TaughtNet: Learning Multi-Task Biomedical Named Entity Recognition From Single-Task Teachers

Vincenzo Moscato, Marco Postiglione, Carlo Sansone, *Senior Member, IEEE*, and Giancarlo Sperlí

***Abstract*—In Biomedical Named Entity Recognition (BioNER), the use of current cutting-edge deep learning-based methods, such as deep bidirectional transformers (e.g. BERT, GPT-3), can be substantially hampered by the absence of publicly accessible annotated datasets. When the BioNER system is required to annotate multiple entity types, various challenges arise because the majority of current publicly available datasets contain annotations for just one entity type: for example, mentions of *disease* entities may not be annotated in a dataset specialized in the recognition of *drugs*, resulting in a poor ground truth when using the two datasets to train a single multi-task model. In this work, we propose *TaughtNet*, a knowledge distillation-based framework allowing us to fine-tune a single multi-task *student* model by leveraging both the ground truth and the knowledge of single-task *teachers*. Our experiments on the recognition of mentions of diseases, chemical compounds and genes show the appropriateness and relevance of our approach w.r.t. strong state-of-the-art baselines in terms of precision, recall and F1 scores. Moreover, TaughtNet allows us to train smaller and lighter student models, which may be easier to be used in real-world scenarios, where they have to be deployed on limited-memory hardware devices and guarantee fast inferences, and shows a high potential to provide explainability. We publicly release both our code on github[1] and our multi-task model on the huggingface repository.[2]**

***Index Terms*—Knowledge distillation, multi-task learning, transformers, named entity recognition.**

I. Introduction

NUMEROUS industrial sectors, including healthcare, are being revolutionised by the uncontrolled growth of data produced by humans and machines as well as the availability of computing resources and algorithms able to handle and analyse it. In 2022, PubMed Central[3] provides open online access to 7.8 million full-text articles. Concomitantly, efforts are being made to collect and make available the unstructured health information associated with hospital admissions (e.g. EHRs, laboratory tests, medications). As a result, the field of biomedical text understanding can profitably benefit from the current advancements in Deep Learning and Natural Language Processing techniques.

Biomedical Named Entity Recognition (BioNER) consists in identifying mentions of biomedical entities (e.g. disorders, chemical compounds, genetic information) from unstructured text data. It is the first and essential step of many text understanding applications, such as the construction of knowledge graphs for data representation and analysis or conversational agents including research assistants and medical chatbots.

It is extremely difficult to develop a BioNER system that can recognise a wide range of entity types with high precision and recall for a number of reasons, including:

- *Presence of synonyms, alternate spellings, polysemous words:* Biomedical datasets are characterized by a large number of synonyms or alternate spellings of entities, which are often referred to with non-standard abbreviations; polysemy is very common, i.e. the same token could represent different entities based on its context (e.g. the token "VHL" may refer to the Von Hippel-Lindau disease or to the gene name which causes the disease).
- *Lack of annotated data:* To guarantee high quality, the labeling process of healthcare datasets requires time, effort and domain knowledge. As a result, there is a lack of publicly available training data. Furthermore, the majority of datasets covers only one or two entity types, making it necessary to integrate different data sources.
- *Inference time and memory constraints:* Being (usually) a component of a larger pipeline architecture, the BioNER system has to be able to promptly provide its results when required. Moreover, in conversational agents, it may be necessary to deploy the system on devices with a limited amount of memory.

NER systems for biological text mining were used to be primarily dictionary- and rule-based, but they had a number of issues, including the *out-of-vocabulary* problem, i.e. they

Manuscript received 9 May 2022; revised 25 November 2022 and 5 January 2023; accepted 5 February 2023. Date of publication 10 February 2023; date of current version 5 May 2023. This work was supported in part by PNRR MUR Project PE0000013-FAIR, and in part by the EU through the CEF-TC Programme, European Platform for Digital Skills and Jobs funding call under Grant INEA/CEF/ICT/A2020/2267304. *(Corresponding author: Marco Postiglione.)*

The authors are with the Department of Electrical and Information Technology, University of Naples Federico II, 80125 Naples, Italy (e-mail: vincenzo.moscato@unina.it; marco.postiglione@unina.it; carlo.sansone@unina.it; giancarlo.sperli@unina.it).

