Learning to Learn Recognising Biomedical Entities from Multiple Domains with Task Hardness

Anonymous ACL submission

Abstract

Few-shot learning has been a big challenge for many classification tasks, where the final classifier is trained only with a few examples. This problem amplifies when we apply the few-shot setup to recognising named entity from different domains, i.e., few-shot domain adaption for NER. In this paper, we present a simple yet effective MAML-based NER model that can effectively leverage the task hardness information to improve the adaptability of the learnt model in the few-shot setting. Experimental results on biomedical datasets show that our model can achieve significant performance improvement over the recently published MetaNER model.

1 Introduction

To assist clinicians in their decision making, information needs to be extracted correctly and appropriately from patients' data such as Electronic Medical Records (EMRs). This paper focuses on one of the key tasks in this extraction process, named entity recognition (NER), specifically, biomedical named entity recognition (BioNER). Correct recognition of biomedical named entities (NEs) can lead to a reliable detection/extraction system, providing a comprehensive picture of the patient's health to assist medical practitioners.

An optimal BioNER method should be robust enough to perform well on unseen tasks across different domains in a lower-resource setting, as annotating medical texts is extremely expensive, requiring medical expertise. Existing BioNER methods could struggle with this setting, as they are based on powerful structures such as BiLSTM-CRF (Xu et al., 2019, 2018) or deep transformer (Lee et al., 2019; Alsentzer et al., 2019). Those structures often have numerous trainable parameters and require a large training dataset, and consequently, this impedes the model's generalizability and adaptability to new tasks with very limited training samples, preventing it from achieving good result outside of the corpus and domain it was trained on.

A potential solution is to inject the prior "experience" to the adaptation process. Few works have explored this area such as Li et al. (2020a) and Li et al. (2020b), where the former followed the optimization meta-learning strategy by (Finn et al., 2017) and the latter introduced a feature critic module similar to the work of Li et al. (2019).

Furthermore, it is sub-optimal to assume tasks are equally important (*i.e.*, randomly sample tasks) in learning the meta model in meta training (Yao et al., 2021), particularly when the number of training tasks in small. Indeed, the importance/hardness of BioNER tasks can vary significantly, as shown in Table 1. While some tasks might contain NEs, many tasks do not (*e.g.*, the last row in Table 1), and hence, contribute little to learning the NER model. Meanwhile, the task difficulty ties not only to the number of entities in the task but also to the length of those entities. Therefore, we argue that bioNER tasks sampled in meta training should be treated differently accroding to their hardness in order to improve learning efficiency and the model performance. To bridge the gap, we present a simple but effective way of incorporating task hardness into a meta-learning framework for BioNER tasks. We show that our task hardness driven meta learning approach for BioNER outperforms recently published meta-learning based NER methods.

2 Related Works

For meta learning in NER tasks, both Li et al. (2020a) and Li et al. (2020b) seek a robust representation for the sequence labeling function BiLSTM-CRF by applying the meta-learning framework (Finn et al., 2017), and the latter further includes an auxiliary network to promote adversarial learning during the training process. However, different tasks can have different level of hardness which both works have not yet addressed. Exist-

Example	Score
Stimulation of human neutrophils with **[chemoattractants] [FMLP]** or **[platelet activating factor (PAF)]** results in different but overlapping functional responses. Of even more interest, **[IkappaBalpha]** overexpression inhibited the production of **[matrix metalloproteinases 1 and 3]** while not affecting their tissue inhibitor. ...more durable inhibition of HIV - 1 replication than was seen with the **[NF-kappa B]** inhibitors alone or the **[anti-Tat sFv intrabodies]** alone. Spontaneous occurrence of early region 1A reiteration mutants of type 5 adenovirus in persistently infected human T-lymphocytes.	0.46
Here we report the fabrication of single-molecule transistors based on individual C60 molecules connected to gold electrodes. The contractile effects of **[oxytocin]**, prostaglandin F2 alpha and their combined use on human pregnant myometrium were studied in vitro. Transcriptional activation of the **[proopiomelanocortin gene]** by **[cyclic AMP-responsive element binding protein]**. The difference between the effects of the two dose levels of Z.	0.18
She was monitored for one more day and then discharged with instructions to discontinue her diet pills The Raf/Ras/ERK/MAPK pathway is known to be involved in NGF-induced outgrowth Our analysis reveals that the oviduct is lined, along its entire length, by a monolayered epithelium comprised of squamous-type cells. In one case study, Bramson et al.	0.01

