Knowledge Guided Named Entity Recognition for BioMedical Text

Pratyay Banerjee and **Kuntal Kumar Pal** and **Murthy Devarakonda** and **Chitta Baral**
Department of Computer Science, Arizona State University
pbanerj6,kkpal,mdevarak,chitta@asu.edu

Abstract

In this work, we formulate the NER task as a multi-answer question answering (MAQA) task, and provide different knowledge contexts, such as, entity types, questions, definitions and definitions with examples. This formulation (a) enables systems to jointly learn from varied NER datasets, enabling systems to learn more NER specific features, (b) can use knowledge-text attention to identify words having higher similarity to provided knowledge, improving performance, (c) reduces system confusion by reducing the classes to be predicted to B, I, O only, and (d) Makes detection of nested entities easier. We perform extensive experiments of this Knowledge Guided NER (KGNER) formulation on 18 Biomedical NER datasets, and through these experiments we note that knowledge helps. Our problem formulation is able to achieve state-of-the-art results in 16 out of 18 datasets.

1 Introduction

There are several tasks in Natural Language Processing and Understanding which require extensive external knowledge for systems to perform reasonably well. The external knowledge can be about entities and their relations, such as in Named Entity Recognition (CoNLL-2003 (Sang and De Meulder, 2003), OntoNotes (Weischedel et al., 2011), etc) and Relation Extraction (SEMEval-2010 (Hendrickx et al., 2010), TACRED (Zhang et al., 2017), etc). External knowledge can also be about commonsense or science, such as in Question Answering tasks (RACE (Lai et al., 2017), OpenBookQA (Mihaylov et al., 2018), SocialIQA (Sap et al., 2019)) etc.

In this work, we focus on Named Entity Recognition (NER) for biomedical texts. In biomedical domain, "NER is difficult because the target words are mainly proper nouns or unregistered words. In addition, new words can be generated frequently, and even same word stream could be recognized as diverse named entities in terms of their current context" (Song et al., 2018; Cohen and Hunter, 2004; Liu et al., 2006). The entities sometimes differ subtly, and hence require even more precise knowledge which we incorporate through sentences or words as shown in Table 1.

Text: It was noted that she became symptomatic when she was not on a beta blocker but that on a beta blocker she had significant pacemaker failure.
Knowledge : **Entity** : Problem **Question** : What is the problem mentioned in the text? **Definition** : Problem is a difficulty, disorder, or condition needing resolution. **Examples** : hypertension, pain, shortness of breath, chest pain,nausea, afebrile, coronary artery disease, vomiting, edema, fever.
Entities : symptomatic, significant pacemaker failure **Entity Span** : (6, 6), (23, 25)

Table 1: NER data with Entities and Knowledge. We hand craft knowledge for each type, for example, here the question, the definition and the examples are provided as knowledge to extract the entity "Problem".

Most of the NER systems, formulate the problem as a classification task. A token T_i is classified to be one of the three tags B-E_k, I-E_k, O in the BIO-Tagging scheme, where $k = 1..K$, K is the number of entity types and E is the entity type. The performance of the problems formulated in this way degrades due to multiple challenges: (a) *Labelling error*, when a token is classified as B-E_k or I-E_k but the token is actually a B-E_j or I-E_j where ($j! = k$), means even though a system was able to identify the location of an entity, it fails to identify the type of the entity, (b) inability to leverage more information for a particular entity type, since their task formulation only allows them to predict all entity types jointly, (c) lack of labelled

data for each entity type, especially in the biomedical domain. Challenge (a) and (b) are even more profound in the presence of nested named entities. Challenge (c) affects low resource languages and other low resource scientific domains.

We attempt to address these challenges through our following contributions:

- (a) and (c) by modelling the task as a multi-answer question answering task, where we predict only one type of entity, given a context. This formulation allows us to avert the issue of nested named entities and allows us to jointly learn from multiple different datasets having similar entities.
- We address challenge (b) by providing various types of knowledge, and do an empirical study of which knowledge types are better.
- We create a considerably large dataset combining 18 source datasets having in total 398495 training data, 148166 validation data and 502306 test data.
- We push the state-of-the-art exact match F1 scores for 16 publicly available biomedical NER datasets.

