Analyzing the Effect of Multi-task Learning for Biomedical Named Entity Recognition

Arda Akdemir  
University of Tokyo  
Tokyo, Japan  
aakdemir@hgc.jp

Tetsuo Shibuya  
University of Tokyo  
Tokyo, Japan  
tshibuya@hgc.jp

Abstract—Developing high-performing systems for detecting biomedical named entities has major implications. State-of-the-art deep-learning based solutions for entity recognition often require large annotated datasets, which is not available in the biomedical domain. Transfer learning and multi-task learning have been shown to improve performance for low-resource domains. However, the applications of these methods are relatively scarce in the biomedical domain, and a theoretical understanding of why these methods improve the performance is lacking. In this study, we performed an extensive analysis to understand the transferability between different biomedical entity datasets. We found useful measures to predict transferability between these datasets. Besides, we propose combining transfer learning and multi-task learning to improve the performance of biomedical named entity recognition systems, which is not applied before to the best of our knowledge.

I. INTRODUCTION

Publicly available biomedical articles keep increasing rapidly. Automated systems that utilize biomedical text mining methods are necessary to be handle this large amount of data with minimal manual effort. An important first step of any biomedical text mining method is the detection and classification of biomedical entities such as disease, drug or chemical mentions in biomedical articles. This task is referred to as biomedical named entity recognition (BioNER).

BioNER has seen remarkable progress with the advents in machine learning and deep learning methods [1]. These methods require labeled datasets and benefit from increasing the amount of labeled examples. Artificial neural networks form the core of almost all state-of-the-art bioNER systems. The main drawback of these methods is that the networks must be trained from scratch for each dataset, separately. Even though recent progress in BioNER is promising, the overall performance is significantly lower than general domain NER. This is mainly due to the scarcity and sub-optimal utilization of the labeled datasets in the biomedical domain.

Transfer learning is a training paradigm that mitigates the above mentioned issues of current approaches. It attempts to utilize the information obtained from a source task to improve the performance on a target task. Transfer learning is shown to be especially useful when the size of the labeled data is limited for the target task [2], making BioNER a very suitable target task. Multi-task learning is a special case of transfer learning where multiple tasks are learned simultaneously [3].

In this context, this corresponds to learning multiple biomedical named entity datasets using a single neural network.

Recently, the seminal work of Devlin et al. [4], i.e. the BERT model, enabled progress in various NLP tasks, including NER. BERT uses self-supervised learning which relieves the need for having labeled examples to train the neural network. Lee et al. [1] proposed BioBERT, which is the BERT model pretrained on a large unlabeled biomedical dataset. They finetuned the BioBERT model on labeled datasets using supervised learning and obtained improvements on several downstream biomedical NLP tasks. Yet, BioBERT has not been applied before in the context of multi-task learning, to the best of our knowledge. This motivated us to use BioBERT as the shared network across all biomedical datasets. We claim that sharing information across datasets help improve the overall performance as the representations obtained on one biomedical dataset is relevant to others, even though the annotated entities are different.

Multi-task learning is also used recently to improve the performance on BioNER datasets [5], [6]. Yet, the analysis of where the improvements come from is limited and the effect of transfer learning is unclear. Thus, there is a lack of theoretical understanding of when and why transfer learning and multi-task learning bring improvements.

In this study, we analyze the effect of multi-task learning for biomedical named entity recognition. To this end, we experimented on seven BioNER benchmark datasets and analyzed the effect of multi-task learning by using ten different measures. We evaluate the usefulness of these measures with three different metrics. Besides, we propose combining transfer learning and multi-task learning for BioNER which is not employed before to the best of our knowledge. The main contributions of this study are as follows:

- We analyzed the effect of multi-task learning by using various dataset characteristics, similarity measures, and evaluation metrics. This allowed us to find useful features and measures to predict transferability, which we believe can help reduce the required manual effort for finding useful auxiliary datasets in future.
- We propose a novel BioBERT-CRF based single task model that outperformed the state-of-the-art on four datasets.
• We incorporated multi-task learning to the proposed neural network which brought performance improvements on six out of eight cases.

