to these challenges, the majority of the existing methods rely on hand-crafted complex rules and architectures including semi-Markov methods (Leaman and Lu, 2016), approximate dictionary matching (Wang et al., 2019) or use a set of external domain-specific tools with manually curated ontologies (Kim et al., 2019). These methods often include multiple steps, each of these steps carrying over the errors to the subsequent stages. Nevertheless, these tasks are usually interdependent and have been proven to often benefit from a joint objective (Durrett and Klein, 2014). Recently, both in the general and biomedical domain, there has been a steady shift to neural methods to solve EL (Kolitsas et al., 2018; Habibi et al., 2017), leveraging a range of methods including the use of entity embeddings (Yamada et al., 2016), multi-task learning (Mulyar and McInnes, 2020; Khan et al., 2020), and others (Radhakrishnan et al., 2018). There have also been a plethora of mixed methods combining heuristic approaches such as approximate dictionary matching with language models (Loureiro and Jorge, 2020).

This work focuses on multi-task approaches to end-to-end entity linking, which has already been studied in the biomedical domain. These include ones leveraging pre-trained language models (Peng et al., 2020; Crichton et al., 2017; Khan et al., 2020), model dependency (Crichton et al., 2017) and building out a cross-sharing model structure

Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pages 12–17
November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

(Wang et al., 2019). An interesting approach has been proposed by Zhao et al. (2019), where authors established a multi-task deep learning model that trained NER and EL models in parallel, with each task leveraging feedback from the other. A model with a similar setup and architecture to the one here, casting the EL problem as a simple per token classification problem has been outlined by Broscheit (2019). Nevertheless, its application domain, architecture, and training regime strongly differ from the one proposed here.

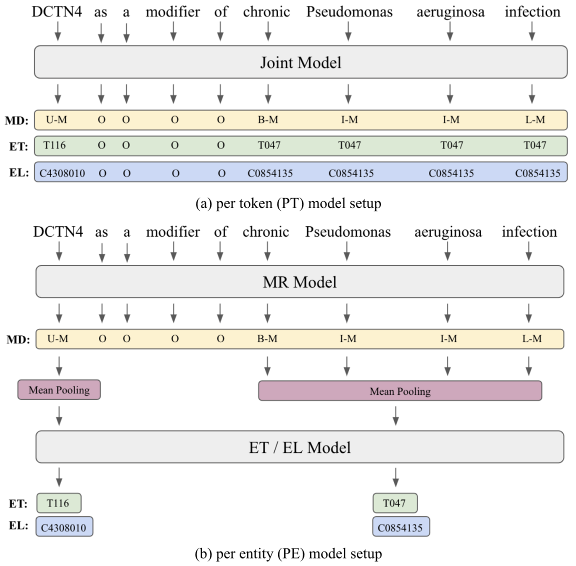
In this study, we investigate the use of a significantly simpler model, drawing on a set of recent developments in NLP, such as pre-trained language models, hierarchical and multi-task learning to outline a simple, yet effective approach for biomedical end-to-end entity linking. We evaluate our models on three tasks, mention recognition, entity typing, and entity linking, investigating different task setups and architectures on the MedMentions and BioCreative V CDR corpora.

Our contributions are as follows: (1) we propose and evaluate two simple setups using fully neural end-to-end entity linking models for biomedical literature. We treat the problem as a per token classification or per entity classification problem over the entire entity vocabulary. All the steps included in the entity linking task are performed in a single or two steps. (2) We examine the use of mention recognition and entity typing as auxiliary tasks in both multi-task and hierarchical multi-task learning scenario, proving that hierarchical multi-task models outperform single-task models when tasks are homogeneous. (3) We outline the optimal training regime including adapting the loss for the extreme classification problem.

2 Methods

2.1 Tasks

Our main task, which we refer to as **Entity Linking** (EL) aims at classifying each token or a mention to an appropriate entity concept unique identifier (CUI). In order for the mention to be correctly identified, all tokens for the mention need to have the correct golden annotation. If the model has wrongly predicted the token right after or before the entity's golden annotated span, the entity prediction is wrong at the mention-level (Mohan and Li, 2019). For the per entity setup, where the entity representation is derived through mean pooling of all tokens spanning a predicted entity, both the final EL and the MR predictions need to be correct. Figure one provides more information on both setups.



(a) per token (PT) model setup

(b) per entity (PE) model setup

Figure 1: The entity linking setup as a (a) per token (PT) classification and (b) per entity (PE) classification problem with a sentence and corresponding labels for EL, ET and MR, which uses a BILOU scheme for annotations. Here, ”O” denotes a *Nil* and ”M” a *Mention* prediction.