Table 1: Examples of task hardness scores (computed from our method) for three tasks during the meta-training procedure, The score is based on a scale from 0 to 1, the higher the score, the more challenging the task is. The NEs are put in brackets with red color for each sentence.

ing meta learning works deal with task hardness via 1) actively ranking the tasks in term of difficulty level (Yao et al., 2021; Zhou et al., 2020; Liu et al., 2020; Achille et al., 2019); 2) designing an adaptive task scheduler (Yao et al., 2021); or 3) relying on generative approaches to quantify the uncertainties of tasks Kaddour et al. (2020); Nguyen et al. (2021). To our knowledge, we are the first to incorporate the concept of task hardness into meta-learning NER for Biomedical tasks.

3 BioNER with Task Hardness

Problem Setup Given a set of biomedical corpora from multiple source domains (*e.g.*, Drug, Gene, Species, etc), and let $p(\mathcal{T})$ be the underline distribution of tasks, i.e., recognising biomedical named entities from different domains. We aim to meta-learn a sequence labelling function $h : \mathcal{X} \rightarrow \mathcal{Y}$[1] from a set of tasks sampled from $p(\mathcal{T})$ so that it can be generalised to a new task $\mathcal{T}^{'}$ sampled from an unseen target domain (e.g., Disease). The labelling function h contains 1) a sentence encoder parameterized with $\boldsymbol{\theta}$ (*e.g.*, BiLSTM-CNN) that captures the contextual information about words, and 2) a tag decoder parameterized with $\boldsymbol{\phi}$ (*e.g.*, CRF) that assigns the entity tags to these words[2]. Thus, the learning objective is to search for the optimal $\boldsymbol{\Theta}^* \equiv \{\boldsymbol{\theta}, \boldsymbol{\phi}\}$ from the source domains with a bi-level optimization framework commonly used in meta-learning. Finally, this optimal $\boldsymbol{\Theta}^*$ should minimise the risk of transferring h from the source domains to a new task $\mathcal{T}^{'}$ from the target domain.

Task Generation To optimize for $\boldsymbol{\Theta}^*$ with stochastic optimization, one first need to sample from $p(\mathcal{T})$, *i.e.*, task generation. Each BioNER task $\mathcal{T}_i$ in our setting is divided into a support set $\mathcal{T}_i^S$ and a query set $\mathcal{T}_i^Q$ with $\mathcal{T}_i^S \cap \mathcal{T}_i^Q = \emptyset$. Both $\mathcal{T}_i^S$ and $\mathcal{T}_i^Q$ contain only K sentences respectively sampled from the same domain, and K can be as small as 5 or 10. Different from Li et al. (2020b), we are interested in the scenarios where there are a small handful of annotated sentences from the target domain. Mimicking the same few-shot setting in meta training has been shown to reduce the PAC-Bayesian meta-learning error bound (Ding et al., 2021). Meanwhile, randomly sampling $\mathcal{T}_i$ from all the source domains will then allow us to learn a good initialization of $\boldsymbol{\Theta}^*$ that can be quickly adapted to a new unseen task, similar to (Li et al., 2020a). We further consider the imbalance issue caused by the NER few-shot setting. As shown in table 2, the majority of the sentences in the biomedical corpora does not contain any entities. Thus, it is highly likely that the K randomly-sampled sentences contain no entity, which can result in a biased sequence labeller that always predict "O" in the adaption phrase. To avoid this issue, we choose sentences in $\mathcal{T}_i^S$ to be those contains at least one biomedical entity, which is shown to be effective.

Bilevel Optimization Following Li et al. (2020b, 2019), we also include a domain classifier as a critic network to regularize the meta-learning process and promote domain generalization. This critic network, parameterized by $\boldsymbol{\omega}$, consists of a fully connected layer used to predict which domain a sentence in a task $\mathcal{T}_i$ belongs to. The classification function f will henceforth be used to represent the composition of the sentence encoder network and the critic network. The overall meta-learning objective is

$$\mathcal{L}_i = \mathcal{L}^{\text{lab}}\left(h\left(\boldsymbol{\theta}, \boldsymbol{\phi}\right), \mathcal{T}_i\right) + \lambda \mathcal{L}^{\text{cls}}\left(f\left(\boldsymbol{\theta}, \boldsymbol{\omega}\right), \mathcal{T}_i\right), \quad (1)$$

where λ control the trade-off between the two loss functions. In meta-training time, we first generate a batch of task from $p(\mathcal{T})$, and for each $\mathcal{T}_i$, we train the model on $\mathcal{T}_i^S$ then validate the performance on $\mathcal{T}_i^Q$ using our learning objective. Consequently,

[1] $\mathcal{X}$ consists of a set of sentences, while $\mathcal{Y}$ indicates the sequence label sets corresponding to these sentences.