II. RELATED WORK

Multi-task learning and transfer learning have both been applied for BioNER before. Lee et al. [1] obtained state-of-the-art results on several biomedical entity recognition datasets by using the BERT language model pretrained on the Biomedical abstracts and papers. Crichton et al. [6] is the first work that attempts to apply multi-task learning for detecting biomedical named entities. They used pretrained biomedical word embeddings and used a CNN based neural network to detect the named entities. They also use Viterbi decoding for the tag transition probabilities. They only calculate binary probabilities for tag transitions (appears or does not appear inside the training dataset), whereas we use a CRF layer with trainable parameters to output a score for each possible transition. Wang et al. [5] proposed a BiLSTM-CRF based neural network where character and word embeddings are shared by all dataset specific components. Unlike Crichton et al., who propose sharing a convolutional layer, they propose sharing only word/character embeddings across different datasets. Mehmoood et al. [2] also analyzed the effect of multi-task learning for BioNER datasets. They propose using stack LSTMs for multi-task learning of BioNER datasets. Their analysis of multi-task learning focus on architectural variations, whereas we focus on analyzing dataset characteristics for transferability. Yoon et al. [8] proposed using an expert BiLSTM-CRF based network for each dataset, separately. During training, they use the output of all other trained models as an additional input to the neural network. They claim that the output of each expert model helps prevent mislabelings caused by polysemy.

Even though both transfer learning and multi-task learning have already been used for BioNER, the analysis of the effect of using these methods is highly limited. A similar analysis to ours is done before on general domain NLP datasets. Bingel et al. [11] analyzed the task relations for multi-task learning on ten NLP tasks. Alonzo et al. [10] analyzed the task relations for multi-task learning on several sequence labeling tasks. Different from these previous studies, we focus on the datasets from the biomedical domain. These datasets have unique characteristics and we claim that findings on the general domain datasets might not be directly applicable to them.

III. DATASETS

In our experiments we used seven biomedical entity datasets. BC5CDR contains both disease and chemical entity mentions. We follow the same convention with previous work to have two identical copies of the dataset that are only annotated for each entity type (BC5CDR-chem and BC5CDR-disease). The details about the datasets are given in Table I.

| Entity type          | Number of sentences | Entity/token ratio |
|----------------------|---------------------|--------------------|
| Species/Protein NE   | 12,737              | 0.042              |
| Chemical NE          | 30,681              | 0.033              |
| Disease NE           | 4,559               | 0.044              |
| Gene/Protein NE      | 14,689              | 0.073              |
| Disease NE           | 5,423               | 0.058              |
| Species NE           | 11,914              | 0.008              |
| Species NE           | 5,732               | 0.017              |

IV. DATASET MEASURES AND EVALUATION METRICS FOR ANALYZING TRANSFERABILITY

In order to understand under which conditions multi-task learning brings more benefits, we used a total of 12 dataset measures. To take into account the asymmetric characteristics of MTL gains we included directed similarity measures.

A. Dataset measures

**Shared Vocabulary** (vocab) (directed): Rate of the target vocabulary words found inside the auxiliary dataset vocabulary [13]. This is a directed metric as the vocabulary sizes differ among datasets.

\[ s = \frac{|V_t \cap V_a|}{|V_t|} \]

where \( s \) is the shared vocabulary ratio for a target dataset \( t \) and an auxiliary dataset \( a \), where \( V \) represents the vocabulary.

**Topic Distribution similarity** (topic) (undirected): Following Ruder et al. [19], we used LDA topic modeling [20] to represent each dataset as \( n \) dimensional vectors, where each index represents a topic. We calculated the cosine similarity between all dataset pairs:

\[ t_{i,j} = t_{j,i} = \frac{dot(v_i, v_j)}{|v_i||v_j|} \]

where \( v_i \) represents the \( n \) dimensional topic vector of the \( i^{th} \) dataset, and \( t_{i,j} = t_{j,i} \) is the topic based similarity score.