2 Related Work

External Knowledge : In the past, there have been several attempts to incorporate external knowledge through feature engineering and lexicons (Liu et al., 2019; Borthwick et al., 1998; Ciaramita and Altun, 2005; Kazama and Torisawa, 2007), or incorporating knowledge in the feature extraction stage (Crichton et al., 2017; Yadav and Bethard, 2018), or using document context (Devlin et al., 2018). There have been some attempts to use simple textual knowledge sentences for solving question answering tasks, such as in OpenBookQA (Mihaylov et al., 2018) and SocialIQA (Sap et al., 2019) by (Banerjee et al., 2019; Mitra et al., 2019). In our work, we incorporate simple textual knowledge sentences, similar to the attempts done for incorporating knowledge in question-answering tasks.

Multi-Task Learning : There have been multiple attempts to use multi-task learning to tackle the labelling problem of NER. For example, multi-task learning with simple word embedding and CNN (Crichton et al., 2017), cross-type NER with Bi-LSTM and CRF (Wang et al., 2018), MTL with private and shared Bi-LSTM-CRF using character and word2Vec word embeddings (Wang et al., 2019). In our work, we do multi-task learning by reducing all different NER tasks to the same generic format.

Language Models and Transfer Learning : There have been other attempts to reduce the labelling confusion by using a single model to predict each entity-type (Lee et al., 2019) and also using transfer-learning (Lee et al., 2019; Beltagy et al., 2019; Si et al., 2019). Our work is similar to them, which also use pre-trained language models (BERT), and/or predict different types of entities separately, but differs in task formulation and use of explicit external knowledge.

NER as a Question Answering Task : In general domain, researchers have formulated multiple NLP tasks as question-answering format in DecaNLP (McCann et al., 2018), semantic-role labelling as in QASRL (He et al., 2015) and others have argued that question-answering is a format not a task (Gardner et al., 2019). We also use question-answering format as a part of our task, to address the aforementioned challenges.

3 Our Approach

In our approach, we attempt to tackle each of the aforementioned challenges by formulating the NER task in the following way. Given a text T_i and entity type E_k we create contexts C_j. We then use C_j to find the entities and their entity types. We use four types of context. (a) *entity types*, E_k (b) separate *question* created using each entity type, Q (c) *definition* of each entity type, D (d) *definition with example*, $D \cup Eg$. For the example mentioned in Table 1, E_k is "*Problem*", Q is "*What are the problems mentioned in the text?*", D is the definition text, "*Problem is a difficulty, disorder, or condition needing resolution*", and Eg are the examples "*hypertension, pain, shortness of breath, chest pain, nausea, afebrile, coronary artery disease, vomiting, edema, fever*".

In the conventional NER task formulation, each token of the text would have been asked to be classified as either B_{E_i}, I_{E_i}, and O, where i is different for different types of entities. For example, "she(O), became(O), symptomatic(B_{E_k}) significant(B_{E_k}), pacemaker(I_{E_k}), failure(I_{E_k})".

We reformulate the task, to classify each token T_i only to three classes, B_{Ans}, I_{Ans} and O even if there are multiple entities in a text of same type. If there are multiple entity types in a text, there will be a question for each entity type. Those tokens which should answer the query using the given knowledge,

should be classified as B_{Ans} or I_{Ans} depending on they being the first token of the answer or the intermediate tokens. All other tokens are to be predicted as O.

4 Dataset Preparation

We create the dataset for NER using fifteen publicly available biomedical datasets and four datasets from previous i2b2 challenges (Sun et al., 2013; Uzuner et al., 2011, 2010, 2012) A sample data for NER can be seen in Table 1. Given a text T_i and its entities with entity types(E_k), we create four contexts serving extra-information. $Context_1$ is the entity type itself. We create a question, using simple rules, like *What are the* $[E_k]$ *mentioned in the text ?* This serves as $Context_2$. For $Context_3$, we create definition of each entity type using task description, UMLS (Bodenreider, 2004) and online sources. We add ten most frequently occurring entities across each entity type from the train dataset as the final source of information and create $Context_4$. The distribution of each of the entities across each of the dataset for Train, Dev and Test sets(both positive and negative samples) and more details about the data preparation can be found in the Supplemental Material. For those dataset where the validation data is less or not present, we used some samples from the training data of those datasets to create our validation data. Our dataset have in total 398495 training data, 148166 validation data and 502306 test data.

