# Bio-Medical Named Entity Recognition via Knowledge Guidance and Question Answering

PRATYAY BANERJEE*, KUNTAL KUMAR PAL*, MURTHY DEVARAKONDA, and CHITTA BARAL,

Department of Computer Science, Arizona State University, U.S.A

In this work, we formulated the named entity recognition (NER) task as a multi-answer knowledge guided question-answer task (KGQA) and showed that the knowledge guidance helps to achieve state-of-the-art results for 12 out of 18 biomedical NER datasets. We prepended four different knowledge contexts – namely, entity types, questions, definitions, and examples – to the input text and trained and tested BERT-based neural models on such input sequences from a combined dataset of the 18 different datasets. This novel formulation of the task (a) improved named entity recognition by relating words having high similarity to the knowledge provided through the attention mechanism, (b) reduced system confusion by limiting prediction to a single class for each input (i.e. B, I, O only), (c) made detection of nested entities easier (d) enabled the models to jointly learn NER specific features from a large number of datasets. We performed extensive experiments of this KGQA formulation on the biomedical datasets, and through the experiments we showed how knowledge improved named entity recognition.

CCS Concepts: • **Computing methodologies** → **Information extraction**; Supervised learning by classification; • **Information systems** → *Question answering*.

Additional Key Words and Phrases: named entity recognition, NER, Question Answering, text tagging, BIO Tagging, Multitask training, BERT-CNN, biomedical, transfer learning

**ACM Reference Format:**

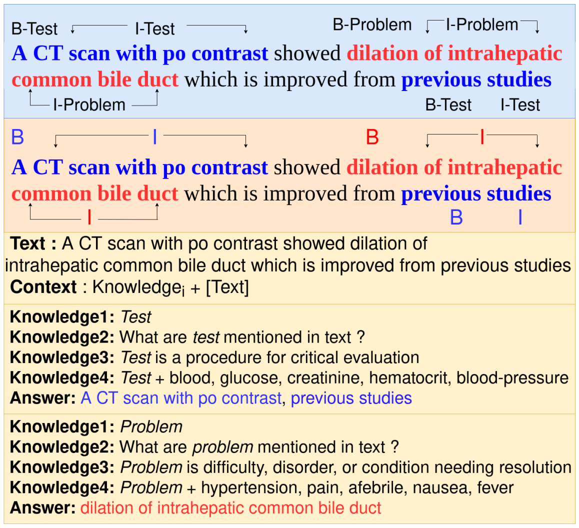
Pratyay Banerjee, Kuntal Kumar Pal, Murthy Devarakonda, and Chitta Baral. 2018. Bio-Medical Named Entity Recognition via Knowledge Guidance and Question Answering. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/1122445.1122456

## 1 INTRODUCTION

Named Entity Recognition (NER) has been considered as a relatively difficult task in biomedical domain due to the stylized writing and domain-specific terminology. Moreover, the target entities are usually proper nouns or unregistered words, with new words for drugs, diseases, and chemicals being generated frequently. Also, the same word phrases can be recognized as different named entities in terms of current context [11, 26, 39]. For these reasons, external knowledge can be helpful to guide automated systems to identify the entities in biomedical domain.

In general domain, use of extensive external knowledge has helped systems in multiple natural language tasks like commonsense question answering [35, 41] and science question answering [23, 29]. In biomedical named entity recognition, external knowledge can be about entities and their relations. Knowledge like entity-types along with their definition and examples can allow attention mechanism to compare and learn to detect new entities, especially if the

*Both authors contributed equally to this research.


B-Test
I-Test
B-Problem
I-Problem
A CT scan with po contrast showed dilation of intrahepatic common bile duct which is improved from previous studies
I-Problem
B-Test
I-Test
B
I
B
I
A CT scan with po contrast showed dilation of intrahepatic common bile duct which is improved from previous studies
I
B
I

**Text :** A CT scan with po contrast showed dilation of intrahepatic common bile duct which is improved from previous studies
**Context** : Knowledge$_i$ + [Text]

**Knowledge1:** *Test*
**Knowledge2:** What are *test* mentioned in text ?
**Knowledge3:** *Test* is a procedure for critical evaluation
**Knowledge4:** *Test* + blood, glucose, creatinine, hematocrit, blood-pressure
**Answer:** A CT scan with po contrast, previous studies

**Knowledge1:** *Problem*
**Knowledge2:** What are *problem* mentioned in text ?
**Knowledge3:** *Problem* is difficulty, disorder, or condition needing resolution
**Knowledge4:** *Problem* + hypertension, pain, afebrile, nausea, fever
**Answer:** dilation of intrahepatic common bile duct

Fig. 1. The top block shows the traditional way of NER. In our method, we predict only B, I and O tags for a given context, i.e, only red tags are predicted if the context is about entity Problem. O tags are not shown, but are predicted for non-answer words.

entity is newly generated and has infrequent mentions in common biomedical texts. Apart from the use of external knowledge, framing an NLP task as a question answering task can lead to better performance [28]. Motivated by this approach, we hypothesize that a Knowledge guided QA framework may be helpful in biomedical NER.

In this paper, we focus on NER in biomedical text and we test our hypothesis using different kinds of knowledge. The ways of expressing knowledge include asking a question about the entity, giving the entity type, providing a definition of the entity type and mentioning some examples of the entity types, as seen in Figure 1.

Figure 1 also shows, how traditional NER systems formulate the problem as a classification task. This traditional task formulation leads to the following challenges: (a) *labelling error*, i.e., even though a system is able to identify the location of an entity correctly, it fails to predict the correct type; (b) inability to leverage more information for a particular entity type, since the conventional task formulation only allows 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.

We can avoid challenges (a) and (c) by modeling the task as multi-answer extraction task, where we predict only one type of entity at a time, given a context determining which entity is being extracted at the current time. This formulation allows us to avert the issue of nested named entities and helps us to jointly learn from multiple biomedical datasets having similar entities. We specifically address the challenge (b) by providing four different types of knowledge context as shown in Figure 1. We perform an empirical study of which knowledge type has the most significant impact in a NER task.

Our task formulation enables us to create a considerably large dataset with knowledge context utilizing 18 biomedical datasets. The goal is to learn jointly from multiple domains containing different target entities. We also propose a new NER model over BERT using a re-contextualization layer called BERT-CNN. This layer uses token-local features to recompute token encoding to enable the model to better understand the start and end locations of an entity. We use the BERT-base model and show our task formulation and model performs better than a strong baseline of a BERT-large model pre-trained on medical corpus, and finetuned using the traditional NER task. We perform extensive experiments to analyze the impact of each of our contributions. We also study the transfer learning ability of our knowledge guided BERT-CNN model, as one of the major challenges currently faced by the biomedical community is the poor ability of the models to transfer in real-life applications.

To summarize our contributions:

- We reformulate the task of named entity recognition as a multi-answer question answering task using knowledge as a context.
- We make available a significantly large, cleaned and pre-processed dataset with knowledge context utilizing 18 biomedical datasets having in total 398495 training, 148166 validation and 502306 test samples.
- We propose a BIO tagging based model for the knowledge guided named entity recognition task, with a re-contextualization layer.
- We perform extensive experiments to evaluate our models, including the ability of the model to adapt to new domains.
- Finally, all our contributions together further push the state-of-the-art exact match F1 scores by 1.5-11% for 12 publicly available biomedical NER datasets.

## 2 OUR APPROACH

### 2.1 Task Formulation

Traditional systems define named entity recognition as a multi-class classification task. Given a context $C = \{c_1, c_2, ..., c_n\}$, any token $c_i$ is classified as one of the three tags $B\text{-}e_k$, $I\text{-}e_k$, $O$ in the BIO-Tagging scheme, where $e_k \in E$ (the set of entity types for a dataset). This formulation leads to *labelling error.* A token $c_i$ is classified as $B\text{-}e_k$ or $I\text{-}e_k$ when the token is actually a $B\text{-}e_j$ or $I\text{-}e_j$ where $j \neq k$. This means that even though a system was able to identify the location of an entity correctly, it fails to identify the correct type.