Digital Object Identifier 10.1109/JBHI.2023.3244044

[1] TaughtNet code: https://github.com/marcopost-it/TaughtNet

[2] TaughtNet model (diseases, chemical, genes): https://huggingface.co/marcopost-it/TaughtNet-disease-chem-gene

[3] https://www.ncbi.nlm.nih.gov/pmc/

TABLE I
EXAMPLE OF TAUGHTNET OUTPUT FOR THE IDENTIFICATION OF DISEASE, CHEMICAL AND GENE MENTIONS, COMPARED TO THE GROUND TRUTH.[4]

Ground truth	Cycloheximide facilitates the identification of aberrant transcripts resulting from a novel splice-site mutation in COL17A1 in a patient with generalized athrophic benign epidemolysis bullosa .
TaughtNet	Cycloheximide facilitates the identification of aberrant transcripts resulting from a novel splice-site mutation in COL17A1 in a patient with generalized athrophic benign epidemolysis bullosa .

struggled to deal with unseen and/or polysemous words and had a low recall.

As a result of the availability of an increasing number of human-labeled datasets, BioNER systems evolved over time by means of deep learning techniques able to infer features from sentence contexts. These methods were typically based on Bidirectional Long-Short Term Memory networks with Conditional Random Fields (BiLSTM-CRF) [1], [2] and/or trying to capture character-level features of words [3], [4], [5], [6]. Recently, large-scale language models pre-trained on biomedical corpora and fine-tuned over BioNER datasets [7], [8], [9], [10], [11] have shown their remarkable potential to enhance the state-of-the-art of biomedical entity recognition and their promising prospects for improvement as the availability of training data increases [12].

Nevertheless, the above-mentioned models usually have hundreds of millions of parameters, and recent research demonstrates that as the training parameters are increased, performance on downstream tasks improves [13]. The expansion of model parameters implies computational and memory limitations, which may make it more difficult to use these systems in real-world settings.

In this paper, we aim to use the technological advances brought about by Transformer models to train a multi-task BioNER system capable of recognizing multiple entity types from its inputs while dealing with the data shortage affecting the biomedical field, where most publicly available datasets contain tags for only one entity type. Based on the premise that designing, training and deploying a single Transformer-based BioNER model for each available dataset is extremely impractical, owed both to their constraining memory requirements and to the problems which would arise due to overlapping predictions — e.g. two models assigning two different entity types to the same mention —, we propose *TaughtNet*, a multi-task framework based on knowledge distillation that allows us to fine-tune a single transformer architecture to recognize multiple entity types (an output example is shown in Table I.

Similar works from Khan et al. [14] and Yoon et al. [15] propose changes in the model architecture and training procedure to accomplish the task: the former trains multiple models sharing some layers in order to build a "shared knowledge" across the datasets, while the latter leverages an ensemble of single-task models. In contrast, TaughtNet produces a single, independent Student Transformer model that is capable of recognising a variety of entity types. Our Teachers do not create an ensemble that works together to make predictions; rather, they merely impart their knowledge to the student during the training phase.

In our experiments, we demonstrate that *TaughtNet* not only allows us to efficiently individuate multiple entity types by ensuring state-of-the-art performance on three benchmark datasets, but it can also be applied to smaller and lighter students, which may be more easily used in real-world scenarios where they must be deployed on hardware with limited memory and/or to ensure quick inferences. Additionally, we show the potential of TaughtNet to easily provide explainability for its predictions, which is not always possible when utilising multiple models or intricately adjusted architectures.

The rest of the paper is structured as follows. We recall some background on Biomedical Named Entity Recognition, Pretrained Language Models, Multitask Learning and Knowledge Distillation and describe the main Related Works in Section 2. Next, we introduce the training framework of *TaughtNet* in Section 3. Experiments are described in Section 4. Finally, we conclude our paper and discuss future directions in Section 5.