5 Model Description

We use different pre-trained language models on biomedical texts, BioBERT (Lee et al., 2019) and MimicBERT (Si et al., 2019), both of which are the current stat-of-the-art models for NER on multiple different datasets. We use these different variants of BERT for the token classification task. We choose the BERT base cased version of the models. We define the input to the BERT model as follows, the knowledge Context tokens C_j is prepended to the text tokens, T_i. The sequence of tokens, $\{[CLS], C_j, [SEP], T_i, [SEP]\}$ is given as input to the BERT model, and for each token we predict using a simple feed-forward layer. Figure 1 represents our model for multi-answer knowledge guided NER (KGNER).

BERT-CNN : In this model we apply a two-dimensional convolution layer on top of BERT contextual word embeddings. The convolution layer

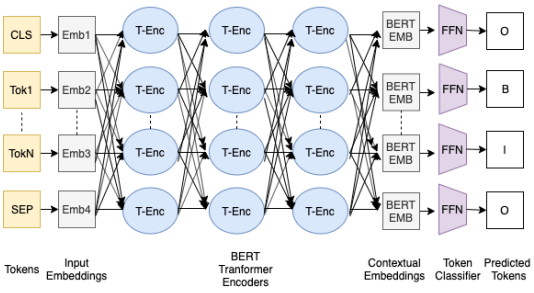

Figure 1: BERT for Multi-Answer KGNER

uses a 5x5 size kernel. The stride size is (1,2), where 1 is across sentence dimension, and 2 is across word embedding dimension. We also do circular padding. It takes input BERT token embedding and predicts the NER Tags. We choose a CNN over BERT approach similar to in (Chiu and Nichols, 2016), where they use CNNs over LSTMs.

6 Experiments

The training and validation dataset comprises of all the 18 datasets. We use a batch size of 32 and a learning rate of 3e-5. The maximum sequence length of 128/256 depends on the 99th percentile of the input token lengths. We train using 4 NVIDIA V100 16GB GPUs, with a patience of 5 epochs.

We first compare the performance of our problem formulation with other models performing NER tasks on a subset of common NER datasets. The complete results are present in Supplemental Materials.

7 Discussion and Error Analysis

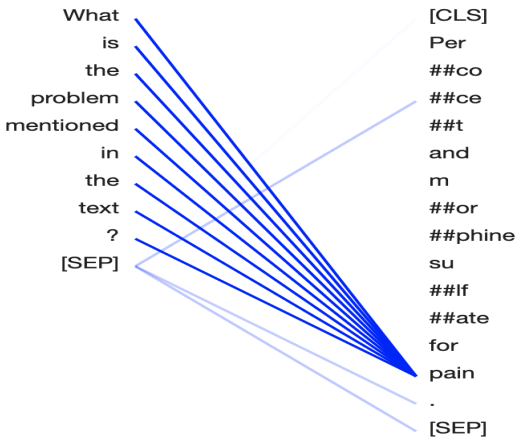

Figure 2: Question attending to the correct answer *pain* in the text with Bio-BERT model trained with $Context_2$

Datasets	Entity			Question			Definition			Example		
	P	**R**	**F**	P	R	F	P	R	F	P	R	F
BC4CHEMD	89.83	88.89	89.36	92.07	91.01	**91.54**	89.77	87.15	88.44	90.57	89.27	89.92
BC5CDR	88.25	86.13	87.18	90.09	89.16	**89.62**	88.07	86.12	87.08	88.49	86.62	87.55
CRAFT	85.26	85.57	85.41	88.07	88.19	88.13	80.72	84.19	82.42	88.00	89.18	**88.58**
ExPTM	84.64	82.67	83.65	83.71	85.73	**84.71**	74.01	83.96	78.67	83.43	82.60	83.01
NCBI Disease	86.19	88.94	88.03	86.66	90.88	**88.72**	87.75	88.91	87.81	86.59	88.68	87.62
2010 i2b2	93.42	94.13	93.77	95.27	95.91	**95.59**	92.43	93.96	93.19	93.43	94.60	94.01
2012 i2b2	76.26	82.20	79.12	81.84	84.52	**83.14**	73.11	82.30	77.44	75.63	83.96	79.58

Table 2: Precision(P), Recall(R) and F-Measure(F) for selected datasets using the best performing BERT-CNN model using different knowledge types, which are, Entity Type, Question, Definition and Examples. Best Exact Match F1 scores are in Bold.