**Embedding similarity** (bert) (undirected): Word embedding based distance is frequently used to measure similarity between datasets [19]. Recently, Vu et al. [21] used BERT-based representations to measure similarity between datasets. Similarly, we leverage the BioBERT model to obtain the vector representations of all sentences of a dataset. Then we average over all sentences to get the BioBERT-based dataset representations:

\[ v_i^b = \frac{1}{|D_i|} \sum_{k=1}^{D_i} BioBERT(D_i(k)) \]

\[ b_{i,j} = \frac{dot(v_i^b, v_j^b)}{|v_i^b||v_j^b|} \]

where \( D_i(k) \) is the \( k^{th} \) sentence of the \( i^{th} \) dataset, \( v_i^b \) is the BioBERT representation of \( i^{th} \) dataset, and \( b_{i,j} \) is the embedding-based similarity.

**Cooccurring entity ratio** (cooccur, directed): Rate of the target entity tokens found and annotated inside the auxiliary dataset. Schroder et al. [13] show that the rate of cooccurring
entities is a good predictor for MTL gains. We follow them to include this value to predict transferability.

We first used each of the above measures to rank all auxiliary datasets for a specific target dataset, separately. Then, we considered pairwise combinations of these four measures. In total this gave us 10 different scores about the transferability we considered pairwise combinations of these four measures. To include this value to predict transferability.

Next, we give the results for the multi-task learning models. As the baseline models, we trained BioBERT based single-task learning models for each dataset, separately. Next, we incorporated a CRF layer to this baseline model. This improved the results obtained on each dataset significantly. Table II shows the results for STL. The results show that incorporating CRF layer improves the performance across all datasets. These results also show that our proposed BioBERT+CRF single task model achieves SOTA results for four out of eight biomedical datasets. For these experiments, we follow Lee et al. [1] to combine the validation and training datasets to train our models, in order to have a fair comparison.

**B. Multi Task Learning Results**

Next, we give the results for the multi-task learning models. For these experiments, we did not combine the training and validation sets. Validation sets are used to determine when to stop the training. Yet, we still report the SOTA results on the same table for reference. In total we used eight datasets and we trained multi-task learning models for all pairs. We used a flat architecture that uses the pretrained BioBERT

**V. Experiment Details**

All our experiments are conducted using a single-thread V100 GPU. We use the PyTorch library to implement the proposed neural networks.

For the initial single-task learning experiments, we applied early stopping on the validation split to determine the number of steps on each dataset. In order to have a fair comparison, we train the single-task and multi-task models for equal number of iterations on each target dataset. We keep the number of iterations on the auxiliary and target datasets the same during the multi-task learning experiments. So, if a single-task model is trained for X steps on a target dataset, the multi-task counterpart is trained for X steps on the target and X steps on the auxiliary dataset.

We conducted single-task experiments on eight different datasets. We also conducted multi-task related experiments for all pairs of datasets (36 pairs). In total, we ran experiments on 44 different settings.

**VI. Results**

**A. Single Task Learning Results**

As the baseline models, we trained BioBERT based single-task learning models for each dataset, separately. Next, we incorporated a CRF layer to this baseline model. This improved the results obtained on each dataset significantly. Table II shows the results for STL. The results show that incorporating CRF layer improves the performance across all datasets. These results also show that our proposed BioBERT+CRF single task model achieves SOTA results for four out of eight biomedical datasets. For these experiments, we follow Lee et al. [1] to combine the validation and training datasets to train our models, in order to have a fair comparison.