We also make use of two other tasks: **Entity Typing** (ET) and **Mention Recognition** (MR), with the former predicting entity Type Unique Identifier (TUI) for each token and the latter predicting whether a token is a part of the mention. We always use the BILOU scheme for mention recognition token annotation, and due to the low number of types in the BC5CDR dataset, also for the ET task on this corpora. We evaluate the entity prediction at mention-level similarly as in the EL and ET. In per token setup, all three tasks are essentially sequence labelling problems, while in per entity setup, only the MR is a sequence labelling problem and both ET and EL are classification problems leveraging the predictions produced by the MR model.

The reason behind employing ET and MR tasks is for investigating the multi-task learning methods, where we treat ET and MR as auxiliary tasks aimed at regularising and providing additional information to the main EL task leveraging its inherently hierarchical structure. Correspondingly, we also look at the performance impact of the two other tasks on EL task.

2.2 Models

We outline three models: single-task model, multi-task model, and hierarchical multi-task model. The model architecture for the latter two models is depicted on Figure 2. All models take a sentence with

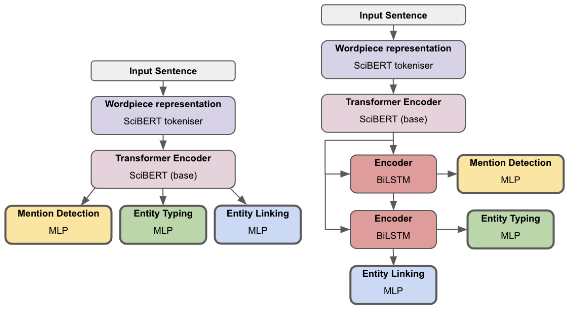


Figure 2: The architectures of the multi-task model (left) and hierarchical multi-task model (right) with hierarchical structure of the tasks and task-specific encoders.

the surrounding context as their input and output a prediction for a token (PT setup) or an average of token embeddings spanning an entity (PE setup). For tokenisation, embedding layer and encoder we use SciBERT (base).

The single-task model only adds a feedforward neural network at the top of the encoder transformer, which acts as a decoder. In the multi-task scenario, three feedforward layers are added on the top of the transformer, each corresponding to a specific task, namely MR, ET, and EL. All of these tasks share the encoder and during a forward pass, the encoder output is fed into each task-specific layers separately, after which the cumulative loss is summed and backpropagated through the model. The intuition behind sharing the encoder is that training on multiple interdependent tasks will act as a regularisation method, thus improving the overall performance and speed of convergence.

The last model is a hierarchical multi-task model that leverages the natural hierarchy between the 3 tasks by introducing an inductive bias by supervising lower level tasks at the bottom layers of the model (MR, ET) and higher level task (EL) at the top layer. Similarly, as in (Sanh et al., 2019), we add task-specific encoders and shortcut connections to process the information from lower to higher level tasks. The higher level tasks take the concatenation of the general transformer encoder output and lower-level task encoder specific output as their input. Here, we use multi-layer BiLSTMs as task-specific encoders.

We experiment with all three models in the per token scenario, as all tasks in this setup are sequence labelling problems. For the per entity framework, we look at a single-task and hierarchical multi-task model, where only the MR step is a sequence labelling task and ET and EL are both classification tasks.

3 Experiments

3.1 Training details

We treat both PE and PT setups as multi-class classification problems over the entire entity vocabulary. In both cases, we use categorical cross-entropy to compute the loss. To address the class imbalance problem in the PT framework, we apply a lower weight to the *Nil* token's output class, keeping other class weights equal. To improve convergence speed and memory efficiency we compute the loss only through the entity classes present in the batch. Therefore, for token t_i in a sequence T, (or correspondingly the mean pooled entity representation from a set of tokens) with a label y_i and its assigned class weight w_k in a minibatch B and entity labels derived from this batch $\hat{E} = E(B)$, the loss is computed by

$$L = -\frac{1}{|B| * |T|} \sum_{k}^{|\hat{E}|} \sum_{j}^{|B|} \sum_{i}^{|T|} w_k y_{ij}^k \log(h_\theta(t_{ij}, k)).$$

Here, y_{ij}^k represents the target label for token i in a sequence j for class k, and $h_\theta(t_{ij}, k)$ represents the model prediction for token t_{ij} and class k, where the parameters θ are defined by the encoder and decoder layers in the model.

We found using the context, namely the sentence after and before the sentence of interest beneficial for the encoder. After encoder, the context sentences are discarded from further steps. For the encoder, we use the SciBERT (base) transformer, and we fine tune the model parameters during training. For the hierarchical multi-task model, we follow the training regime outlined in (Sanh et al., 2019) and found tuning the encoder only on the EL task marginally outperforming sharing it across all three tasks. We treated the *Nil* output class weight as an additional hyperparameter that we set to 0.125 for MedMentions (full) and BC5CDR datasets, and 0.01 for MedMentions st21pv. All trainings were performed using Adam (Kingma and Ba, 2015) with $1e-4$ weight decay, $2-e5$ learning rate, batch size of 32 and max sequence length of 128.