II. BACKGROUND AND RELATED WORK

A. Biomedical Named Entity Recognition

The Named Entity Recognition (NER) task has been introduced in [16] with the aim to identify mentions of interest in unstructured texts. Biomedical Named Entity Recognition (BioNER) differs from general NER under several different points of view [17]: (1) datasets are characterized by a large number of synonyms or alternate spellings of entities, which are often referred to with (even non-standard) abbreviations; (2) entities often consist of long sequences of tokens, making it difficult to detect their boundaries; (3) entities are sometimes nested, e.g. an entity of class *"species"* can be part of a longer entity of class *"disease"*; (4) polysemy is very common, i.e. the same token may refer to different entity types, but the right one has to be chosen based on its context.

While neural networks have demonstrated to generally outperform other approaches because of their capacity to analyse the syntactic and semantic structure of sentences [13], [18], annotating training data to train them is a laborious and time-consuming task that requires knowledge from domain experts. Furthermore, the lack of resources affecting the healthcare industry is primarily caused by privacy concerns surrounding the sharing of personal information. It would be preferable for a recognition system to maximise the usage of publicly accessible datasets unless it is feasible to employ a significant quantity of data given by private entities (such as hospitals and organisations)

B. Pretrained Language Models in the Healthcare Domain

In recent years, the research interest in the area of *Natural Language Processing* is rapidly growing, especially thanks to the

[4] Sample Taken From the Test Set of the Dataset *NCBI-Disease*

pretrain-and-finetune approach which has brought significant improvements in many downstream tasks [13], [18], [19], [20]. A broad variety of pretrained models have been presented in the healthcare industry, driven by the well-established fact that pretraining the language model using domain-dependent training data significantly enhances performance [11]. Lewis et al. [12] provide an accurate comparison of the current landscape of pretrained healthcare models, highlighting the main training choices affecting downstream performance.

Techniques mostly vary in terms of the training data, which is either acquired from medical records or scientific literature (e.g., PubMed, Semantic Scholar, PMC) (e.g. MIMIC-III or other private datasets). In the first scenario, data can be easily retrieved (at least for the English language), allowing for the collection of enormous amounts of raw text to train the model; in the second scenario, data is more difficult to gather and share due to privacy concerns, but is closer to the real world of medical practise than the idealised information found in textbooks and journals [21].

The focus of this paper is not to pretrain a novel language model, but rather to design a fine-tuning framework which, based on knowledge distillation, allows us to accomplish the NER task for multiple entities by exploiting pretrained language models and heterogeneous publicly available healthcare datasets, each of them referring to a different entity type.

C. Multi-Task Learning

Multi-Task Learning (MTL) aims to leverage multiple datasets that are similar to one another yet address various tasks [22]. The key idea is that the knowledge acquired by the model for solving a task (e.g., disease extraction) can help it in solving similar tasks (e.g. drug extraction).

In biomedical text mining, the first approaches (e.g. [23]) ignored the information of subwords which can be crucial to obtain high performance. Wang et al. [6] propose the combination of a multi-task BiLSTM-CRF model and a BiLSTM layer for modeling character sequences, obtaining promising results. To the best of our knowledge, [14] is the first work adopting the multi-task learning framework with a pre-trained language model.

Yoon et al. [15] highlight that despite the high recall obtained by MTL models, their precision is relatively low, i.e. they have difficulties in differentiating between entity types, primarily due to the presence of polysemous words in text which confuse the model. To solve such false-positive problem, the authors propose *CollaboNet*, a network composed of multiple models, each one built on a different dataset for a different task, which collaborates during training and inferences to output the final prediction. Despite the promising results, this framework requires "collaborator" models to be stored in memory at inference time and to provide their outputs when a prediction is required, resulting in low efficiency in computational and memory consumption terms.

To overcome the low-precision and the computational and memory consumption challenges, inspired by CollaboNet, we developed *TaughtNet*, a training framework which allows us to fine-tune a single transformer language model for multi-task BioNER based on Knowledge Distillation. In simple terms, we train single-task models on different datasets, but they do not collaborate to provide the outputs of predictions, but rather to "teach" to a single multi-task "student" how to predict the entity types in which they are experts.

D. Knowledge Distillation

Knowledge Distillation (KD) has been originally proposed in [24] as a *teacher-student* framework which allows the knowledge embedded in a large "teacher" model to be shared with its small "student". Modeling the behavior of teacher and student with functions $f_T(\cdot)$ and $f_S(\cdot)$, respectively, the objective of KD is to minimize the following objective function:

$$\mathcal{L} = \sum_{x \in \mathcal{X}} L\left(f_S(x), f_T(x)\right), \tag{1}$$

where $\mathcal{X}$ is the training dataset and $L(\cdot)$ denotes the loss function computing the difference between the two behavior function outputs for the input $x \in \mathcal{X}$.