In our approach, we tackle this issue by formulating the NER task in the following way. Given a context $C = \{c_1, c_2, ..., c_n\}$, any token $c_i$ is classified as $B$, $I$ and $O$. To identify which entity, type the token belongs to, we provide external knowledge $K$ to the context which in turn contains the entity type information. For example, if we want to extract two entities $e_1$ and $e_2$ from context $C$, we first provide $K_{e_1}$ and $C$ as input to our model to extract $e_1$ entities, then provide $K_{e_2}$ and $C$ as input to extract $e_2$ entities. This formulation decouples the classification and the entity location tasks, enabling the model to learn from multiple datasets and overcoming *labelling error.*

### 2.2 Knowledge Context Generation

We experiment with five types of knowledge context ($K$) to identify entities and their types. These are: (a) *Entity types* ($e_k \in E$) (b) separate *Question* ($Q_k$) created using each entity type, (c) *Definition* ($D_k$) of each entity type along with the entity type itself, (d) *Examples* ($Eg_k$) along with entity type and (e) *All* of the above. If there are entities of $n$ entity types in a text, during training we create a set of five knowledge context for each different entity type. Since the approach works one entity type at a time, we make sure that the entity type is mentioned in each of the five contexts. During inference, only the best knowledge context is used, i.e, if *Question* performs best for a dataset, we use only that context. For the example mentioned in Figure 1, $E$ is {"*Problem*", "*Test*"} , $Q$ is {"*What are problem mentioned in text?*", "*What are test mentioned in text?*"}, $D$ is the definition text, {"*Problem is a difficulty, disorder, or condition needing resolution*", "*Test is* *a procedure for critical evaluation*"}, and $Eg$ are the examples {"*hypertension, pain, afebrile, nausea, fever*", "*blood, glucose, creatinine, hematocrit, blood-pressure*"}.

### 2.3 Datasets

We create the dataset for NER using fifteen publicly available biomedical datasets[1][12] and three datasets from previous i2b2 challenges [40, 42–44]. One of the samples is shown in Figure 1. Our task formulation enables us to combine the datasets and a create a significantly large dataset, that enables deep neural model learning. Moreover, the multi-task learning for different entity types enables the model to generalize better.

**Bionlp Shared Task and Workshop**: Six of the datasets Bionlp09 [20], Bionlp11ID [31], Bionlp11EPI [31], Bionlp13PC [30], Bionlp13CG [30], Bionlp13GE [30] are from the Biomedical Natural Language Processing Workshops. Some of the basic entities of these datasets are gene or gene products, protein, chemicals and organisms.

**i2b2 Shared Task and Workshop**: We use three datasets from i2b2 shared task and Workshop Challenges in Natural Language Processing for Clinical Data. We only use training and testing data from 2010 Relations Challenge [44], 2011 Coreference Challenge [42] and 2012 Temporal Relations Challenge [40]. These datasets primarily contain entities like problems, tests and treatments.

**Bio-Creative Challenge and Workshop**: These workshops provide datasets for information extraction task in biological domain. We only use three datasets namely BC4CHEMD (Chemical) [22], BC5CDR (Chemical and disease) [48] and BC2GM (gene or protein) [38]. We consider these datasets since they are similar to biomedical texts and can be augmented to be trained together to generalize on extraction of some of the entities.

**Others**: Apart from these 12 datasets we also include CRAFT [3], AnatEM [33], Linnaeus [16], JNLPBA [21], Ex-PTM [34] and NCBI-Disease [14] to increase our training and evaluation set. They include entities such as anatomy, species, diseases, cell-line, DNA, RNA, gene or protein and chemicals.

### 2.4 Rule-based Template Creation

We use the following rules to create contexts for each knowledge type.

**Entity:** The first and the simplest context, is the Entity type name itself.

**Question:** We create a knowledge context Question ($Q_k$), using simple rules, like:

$Q_k$ = What are the $[e_k]$ mentioned in the text ?

**Definition:** To get knowledge context ($D_k$), we find the corresponding scientific definition of each entity type from UMLS Meta-thesaurus by considering the entity type as concept [6]. Other sources include challenge dataset definitions and online resources.

$D_k \in$ {UMLS | Challenge | Online Resource}

**Examples:** In order to determine representative examples of an entity type, we find the top ten most frequent entities for each type from the entire training dataset. We concatenate these ten entities as the final knowledge context ($Eg_k$) and prepend this in front of the text.

**All:** is just the concatenation of all of the above knowledge context.

The motivation behind using the entity type and definition as knowledge context is that the neural model can leverage the information present in the knowledge to attend to correct entities. If we use question as a context then the

[1] https://github.com/cambridgeltl/MTL-Bioinformatics-2016

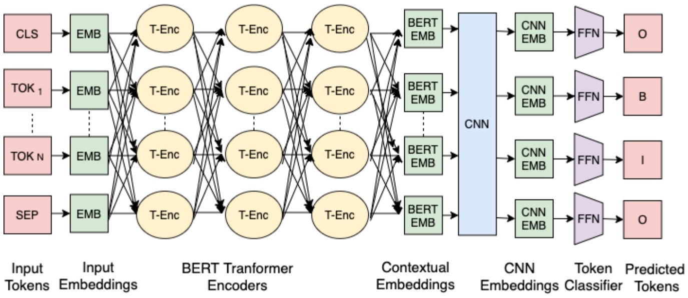


Fig. 2. BERT-CNN for Multi-Answer KGQA

task becomes a multi-answer question-answering task. We use examples as knowledge with the hypothesis that our model will be able to choose the entities that can belong to same categories as the examples.

The distribution of each of the entities across each of the dataset for Training, Validation and Test splits (both positive and negative samples) and more details about the dataset preparation can be found in the Appendix.

We treat each individual sentence in a medical document or paragraph as an individual sample. If a sentence has an entity corresponding to a context, we consider that as a positive sample for that context. Similarly, we treat a sentence that does not have an entity for a corresponding context as a negative sample for that context. Although these sentences can contain entities for other entity types. Since many datasets do not provide a validation split, we randomly sample from the train split to create our validation data. Overall, our dataset has 398495 train, 148166 dev and 502306 test samples.

## 3 MODEL DESCRIPTION

### 3.1 Knowledge Guided NER

We choose the BERT-base cased version [13] as our base model. In our approach, given a text $C$, we create a knowledge context $K_i$ for each context type. We need to find the spans of entities $S_{start}$ and $S_{end}$. So, we define the input to the BERT model as follows, the knowledge context tokens $K_i = \{k_{ij}\}$ are prepended to the text tokens, $C = \{c_j\}$. The sequence of tokens, $\{[CLS], k_{i1}, ..k_{im}, [SEP], c_1, ..c_n, [SEP]\}$ is given as input to the BERT model where $m$ is the size of knowledge context $K_i$ and $n$ is the size of text $C$. In our baseline model, for each token we predict $B$, $I$ and $O$ using a feed-forward layer.

### 3.2 Re-contextualization

We modify the BERT-base model by adding a re-contextualization layer consisting of a two-dimensional convolution layer. The purpose of this layer is to leverage information from adjacent or token-local embeddings and help in better start and end prediction of the entities. As BERT uses multiple layers of attention which jointly focuses on all the tokens, we add this CNN layers with a window of $W < 5$ to focus on nearby tokens only. We take the outputs of the CNN layer and feed it to the final feed-forward layer to predict the tags. Figure 2 represents the end-to-end architecture of our BERT-CNN model.