Dataset	#Docs	#Mentions	#Unq TUI	#Unq CUI
MedMentions (full)	4,392	352,496	126	34,724
MedMentions (st21pv)	4,392	203,282	21	25,419
Bio CDR	1,500	28,559	2	5,818

Table 1: Details of biomedical entity linking datasets used in our experiments.

	MedMentions(full)				**MedMentions (st21pv)**				**BC5CDR**			
	Mention Recognition											
Model	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
SciSpacy	N/A	69.61	68.56	69.08	N/A	41.23	70.57	52.05	N/A	81.47	73.47	77.81
BiLSTM-CRF	82.47	64.09	65.03	64.56	84.86	60.53	61.7	61.11	94.00	72.09	78.65	75.23
Single-task	85.56	73.4	69.38	**71.33**	87.72	73.55	**66.92**	**70.05**	**97.04**	89.64	**88.25**	**88.94**
Multi-task	**85.62**	72.62	**69.72**	71.14	**87.84**	73.34	66.53	69.76	96.93	**90.56**	87.24	88.87
Hier. Multi-task	85.40	**73.13**	68.93	70.97	85.59	**74.19**	59.25	65.88	96.68	89.31	84.91	87.05

	Entity Typing											
Model	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
SciSpacy	N/A	39.67	39.08	39.37	N/A	10.14	31.68	15.26	N/A	N/A	N/A	N/A
BiLSTM-CRF	72.26	45.14	44.98	45.06	82.46	47.15	52.29	49.59	94.03	72.08	78.70	75.24
PT-Single-task	78.27	55.79	51.65	53.64	86.67	63.10	58.26	60.59	**96.96**	89.52	87.48	88.45
PE-Single-task	N/A	**57.5**	52.62	**54.95**	N/A	**65.05**	**60.43**	**62.65**	N/A	**90.53**	**87.65**	**89.07**
PT-Multi-task	**78.3**	55.39	**52.66**	53.99	**86.72**	63.77	58.86	61.21	96.90	90.33	87.04	88.65
PT-Hier. Multi-task	76.7	61.94	49.41	50.61	80.87	46.22	40.76	43.32	96.57	88.40	84.24	86.27
PE-Hier. Multi-task	N/A	50.91	46.49	48.65	N/A	59.44	55.27	57.30	N/A	88.15	85.34	86.72

	Entity Linking											
Model	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
SciSpacy	N/A	34.14	33.63	33.88	N/A	25.17	53.52	34.24	N/A	58.43	52.70	55.42
BiLSTM-CRF*	62.73	39.89	30.25	32.22	71.35	33.65	25.46	28.99	89.52	52.72	47.59	50.02
PT-Single-task	67.98	46.41	39.46	42.65	75.57	44.09	35.58	39.36	91.62	64.14	57.56	60.67
PE-Single-task	N/A	46.3	**42.37**	**44.25**	N/A	43.03	39.97	41.45	N/A	**64.98**	**62.91**	**63.93**
PT-Multi-task	**68.23**	45.88	40.13	42.81	**76.43**	44.03	37.85	40.71	91.45	63.51	54.35	58.62
PT-Hier. Multi-task	68.13	**46.89**	39.93	43.13	76.14	**44.32**	37.69	40.74	**91.65**	64.35	59.27	63.15
PE-Hier. Multi-task	N/A	46.21	42.29	44.16	N/A	43.12	**40.06**	**41.53**	N/A	64.54	62.49	63.5

Table 2: Results: performance of various models on MR, EL and ET tasks on the test sets. Here *Acc-pt* denotes per token accuracy. * for EL task on MedMentions full and st21pv we used a MLP layer on top of BiLSTM instead of CRF due to the lower performance of CRF on large number of output classes.

The models were trained on a single NVIDIA V100 GPU until convergence.

3.2 Datasets and Evaluation metrics

We evaluate our models on three datasets; two versions of the recently released MedMentions dataset; (1) full set and (2) and st21pv subset of it (Mohan and Li, 2019) and BioCreative V CDR task corpus (Li et al., 2016). Each mention in the dataset is labelled with a concept unique identifier (CUI) and type unique identifier (TUI). Both MedMentions datasets target UMLS ontology but vary in terms of number of types and mentions, while the BioCreative V corpora is normalised with MeSH identifiers. The datasets details are summarised in Table 1.