With the primary aim to "compress" the knowledge embedded in a large model — which shows good performance but is too large to be used in real scenarios — into a smaller one, the application of KG in NLP and pre-trained models has been extensively studied [25], [26], [27], [28], [29], [30], [31].

Research on the application of the KD framework for purposes other than model compression is restricted to a few works. Reimers et al. [32] try to transfer the knowledge embedded in an English BERT model to the German language. In [33], a fine-tuned BERT teacher is used as extra supervision to improve the text generation performance of conventional Seq2Seq student models.

To the best of our knowledge, TaughtNet is the first approach exploiting KD in a NER scenario to transfer the knowledge encoded in a variety of teachers, specialized in single entity types, into a single student, which learns to recognize all the entity types.

The *multi-teacher* scenario in the application of the KD approach has been thoroughly investigated [34], [35], [36], [37]. Fukuda et al. [35] hypothesize that the different "views" provided by various teacher distributions may help the student generalizing better while also capturing the complementary information embedded in each teacher stream. In [34], the teacher is an ensemble of models whose outputs are determined by the combination of the individual model predictions and the student learns to imitate its behavior by minimizing the Kullback-Leibler (KL) divergence [38] between student and teacher distributions (which the authors prove to be equal to minimizing the cross-entropy error between the two distributions). The use of an ensemble knowledge distillation framework in [36] results in better student accuracy thanks to the encouragement of heterogeneity in feature learning. [37] highlights the importance of assigning the proper weights to teachers when distilling their knowledge.

In contrast to traditional KD approaches, where teachers and students share the same tasks, we aim to design a student able to handle all the tasks learned from teachers in a single model. Tan et al. [39] propose a similar approach, designing a multilingual translation system based on knowledge distillation from multiple