Dataset	Models	P	R	F
BC4CHEMD	BioBERT	91.93	91.11	91.52
	MimicBERT	89.47	88.86	89.16
	SOTA	-	-	89.37
	BERTCNN	92.07	91.01	**91.54**
BC5CDR	BioBERT	89.63	88.80	89.21
	MimicBERT	88.25	86.78	87.51
	SOTA	-	-	86.23
	BERTCNN	90.09	89.16	**89.62**
CRAFT	BioBERT	86.99	87.11	87.05
	MimicBERT	86.12	84.74	85.43
	SOTA	-	-	79.55
	BERTCNN	88.07	88.19	**88.13**
ExPTM	BioBERT	85.97	85.30	**85.64**
	MimicBERT	84.09	81.34	82.69
	SOTA	-	-	74.90
	BERTCNN	83.71	85.73	84.71
NCBI-Disease	BioBERT	87.55	90.67	**89.05**
	MimicBERT	86.82	88.80	87.80
	SOTA	-	-	74.90
	BERTCNN	86.66	90.88	88.72
2010-i2b2	BioBERT	89.16	92.47	90.79
	MimicBERT	94.85	95.76	95.30
	SOTA	-	-	90.25
	BERTCNN	95.27	95.91	**95.59**
2012-i2b2	BioBERT	74.00	70.90	72.42
	MimicBERT	81.57	84.76	83.13
	SOTA	-	-	80.91
	BERTCNN	81.84	84.52	**83.14**

Table 3: Precision(P), Recall(R) and F-Measure(F) for selected datasets using multiple models. SOTA scores are from (Si et al., 2019; Lee et al., 2019; Beltagy et al., 2019). Few current SOTA scores are for BERT Large, ours though use BERT Base. i2b2 scores are compared with (Si et al., 2019).

The performance of BERT-CNN model with four different types of knowledge across the test set of seven datasets is shown in Table 2. The scores shown are entity exact match F1 scores. This shows that our problem formulation and addition of knowledge produces significant improvements. Overall, the knowledge in the form of question helps in better NER task across all datasets. This may be because the presence of "what" helps the model to find entities much better than given just a text. We analyze for a few sample, how the attention heads help the BioBERT+CNN model with question as knowledge to choose the answer using the BERT visualization tool (Vig, 2019). In Figure 2, for a text, "*Percocet and Percocet sulfate for pain.*" and knowledge "*What is the problem mentioned in the text ?*", it can be seen that each of the words of the question attends to the correct answer option "*pain*" with high confidence although there are other entities present in the text. We also observe for CRAFT, the knowledge in the form of Examples help, but definitions under perform. This indicates better knowledge formulation can improve performance even further. In Table 3 we compare our models with baseline models of BioBERT and MimicBERT trained using our task formulation, and the current state-of-the-art models. It shows our task formulation considerably improves the state-of-the-art for these tasks. We identify four reasons for improvements in accuracy. First, making the task to have no overlapping entities. Second, reducing the number of classes for prediction. Third, the sheer size of the dataset which enables the system to learn more NER specific features. Last, a CNN model over the BERT language model which enables to combine contextual features to perform better NER predictions.

8 Conclusion

We reformulated the NER task as a knowledge guided, context driven NER task and showed it has considerable promise. We attempt to solve the major challenges faced by current NER systems. Our approach has achieved above state-of-the-art F1 measures for 16 of the common biomedical NER datasets. In future, we plan to perform more experiments, such as few-shot learning between different entity groups, adding specific loss functions and logical constraints for NER tasks, and a deeper study of where our current model fails.

References

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2019. Careful selection of knowledge to solve open book question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6120–6129, Florence, Italy. Association for Computational Linguistics.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Sixth Workshop on Very Large Corpora*.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Massimiliano Ciaramita and Yasemin Altun. 2005. Named-entity recognition in novel domains with external lexical knowledge. In *Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*, volume 2005.