**B. Multi Task Learning Results**

Next, we give the results for the multi-task learning models. For these experiments, we did not combine the training and validation sets. Validation sets are used to determine when to stop the training. Yet, we still report the SOTA results on the same table for reference. In total we used eight datasets and we trained multi-task learning models for all pairs. We used a flat architecture that uses the pretrained BioBERT
TABLE II: Results for single task learning of Biomedical NER.

| DATASET      | BioBERT baseline | BioBERT+CRF | SOTA     |
|--------------|------------------|-------------|----------|
|              | PRE  | REC  | F-1  | PRE  | REC  | F-1  | PRE  | REC  | F-1  |
| BC2GM        | 82.19 | 83.93 | 83.05 | 82.86 | 84.61 | 83.73 | 84.32 | 85.12 | 84.72 |
| BC4CHEMD     | 90.12 | 90.74 | 90.43 | 91.74 | 90.69 | 91.21 | 92.80 | 91.92 | 92.36 |
| BC5CDR-disease | 93.08 | 93.87 | 93.47 | 93.29 | 93.72 | 93.76 | 93.68 | 93.26 | 93.47 |
| JNLPBA       | 83.78 | 86.87 | 85.3  | 85.56 | 86.69 | 86.12 | 86.47 | 87.84 | 87.15 |
| NCBIdisease  | 70.9  | 82.65 | 76.32 | 71.79 | 85.11 | 78.71 | 72.24 | 83.56 | 77.49 |
| linnaeus     | 86.43 | 88.23 | 87.32 | 87.21 | 90.21 | 88.89 | 88.22 | 91.25 | 89.71 |
| s800         | 91.36 | 87.97 | 89.63 | 91.72 | 85.94 | 89.16 | 90.77 | 85.83 | 88.24 |
|              | 70.11 | 76.14 | 73.0  | 72.0  | 80.44 | 75.99 | 72.80 | 75.36 | 74.06 |

We analyzed the effect of the entity/token ratio for multi-task learning. We observed that the target entity/token ratio is negatively correlated (-0.382 and -0.5 for pearson and spearman correlations, respectively) whereas auxiliary entity/token ratio is positively correlated (0.698 and 0.833 for pearson and spearman correlations, respectively). In other words, multi-task gains are more likely to occur when the target dataset entities are sparse and the auxiliary dataset entities are abundant. The correlation for the auxiliary entity/token ratio is stronger and seems to be a better predictor for multi-task gains in this setting. We speculate that the abundance of auxiliary entities may help capturing important representations of the biomedical entities without simply memorizing them.

C. Learning All Datasets at Once

Next, we attempted to train a single multi-task learning model to learn all datasets at once. This methodology is different than training a single model on the concatenation of all datasets, as the architecture contains task-specific sequence labeling layers for each dataset. The rightmost column of Table IV shows the results of this experiment. We observe that learning all datasets at once only brought improvement over single-task learning for the NCBI-disease dataset. We observe that except for BC2GM, BC4CHEMD, and BC5CDR-chem datasets, learning all datasets at once performed better than some pairwise multi-task learning setting. Yet, for all target datasets, pairwise multi-task learning outperformed learning all datasets at once for at least one setting. These results strongly suggest that the model struggles to learn good representations on the common layer when it is exposed to all datasets at once. The results also show that naively exposing the neural network to more data is not guaranteed to bring performance improvements.
Fig. 1: Plots showing the relation between dataset characteristics and change in accuracy. (a-b) show the relation between dataset sizes and change in target task accuracy. (c-d) show the relation between the target accuracy change and the entity/token ratios of auxiliary and target datasets, respectively. The correlation of the auxiliary entity/token ratio and the correlation of the target dataset size with the change in accuracy is stronger.

TABLE IV: Comparison of all multi-task learning experiments with the single-task learning counterparts and the current state-of-the-art. All results are given in percentage (%) F-1.