[2] We consider the BIO tagging schema containing three labels: B-Begin, I-Inside, and O-Outside.

Corpora	Entity Type	No. Unique Tokens	% sentences with NEs
NCBI (Doğan et al., 2014)	Disease	$12,128$	55
BC5CDR (Li et al., 2016)	Disease	$23,068$	59
BC5CDR (Li et al., 2016)	Drug	$23,068$	65
BC4CHEMD (Krallinger et al., 2015)	Drug	$114,837$	48
JNLPBA (Collier and Kim, 2004)	Gene	$25,046$	81
BC2GM (Smith et al., 2008)	Gene	$50,864$	51
LINNAEUS (Gerner et al., 2010)	Species	$34,396$	13
S800 (Pafilis et al., 2013)	Species	$205,26$	30

Table 2: Biomedical corpora used in our experiments (Habibi et al., 2017; Lee et al., 2019; Zhu et al., 2018).

we gather the gradients from each $\mathcal{T}_i$ in the current batch of task and make the update to the parameters, finishing one iteration of the training process. This procedure runs until no further improvement can be made. The full meta-learning algorithm is summarized in Algorithm 1 and 2 in the appendix.

Task Hardness We develop a simple but effective way of computing NER task hardness based on the losses. Each task is re-weighted according to the hardness while being used in the gradient update. Specifically, we define the task difficulty $\Gamma_i = \{\gamma_i^\theta, \gamma_i^\phi, \gamma_i^\omega\}$ for task $\mathcal{T}_i$ with its corresponding objective values as

$$\gamma_i^\theta = \frac{\mathcal{L}_i}{\sum \mathcal{L}_j};\ \gamma_i^\phi = \frac{\mathcal{L}_i^{\text{lab}}}{\sum \mathcal{L}_j^{\text{lab}}};\ \gamma_i^\omega = \frac{\mathcal{L}_i^{\text{cls}}}{\sum \mathcal{L}_j^{\text{cls}}}, \qquad (2)$$

where $\{\gamma_i^\theta, \gamma_i^\phi, \gamma_i^\omega\}$ represent the task hardness scores to update parameters $\{\theta, \phi, \omega\}$, respectively. By incorporating task hardness into the optimization process, MetaBioNER would gradually shift the focus to more challenging tasks rather than the ones that contribute little to no learning value, *e.g.*, a task that contains short sentences without biomedical named entities, as multiplying the hardness score with the corresponding gradient value will force the gradient update to zero for those sentences. Table 1 shows how our learning algorithm ranks the contribution of each task towards the gradient update.

4 Experimental Results

Datasets We used the pre-processed version of those benchmark corpora used by BioBERT (Lee et al., 2019), which are publicly available at BioBERT's github website[3]. They are shown in Table 2. They are from four domains, Disease, Drug, Gene and Species, each of which will be used as the target domain. Rather than the general pre-trained GloVe embeddings, we used BioWordVec embeddings with with 200 dimensions (Chen et al., 2018; Yijia et al., 2019), which is pre-trained based on both PubMed database and clinical notes from MIMIC-III. serves the purpose of the project well.

Experimental Settings & Baselines To analyze the generalization ability of the learning framework under a low-resource setting, we considered the following experimental settings:

- The size of supporting set K: We used $K \in \{5, 10, 20, 50, 100\}$, since annotating medical corpora is both expensive and time consuming, requiring much domain expertise.
- Heterogeneous adaptation: Following Li et al. (2020b), we considered the more hard task, *i.e.*, heterogeneous adaptation that assumes each domain has a domain-specific decoder $\boldsymbol{\phi}$ and only the sentence encoder $\boldsymbol{\theta}$ is shared across domain and meta-learned. In other words, $\boldsymbol{\phi}$ is randomly initialised in meta testing for each corpus in the target domain and only $\boldsymbol{\theta}$ is adapted.

We implemented two variants of our MetaBioNER and compared them with MetaNER (Li et al., 2020b) and its variant without feature critic network.