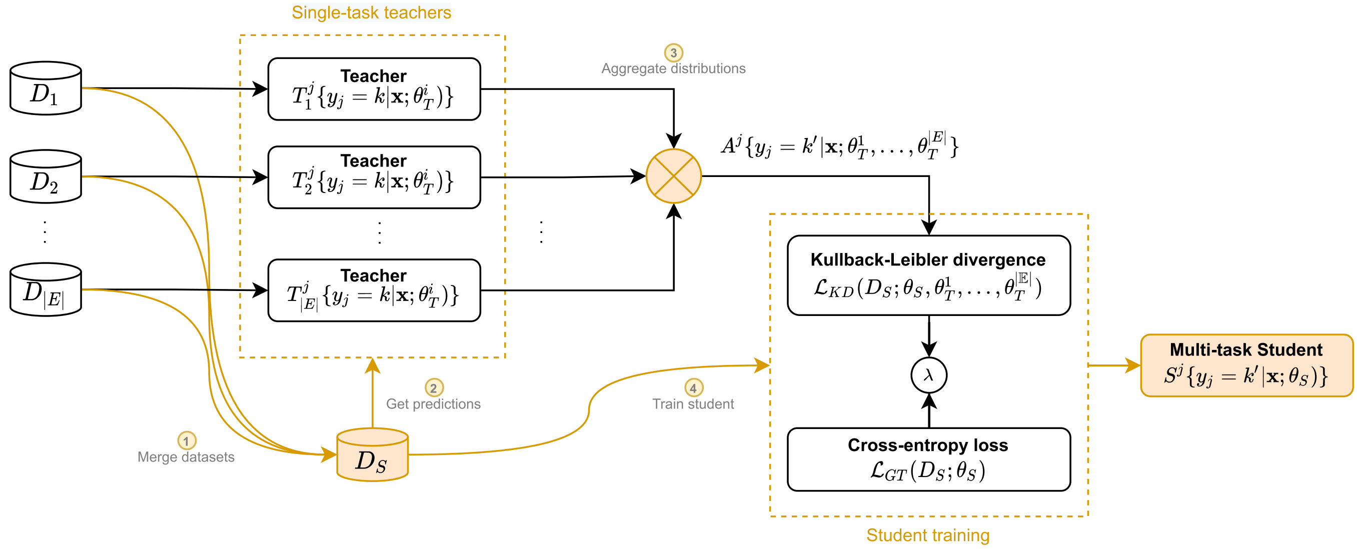

Single-task teachers
D_1
D_2
$D_{|E|}$
Teacher $T_1^j\{y_j = k|\mathbf{x}; \theta_T^i)\}$
Teacher $T_2^j\{y_j = k|\mathbf{x}; \theta_T^i)\}$
Teacher $T_{|E|}^j\{y_j = k|\mathbf{x}; \theta_T^i)\}$
3
Aggregate distributions
$A^j\{y_j = k'|\mathbf{x}; \theta_T^1, \ldots, \theta_T^{|E|}\}$
Kullback-Leibler divergence
$\mathcal{L}_{KD}(D_S; \theta_S, \theta_T^1, \ldots, \theta_T^{|\mathbb{E}|})$
λ
Cross-entropy loss
$\mathcal{L}_{GT}(D_S; \theta_S)$
Multi-task Student
$S^j\{y_j = k'|\mathbf{x}; \theta_S)\}$
2
Get predictions
4
Train student
1
Merge datasets
D_S
Student training

Fig. 1. Overview of *TaughtNet* training framework. (1) First, single-entity datasets are merged together to build a multi-entity corpus $\mathcal{D}_S$ used as a ground truth reference during the training of the Student; (2) then, each Teacher provides its predictions for each sample in $\mathcal{D}_S$ and (3) their output distributions are aggregated so as to build the corpus $\mathcal{A}$ used to distill the knowledge from Teachers. (4) Finally, the Student is trained to minimize two loss components referring to Teachers' knowledge and ground truth, respectively.

individual teachers handling separate language pairs. Their experimental results, showing that the multilingual model reaches comparable performance with teachers — even outperforming them in many cases — further encourage our work.

III. Method

In this work, we aim to leverage a set of publicly available healthcare datasets to train a single multi-task BioNER model. A comprehensive overview of our framework is shown in Fig. 1. To facilitate the reader in the understanding of our methodology, we summarize the adopted notation in Table II and support every methodological step with a running example.

A. Problem Formulation

Let $\mathbb{E}$ the set of entity types we aim to individuate, $e_i \in \mathbb{E}$ representing the i-th entity type (with $i \in \{1, \ldots, |\mathbb{E}|\}$). A corpus of annotated sentences $\mathcal{D}_i$ is associated to each entity type, $\mathcal{D}_i = \{(\mathbf{x}, \mathbf{y}) \in \mathcal{X}_i \times \mathcal{Y}_i\}$, $\mathcal{X}_i$ being the set of sentences $\mathbf{x}$ (sequences of tokens $x_j \in \mathbf{x}$, $j \in \{1, \ldots, \mathrm{H}\}$, where H represents the maximum sequence length) and $\mathcal{Y}_i$ being the relative set of labels. In this work, we will refer to the IOB2 annotation schema [40], assigning the "B" label to the *beginning*, the "I" label to the *inside* and the "O" to the *outside* of an entity mention.

Based on such datasets, our aim is to learn a model $f(\cdot)$ able to map each token x_j in a sentence $\mathbf{x}$ to its label $y_j \in \mathcal{Y}^{multi}$, where:

$$\mathcal{Y}^{multi} = \{B\text{-}e_1, I\text{-}e_1, B\text{-}e_2, I\text{-}e_2, \ldots, B\text{-}e_{|\mathbb{E}|}, I\text{-}e_{|\mathbb{E}|}, O\} \quad (2)$$