|       | STL  | SOTA | BC2GM | BC4CHEMD | BC5CDR-chem | BC5CDR-disease | JNLPBA | NCBI-disease | linnaeus | s800 | conll-eng | All  |
|-------|------|------|-------|----------|--------------|----------------|--------|--------------|----------|------|-----------|------|
| BC2GM | 82.31| 84.72| -     | 82.99    | 81.86        | 81.85          | 81.06  | 82.13        | 82.82    | 82.2 | 82.35     | 80.76|
| BC4CHEMD| 88.43| 92.36| -     | 92.42    | 92.58        | -              | 89.15  | 89.23        | 89.47    | 88.82 | 89.31     | 87.71|
| BC5CDR-disease | 92.9 | 94.7 | -     | 92.47    | 92.58        | -              | 92.29  | 91.98        | 92.1     | 91.36 | 92.37     | 91.3 |
| JNLPBA | 85.57| 87.15| -     | 84.49    | 84.86        | 83.78          | -      | 84.24        | 84.0     | 83.47 | 83.93     | 84.49|
| NCBI-disease | 77.81| 77.49| -     | 77.88    | 77.76        | 76.8           | 77.24  | -            | 77.69    | 77.52 | 76.78     | 77.56|
| linnaeus | 86.98| 89.71| -     | 88.1     | 87.81        | 86.8           | 87.92  | 88.0         | -        | 87.5  | 87.23     | 88.09|
| s800   | 75.33| 74.06| -     | 73.27    | 74.17        | 72.34          | 71.26  | 73.54        | 74.64    | 73.55 | 75.62     | 74.15|
| conll-eng | -    | 93.09| -     | 88.4     | 88.32        | 88.16          | 88.3   | 88.53        | 88.39    | 87.87 | 88.74     | -   |

Fig. 2: Heatmaps showing the relative performance of multi-task learning over single-task learning(left) and the relative performance of using different auxiliary dataset (right). For each row on the right figure, the auxiliary dataset (column) that brought the least improvement is denoted with zero.

D. Analysis of Transferring Information between Datasets

Finally, we analyze the transferability between datasets. The aim of this analysis is to understand when and why transfer learning helps improve performance on a target dataset. For this reason, we also included Conll-2003 English Named Entity Recognition dataset as an out-of-domain dataset. The results are given in Table [IV]. Surprisingly, we did not observe that transferring information from the Conll dataset (out-of-domain) performs worse than transferring information from all other biomedical entity datasets (in-domain). Only for the linnaeus and BC5CDR-disease datasets, we observe that using Conll as the auxiliary dataset has the lowest performance.

Right heatmap of Fig[2] shows the relative performance on each target dataset. For each row (target dataset) the lowest scoring auxiliary dataset has the value zero. We see that the benefits of multi-task learning is not symmetric. This motivated us to incorporate directed similarity measures to analyze the multi-task gains (explained in section IV).

Multi-task learning in grouped datasets. The seven biomedical datasets contained four types of entities (gene/protein, chemical, disease and species). JNLPBA and BC2GM contain gene/protein mentions. For the gene/protein group, we see that in-group transfer is very useful for JNLPBA but very harmful for BC2GM(lowest performance for BC2GM, and highest improvement for JNLPBA). BC4CHEMD and BC5CDR-chem contain chemical entity mentions. For the chemical entity group, we see that in-group transfer gives the second lowest performance for BC4CHEMD (transferring from BC5CDR-chem to BC4CHEMD), but highest performance for BC5CDR-chem. For disease mention datasets (NCBI-disease and BC5CDR-disease) in-group transfer brought performance degradation for BC5CDR-disease (-1.6 F-1 Score over single-task learning), whereas it brought the third highest improvement +0.94 F-1 score over single task learning) for the NCBI-disease dataset. For s800 and linnaeus (species group) datasets, we see that in-group transfer gives the 4th highest performance out of 9 datasets in both cases. The non-symmetry between the mtl gains in these pairwise setting is interesting and a detailed analysis might yield important findings.

Next, we used the similarity measures and evaluated the predictability of the MTL gains. Table [V] shows the results for measuring the goodness of all similarity measures. For NDCG, higher scores are better, whereas for the other two lower scores indicate a more accurate ranking. The first row (random) shows the values for randomly selecting the
TABLE V: Evaluation of different similarity measures for predicting MTL gains.