- **MetaNER** acts as the major baseline. It is the latest and most related work to ours, showing the state-of-the-art performance. We followed the parameter settings that the authors detailed in their paper and tried to replicate the model based on our understandings. We validated our implementation by comparing its performance to the multi-tasking method used in MetaNER.
- **MetaNER w/o critic** excludes the feature critic network used in MetaNER, which indeed degenerate to MAML. We used the same parameter settings as those in MetaNER.
- **MetaBioNER** uses similar parameter settings to MetaNER, except that we re-calibrate the update of $\{\theta, \phi, \omega\}$ using equation (2), and weigh the gradient update based on the task hardness scores in meta training. [4]
- **MetaBioNER-NEs** makes use of clean sentences, each of which contains at least one biomedical entity. Its other parameter settings as the same as MetaBioNER without task hardness.

Note that corpora from the target domain are unseen by the meta learner in the meta-training phase. For instance, if the "Disease" domain is treated as the target domain for adaptation, we only perform meta-learning for Θ^* using the remaining

[3] https://github.com/dmis-lab/biobert

[4] As second-order gradients cannot be obtained for the recurrent neural net unit; thus, we use both first-order approximations (Nichol et al., 2018) and implicit gradients (Rajeswaran et al., 2019) to perform the update to $\{\theta, \phi, \omega\}$

		Disease		Drug		Gene		Species		Overall
		NCBI	BC5CDR	BC5CDR	BC4CHEMD	JNLPBA	BC2GM	LINNAEUS	S800	
5 shots	**MetaNER**	0.2729	0.2171	0.5784	0.2212	0.2175	0.2443	0.1214	0.1516	0.2530
	MetaNER w/o critic	0.2750	0.2327	**0.6136**	0.2347	0.2535	0.2374	0.1262	0.2216	0.2744
	MetaBioNER	**0.3001**	**0.2698**	0.6102	0.2464	0.3687	0.3326	**0.1753**	**0.2840**	**0.3234**
	MetaBioNER-NEs	0.2825	0.2530	0.5517	**0.2571**	**0.3776**	**0.3573**	0.1557	0.2615	0.3121
10 shots	**MetaNER**	0.3330	0.3688	0.6659	0.3360	0.3374	0.3265	0.3038	0.3164	0.3735
	MetaNER w/o critic	0.3785	0.3689	**0.6880**	0.3261	0.3394	0.3171	**0.3441**	0.2675	0.3787
	MetaBioNER	0.3953	0.4178	0.6798	**0.4227**	**0.4790**	**0.4489**	0.2939	**0.3703**	**0.4385**
	MetaBioNER-NEs	**0.4386**	**0.4222**	0.6605	0.3933	0.4371	0.4086	0.2474	0.3225	0.4163
20 shots	**MetaNER**	0.4612	0.4722	0.7301	0.4383	0.4167	0.3926	0.4952	0.2977	0.4630
	MetaNER w/o critic	0.4746	0.5115	0.6979	0.4200	0.3783	0.3680	**0.5309**	0.4173	0.4748
	MetaBioNER	**0.5631**	**0.5529**	**0.7472**	**0.4935**	**0.5466**	**0.5114**	0.3657	0.4432	0.5280
	MetaBioNER-NEs	0.5540	0.5098	0.7305	0.4694	0.5375	0.5097	0.4843	**0.5205**	**0.5394**
50 shots	**MetaNER**	0.5731	**0.6106**	0.7478	0.5082	0.5337	0.5058	0.6125	0.3607	0.5565
	MetaNER w/o critic	0.5890	0.6052	0.7364	0.4788	0.4681	0.4540	**0.6493**	0.4040	0.5481
	MetaBioNER	0.6250	0.5939	**0.7737**	0.5728	0.5666	0.5442	0.6369	**0.5855**	0.6123
	MetaBioNER-NEs	**0.6208**	0.5847	0.7612	**0.5781**	**0.6146**	**0.6016**	0.6373	0.5445	**0.6179**
100 shots	**MetaNER**	0.6005	0.6484	**0.7975**	0.5852	0.4993	0.4747	0.6581	0.3854	0.5811
	MetaNER w/o critic	0.5637	0.6447	0.7943	0.5904	0.5257	0.4985	**0.6880**	0.4159	0.5902
	MetaBioNER	0.6433	0.6489	0.7822	**0.6395**	0.6210	0.6064	0.6268	0.5416	0.6387
	MetaBioNER-NEs	**0.6699**	**0.6531**	0.7773	0.6093	**0.6338**	**0.6163**	0.5956	**0.5755**	**0.6413**

Table 3: Average performance (F1 Scores) of the heterogeneous domain adaptation for BioNER. The best performance is in boldface and the second best is underlined. All settings results are averaged from 20 distinct samples, *e.g.*, 5-shots score of MetaNER for NCBI corpus are averaged from 20 samples with each sample having 5 distinct sentences used as $\mathcal{T}_{\text{tr}}^{'}$ to optimize for the model initialized with parameters set learnt under MetaNER framework.

source domains "Drug", "Gene", "Species". We report the average F1 scores of five random runs of each method considered. More detailed settings to reproduce this work can be found in the appendix.