K Bretonnel Cohen and Lawrence Hunter. 2004. Natural language processing and systems biology. In *Artificial intelligence methods and tools for systems biology*, pages 147–173. Springer.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. Question answering is a format; when is it useful? *arXiv preprint arXiv:1909.11291*.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 643–653.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Junichi Kazama and Kentaro Torisawa. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 698–707.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Hongfang Liu, Zhang-Zhi Hu, Manabu Torii, Cathy Wu, and Carol Friedman. 2006. Quantitative assessment of dictionary-based protein named entity tagging. *Journal of the American Medical Informatics Association*, 13(5):497–507.

Tianyu Liu, Jin-Ge Yao, and Chin-Yew Lin. 2019. Towards improving neural named entity recognition with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5301–5307, Florence, Italy. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. How additional knowledge can improve natural language commonsense question answering? *arXiv preprint arXiv:1909.08855*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle,

A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *EMNLP*.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embedding. *arXiv preprint arXiv:1902.08691*.

Hye-Jeong Song, Byeong-Cheol Jo, Chan-Young Park, Jong-Dae Kim, and Yu-Seop Kim. 2018. Comparison of named entity recognition methodologies in biomedical documents. *Biomedical engineering online*, 17(2):158.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

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):427.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2018. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

9 Supplemental Materials

9.1 Model Training

We use the HuggingFace (Wolf et al., 2019) and Pytorch Deep learning framework (Paszke et al., 2019). We train the model with following hyper-parameters, learning rates in the range [1e-6,5e-5], batch sizes of [16,32,48,64], linear weight-decay in range [0.001,0.1] and warm-up steps in range of [100,1000].

9.2 Source of our Datasets

We use the following github as the source 15 publicly available datasets: https://github.com/cambridgeltl/MTL-Bioinformatics-2016 .

9.3 Entity Distribution in the Dataset

The number of samples present in the Table 4 are created directly from the train, validation and test samples present in these datasets. So there are no samples for valiidation data for last four datasets. We use some samples from training data as validation data in our created datasets.

The entity mentions represents the total number of each entities present for the datasets including train, dev and test samples. Since each sample data can have multiple entities , the number is higher than the total samples for the dataset.

9.4 Performance of Models on Test Data

The Table 5 shows the performance of each of the models on the test data for each of the datasets. The state-of-the-art for Linnaeus and AnatEM datasets uses dictionaries developed without a clear train/test split, hence our scores are not directly comparable.

Transfer Learning : To check the transfer learning ability of the model, we choose three major entity groups disease, chemical, gene or protein which is present in more than one datasets. For each entity group, We train the models with train data and validation data from a few datasets and test on a completely different dataset for the entity group. For each of the chemical, disease and gene or protein, we test on BC4CHEMD, NCBI-Disease and JNLPBA with the best performing model trained on rest of the data.