| Measure | NDCG Score | Best auxiliary dataset rank | Rank of best auxiliary prediction |
|---------|------------|----------------------------|----------------------------------|
| topic   | 0.897      | 3.333                      | 3.778                            |
| vocab   | 0.906      | 3.889                      | 3.667                            |
| cooccur | 0.918      | 3.778                      | 3.889                            |
| bert    | 0.886      | 4.444                      | 4.667                            |
| topic_vocab | 0.896   | 3.778                      | 4.333                            |
| topic_cooccur | 0.905  | 3.444                      | 3.778                            |
| topic_bert | 0.895  | 3.556                      | 4.333                            |
| vocab_bert | 0.895  | 3.889                      | 4.444                            |
| cooccur_bert | 0.906  | 3.778                      | 3.556                            |

auxiliary datasets. NDCG score for random is generated by randomly scoring auxiliary datasets for 10,000 iterations. For all similarity measures, the NDCG score is significantly higher (at least 0.207) than selecting the auxiliary dataset randomly.

Observations. The first observation is that, we see a strong signal in the similarity measures we use. Using any of the similarity measures to pick an auxiliary dataset is better than making a random selection. We observe that in our setting, cooccur produced the best ranking according to the NDCG score. Similarity measures using topic (topic alone or its combination with other similarity measures) occupies the top 4 best average ranking for the best auxiliary dataset ($\rho$ metric). This result strongly suggests that utilizing the topic-based similarity information is very useful for finding beneficial auxiliary datasets. In other words, topic based similarity scores obtained using an LDA model is a good predictor for MTL gains.

Moreover, we observed that combining topic similarity with any other similarity measure always improved the $\rho$ score. Another critical observation is related to the BERT-based similarity. We see that BERT-based similarity performed the lowest according to all three evaluation methods. This result strongly suggests that using the BERT-based dataset vector similarities is not a good predictor for mtl gains. We suspect that this might be caused by large number of entities that occur but are not annotated inside the auxiliary datasets with similar BioBERT vectors. These unlabelled entities might hinder the target dataset performance.

VII. CONCLUSION

In this work, we proposed combining transfer learning and multi-task learning for BioNER, which is not done before to the best of our knowledge. The proposed method achieved state-of-the-art results on several biomedical named entity datasets. The main purpose of this study was to analyze and understand under which conditions transferring information from an auxiliary dataset helps improve performance on a target dataset. To this end, we used various dataset measures and evaluated their ability to predict the MTL gains using three different evaluation metrics. The analysis showed that the dataset measures contain strong signals about the benefits of multi-task learning.

VIII. FUTURE WORK

First, we plan to expand the analysis of predictability of MTL gains by including more dataset measures. Next, we plan to focus on data selection. We aim to focus on auxiliary datasets that brought the least MTL gains for a target dataset, and find subsets of these datasets that will bring higher MTL gains.