Results Table 3 presents the NER performance of our MetaBioNER, MetaNER and their variants under the heterogeneous adaptation setting. We have the following observations:

- MetaNER w/o critics v.s. MetaNER: The performance gap between MetaNER w/o critics and MetaNER is negligible in most cases. This could imply that learning a good feature-critic network, (*i.e.*, a domain classifier), needs more training samples in meta training than what we have in our experiments. Li et al. (2020b) assumed the support set in each iteration contains all sentences in the selected source domains in meta training.
- MetaNER v.s. MetaBioNER: By simply including the task hardness in the gradient update, MetaBioNER achieves a significant performance improvement over MetaNER with an average of $4-5\%$ improvement in terms of F1 score. In multiple cases (*e.g.*, JNLPBA 5 shots, BC2GM 10 shots, S800 100 shots, etc), the performance gain of MetaBioNER goes up to 15% in terms of F1-score. This comparison demonstrates that using the task hardness to differentiate the importance of each task in the gradient update is beneficial, contributing largely to the NER performance.
- MetaBioNER v.s. MetaBioNER-NEs: Both MetaBioNER and MetaBioNER-NEs work well in our experiments, outperforming the strong baseline with a large margin. MetaBioNER reweights the gradient update based on the task difficulty and MetaBioNER-NEs trains the meta learner exclusively on only sentences containing NEs. It is not surprising to see both approaches perform similarly when K is large, as the task hardness basically tries to automatically downweight tasks with sentences containing less or no NEs in a dynamic fashion based on the loss.
- It is interesting that both MetaBioNER and MetaBioNER-NEs performs worse than MetaNER w/o critic on the LINNAEUS corpus. The statistics in Table 2 show that 87% (table 2) of its sentences contains no NEs. Both MetaBioNER and MetaBioNER-NEs simply toss out those sentences implicitly and explicitly respectively during training, which can attribute to the performance loss.

5 Conclusion

We have proposed a simple yet effective method that can effectively leverage the task hardness information to improve the effectiveness of the learnt model in the few-shot NER settings. Experiments on biomedical corpora have shown that the sequence labelling function derived from our techniques have achieved substantial performance improvements compared to current baselines. As future work, we will further investigate task hardness strategies in the NLP settings, and apply the idea to other sequence labelling architectures, *e.g.*, deep transformers.

References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6430–6439.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. 2018. Metareg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems*, 31:998–1008.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2018. Biosentvec: creating sentence embeddings for biomedical texts. *CoRR*, abs/1810.09302.

Nigel Collier and Jin-Dong Kim. 2004. 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 (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Nan Ding, Xi Chen, Tomer Levinboim, Sebastian Goodman, and Radu Soricut. 2021. Bridging the gap between practice and pac-bayes theory in few-shot meta-learning. *arXiv preprint arXiv:2105.14099*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Special report: Ncbi disease corpus: A resource for disease name recognition and concept normalization. *J. of Biomedical Informatics*, 47:1–10.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

Martin Gerner, Goran Nenadic, and Casey Bergman. 2010. Linnaeus: A species name identification system for biomedical literature. *BMC bioinformatics*, 11:85.

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.

Jean Kaddour, Steindór Sæmundsson, and Marc Peter Deisenroth. 2020. Probabilistic active meta-learning. *arXiv preprint arXiv:2007.08949*.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel Lowe, Roger Sayle, Riza Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, and Alfonso Valencia. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:S2.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn Mattingly, Thomas Wiegers, and Zhiyong lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Jing Li, Billy Chiu, Shanshan Feng, and Hao Wang. 2020a. Few-shot named entity recognition via meta-learning. *IEEE Transactions on Knowledge and Data Engineering*.

Jing Li, Shuo Shang, and Ling Shao. 2020b. Metaner: Named entity recognition with meta-learning. In *Proceedings of The Web Conference 2020*, WWW '20, page 429–440, New York, NY, USA. Association for Computing Machinery.

Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. 2019. Feature-critic networks for heterogeneous domain generalization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3915–3924, Long Beach, California, USA. PMLR.

Chenghao Liu, Zhihao Wang, Doyen Sahoo, Yuan Fang, Kun Zhang, and Steven CH Hoi. 2020. Adaptive task sampling for meta-learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 752–769. Springer.

Cuong C Nguyen, Thanh-Toan Do, and Gustavo Carneiro. 2021. Probabilistic task modelling for meta-learning. *arXiv preprint arXiv:2106.04802*.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999.

Evangelos Pafilis, Sune P. Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PLOS ONE*, 8(6):1–6.

Aravind Rajeswaran, Chelsea Finn, Sham Kakade, and Sergey Levine. 2019. Meta-learning with implicit gradients.

L. Smith, L. Tanabe, R. Ando, C. Kuo, I-Fang Chung, C. Hsu, Y. Lin, R. Klinger, Christoph Friedrich, K. Ganchev, M. Torii, Hongfang Liu, Barry Haddow, Craig Struble, Richard Povinelli, Andreas Vlachos, William Baumgartner Jr, Lawrence Hunter, B. Carpenter, and W. Wilbur. 2008. Overview of biocreative ii gene mention recognition. *Genome Biology*, 9.

Max Woolf. 2018. char-embeddings.

Kai Xu, Zhenguo Yang, Peipei Kang, Qi Wang, and Wenyin Liu. 2019. Document-level attention-based bilstm-crf incorporating disease dictionary for disease named entity recognition. *Computers in Biology and Medicine*, 108:122 – 132.

Kai Xu, Zhanfan Zhou, Tianyong Hao, and Wenyin Liu. 2018. A bidirectional lstm and conditional random fields approach to medical named entity recognition. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017*, pages 355–365, Cham. Springer International Publishing.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*.

Huaxiu Yao, Yu Wang, Ying Wei, Peilin Zhao, Mehrdad Mahdavi, Defu Lian, and Chelsea Finn. 2021. Meta-learning with an adaptive task scheduler. *arXiv preprint arXiv:2110.14057*.

Zhang Yijia, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific Data*, 6.

Yucan Zhou, Yu Wang, Jianfei Cai, Yu Zhou, Qinghua Hu, and Weiping Wang. 2020. Expert training: Task hardness aware meta-learning for few-shot classification. *arXiv preprint arXiv:2007.06240*.

Henghui Zhu, Ioannis Ch. Paschalidis, and Amir Tahmasebi. 2018. Clinical concept extraction with contextual word embedding. *CoRR*, abs/1810.10566.

Algorithm 1 MetaBioNER

```
Require: $p(\mathcal{T})$ from source domains
Require: $\alpha, \beta, \lambda$ hyper-parameters
Require: m tasks batch size
Initialize $\theta$, $\phi$, $\omega$
while not converge do
    for $i = 1, \ldots,$ m do
        $\mathcal{T}_i \sim p(\mathcal{T})$
        $\mathcal{T}_i^S, \mathcal{T}_i^Q = \mathcal{T}_i$ s.t. $\mathcal{T}_i^S \cap \mathcal{T}_i^Q = \emptyset$
        $\mathcal{L}_i^{\text{lab}}, \mathcal{L}_i^{\text{cls}} =$ algorithm 2
        $\mathcal{L}_i = \mathcal{L}_i^{\text{lab}} + \lambda \mathcal{L}_i^{\text{cls}}$
    end for
    $\Gamma_1, \ldots, \Gamma_m =$ equation 2
    $\theta \leftarrow \theta - \alpha \sum_i \gamma_i^\theta \nabla_\theta \mathcal{L}_i$
    $\phi \leftarrow \phi - \alpha \sum_i \gamma_i^\phi \nabla_\phi \mathcal{L}_i^{\text{lab}}$
    $\omega \leftarrow \omega - \alpha \sum_i \gamma_i^\omega \nabla_\omega \mathcal{L}_i^{\text{cls}}$
end while
return $\Theta = (\theta, \phi)$
```

