Knowledge Guided Named Entity Recognition

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 try to perform Named Entity Recognition (NER) with external knowledge. 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. Moreover, the formulation of the task as a MAQA task, helps reducing other errors. 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 entity type mentioned in the knowledge, improving performance, (c) reduces confusion in systems by reducing the classes to be predicted to be limited to only three (B,I,O), (d) Makes detection of Nested Entities easier. We perform extensive experiments of this Knowledge Guided NER (KGNER) formulation on 15 Biomedical NER datasets, and through these experiments we see external knowledge helps. We will release the code for dataset conversion and our trained models for replicating experiments.

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.

Text: Mice lacking the **myotonic dystrophy** protein kinase develop a late onset progressive **myopathy**.
Entity : Disease **Question** : What are the diseases mentioned in the text? **Definition** : In medicine, a health problem with certain characteristics or symptoms **Examples** : Diabetes, Malaria, Measles.
Entities : myotonic dystrophy, myopathy **Entity Span** : (3, 4), (12, 12)

Table 1: An example of different kind of knowledge for a given entity type.

In this work, we focus on Named Entity Recognition (NER) for biomedical texts. NER is one of the core Natural language processing tasks, in which given a text, systems identify entities such as *Person*, *Organization*, *Location*, etc. In biomedical domain, we need to identify different entities, such as *Disease*, *Treatment*, *Test*, *Chemical*, *Gene*, etc. In biomedical domain, the entities sometimes differ subtly, and hence require even more precise knowledge. We incorporate such required knowledge through sentences or words as shown in Table 1.

Most of the NER systems, formulate the problem as a classification task. Given the token T_i, it is classified to be one of the three tags B-E_k, I-E_k, O in the BIO-Tagging scheme (Begin-Intermediate-Other), where $k = 1..K$, K is the number of entities and E is the type of entity. The performance of the problem 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 regarding 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 try to address these challenges through our following contributions:

- We address challenge (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, as we predict only one type of entities at a time. This generic format, allows us to jointly learn from multiple different datasets.
- We address challenge (b) by providing various types of external knowledge, and do an empirical study of which knowledge types are better.

2 Related Work

2.1 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, like OpenBookQA (Mihaylov et al., 2018) and SocialIQA (Sap et al., 2019) in works like (Mitra et al., 2019; Banerjee et al., 2019). In our work, we incorporate simple textual knowledge sentences, similar to the attempts done for incorporating knowledge in question-answering tasks.

2.2 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.

2.3 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 these works, which also use pre-trained language models (BERT), and/or predicts different types of entities separately, but differs in task formulation and use of explicit external knowledge.

2.4 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 try 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* or E_k (b) separate *question* created using each entity type or Q (c) *definition* of each entity type or D (d) *definition with example* or $D \cup Eg$. For the example mentioned in Table 1, E_k is *Disease*, Q is *What are the diseases mentioned in the text?*, D is the definition text, and Eg are the examples *Diabetes, Malaria, Measles*.

In the original task formulations, 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, "Mice: O, lacking: O, myotonic: B_{E_k}, dystrophy: I_{E_k} ..".

We reformulate the task, to classify each token T_i only to three classes, B_{Ans}, I_{Ans} and O. Those tokens which should answer the query using the given knowledge, should be classified as B_{Ans} or

I_{Ans} depending on if they are the first token of the answer or the intermediate tokens. All other tokens are to predicted O.

4 Dataset Preparation

We created the dataset using 15 publicly available biomedical datasets as mentioned in the Table 2. The types of entities can be seen in Table 2. A sample of data in this dataset can be seen in Table 1. Given a text T_i and its entities along with entity types E_k provided as labels, we create four contexts. $Context_1$ is the entity type itself. We create $Context_2$ that is the question, using simple rules, like *What are the* $[E_k]$ *mentioned in the text ?* We add the definition of the entity type from UMLS(Bodenreider, 2004) and Wikipedia to create $Context_3$.