| | | ANATEM | | | BC2GM | | | BC4CHEMD | | | BC5CDR | | | BIONLP09 | | | BIONLP11EPI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BASE | **BioBERT** | 89.63 | 89.32 | 89.47 | 82.45 | 83.83 | 83.13 | 90.97 | 89.59 | 90.27 | 87.19 | 90.59 | 88.84 | 89.65 | 88.60 | 89.13 | 85.23 | 85.60 | 85.41 |
| | **MimicBERT** | 86.23 | 85.72 | 85.98 | 79.04 | 80.32 | 79.68 | 88.77 | 85.41 | 87.06 | 84.01 | 86.86 | 85.39 | 87.71 | 84.15 | 85.89 | 79.37 | 77.78 | 78.57 |
| | **BERT-MRC** | 72.24 | 75.09 | 73.64 | 73.80 | 74.59 | 74.19 | 86.47 | 85.52 | 85.99 | 71.22 | 73.68 | 72.43 | 74.62 | 70.69 | 72.60 | 77.81 | 67.01 | 72.01 |
| | **SOTA** | - | - | **91.61*** | - | - | 81.69* | - | - | **92.36** | - | - | **90.01** | - | - | 84.20* | - | - | 78.86* |
| OURS | **BioBERT** | 90.29 | 89.43 | <u>89.85</u> | 82.47 | 83.36 | 82.91 | 91.93 | 91.11 | 91.52 | 89.63 | 88.80 | 89.21 | 91.35 | 92.21 | 91.78 | 88.26 | 86.77 | 87.51 |
| | **MimicBERT** | 87.05 | 86.50 | 86.80 | 81.22 | 81.40 | 81.31 | 89.47 | 88.86 | 89.16 | 88.25 | 86.78 | 87.51 | 89.19 | 91.41 | 90.29 | 88.01 | 82.19 | 85.00 |
| | **BERT-CNN** | 89.78 | 89.24 | 89.51 | 82.89 | 83.39 | **83.14**† | 92.56 | 91.10 | <u>91.82</u> | 90.09 | 89.16 | <u>89.62</u> | 91.55 | 92.95 | **92.25**† | 88.58 | 87.40 | **87.99**† |

| | | BIONLP11ID | | | BIONLP13CG | | | BIONLP13GE | | | BIONLP13PC | | | CRAFT | | | EXPTM | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BASE | **BioBERT** | 84.23 | 85.77 | 84.70 | 84.82 | 86.42 | 85.56 | 72.92 | 85.42 | 78.68 | 87.64 | 90.56 | 89.06 | 84.92 | 86.56 | 85.70 | 76.12 | 79.81 | 77.92 |
| | **MimicBERT** | 83.93 | 81.84 | 82.35 | 77.52 | 80.32 | 78.71 | 64.72 | 65.53 | 65.12 | 81.59 | 85.45 | 83.43 | 81.27 | 79.08 | 80.04 | 66.74 | 67.29 | 67.01 |
| | **BERT-MRC** | 80.25 | 73.26 | 76.60 | 74.57 | 69.25 | 71.81 | 77.79 | 75.54 | 76.65 | 76.51 | 76.95 | 76.73 | 75.79 | 72.25 | 73.98 | 76.61 | 76.93 | 76.77 |
| | **SOTA** | - | - | 81.73* | - | - | 78.90* | - | - | 78.58* | - | - | 81.92* | - | - | 79.56* | - | - | 74.90* |
| OURS | **BioBERT** | 86.34 | 85.58 | 85.96 | 87.18 | 87.28 | 87.23 | 82.28 | 86.58 | <u>84.38</u> | 90.14 | 92.09 | **91.11**† | 88.18 | 88.61 | 88.39 | 85.97 | 85.30 | **85.64**† |
| | **MimicBERT** | 83.12 | 81.78 | 82.45 | 85.08 | 85.37 | 85.23 | 81.61 | 86.28 | 83.88 | 87.62 | 89.63 | 88.61 | 85.01 | 87.14 | 86.06 | 84.09 | 81.34 | 82.69 |
| | **BERT-CNN** | 87.98 | 84.64 | **86.27**† | 90.62 | 88.56 | **89.58**† | 83.77 | 88.01 | **85.84**† | 89.03 | 91.87 | 90.43 | 90.54 | 89.19 | **89.86**† | 85.08 | 84.79 | 84.94 |

| | | JNLPBA | | | LINNAEUS | | | NCBIDISEASE | | | 2010-i2b2 | | | 2011-i2b2 | | | 2012-i2b2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BASE | **BioBERT** | 69.96 | 78.19 | 73.63 | 92.30 | 86.42 | 89.27 | 86.67 | 89.38 | 88.00 | 85.32 | 83.23 | 84.26 | 91.24 | 90.32 | 90.78 | 79.31 | 75.89 | 77.56 |
| | **MimicBERT** | 67.99 | 76.32 | 71.66 | 91.69 | 81.81 | 86.46 | 84.04 | 88.23 | 86.08 | 90.37 | 88.29 | 89.32 | 92.83 | 91.22 | 92.02 | 79.78 | 81.01 | 80.39 |
| | **BERT-MRC** | 70.52 | 69.38 | 69.95 | 74.13 | 73.56 | 73.84 | 77.25 | 73.23 | 75.19 | 75.32 | 73.23 | 74.26 | 81.24 | 80.32 | 80.78 | 69.31 | 65.89 | 67.56 |
| | **SOTA** | - | - | 78.58* | - | - | **95.68*** | - | - | 89.36 | - | - | 90.25# | - | - | - | - | - | 80.91# |
| OURS | **BioBERT** | 76.12 | 82.15 | 79.02 | 90.32 | 89.88 | 90.10 | 87.50 | 90.67 | 89.05 | 93.29 | 94.41 | 93.84 | 93.88 | 94.15 | 94.02 | 73.53 | 83.21 | 78.07 |
| | **MimicBERT** | 74.97 | 80.79 | 77.77 | 86.31 | 85.10 | 85.70 | 86.82 | 88.80 | 87.80 | 94.85 | 95.76 | 95.30 | 94.18 | 94.30 | 94.24 | 81.57 | 84.76 | 83.13 |
| | **BERT-CNN** | 76.85 | 81.79 | **79.24** | 90.69 | 90.53 | <u>90.61</u> | 87.89 | 91.56 | **89.69** | 95.27 | 95.91 | **95.59**† | 94.70 | 94.94 | **94.82** | 84.83 | 85.25 | **85.04**† |

Table 1. Precision, Recall and F-Measure (in order) for 18 datasets compared with multiple models. * tagged scores are non-BERT systems, # BERT-Large and rest are BERT-Base systems. Our models use knowledge type All or Question, whichever is observed best on validation accuracy. Best F1-scores are in bold. Underlined are our best scores where our models are not SOTA. † tagged scores are statistically significantly better than SOTA ($p \leq 0.05$ based on Wilson score intervals [49]). Dataset statistics are in the Appendix.

### 3.3 Training and Testing

During training, the context, $X$ (combination of knowledge, $K_i$ and given text, $C$) has gold annotations ($y_i$) of $B$, $I$ and $O$ for each token ($x_i$). We calculate cross-entropy loss for each token $x_i$ as:

$$L_{token} = -\sum_{c=1}^{M} y_{x_i,c} log(P_{x_i,c})$$

where $M$ is the total number of classes (B, I, O), $y_{x_i,c}$ is a binary indicator whether the label $c$ is the correct classification of token $x_i$, $P_{x_i,c}$ is the predicted probability of $x_i$ belonging to class $c$.

The model is trained end-to-end with the above loss. During inference we consider the tokens $x_i$ present only in the text $C$. Text chunks that start from label $B$ and continue till last $I$ tag are predicted as entities. For each entity type we feed the text with a separate context and text input. We train our model jointly on a processed combined dataset of 18 common biomedical datasets and compare the performance when trained individually.

## 4 EXPERIMENTS

### 4.1 Experimental Setup and Training Parameters

We use a batch size of 32 and a learning rate of 5e-5 for all our experiments. 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 report the mean F1 scores for three random seeds, the deviation is reported in Appendix. For BERT-CNN model, we apply a two-dimensional convolution layer on top of BERT contextual token embeddings. The convolution layer uses a 5 × 5 size kernel. The stride size is (1,2), where 1 is across sentence dimension, and 2 is across word embedding dimension. We also perform circular padding.