Algorithm 2 Support and query loss for $\mathcal{T}_i$

```
Require: $\mathcal{T}_i = \left(\mathcal{T}_i^S, \mathcal{T}_i^Q\right)$ s.t. $\mathcal{T}_i^S \cap \mathcal{T}_i^Q = \emptyset$
Require: $\theta, \phi, \omega$ current iteration parameters
Require: $\beta, \lambda$ hyper-parameters
Initialize $\theta_i$, $\phi_i$, $\omega_i$ with $\theta, \phi, \omega$
for $i = 1, \ldots,$ adaptation steps do
    $\mathcal{L}_i^{\text{lab}} = \mathcal{L}\left(h\left(\theta_i, \phi_i\right), \mathcal{T}_i^S\right)$
    $\mathcal{L}_i^{\text{cls}} = \mathcal{L}\left(f\left(\theta_i, \omega_i\right), \mathcal{T}_i^S\right)$
    $\mathcal{L}_i = \mathcal{L}_i^{\text{lab}} + \lambda \mathcal{L}_i^{\text{cls}}$
    $\theta_i \leftarrow \theta_i - \beta \nabla_{\theta_i} \mathcal{L}_i$
    $\phi_i \leftarrow \phi_i - \beta \nabla_{\phi_i} \mathcal{L}_i^{\text{lab}}$
    $\omega_i \leftarrow \omega_i - \beta \nabla_{\omega_i} \mathcal{L}_i^{\text{cls}}$
end for
$\mathcal{L}_i^{\text{lab}} = \mathcal{L}\left(h\left(\theta_i, \phi_i\right), \mathcal{T}_i^Q\right)$
$\mathcal{L}_i^{\text{cls}} = \mathcal{L}\left(f\left(\theta_i, \omega_i\right), \mathcal{T}_i^Q\right)$
return $\mathcal{L}_i^{\text{lab}}, \mathcal{L}_i^{\text{cls}}$
```

A Appendix

BiLevel Optimization Inspired by the meta-learning set-up in Li et al. (2019) and Balaji et al. (2018), MetaBioNER also includes an auxiliary network, named domain classifier network, to regularize the meta-learning process and promote domain generalization. The learning objective for the optimization process, similar to that of Li et al. (2019), can be described as

$$\mathcal{L}_i = \mathcal{L}^{\text{lab}}\left(h\left(\theta, \phi\right), \mathcal{T}_i\right) + \lambda \mathcal{L}^{\text{cls}}\left(f\left(\theta, \omega\right), \mathcal{T}_i\right) \tag{3}$$

where $\mathcal{L}^{\text{lab}}\left(h\left(\theta, \phi\right), \mathcal{T}_i\right)$ denotes the sequence labelling cross entropy loss obtained from the context encoder and tag decoder networks for the current task $\mathcal{T}_i$, replicating the finetuning process; and the auxiliary classification loss $\mathcal{L}^{\text{cls}}\left(f\left(\theta, \omega\right), \mathcal{T}_i\right)$ serves as a regularizer to balance the contribution of this auxiliary loss to the learnt representations.

The choice of the auxiliary network is vital to a successful implementation of the learning strategy. This network needs to satisfy two conditions: 1) the input is the output of the context encoder; and 2) the output is a non-negative scalar so thatt one can backpropagate the gradients properly Li et al. (2019). The purpose of this auxiliary objective is to introduce a sub-task to the base meta-training, which regulates the learning process to ensure agnostic representations of Θ^* are attained to minimize the labelling function transfer risk. In our implementation, this auxiliary network, parameterized by ω, consists of a fully connected layer used to predict which domain a sentence in a task $\mathcal{T}_i$ belongs to. The classification function f will henceforth be used to represent the composition of the context encoder network and the domain classifier network.

The full meta-learning algorithm is summarized in Alg. 1 and Alg. 2. It is apparent that the support set $\mathcal{T}_i^S$ and query set $\mathcal{T}_i^Q$ serve to imitate the finetuning process for the unobserved task $\mathcal{T}^{'}$. By training the labeling function h repeatedly with multiple tasks under the same constraints faced by the function during the finetuning process, MetaBioNER searches for the optimal parameters Θ^* that will minimize the transfer risk to tasks from the new domain. **Detailed experimental setups** The word embeddings used are taken directly from BioWordVec without any modifications (Chen et al., 2018; Yijia et al., 2019). This project uses the method created by Woolf (2018) to learn character embedding from the word embeddings statistics.

For each of the variants that we considered, the learnt parameters are trained with 10000 iterations or until convergence. The learning rates α and β are set at 1^{e-5}, the depth of the BiLSTM is 1 with a hidden size of 128; the gradient clip is set at 5; momentum at 0.9; the optimizer is adamW with a linear learning scheduler. Although MetaNER (Li et al., 2020b) claims that $\lambda = 0.8$ yields the best performance, we use $\lambda = 1$ as there are no significant difference. The size for each task will depend on the K-shots setting, *e.g.*, if we are interested in measuring the performance for 5-shots, we will the support set size and query set size to be both set to 5 during meta-training.