The distribution of each of the entities across each of the datasets for Train, Dev and Test sets can be found in the Table 2. The counts of positive and negative samples created from the original datasets can also be found from the distribution table. 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.

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 state 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).

6 Experiments

The training and validation dataset comprises of all the 18 datasets. The total train samples are around, 430K and validation samples are 120K. We use a batch size of 32 and a learning rate of 3e-5. The maximum sequence length of 128/256

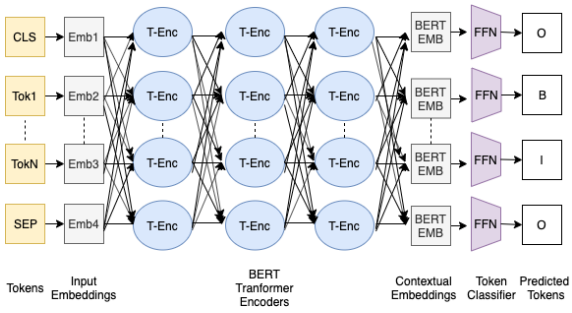

Figure 1: BERT for Multi-Answer KGNER

depends on the 99th percentile of the input token lengths. We train using 4 NVIDIA V100 16GB GPUs, with a patience of 5 epochs. The training usually lasts for 14-15 epochs.

Table 3 are our Test set results for the different datasets.

7 Error Analysis and Discussion

Our preliminary results are present in Table 3. The scores shown are entity exact match F1 scores. These show that our approach produces significant improvements. Even a simple external knowledge of entity name provides significant improvements to baselines. Moreover the ability to use such a large and varied corpus leads to further improvements. The F1-Measure drops for definitions considerably for few datasets, showing the performance is tightly coupled with the preciseness of the knowledge provided. For example, for AnatEM dataset, the knowledge definition is vague, leading to a large drop in performance. The SOTA for Linnaeus and AnatEM datasets uses dictionaries developed without a clear train/test split, hence our scores are not directly comparable.

8 Conclusion and Future Work

We reformulated the NER task as a knowledge guided, context driven NER task and showed it has considerable promise. We have tried to solve the major challenges faced by current NER systems. Our approach has achieved above state of the art F1 measures for some of the Biomedical NER datasets.

In future, we plan to perform more experiments, such as transfer learning and few-shot learning between different entity groups, adding specific loss functions and constraints for NER tasks, and a deeper study of where our current model fails.

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		2951	1595	3017	1551	3090	1688
	Disease		2658	1888	2727	1841	3090	1688
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

Table 2: Data Distribution, with counts of entities, number of positive samples with atleast 1 entity mentions, and negative samples with no target entity mention

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*.

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.