Dataset	Entity	Entity Mentions	Train +	Train -	Dev +	Dev -	Test +	Test -
AnatEM	Anatomy	13701	3514	2169	1122	959	2308	1405
BC2GM	Gene/Protein	24516	6404	6071	1283	1214	2568	2424
BC4CHEMD	Chemical	84249	14488	16002	14554	15909	12415	13738
BC5CDR	Chemical	14913	2951	1595	3017	1551	3090	1688
	Disease	12852	2658	1888	2727	1841	2842	1936
BioNLP09	Gene/Protein	14963	4711	2716	1014	433	1700	739
BioNLP11EPI	Gene/Protein	15881	3797	1896	1241	714	2836	1282
BioNLP11ID	Gene/Protein	6551	1255	1193	446	265	955	977
	Organism	3469	1120	1328	270	441	779	1153
	Chemical	973	334	2114	77	634	151	1781
	Regulon-Operon	87	9	2439	19	692	43	1889
BioNLP13CG	Gene/Protein	7908	1956	1077	393	610	1185	721
	Cell	4061	1388	1645	399	604	714	1192
	Chemical	2270	645	2388	274	729	431	1475
	Cancer	2582	908	2125	324	679	665	1241
	Organ	2517	919	2114	305	698	565	1341
	Organism	2093	827	2206	267	736	486	1420
	Tissue	587	259	2774	77	926	153	1753
	Amino Acid	135	38	2995	17	986	34	1872
	Cellular Component	569	247	2786	78	925	138	1768
	Organism Substance	283	131	2902	33	970	81	1825
	Pathological Formation	228	91	2952	35	968	73	1833
	Anatomical System	41	16	3017	3	1000	17	1889
	Immaterial Anatomical	102	47	2986	18	985	29	1877
	Organism Subdivision	98	42	2991	12	991	35	1871
	Multi-Tissue Structure	857	345	2688	114	889	236	1670
	Developing Anatomical Structure	35	13	3020	5	998	17	1889
BioNLP13GE	Gene/Protein	12031	1499	901	1655	1010	1936	1376
BioNLP13PC	Gene/Protein	10891	2153	346	723	134	1396	298
	Complex	1502	542	1957	178	679	398	1296
	Chemical	2487	596	1903	244	613	450	1244
	Cellular/ Component	1013	373	2126	144	713	263	1431
CRAFT	Gene/Protein	16108	4458	5539	1358	2105	3140	3634
	Taxonomy	6835	2511	7486	994	2469	1710	5064
	Chemical	6018	1908	8089	586	2877	1344	5430
	Cell Line	5487	2058	7939	540	2923	1257	5517
	Sequence Ontology	18856	4303	5694	1711	1752	3023	3751
	Gene Ontology	4166	1499	8498	336	3127	1344	5430
Ex-PTM	Gene/Protein	4698	857	520	279	158	1160	679
JNLPBA	DNA	10550	4670	12146	553	1218	624	3226
	RNA	1061	713	16103	89	1682	102	3748
	Cell Line	4315	2591	14225	285	1486	378	3472
	Cell Type	8584	4735	12081	415	1356	1403	2447
	Gene/Protein	35234	11840	4976	1137	634	2368	1482
Linnaeus	Species	4242	1546	9173	520	3300	1029	5381
NCBI-Disease	Disease	6871	2921	2473	489	434	538	398
2010RelationsChallege	Problem	18979	4213	4226	-	-	5802	6590
	Treatment	13809	3126	4226	-	-	7234	6590
	Test	13576	2426	4226	-	-	4591	6590
2011CoreferenceResolution	Person	17744	7207	3990	-	-	4715	2971
	Problem	18869	7003	3990	-	-	4384	2971
	Treatment	17708	5300	3990	-	-	3565	2971
	Test	13514	4191	3990	-	-	2786	2971
2012TemporalRelationsEvent	Problem	4754	2832	3597	-	-	2326	2683
	Treatment	7076	2341	4088	-	-	1976	3033
	Test	4754	1786	4643	-	-	1465	3544
	Occurance	5126	2086	4343	-	-	1677	3332
	Clinical-Department Event	1716	852	5577	-	-	655	4354
	Evidential Event	1334	706	5723	-	-	560	4449

Table 4: Data Distribution, with counts of entities, number of positive samples with at least one entity mentions, and negative samples with no target entity mention