| DATASET | PRECISION | | | | | RECALL | | | | | F-MEASURE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T | Q | D | E | A | T | Q | D | E | A | T | Q | D | E | A |
| ANATEM | 89.43 | 88.81 | 89.03 | 89.41 | **89.78** | 87.54 | 89.12 | 87.23 | 88.69 | **89.24** | 88.47 | 88.96 | 88.13 | 89.05 | **89.51** |
| BC2GM | 81.79 | **82.89** | 80.96 | 82.37 | 81.17 | 82.32 | **83.39** | 81.45 | 82.29 | 82.06 | 82.05 | **83.14** | 81.21 | 82.33 | 81.61 |
| BC4CHEMD | 90.25 | **92.07** | 89.81 | 90.57 | 91.19 | 88.48 | **91.01** | 88.13 | 89.27 | 90.49 | 89.36 | **91.54** | 88.96 | 89.92 | 90.84 |
| BC5CDR | 87.93 | **90.09** | 88.07 | 88.49 | 89.01 | 86.72 | **89.16** | 86.12 | 86.62 | 87.83 | 87.32 | **89.62** | 87.08 | 87.55 | 88.42 |
| BIONLP09 | 88.86 | 90.85 | 51.14 | 89.75 | **91.55** | 89.35 | 92.75 | 69.14 | 89.05 | **92.95** | 89.11 | 91.78 | 58.79 | 89.40 | **92.25** |
| BIONLP11EPI | 86.49 | **88.58** | 77.56 | 87.55 | 87.94 | 83.74 | **87.40** | 83.53 | 83.32 | 85.75 | 85.09 | **87.99** | 80.44 | 85.38 | 86.83 |
| BIONLP11ID | 86.19 | **86.60** | 84.29 | 85.14 | 85.56 | 81.09 | **85.35** | 81.24 | 81.05 | 83.38 | 83.56 | **85.97** | 82.74 | 83.04 | 84.46 |
| BIONLP13CG | 88.21 | 89.49 | 87.45 | 89.04 | **90.62** | 83.23 | 86.45 | 82.71 | 85.57 | **88.56** | 85.65 | 87.94 | 85.01 | 87.27 | **89.58** |
| BIONLP13GE | 80.86 | **83.77** | 68.64 | 80.26 | 83.25 | 84.72 | **88.01** | 81.32 | 84.48 | 85.36 | 82.74 | **85.84** | 74.44 | 82.32 | 84.29 |
| BIONLP13PC | 89.79 | 89.03 | 88.43 | 88.95 | **89.87** | 89.55 | **91.87** | 88.33 | 90.58 | 90.60 | 89.67 | **90.43** | 88.38 | 89.76 | 90.23 |
| CRAFT | 86.81 | 88.07 | 82.79 | 88.00 | **90.54** | 84.31 | 88.19 | 84.67 | 89.18 | **89.19** | 85.54 | 88.13 | 83.72 | 88.58 | **89.86** |
| EXPTM | 84.27 | 84.26 | 74.01 | 84.06 | **85.08** | 82.67 | **85.39** | 83.96 | 82.72 | 84.79 | 83.65 | 84.83 | 78.67 | 83.38 | **84.94** |
| JNLPBA | 71.56 | **76.64** | 68.07 | 71.68 | 75.79 | 77.63 | **80.97** | 70.41 | 78.42 | 80.36 | 74.48 | **78.75** | 69.22 | 74.89 | 78.01 |
| LINNAEUS | 91.34 | 88.47 | 86.33 | **92.37** | 90.69 | 86.01 | 88.47 | 86.52 | 88.30 | **90.53** | 88.59 | 88.47 | 86.43 | 90.29 | **90.61** |
| NCBIDISEASE | 86.64 | **87.94** | 86.99 | 86.33 | 87.89 | 90.05 | 89.94 | 89.43 | 89.84 | **91.56** | 88.31 | 88.93 | 88.19 | 88.05 | **89.69** |
| 2010-i2b2 | 93.42 | **95.27** | 93.06 | 93.43 | 94.87 | 94.13 | **95.91** | 93.88 | 94.60 | 95.66 | 93.77 | **95.59** | 93.47 | 94.01 | 95.26 |
| 2011-i2b2 | 93.10 | 94.42 | 92.66 | 94.23 | **94.70** | 92.04 | 94.37 | 92.05 | 93.67 | **94.94** | 92.57 | 94.40 | 92.35 | 93.95 | **94.82** |
| 2012-i2b2 | 82.27 | **84.83** | 81.00 | 75.63 | 67.23 | 81.27 | **85.25** | 81.33 | 83.96 | 84.00 | 81.77 | **85.04** | 81.17 | 79.58 | 74.68 |

Table 2. Precision, Recall and F-Measure of BERT-CNN model using different knowledge types: Entity Type (T), Question (Q), Definition (D), Examples (E) and all of them together (A). Best scores are in bold, second best are underlined. Mean of three random seed runs are reported.

### 4.2 Baseline Models

We consider the following models as strong baselines for our work. The first set of baselines are the BERT models pre-trained on biomedical text BioBERT [24] and MimicBERT [37] finetuned using traditional NER task. BioBERT and MimicBERT are the current state-of-the-art (SOTA) models for NER on multiple biomedical datasets. The second set of baselines are BioBERT, MimicBERT and BERT-MRC finetuned on the knowledge guided NER task. BERT-MRC is initialized with BioBERT base weights, same as our BERT-CNN model. BERT-MRC model is another concurrent query-driven NER model [25], that models the task as a machine reading comprehension task. It predicts all possible start and end positions and predicts valid start-end spans through another feed-forward layer that takes input the predicted start-ends. This model shows considerable improvements in general domain query based NER tasks. The baselines are trained on individual datasets as each dataset has a separate set of entities.

## 5 RESULTS AND DISCUSSION

### 5.1 Biomedical NER

Table 1 compares our method with our baselines on the 18 biomedical NER datasets. Our methods use the best knowledge context identified on the validation set performance. Current state-of-the-art for AnatEM and Linnaeus use specific lexicons and entity specific rules that do not generalize, and hence are not directly comparable to neural models, although our methods approach their performance. On BC4CHEMD and BC5CDR the state-of-the-art methods are BERT-Base models finetuned specifically on chemical and other science corpus, whereas our methods use BioBERT as backbone. Still our models are within 1% F1 score. 2011-i2b2 does not have a task specific to NER, therefore does not have current state-of-the-art methods, but still has annotations for the named entities which we use for joint training. On the rest 12 datasets, we achieve state-of-the-art using BERT-Base and beat methods that use BERT-Large. On JNLPBA

| DATASET | P | ΔP | R | ΔR | F | ΔF |
|---|---|---|---|---|---|---|
| ANATEM | 87.34 | **-1.47** | 89.31 | +0.19 | 88.31 | **-0.65** |
| BC2GM | 79.91 | **-2.9**8 | 81.63 | **-1.76** | 80.76 | **-2.38** |
| BC4CHEMD | 92.56 | +0.49 | 91.10 | +0.09 | 91.82 | +0.28 |
| BC5CDR | 87.67 | **-2.42** | 90.13 | +0.97 | 88.88 | **-0.74** |
| BIONLP09 | 86.92 | **-3.24** | 84.91 | **-6.61** | 85.90 | **-4.93** |
| BIONLP11EPI | 84.57 | **-4.01** | 86.97 | **-0.43** | 85.75 | **-2.24** |
| BIONLP11ID | 87.98 | +1.38 | 84.64 | **-0.71** | 86.27 | +0.30 |
| BIONLP13CG | 84.01 | **-3.97** | 80.84 | **-6.42** | 82.39 | **-5.23** |
| BIONLP13GE | 72.33 | **-9.49** | 86.44 | +0.18 | 78.76 | **-5.22** |
| BIONLP13PC | 86.40 | **-2.63** | 87.21 | **-4.66** | 86.80 | **-3.63** |
| CRAFT | 86.35 | **-1.72** | 85.65 | **-2.54** | 86.00 | **-2.13** |
| EXPTM | 75.28 | **-8.44** | 81.51 | **-4.23** | 78.27 | **-6.44** |
| JNLPBA | 76.85 | +0.81 | 81.79 | +0.16 | 79.24 | +0.51 |
| LINNAEUS | 91.28 | +2.81 | 86.15 | **-2.32** | 88.64 | +0.17 |
| NCBI-DISEASE | 83.86 | **-2.80** | 87.25 | **-3.63** | 85.52 | **-3.20** |
| 2010-i2b2 | 89.87 | **-5.40** | 90.75 | **-5.16** | 90.31 | **-5.28** |
| 2011-i2b2 | 91.49 | **-2.93** | 92.25 | **-2.12** | 91.87 | **-2.53** |
| 2012-i2b2 | 82.05 | +0.72 | 82.31 | **-2.21** | 82.18 | **-0.71** |

Table 3. Change in performance when BERT-CNN model is trained individually on respective datasets with **Question** context. Negative Δ indicates Multi-task is better and are in bold. Precision (P), Recall (R), F-Measure (F).

and NCBI-Disease datasets we improve, but our improvement is not statistically significant. The SOTA scores are F1 values from the following work [4, 12, 24, 37].