REFERENCES

[1] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: pre-trained biomedical language representation model for biomedical text mining,” arXiv preprint arXiv:1901.08746, 2019.
[2] S. Ruder, “Neural Transfer Learning for Natural Language Processing,” Ph.D. dissertation, National University of Ireland, Galway, 2019.
[3] R. Caruana, “Multi-task learning,” Machine learning, vol. 28, no. 1, pp. 41–75, 1997.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 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), 2019, pp. 4171–4186.
[5] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, and J. Han, “Cross-type biomedical named entity recognition with deep multi-task learning,” Bioinformatics, vol. 35, no. 10, pp. 1745–1752, 2019.
[6] G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen, “A neural network multi-task learning approach to biomedical named entity recognition,” BMC bioinformatics, vol. 18, no. 1, p. 368, 2017.
[7] T. Mehmood, A. Gerevini, A. Lavelli, and I. Serina, “Multi-task learning applied to biomedical named entity recognition task,” in CLIC-it, 2019.
[8] W. Yoon, C. H. So, J. Lee, and J. Kang, “Collaboronet: collaboration of deep neural networks for biomedical named entity recognition,” BMC bioinformatics, vol. 20, no. 10, p. 249, 2019.
[9] J. Bingel and A. Søgaard, “Identifying beneficial task relations for multi-task learning in deep neural networks,” arXiv preprint arXiv:1702.08303, 2017.
[10] H. M. Alonso and B. Plank, “When is multitask learning effective? semantic sequence prediction under varying data conditions,” arXiv preprint arXiv:1612.02251, 2016.
[11] L. Smith, L. K. Tanabe, R. J. see Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev et al., “Overview of biocreative ii gene mention recognition,” Genome biology, vol. 9, no. S2, p. S2, 2008.
[12] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, and A. Valencia, “Chemdner: The drugs and chemical names extraction challenge,” Journal of cheminformatics, vol. 7, no. S1, p. S1, 2015.
[13] C. Wei, H. Kao, and Z. Lu, “Proceedings of the biocreative 2012 workshop,” 2012.
[14] N. Collier, P. Ruch, and A. Nazarenko, “Jnlpba’04: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications,” ACL, Morristown, vol. 18, 2004.
[15] R. I. Doğan, R. Leaman, and Z. Lu, “Ncbi disease corpus: a resource for disease name recognition and concept normalization,” Journal of biomedical informatics, vol. 47, pp. 1–10, 2014.
[16] M. Gernert, G. Nenadic, and C. M. Bergman, “Linnaeus: a species name identification system for biomedical literature,” BMC bioinformatics, vol. 11, no. 1, p. 85, 2010.
[17] E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwarter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen, “The species and organisms resources for fast and accurate identification of taxonomic names in text,” PloS one, vol. 8, no. 6, p. e65390, 2013.
[18] F. Schröder and C. Biemann, “Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2971–2985.
[19] S. Ruder and B. Plank, “Learning to select data for transfer learning with bayesian optimization,” arXiv preprint arXiv:1707.05246, 2017.
[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.
[21] T. Vu, T. Wang, T. Munkhdalai, A. Sordoni, A. Trischler, A. Mattarella-Micke, S. Maji, and M. Iyyer, “Exploring and predicting transferability across nlp tasks,” arXiv preprint arXiv:2005.00770, 2020.
[22] K. Jarvelin and J. Kekäläinen, “Cumulated gain-based evaluation of ir techniques,” ACM Transactions on Information Systems (TOIS), vol. 20, no. 4, pp. 422–446, 2002.
APPENDIX

All models are implemented using Python 3.7.2 programming language and the PyTorch library version 1. All the required source code and links to the relevant datasets to replicate the results are included inside the Supplemental Material. Please refer to the README.md document inside the Supplemental Material on how to run the codes and download the necessary materials.

Table VI shows the results on the small datasets. We repeated each experiment five times and report the average score on each target dataset.
|                  | BC2GM | BC4CHEMD | BC5CDR-chem | BC5CDR-disease | JNLPBA | conll-eng | NCBI-disease | linnaeus | s800 |
|------------------|-------|----------|-------------|----------------|--------|----------|--------------|---------|------|
| BC2GM            | -     | 79.45    | 77.32       | 79.13          | 76.11  | 77.03    | 78.81        | 79.04   | 76.71|
| BC4CHEMD         | 85.7  | -        | 84.05       | 84.26          | 84.75  | 86.03    | 84.73        | 85.01   | 85.28|
| BC5CDR-chem      | 91.76 | 93.09    | -           | 92.15          | 91.93  | 92.41    | 91.87        | 92.19   | 92.37|
| BC5CDR-disease   | 81.46 | 82.0     | 82.3        | -              | 81.63  | 82.6     | 81.25        | 81.5    | 81.53|
| JNLPBA           | 76.32 | 75.73    | 76.73       | 75.45          | -      | 75.82    | 75.34        | 75.9    | 76.3 |
| conll-eng        | 83.31 | 84.0     | 84.38       | 84.29          | 83.85  | -        | 84.02        | 83.56   | 85.95|
| NCBI-disease     | 84.85 | 83.64    | 83.64       | 81.01          | 82.21  | 83.95    | -            | 87.43   | 84.15|
| linnaeus         | 84.48 | 82.93    | 86.03       | 86.43          | 86.96  | 82.91    | 83.15        | -       | 83.83|
| s800             | 78.01 | 74.29    | 75.71       | 80.85          | 77.14  | 79.1     | 80.0         | 78.87   | -    |