Dataset	Models	Entity Type			Question			Definition			Example		
		P	R	F	P	R	F	P	R	F	P	R	F
AnatEM	BioBERT	89.49	87.02	88.24	89.68	88.57	89.12	88.42	88.02	88.22	90.29	89.43	89.85
	MimicBERT	86.25	84.53	85.38	86.86	85.73	86.29	85.82	82.13	83.93	87.05	86.5	86.8
	SOTA			**91.61**	-	-	-	-	-	-	-	-	-
	CNNBERT	89.43	87.54	88.47	88.81	89.12	88.96	89.03	87.23	88.13	89.41	88.69	89.05
BC2GM	BioBERT	81.63	82.37	81.99	82.47	83.36	82.91	81.3	81.75	81.52	81.82	82.28	82.05
	MimicBERT	79.56	79.57	79.56	81.22	81.4	81.31	78.53	77.88	78.21	78.44	79.55	78.99
	SOTA	-	-	81.69	-	-	-	-	-	-	-	-	-
	BERTCNN	81.79	82.06	81.93	82.89	83.39	**83.14**	80.62	80.21	80.42	82.37	82.29	82.33
BC4CHEMD	BioBERT	90.14	89.53	89.83	91.93	91.11	91.52	89.64	88.23	88.93	90.15	89.27	89.71
	MimicBERT	87.88	85.79	86.83	89.47	88.86	89.16	86.32	84.21	85.25	86.71	86.01	86.36
	SOTA	-	-	89.37	-	-	-	-	-	-	-	-	-
	CNNBERT	89.83	88.89	89.36	92.07	91.01	**91.54**	89.77	87.15	88.44	90.57	89.27	89.92
BC5CDR	BioBERT	87.55	87.27	87.41	89.63	88.8	89.21	87.66	86.3	86.97	88.4	87.48	87.94
	MimicBERT	86.29	84.85	85.56	88.25	86.78	87.51	85.54	83.97	84.75	86.31	84.35	85.31
	SOTA	-	-	86.23	-	-	-	-	-	-	-	-	-
	CNNBERT	88.25	86.13	87.18	90.09	89.16	**89.62**	88.07	86.12	87.08	88.49	86.62	87.55
BioNLP09	BioBERT	88.56	88.88	88.72	91.35	92.21	**91.78**	50.03	69.19	58.07	89.81	88.92	89.36
	MimicBERT	88.26	88.46	88.36	89.19	91.41	90.29	49.13	70.51	57.91	88.02	89.66	88.83
	SOTA	-	-	84.20	-	-	-	-	-	-	-	-	-
	CNNBERT	88.91	88.99	88.95	90.16	91.52	90.83	49.26	70.73	58.07	89.75	88.97	89.36
BioNLP11EPI	BioBERT	86.95	82.78	84.822	88.26	86.77	87.51	78.49	83.85	81.08	87.55	84.91	86.21
	MimicBERT	84.65	79.59	82.04	88.01	82.19	85	73.16	79.87	76.37	83.84	79.59	81.66
	SOTA	-	-	78.86	-	-	-	-	-	-	-	-	-
	CNNBERT	87	83.15	85.03	88.58	87.4	**87.99**	77.56	83.53	80.44	87.55	83.32	85.38
BioNLP11ID	BioBERT	81.1	81.14	81.12	86.34	85.58	85.96	83.03	81.82	82.42	82.42	83.74	83.08
	MimicBERT	79.98	78.19	79.078	83.12	81.78	82.45	77.63	77.85	77.74	77.38	78.35	77.87
	SOTA	-	-	82.26	-	-	-	-	-	-	-	-	-
	CNNBERT	81.91	83.01	82.45	86.6	85.35	**85.97**	79.43	83.17	81.26	82.85	82.2	82.52
BioNLP13CG	BioBERT	82.99	84.31	83.65	87.18	87.28	87.23	80.73	84.25	82.45	85.38	87.91	86.62
	MimicBERT	77.28	82.4	79.76	85.08	85.37	85.23	75.95	80.17	78	80.54	84.96	82.69
	SOTA	-	-	78.90	-	-	-	-	-	-	-	-	-
	CNNBERT	81.24	84.79	82.98	87.97	87.26	**87.62**	78.47	84.21	81.24	84.57	87.11	85.82
BioNLP13GE	BioBERT	78.27	83.98	81.03	82.28	86.58	**84.38**	71.25	81.65	76.1	78.77	83.55	81.09
	MimicBERT	77.41	82.95	80.08	81.61	86.28	83.88	66.66	77.72	71.77	77.04	84.42	80.56
	SOTA	-	-	78.58	-	-	-	-	-	-	-	-	-
	CNNBERT	80.86	84.72	82.747	81.82	86.26	83.98	68.64	81.32	74.44	80.26	84.48	82.32
BioNLP13PC	BioBERT	87.02	90.10	88.53	90.14	92.09	**91.11**	87.29	88.87	88.07	89.33	91.11	90.21