Our task formulation gives a significant boost in performance, which is observed in the improvements made by our BioBERT and MimicBERT base models compared to the baseline models following traditional NER formulation. Our BERT-CNN model further improves performance over BioBERT knowledge guided QA model on 12 tasks with a margin of 0.5 to 2.2% (173 - 1934 samples). When it under-performs, it is within a margin of 0.5% (less than 100 samples).

BERT-MRC fails to perform strongly using the same knowledge context as BERT-CNN. On analysis, we observe the model fails to predict the correct end locations for majority of the samples. Overall, BERT-MRC suffers in Recall. When we compare our BIO tagging scheme to BERT-MRC start-end prediction method, if $k$ is the number of entities and $N$ is the number of tokens, time complexity wise our method is $O(kN)$ as we classify each token, whereas BERT-MRC is $O(kN^2)$ as they independently predict start and end locations and then match each start location with an end location.

### 5.2 Ablation Studies and Analysis

*Effect of different knowledge contexts:* Since we incorporate four different knowledge contexts to help in NER, here we identify which knowledge context is better for the NER task across all the 18 datasets. The performance of BERT-CNN model with the knowledge contexts across the test set is shown in Table 2. The scores are entity precision, recall and exact match F1 scores. We observe *Question* and *All* contexts to perform consistently on all the datasets. We believe this is because of the presence of "what" that helps the model to find entities much better than given just a text. To verify our hypothesis we probe our BERT-CNN model trained with *Question* context with multiple probes like "what problem?", the complete question, and "problem" for 20 samples and observe the change in attention scores. As the model is trained with a template, the best prediction is observed on the complete question, with the least scores for only the entity type. Attention scores for "what" were consistently high. The question of why BERT attention scores are

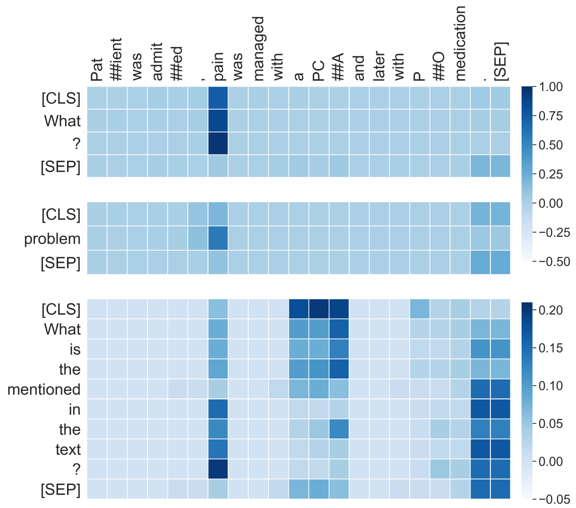


Fig. 3. Attention scores of our BERT-CNN model trained with *Question* context for three knowledge context probes, "What", "problem" and a context where we remove the entity-type. For the last, the model attends to all three entities present **pain** (*problem*), **PCA** and **PO medication** (*treatment*)

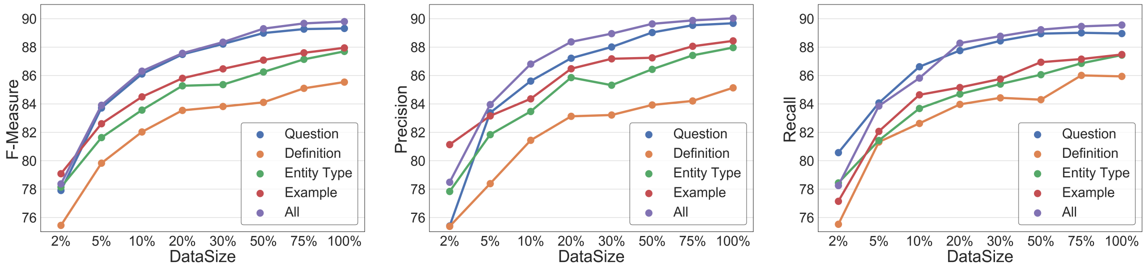


Fig. 4. Effect of the train set size on the three validation set metrics for BERT-CNN model with five contexts.

high for query words is a future research problem, and similar questions about Attention have been raised by others [10]. Figure 3 shows examples of such probes. We can observe our model identifies all the entities, and the entity-type acts as a filter. *Definition* and *Examples* under-perform, we believe the definition might be too generic for an entity type and examples, although representative of class might not be comprehensive.

**Effect of Multi-task training:** To analyze how much multi-task training strategy affects the performance, we define an experiment where we keep the model (BERT-CNN) and the knowledge context *Question* same, but train on each individual datasets, and compare with joint training. Table 3 shows the results of our experiments. Individual dataset training helps improving F1 scores marginally (max 0.5%) on four datasets, whereas joint training substantially improves performance (0.65%-6.44%) on some datasets. This empirically validates our hypothesis that training on combined huge biomedical datasets helps. On further analysis the datasets that have unique entities like AnatEM, 2012-i2b2, Bionlp11ID, Linnaeus and JNLPBA do not show much difference. The datasets that improve is due to presence of common entity-types like Gene/Protein, Chemical, Disease, Problem, Treatment and Test that helps the model to

| **ENTITY↓** | **PRECISION** | | **RECALL** | | **F-MEASURE** | |
|---|---|---|---|---|---|---|
| | **No-K** | **K** | **No-K** | **K** | **No-K** | **K** |
| Gene/Protein | 67.33 | **84.19** | 74.01 | **86.73** | 70.51 | **85.44** |
| Chemical | 89.04 | **91.84** | 88.42 | **91.15** | 88.73 | **91.49** |
| Disease | 83.17 | **83.48** | 86.62 | **88.00** | 84.86 | **85.68** |
| Problem | 91.95 | **93.28** | 92.83 | **94.18** | 92.39 | **93.73** |
| Treatment | 91.78 | **92.91** | 91.98 | **93.47** | 91.88 | **93.19** |
| Test | 92.12 | **94.09** | 93.23 | **94.67** | 92.67 | **94.38** |

Table 4. Comparison of Entity specific (No-K) BERT-CNN model with Question Context provided multi-entity BERT-CNN model (K). Better values are in bold

| **MODEL→** | **BIOBERT(Ours)** | | **MimicBERT(Ours)** | | **BERT-CNN(Ours)** | |
|---|---|---|---|---|---|---|
| **ENTITY ↓** | **SRC F1** | **TGT F1** | **SRC F1** | **TGT F1** | **SRC F1** | **TGT F1** |
| Gene/Protein | 84.83 | 83.27 | 82.46 | 80.75 | 84.99 | **85.63** |
| Chemical | 91.35 | **76.13** | 85.60 | 66.63 | 89.83 | 75.92 |
| Disease | 86.46 | 68.01 | 84.13 | 60.54 | 88.74 | **70.00** |
| Problem | 94.42 | 90.29 | 93.72 | 89.67 | 94.43 | **90.90** |
| Treatment | 94.01 | 89.67 | 93.76 | 89.99 | 94.16 | **90.22** |
| Test | 94.99 | **91.36** | 94.85 | 90.91 | 94.85 | 91.11 |

Table 5. Transfer Learning experiment results. The metric is exact match F1 for source (SRC) and target (TGT) domain. Bold across each of the entities are the best, underlined are the second best.

generalize well and learn better representations. We noticed that the definition of an entity was consistent when it was present in multiple datasets. It ensured that the model was not confused by different definitions. On the other hand, some identical entities had different names in different datasets. For example, disease was called Problem in i2B2, Disease in NCBI-Disease, and by a more specific name Cancer in Bionlp13CG.

**Effect of Knowledge Context compared to Individual entity training:** In this experiment we study the effect of knowledge context over our task formulation. We compare our single BERT-CNN model trained with *Question* context to six different BERT-CNN models each trained only for one specific entity type detection without any context. The entity types are selected such that they are present in multiple datasets and have the most significant number of samples. Table 4 summarizes the results. The results show that knowledge context and training jointly with multiple entity types helps in improving the performance for all entity types, compared to individual entity specific models. The performance improvement for Gene/Protein is the most significant.