We measure the performance of each task using mention-level metrics described in (Mohan and Li, 2019), providing precision, recall, and F1 scores. Additionally, we record the per token accuracy for the per token setup. As benchmarks, we use SciSpacy (Neumann et al., 2019) package, which has been shown to outperform other biomedical text processing tools such as QuickUMLS or MetaMap on full MedMentions and BC5CDR (Vashishth et al., 2020). Due to little results reported on the end-to-end entity linking task on MedMentions, we also use BiLSTM-CRF in per token setup as a benchmark.

3.3 Results and discussion

In Tables 2 and 3 we outline the results on MR, ET, and EL tasks. While the reported results are all optimal for single-task models, it should be noted that all multi-task models optimise for the EL task with MR and ET serving as auxiliary tasks, hence the EL is the focus of the discussion. All of the models outlined here significantly outperform SciSpacy and BiLSTM-CRF, particularly in ET and EL. The per entity setup proves to perform better on EL than the simpler per token framework by 0.87 F1 points on average, yielding particularly better recall results (2.03 points). Error analysis has shown that this is often due to the lexical overlap of some *Nil* tokens with entity tokens, resulting in a model often assigning an entity label for tokens with gold *Nil* token label. Furthermore, in the per token setup, the multi-task models consistently outperform the single-task models on EL, with the hierarchical multi-task model achieving the best results (on average 1.45 F1 points better than single-task models). In contrast, this has not been the case for the per entity framework, where the single-task models have on average performed marginally better on EL. We hypothesise that this is due to the homogeneity of the tasks in the per token setup, with all the tasks being sequence labelling problems, which is not the case for the per entity case. Interestingly, the achieved results are higher for the full MedMentions dataset than for the st21pv subset. This highlights the problem of achieving high macro performance mentioned in (Loureiro and Jorge, 2020) for biomedical entity linking.

4 Conclusion & Future Work

In this work, we have proposed a simple neural approach to end-to-end entity linking for biomedical text which makes no use of heuristic features. We have proven that the problem can benefit from the hierarchical multi-task learning when tasks are homogeneous. We report state-of-the-art results on EL on the full MedMentions dataset and comparable results on the MR and ET tasks on BC5CDR (Zhao et al., 2019). The work could easily be extended by, for example, using the output of the PT setup as features or by further developing the hierarchical multi-task framework of end-to-end entity linking problem. Moreover, the additional parameters such as output class weights or loss scaling which has not been used here could be easily adapted to a particular problem.

Acknowledgments

We thank Theodosia Togia, Felix Kruger, Mikko Vilenius and Jonas Vetterle for helpful feedbacks and the anonymous reviewers for constructive comments on the paper.

References

Samuel Broscheit. 2019. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China. Association for Computational Linguistics.

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):1–14.

Greg Durrett and Dan Klein. 2014. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

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

Muhammad Raza Khan, Morteza Ziyadi, and Mohamed Abdelhady. 2020. MT-BioNER: Multi-task Learning for Biomedical Named Entity Recognition using Deep Bidirectional Transformers. *ArXiv*, abs/2001.08904.

Donghyeon Kim, Jinhyuk Lee, Chan H O So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. 2019. A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. *IEEE Access*, 7:73729–73740.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-End Neural Entity Linking. *ArXiv*, abs/1808.07699.

Robert Leaman and Zhiyong Lu. 2016. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 32(18):2839–2846.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016. Baw068.

Daniel Loureiro and Alípio Jorge. 2020. MedLinker: Medical Entity Linking with Neural Representations and Dictionary Matching. *Advances in Information Retrieval*, 12036:230 – 237.

Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In *In Proceedings of the 2019 Conference on Automated Knowledge Base Construction (AKBC 2019). Amherst, Massachusetts, USA. May 2019*.

Andriy Mulyar and Bridget T. McInnes. 2020. MT-Clinical BERT: Scaling Clinical Information Extraction with Multitask Learning.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on bert for biomedical text mining. In *In BioNLP 2020 Workshop on Biomedical Natural Language Processing*.

Priya Radhakrishnan, Partha Talukdar, and Vasudeva Varma. 2018. ELDEN: Improved entity linking using densified knowledge graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1844–1853, New Orleans, Louisiana. Association for Computational Linguistics.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *AAAI*.

Shikhar Vashishth, Rishabh Joshi, Ritam Dutt, Denis Newman-Griffis, and Carolyn Rose. 2020. Med-Type: Improving Medical Entity Linking with Semantic Type Prediction.

Xi Wang, Jiagao Lyu, Li Dong, and Ke Xu. 2019. Multitask learning for biomedical named entity recognition with cross-sharing structure. *BMC Bioinformatics*, 20(1):1–13.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259, Berlin, Germany. Association for Computational Linguistics.

Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2019. A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization. In *AAAI*.