	MimicBERT	85.96	88.13	87.03	87.62	89.63	88.61	84.95	85.87	85.4	86.84	87.66	87.25
	SOTA	-	-	81.92	-	-	-	-	-	-	-	-	-
	CNNBERT	88.14	89.77	88.95	89.03	91.87	90.43	86.23	88.94	87.56	88.95	90.58	89.76
CRAFT	BioBERT	85.2	85.59	85.4	86.99	87.11	87.05	82.84	84.1	83.47	88.18	88.61	88.39
	MimicBERT	82.77	82.42	82.6	86.12	84.74	85.43	78.27	80.12	79.19	85.01	87.14	86.06
	SOTA	-	-	79.55	-	-	-	-	-	-	-	-	-
	CNNBERT	85.26	85.57	85.41	88.07	88.19	88.13	80.72	84.19	82.42	88.00	89.18	**88.58**
Ex-PTM	BioBERT	82.74	84.31	83.52	85.97	85.3	**85.64**	74.52	83.57	78.79	84.02	84.44	84.23
	MimicBERT	82.007	78.15	80.03	84.09	81.34	82.69	68.92	77.32	72.88	81.04	79.75	80.39
	SOTA	-	-	74.90	-	-	-	-	-	-	-	-	-
	CNNBERT	84.64	82.67	83.65	83.718	85.738	84.71	74.012	83.96	78.67	83.43	82.6	83.01
JNLPBA	BioBERT	70.12	77.88	73.8	76.127	82.15	**79.02**	66.11	70.33	68.16	69.14	79.54	73.98
	MimicBERT	68.63	76.78	72.48	74.97	80.79	77.77	65.42	71.41	68.28	67.17	78.06	72.21
	SOTA	-	-	78.58	-	-	-	-	-	-	-	-	-
	CNNBERT	69.78	77.77	73.55	76.04	81.632	78.73	64.45	73.38	68.63	69.29	79.25	73.94
Linnaeus	BioBERT	87	87.49	87.24	88.34	88.4	88.37	86.8	86.94	86.88	90.32	89.88	90.10
	MimicBERT	80.27	83.48	81.84	86.31	85.1	85.7	81.41	82.56	81.99	84.8	85.04	84.92
	SOTA	-	-	**95.68**	-	-	-	-	-	-	-	-	-
	CNNBERT	87.97	87.42	87.69	88.47	88.47	88.47	86.33	86.52	86.429	88.48	89.06	88.77
NCBI-Disease	BioBERT	86.92	89.53	88.2	87.5	90.67	**89.05**	85.75	88.6	87.15	86.32	89.16	87.72
	MimicBERT	84.22	86.83	85.51	86.82	88.8	87.80	85.69	85.69	85.69	84.43	85	84.72
	SOTA	-	-	88.60	-	-	-	-	-	-	-	-	-
	CNNBERT	86.19	89.94	88.03	86.66	90.88	88.72	86.75	88.91	87.81	86.59	88.68	87.62
2010 Relations	BioBERT	93.41	94.12	93.76	89.16	92.47	90.79	93.21	94.09	93.65	93.29	94.41	93.84
	MimicBERT	93.09	94.38	93.73	94.85	95.76	95.3	92.69	94.07	93.38	93.1	94.19	93.64
	SOTA	-	-	90.25	-	-	-	-	-	-	-	-	-
	CNNBERT	93.42	94.13	93.77	95.27	95.91	**95.59**	92.43	93.96	93.19	93.43	94.6	94.01
2011 Coreference	BioBERT	93.14	92.2	92.67	93.88	94.15	94.02	92.93	92.12	92.52	93.89	93.36	93.62
	MimicBERT	93.028	92.54	92.78	94.18	94.3	94.24	92.69	92.14	92.42	93.77	93.25	93.5
	SOTA	-	-	-	-	-	-	-	-	-	-	-	-
	CNNBERT	93.1	92.04	92.57	94.42	94.37	**94.40**	92.32	92.09	92.21	94.23	93.67	93.95
2012 Temporal	BioBERT	73.61	81.76	77.47	74	70.9	72.42	73.47	81.97	77.49	73.53	83.21	78.07
	MimicBERT	73.89	82.27	77.86	81.57	84.76	83.13	71.01	82.27	76.22	72.44	83.74	77.68
	SOTA	-	-	80.91	-	-	-	-	-	-	-	-	-
	CNNBERT	76.26	82.2	79.12	81.84	84.52	**83.14**	73.11	82.3	77.44	75.63	83.96	79.58

Table 5: Precision(P), Recall(R) and F-Measure(F) for all the mentioned datasets using BioBERT and Mimic-TrainedBERT for Context_1 (Entity Name), Context_2 (Question), Context_3 (Definition) and Context_4 (Examples). Bold represents state-of-the-art.