**Effect of Train Set Size:** In this experiment we study how the train set size affects the model performance. We sample different percentage of training samples from the total train dataset as seen in Figure 4. We ensure a balanced sampled train set with equal number of positive and negative samples and evaluate across all entities. The training samples are chosen from each of the datasets to ensure the model is not biased towards a dataset or entity type. We do not change any parameters of the model. It can be observed that the performance of the model increases rapidly and then tapers down for each of the context types. We can infer that the model can achieve quite a good performance (84%) with just 5% of training samples but needs much more samples to achieve the state-of-the-art performance. As training samples goes beyond 5%, the precision, recall and the F1-scores for knowledge types *Question* and *All* clearly separate themselves from the other contexts.

**Transfer Learning:** We also examine the transfer learning capability of our system on the six entity sets: gene/protein, chemical, disease, problem, test and treatment, since they are present in multiple datasets. We tested our model on samples of one dataset (target domain) while training and validating on the remaining datasets (source domain). We choose the target domain for each entity to be the dataset that produces the best overall F1-score on full data with BERT-CNN model. We consider Bionlp11ID, BC4CHEMD, BC5CDR datasets as the target domain for gene/protein, chemical and disease respectively and 2010-i2b2 for problem, treatment and test. Table 5 summarizes the results.

The results show a varied degree of transfer learning, losing only 0.5% F1 in some tasks and by as much as 20% in other tasks, which is in line with earlier observed performance loss [5]. The difference in source and target F1-scores is remarkably close for Problem, Treatment and Test entities. The two domains although are close for these entities, they do have different set of entities. Chemical shows a significant drop but still our model achieves 75% F1 despite the target domain containing many prior unseen entities. For Gene/Protein the source and target domain are nearly the same set of entities.

## 6 RELATED WORK

**External Knowledge** : In the past, there have been several attempts to incorporate external knowledge through feature engineering and lexicons [7, 9, 19, 27], or incorporating knowledge in the feature extraction stage [12, 52], or using document context [13]. In our work, we incorporate simple textual knowledge sentences and show how to integrate them in named entity recognition tasks.

**Multi-Task Learning** : Multi-task learning have been used in the past to tackle the labelling problem of NER. For example, multi-task learning with simple word embedding and CNN [12], cross-type NER with Bi-LSTM and CRF [47], MTL with private and shared Bi-LSTM-CRF using character and word2Vec word embeddings [45]. In our work, we do multi-task learning by reducing all different NER tasks to the same generic format and use transformer encoders.

**Language Models and Transfer Learning** : There have been prior attempts to reduce the labelling confusion by using a single model to predict each entity-type [24] and also using transfer-learning [4, 24, 37]. 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 context. We show jointly learned single model is better than per entity-type models.

**NER as a Question Answering Task** : In general domain, researchers have formulated multiple NLP tasks as question-answering format in DecaNLP [28], semantic-role labelling as in QASRL [18] and others have argued that question-answering is a format not a task [15]. We also use QA format as a part of our task, to address previously mentioned challenges . A possibly concurrent work, BERT-MRC [25] also attempts at NER as a QA task in general domain by span predictions for individual entity types in a reading comprehension style approach. We however differ in the task formulation using BIO tagging scheme, our model design and our focus in Biomedical NER. A detailed comparison is in Section 5.1.

**BioMedical NER:** In Biomedical domain, CollaboNet [54] uses multiple expert models for each dataset collaborating to reduce the misclassification error and NER using variational dropout [17]. Dictionary-based distantly supervised methods [46] have been proposed to reduce the need of human-annotations, but are still far from fully-supervised methods. Other methods use more nuanced approaches of adding word and character-level features [53]. The most common approach of using BiLSTM-CRFs (recurrent neural networks) for NER, has been applied to several bio-medical or chemical applications, such as Medline indexing [36], entity extraction for fMRI [1], postpartum depression detection [8], and conversational agents [2]. Recurrent neural networks are prevalent for clinical and biomedical sequence labelling tasks such as NER [51]. A major drawback to such approaches is the need for multiple models for each dataset. We hope our work can motivate adaptation of KGQA and transformer encoders, which not only reduces the need for multiple models, but improves the overall task performance.

## 7 CONCLUSION

We reformulated the NER task as a knowledge guided, context driven QA task and showed it has a significant impact. Our models are more explainable using the query-text attention and address the major challenges faced by current NER systems. Our approach has achieved above state-of-the-art F measures for 14 of the common public 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.

## REFERENCES

[1] Asma Ben Abacha, Alba G Seco de Herrera, Ke Wang, L Rodney Long, Sameer Antani, and Dina Demner-Fushman. 2017. Named entity recognition in functional neuroimaging literature. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2218–2220.

[2] Muhammad Amith, Licong Cui, Kirk Roberts, and Cui Tao. 2020. Towards an Ontology-based Medication Conversational Agent for PrEP and PEP. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*. Association for Computational Linguistics, Online, 31–40. https://doi.org/10.18653/v1/2020.nlpmc-1.5

[3] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. 2012. Concept annotation in the CRAFT corpus. *BMC bioinformatics* 13, 1 (2012), 161.

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

[5] Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 Task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 565–572. https://doi.org/10.18653/v1/S17-2093

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

[7] 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*.

[8] Sanchari Chowdhuri, Sidnei McCrea, Dina Demner Fushman, and Casey Overby Taylor. 2019. Extracting Biomedical Terms from Postpartum Depression Online Health Communities. *AMIA Summits on Translational Science Proceedings* 2019 (2019), 592.

[9] 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*, Vol. 2005.

[10] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. *arXiv preprint arXiv:1906.04341* (2019).

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

[12] 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 (2017), 368.

[13] 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* (2018).

[14] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics* 47 (2014), 1–10.

[15] 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* (2019).

[16] Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. LINNAEUS: a species name identification system for biomedical literature. *BMC bioinformatics* 11, 1 (2010), 85.

[17] John M Giorgi and Gary D Bader. 2020. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics* 36, 1 (2020), 280–286.

[18] 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*. 643–653.

[19] Jun'ichi 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)*. 698–707.

[20] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 workshop companion volume for shared task*. 1–9.

[21] Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 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*. Citeseer, 70–75.

[22] Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. CHEMDNER: The drugs and chemical names extraction challenge. *Journal of cheminformatics* 7, S1 (2015), S1.

[23] 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* (Copenhagen, Denmark). Association for Computational Linguistics, 785–794. https://doi.org/10.18653/v1/D17-1082

[24] 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* (2019).

[25] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019. A Unified MRC Framework for Named Entity Recognition. *arXiv preprint arXiv:1910.11476* (2019).

[26] 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 (2006), 497–507.

[27] 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*. Association for Computational Linguistics, Florence, Italy, 5301–5307. https://doi.org/10.18653/v1/P19-1524

[28] 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* (2018).

[29] 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*.

[30] Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP shared task 2013 workshop*. 1–7.

[31] Tomoko Ohta-Robert Bossy Ngan Nguyen, Junichi Tsujii Jin-Dong Kim, and Sampo Pyysalo. 2011. Overview of bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*. 1–6.

[32] 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 *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[33] Sampo Pyysalo and Sophia Ananiadou. 2014. Anatomical entity mention recognition at literature scale. *Bioinformatics* 30, 6 (2014), 868–875.

[34] Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, and Jun'ichi Tsujii. 2011. Towards exhaustive protein modification event extraction. In *Proceedings of BioNLP 2011 Workshop*. 114–123.

[35] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *EMNLP*.

[36] Max E Savery, Willie J Rogers, Malvika Pillai, James G Mork, and Dina Demner-Fushman. 2020. Chemical Entity Recognition for MEDLINE Indexing. *AMIA Summits on Translational Science Proceedings* 2020 (2020), 561.

[37] Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing Clinical Concept Extraction with Contextual Embedding. *arXiv preprint arXiv:1902.08691* (2019).

[38] Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of BioCreative II gene mention recognition. *Genome biology* 9, S2 (2008), S2.

[39] 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 (2018), 158.

[40] 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 (2013), 806–813.

[41] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4149–4158. https://doi.org/10.18653/v1/N19-1421

[42] 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 (2012), 786–791.