Iter	**Example**	**Score**
1	After discharge, he was finally given a diagnosis of PCH because a Donath-Landsteiner test was positive. The hatcher incubators of both companies were also persistently contaminated with Salmonella livingstone and Salmonella thomasville... ...**[gamma CACCC box binding factors]** mediate **[LCR-gamma]** interactions which normally enhance **[gamma-globin]** and suppress **[beta-globin gene]**... Toxicity was very mild with both regimens, although sedation was significantly higher in arm B (p less than 0.001).	0.13
	This article is part of the Special Issue entitled 'Neurodevelopmental Disorders'. Nuclear factor-kappaB (NF-kappa B) has been reported to regulate various genes involved in cancer and inflammation. Construction of block copolymers for the coordinated delivery of **[doxorubicin]** and **[magnetite]** nanocubes. The mice (10 per sex for each dose) was orally administered with neem oil with the doses of 0 (to serve as a control), 177, 533 and 1600 mg/kg/day for 90 days .	0.34
	The true incidence of nonsteroidal anti-inflammatory drug-induced cystitis in humans must be clarified by prospective clinical trials. There was marked QT prolongation greater than 0.55s in 13 patients, bradycardia less than 40 beats/min in 6 patients, dizziness and general fatigue in 1 patient each . Severe complications developed in four patients. These findings are consistent with the postulated mechanism for this unusual syndrome: acute diffuse crystallization of uric acid in renal tubules.	0.22
450	Stimulation of human neutrophils with **[chemoattractants]** **[FMLP]** or **[platelet activating factor (PAF)]** results in different but overlapping functional responses. Of even more interest, **[IkappaBalpha]** overexpression inhibited the production of **[matrix metalloproteinases 1 and 3]** while not affecting their tissue inhibitor. ...more durable inhibition of HIV - 1 replication than was seen with the **[NF-kappa B]** inhibitors alone or the **[anti-Tat sFv intrabodies]** alone. Spontaneous occurrence of early region 1A reiteration mutants of type 5 adenovirus in persistently infected human T-lymphocytes.	0.46
	Here we report the fabrication of single-molecule transistors based on individual C60 molecules connected to gold electrodes. The contractile effects of **[oxytocin]**, prostaglandin F2 alpha and their combined use on human pregnant myometrium were studied in vitro. Transcriptional activation of the **[proopiomelanocortin gene]** by **[cyclic AMP-responsive element binding protein]**. The difference between the effects of the two dose levels of Z.	0.18
	She was monitored for one more day and then discharged with instructions to discontinue her diet pills The Raf/Ras/ERK/MAPK pathway is known to be involved in NGF-induced outgrowth Our analysis reveals that the oviduct is lined, along its entire length, by a monolayered epithelium comprised of squamous-type cells. In one case study, Bramson et al.	0.01

Table 4: Examples of task hardness scores (computed from our method) for three tasks during the meta-training procedure, this score was recorded for two separate iterations, 1 and 450. The score is based on a scale from 0 to 1, the higher the score, the more challenging the task is. The NEs are put in brackets with red color for each sentence. We can observe that this task hardness ranks all tasks relatively equally in the first few iterations, and as the learning goes on, tasks that contain more NEs with more words will be given higher weights.

	Disease		Drug		Gene		Species		Overall
	NCBI	BC5CDR	BC5CDR	BC4CHEMD	JNLPBA	BC2GM	LINNAEUS	S800	
Multi-task (Yang et al., 2017)	0.829622	**0.821031**	**0.917411**	**0.881651**	**0.754455**	**0.800032**	**0.851441**	0.729993	0.823205
MetaNER (Li et al., 2020b)	**0.843734**	0.820414	0.906567	0.880616	0.752188	0.796445	0.838949	**0.764133**	**0.825381**

Table 5: MetaNER (Li et al., 2020b) vs. Multi-task (Yang et al., 2017). Performance (F1 Scores) for BioNER tasks using 100% training data of the target domain corpus to finetune the model. We acknowledge that although Li et al. (2020b) presented that MetaNER has on average an f1-score that is $2-3\%$ higher than multi-task learning does in their paper, this might not translate to the performance for the domains we use in this study.