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*

Dataset	Models	Concept$_1$			Concept$_2$			Concept$_3$		
		P	R	F	P	R	F	P	R	F
AnatEM	BioBERT	88.32	87.30	87.81	89.54	88.49	*89.01*	78.09	18.18	29.5
	MimicBERT	85.28	85.38	85.33	86.74	85.64	86.19	72.95	13.88	23.32
	SOTA	-	-	**91.61**	-	-	-	-	-	-
BC2GM	BioBERT	81.56	81.73	81.64	82.46	82.71	**82.59**	72.01	68.78	70.36
	MimicBERT	79.39	78.92	79.15	80.57	81.01	80.79	67.9	64.1	65.94
	SOTA	-	-	81.69	-	-	-	-	-	-
BC4CHEMD	BioBERT	90.27	88.78	89.52	91.77	90.92	**91.34**	88.18	89.11	88.64
	MimicBERT	87.32	85.31	86.3	89.51	88.65	89.07	85.69	84.43	85.05
	SOTA	-	-	89.37	-	-	-	-	-	-
BC5CDR	BioBERT	87.87	86.56	87.21	90.5	88.92	**89.70**	87.21	87.33	87.27
	MimicBERT	85.78	85.72	85.75	88.68	87.35	88.01	85.25	84.27	84.76
	SOTA	-	-	86.23	-	-	-	-	-	-
BioNLP09	BioBERT	88.92	89.02	88.97	90.11	91.69	**90.89**	50.63	71.71	59.35
	MimicBERT	88.27	88.1	88.19	90.01	91.66	90.83	50.53	70.85	58.99
	SOTA	-	-	84.20	-	-	-	-	-	-
BioNLP11EPI	BioBERT	86.97	84.14	85.53	88.59	87.11	**87.84**	72.67	83.44	77.69
	MimicBERT	85.14	81.19	83.12	86.78	83.51	85.12	70.43	79.85	74.84
	SOTA	-	-	78.86	-	-	-	-	-	-
BioNLP11ID	BioBERT	84.49	83.33	83.90	86.27	84.88	**85.57**	78.87	80.66	79.75
	MimicBERT	80.34	77.61	78.96	83.17	82.29	82.73	78.28	75.09	76.65
	SOTA	-	-	82.26	-	-	-	-	-	-
BioNLP13CG	BioBERT	87.42	82.36	84.81	89.52	87.22	**88.36**	71.93	52.15	60.46
	MimicBERT	85.47	80.8	83.07	87.12	84.64	85.87	68.16	50.27	57.87
	SOTA	-	-	78.90	-	-	-	-	-	-
BioNLP13GE	BioBERT	77.29	85.66	81.26	82.81	89.02	**85.81**	68.75	83.23	75.30
	MimicBERT	75.63	82.36	78.85	81.89	87.83	84.76	65.33	80.91	72.29
	SOTA	-	-	78.58	-	-	-	-	-	-
BioNLP13PC	BioBERT	88.37	88.3	88.34	89.53	91.46	**90.49**	78.97	83.29	81.07
	MimicBERT	86.69	87.24	86.97	87.79	89.55	88.66	77.74	80.68	79.19
	SOTA	-	-	81.92	-	-	-	-	-	-
CRAFT	BioBERT	85.56	85.06	85.31	88.09	87.61	**87.85**	80.16	79.99	80.08
	MimicBERT	85.01	80.62	82.76	87.27	84.93	86.08	78.24	76.6	77.41
	SOTA	-	-	79.55	-	-	-	-	-	-
Ex-PTM	BioBERT	83.78	84.58	84.18	84.74	85.87	**85.30**	69.77	84.67	76.50
	MimicBERT	80.48	78.72	79.59	82.47	81.69	82.08	67.87	80.06	73.46
	SOTA	-	-	74.90	-	-	-	-	-	-
JNLPBA	BioBERT	71.2	77.38	74.16	77.34	81.06	**79.16**	65.44	70.16	67.72
	MimicBERT	69.89	75.54	72.60	75.76	80.05	77.85	63.43	67.36	65.34
	SOTA	-	-	78.58	-	-	-	-	-	-
Linnaeus	BioBERT	89.54	87.21	88.36	89.37	87.42	*88.38*	86.98	85.51	86.24
	MimicBERT	87.49	83.56	85.48	86.14	85.59	85.87	88.61	82.79	85.6
	SOTA	-	-	**95.68**	-	-	-	-	-	-
NCBI-Disease	BioBERT	86.04	89.43	87.70	88.37	91.39	**89.86**	84.85	91.09	87.86
	MimicBERT	83.91	84.87	84.39	86.73	88.08	87.40	83.98	83.63	83.8
	SOTA	-	-	88.60	-	-	-	-	-	-

Table 3: Precision(P), Recall(R) and F-Measure(F) for all the mentioned datasets using BioBERT and MimicTrainedBERT for Context$_1$ (Entity Name), Context$_2$ (Question) and Context$_3$ (Definition). Bold represents state of the art.

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*.

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 elec-

tricity? a new dataset for open book question answering. In *EMNLP*.

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

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*.

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.

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.

A Appendices

B Supplemental Material