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

[44] 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 (2011), 552–556.

[45] 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 (2019), 427.

[46] Xuan Wang, Yu Zhang, Qi Li, Xiang Ren, Jingbo Shang, and Jiawei Han. 2019. Distantly supervised biomedical named entity recognition with dictionary expansion. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 496–503.

[47] 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 (2018), 1745–1752.

[48] Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2015. Overview of the BioCreative V chemical disease relation (CDR) task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, Vol. 14.

[49] Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212.

[50] 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 (2019).

[51] Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, et al. 2020. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association* 27, 3 (2020), 457–470.

[52] 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*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2145–2158. https://www.aclweb.org/anthology/C18-1182

[53] Vikas Yadav, Rebecca Sharp, and Steven Bethard. 2018. Deep affix features improve neural named entity recognizers. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. 167–172.

[54] Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. 2019. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC bioinformatics* 20, 10 (2019), 249.

## A MODEL EXPLANATION WITH ATTENTION PROBING

We study our BERT-CNN model using attention value heatmaps and try to explain how our model uses the knowledge contexts to extract specific entities from a given dataset. To show this, we choose a sample "*Patient* was *admited* , *pain* was managed with *a PCA* and later with *PO medication*." where there are multiple entities **pain** (problem), **a PCA** and **PO medication** (treatment), **Patient** (person) and **admitted** (occurance).

### A.1 Using Question as Knowledge Context

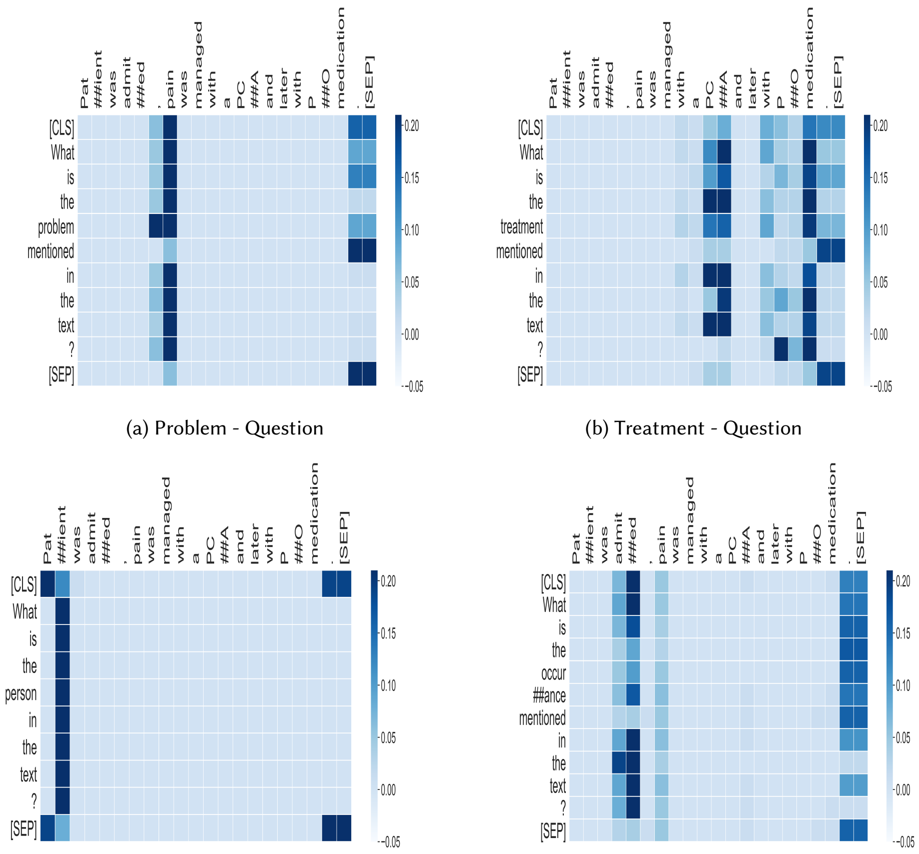
From the Figures 5a, 5b, 5c and 5d it can be seen that, when knowledge context is in question format then, each of the entity types present in the knowledge context guide the model to chose the correct entities. This can be seen from the higher attention values for those specific entities.

### A.2 Using Entity type, Definition, Example and All combined as Knowledge Context

We also probed our model to extract the attention weights for each of the other four knowledge contexts. Here we show this only for problem entity type. In Figure 6a it can be seen that attention weight between *problem* in knowledge context and *pain* in text is highest. Figure 6b shows keyword like *disorder* in the definition representing the meaning of the entity type problem highly attends to the entity *pain*. On using the example as knowledge, it can be seen from Figure 6c, keywords like hypertension, pain, nausea, fever highly attend to the entity pain in the text. This is in line with our hypothesis that providing similar entities belonging to the same entity group might help find the entity in a text. Finally, in Figure 6d, it can be seen that the keywords from question, definition and example all collectively help to predict the entity pain in the text.

## B MODEL TRAINING

We use the HuggingFace [50] and Pytorch Deep learning framework [32]. We train the model with following hyperparameters, 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]



(a) Problem - Question

(b) Treatment - Question

(c) Person - Question

(d) Occurance - Question

Fig. 5. Attention Probes for different knowledge and entity types.

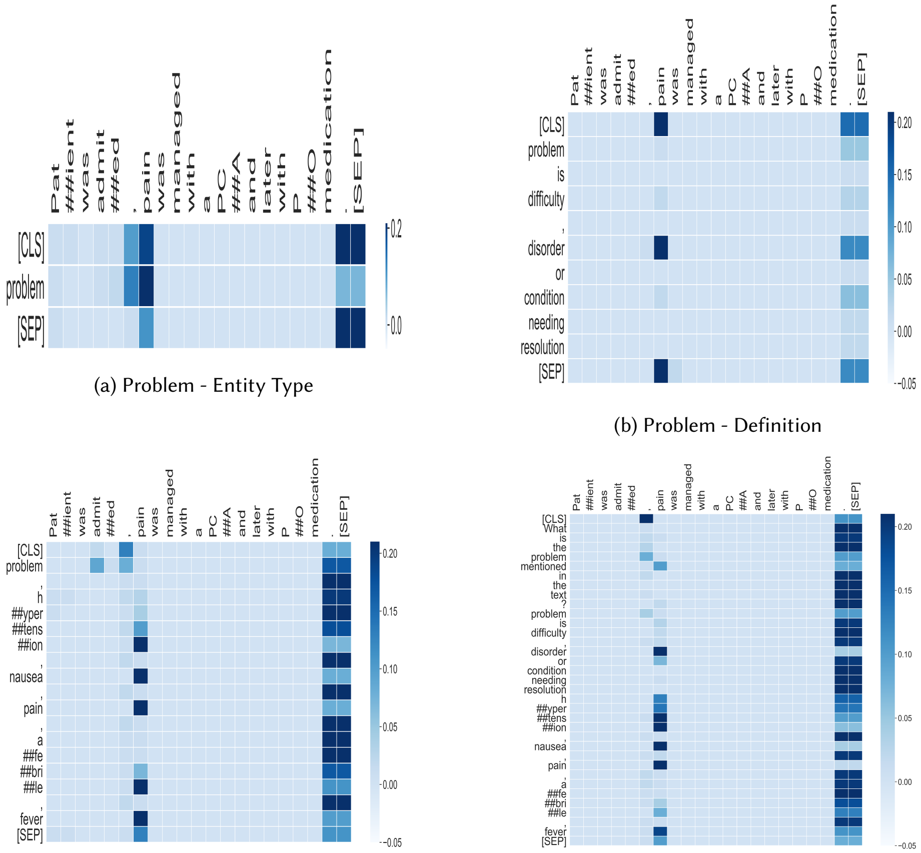
and warm-up steps in range of [100,1000]. We use BERT-base-cased version for all our models. The BERT-base-cased model has nearly 110M parameters.

## C ENTITY DISTRIBUTION IN THE DATASET

The number of samples present in the Table 8 and 9 are created directly from the train, validation and test samples of 18 biomedical datasets. The i2b2 datasets do not have separate validation data splits. We use 30% of the samples from training data as validation data. The *Entity Mentions* represents the total number of entities present for the datasets in all of train, validation and test samples. Since each sample data can have multiple entities, the number is higher than the total positive and negative samples for the dataset.

## D PERFORMANCE COMPARISON ON TEST DATA

Table 6 shows the performance comparison of our best model with SOTA on the test data for each of the 18 datasets. The F-measures in bold are the best performance on each datasets. The state-of-the-art for Linnaeus and AnatEM



(a) Problem - Entity Type

(b) Problem - Definition

(c) Problem - Example

(d) Problem - All knowledge combined

Fig. 6. Attention Probes for different knowledge and entity types.

datasets uses dictionaries developed without a clear train/test split, hence our scores are not directly comparable. Also 2011-i2b2 do not have SOTA concept extraction performance. Improvements in 10 datasets are significant as compared to the SOTA.

## E TRAINING WITH BALANCED DATASET

We generated a sample for each text available in the source data. The text may or may not contain a particular entity. So we generate negative samples for each text and for each available entity types of the dataset making the datasets unbalanced. In Table 7 we show that the negative samples does not have much impact on the performance of our models. The results are taken using BERT-CNN model with *Question* as a context. Negative values in $\Delta P$,$\Delta R$ and $\Delta F$ means training on unbalanced data is better than on balanced data.

| Dataset | Entities | SOTA F1 | OURS P | OURS R | OURS F1 | Significant |
|---|---|---|---|---|---|---|
| ANATEM | 4616 | **91.61** | 90.29 | 89.43 | 89.85 ± 0.48 | No |
| BC2GM | 6322 | 81.69 | 82.89 | 83.39 | **83.14 ± 0.54** | Yes |
| BC4CHEMD | 25331 | **92.36** | 92.56 | 91.10 | 91.82 ± 0.58 | No |
| BC5CDR | 9808 | **90.01** | 90.09 | 89.62 | 89.62 ± 0.71 | No |
| BIONLP09 | 3589 | 84.20 | 91.55 | 92.95 | **92.25 ± 0.57** | Yes |
| BIONLP11EPI | 5730 | 78.86 | 88.58 | 87.40 | **87.99 ± 1.10** | Yes |
| BIONLP11ID | 3810 | 81.73 | 87.98 | 84.64 | **86.27 ± 1.80** | Yes |
| BIONLP13CG | 7861 | 78.90 | 90.62 | 88.56 | **89.58 ± 0.68** | Yes |
| BIONLP13GE | 4354 | 78.58 | 83.77 | 88.01 | **85.84 ± 0.93** | Yes |
| BIONLP13PC | 5306 | 81.92 | 90.14 | 92.09 | **91.11 ± 0.12** | Yes |
| CRAFT | 18770 | 79.56 | 90.54 | 89.19 | **89.86 ± 0.55** | Yes |
| EXPTM | 2308 | 74.90 | 85.97 | 85.30 | **85.64 ± 0.61** | Yes |
| JNLPBA | 8673 | 78.58 | 76.85 | 81.79 | **79.24 ± 0.45** | No |
| LINNAEUS | 1428 | **95.68** | 90.69 | 90.53 | 90.61 ± 0.28 | No |
| NCBIDISEASE | 956 | 89.36 | 87.89 | 91.56 | **89.69 ± 0.37** | No |
| 2010-i2b2 | 30140 | 90.25 | 95.27 | 95.91 | **95.59 ± 0.30** | Yes |
| 2011-i2b2 | 25271 | - | 94.70 | 94.94 | 94.82 ± 0.41 | - |
| 2012-i2b2 | 15301 | 80.91 | 84.83 | 85.25 | **85.04 ± 1.18** | Yes |

Table 6. Precision(P), Recall(R) and F-measure(F1) with our best model measured by running with three seed values. Significant column shows whether our F1-scores are statistically significantly better than SOTA F1 ($p \leq 0.05$, based on Wilson score intervals [49]). Best F-measures are in bold.

## F DATA PREPROCESSING

We have done the experiments on 18 biomedical dataset which are available in different formats. For the 15 publicly available datasets, we used the BIO annotated files and automatically extracted the spans based on the tags. The three i2b2 files has different format. They have individual biomedical reports containing multiple sentences which may or may not contain the entities. So we preprocessed them by considering each statement as a sample without rejecting any sentence. Thus we bring all the datasets into a common format. Each sample in our pre-processed data contains the id(indicating the dataset which is the sample origin), text, answers, spans, number of answers present in the text, entity type, question context, definition of entities, top ten frequently occurring examples with counts. We also grouped together similar entity-types to form entity groups and add entity group definitions which can be used for further research.

| DATASET | P | ΔP | R | ΔR | F | ΔF |
|---|---|---|---|---|---|---|
| ANATEM | 88.88 | **-0.07** | 88.23 | +0.89 | 88.55 | +0.41 |
| BC2GM | 82.93 | **-0.04** | 82.84 | +0.55 | 82.88 | +0.26 |
| BC4CHEMD | 91.64 | +0.43 | 91.03 | **-0.02** | 91.33 | +0.21 |
| BC5CDR | 89.61 | +0.48 | 88.37 | +0.79 | 88.98 | +0.64 |
| BIONLP09 | 90.76 | **-0.60** | 92.33 | **-0.81** | 91.54 | **-0.71** |
| BIONLP11EPI | 87.86 | +0.72 | 86.29 | +1.11 | 87.07 | +0.92 |
| BIONLP11ID | 85.39 | +1.21 | 85.19 | +0.16 | 85.29 | +0.68 |
| BIONLP13CG | 88.61 | **-0.63** | 87.54 | **-0.27** | 88.07 | **-0.45** |
| BIONLP13GE | 81.94 | **-0.12** | 88.77 | **-2.51** | 85.22 | **-1.24** |
| BIONLP13PC | 89.48 | **-0.45** | 90.73 | +1.14 | 90.10 | +0.33 |
| CRAFT | 87.43 | +0.64 | 88.12 | +0.07 | 87.78 | +0.35 |
| EXPTM | 85.12 | **-1.40** | 85.99 | **-0.25** | 85.55 | **-0.84** |
| JNLPBA | 75.85 | +0.19 | 82.37 | **-0.74** | 78.98 | **-0.25** |
| LINNAEUS | 88.16 | +0.31 | 87.91 | +0.56 | 88.03 | +0.44 |
| NCBI-DISEASE | 88.00 | **-1.34** | 90.46 | +0.42 | 89.22 | **-0.50** |
| 2010-i2b2 | 94.96 | +0.31 | 95.71 | +0.20 | 95.33 | +0.26 |
| 2011-i2b2 | 94.01 | +0.41 | 94.09 | +0.28 | 94.05 | +0.35 |
| 2012-i2b2 | 82.97 | **-1.64** | 86.65 | **-2.13** | 84.77 | **-1.88** |

Table 7. Precision (P), Recall (R) and F-Measure (F) using BERT-CNN model trained on **balanced dataset** with **Question** as knowledge. ΔP, ΔR, ΔF represent change in performance when compared to training our model on full datasets. Negative value indicates training on unbalanced dataset is better while positive value indicates balanced dataset training produces better performance. Negative values are in bold.

| 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 |
| EXPTM | 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 8. Data Distribution, with counts of entities, number of positive samples with at least one entity mentions, and negative samples with no target entity.

| Dataset | Entity | Entity Mentions | Train + | Train - | Dev + | Dev - | Test + | Test - |
|---|---|---|---|---|---|---|---|---|
| **2010-i2b2** | Problem | 18979 | 4213 | 4226 | - | - | 5802 | 6590 |
| | Treatment | 13809 | 3126 | 4226 | - | - | 7234 | 6590 |
| | Test | 13576 | 2426 | 4226 | - | - | 4591 | 6590 |
| **2011-i2b2** | Person | 17744 | 7207 | 3990 | - | - | 4715 | 2971 |
| | Problem | 18869 | 7003 | 3990 | - | - | 4384 | 2971 |
| | Treatment | 17708 | 5300 | 3990 | - | - | 3565 | 2971 |
| | Test | 13514 | 4191 | 3990 | - | - | 2786 | 2971 |
| **2012-i2b2** | 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 9. Data Distribution, with counts of entities, number of positive samples with at least one entity mentions, and negative samples with no target entity.