TABLE II
Notations

Symbol	Description
$\mathbb{E}$	set of entities we aim to recognize
$e_i \in \mathbb{E}$	i-th entity
H	maximum sequence length
$\mathbf{x}$	sentence (sequence of tokens)
$x_j \in \mathbf{x}$	j-th token
$\mathbf{y}$	label list associated to a sentence $\mathbf{x}$
$y_j \in \mathbf{y}$	label associated to x_j.
$\mathcal{D}_i$	corpus of annotated sentences associated with the entity e_i or the student s, $\mathcal{D}_i = \{(\mathbf{x}, \mathbf{y}) \in \mathcal{X}_i \times \mathcal{Y}_i\}$
$\mathcal{X}_i$	set of sentences $\mathbf{x}$ (sequences of tokens $x_j \in \mathbf{x}$, $j \in \{1, \ldots, \mathrm{H}\}$) associated to $\mathcal{D}_i$
$\mathcal{Y}_i$	set of label lists $\mathbf{y}$ associated to each sentence $\mathbf{x} \in \mathcal{X}_i$
θ_T^i	model parameters of the teacher associated to the entity e_i
θ_S	model parameters of the student
$T_i^j\{\cdot\}$	output distribution of the teacher related to the entity e_i for the input token x_j
$\mathcal{A}^j\{\cdot\}$	output distribution resulting from the aggregation of teacher distributions for the input token x_j
$S\{\cdot\}$	student output distribution for the input token x_j
$\mathcal{L}(\cdot;\cdot)$	training loss function
$\mathcal{L}_{KD}(\cdot;\cdot)$	loss component based on teacher distributions (knowledge distillation)
$\mathcal{L}_{GT}(\cdot;\cdot)$	loss component based on ground truth
λ	hyper-parameter allowing to control the weight of the two loss components

Running Example: The set of entity types in our running example is $\mathbb{E} = \{e_1, e_2, e_3\} = \{disease, gene, drug\}$. Hence, `TaughtNET` will learn how to predict one of the labels reported below to each token of input samples:

$$\mathcal{Y}^{multi} = \{B\text{-}disease, I\text{-}disease, B\text{-}gene, \\ I\text{-}gene, B\text{-}drug, I\text{-}drug, O\} \quad (3)$$

B. TaughtNet

The structure of this section reflects the procedural steps summarized in Fig. 1 by comprehensively describing the phases involved in the training procedure: (1) datasets aggregation, (2) retrieval of teacher distributions, (3) aggregation of teacher distributions and (4) student training.

1) Datasets Aggregation: Based on the available training datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{\mathbb{E}}$, we build an aggregated dataset:

$$\mathcal{D}_S = \{(\mathbf{x}, \mathbf{y}) \in \mathcal{X}'_S \times \mathcal{Y}'_S\}, \tag{4}$$

where $\mathcal{X}'_S$ results from the concatenation of the sentences contained in each single-task dataset $\mathcal{X}'_S = \mathcal{X}_1 \mathbin{+\!\!+} \mathcal{X}_2 \mathbin{+\!\!+} \ldots \mathbin{+\!\!+} \mathcal{X}_{|\mathbb{E}|}$, and the same goes for labels $\mathcal{Y}'_S$ with the only difference that B and I labels are diversified based on the corresponding entities, as described in Section III-A.

The aggregated dataset $\mathcal{D}_S$ will serve as the data source to obtain the distribution representing the knowledge of teachers (used for knowledge distillation) and as a ground truth reference during student training.

2) Retrieval of Teacher Predictions: Let $\theta_T^1, \theta_T^2, \ldots, \theta_T^{|\mathbb{E}|}$ be the parameters learnt by *teacher* models on their corresponding single-task datasets. For each sentence token $x_j \in \mathbf{x}$, the i-th teacher will be able to provide the distribution T_i^j:

$$T_i^j\{y_j = k|\mathbf{x}; \theta_T^i)\}, k \in \{B, I, O\} \tag{5}$$

3) Distributions Aggregation: Thanks to knowledge distillation, a *student* model learns how to mimic the output distribution of a *teacher* model. Differently from the standard approach, our *student* has to learn from an heterogeneous set of teachers, each of them able to individuate a different entity type. Hence, we need an aggregation phase, where teacher distributions are merged in one single distribution to be used in the knowledge distillation framework.

Let $x_j \in \mathbf{x}$ be a token we have to aggregate distributions for. Let's denote with $p_k^i = T_i^j(y_j = k|\mathbf{x}; \theta_T^i)$ the probability which the i-th teacher assigns to the label k, where $k \in \{B, I, O\}$.

The probability of the token x_j being assigned to the label B-e_i, I-e_i and O can be respectively computed as the probability of the intersection of the events shown as follows:

$$\begin{aligned} P(B{-}e_i) = P\,((T_i \text{ assigns } B) \cap_{j \neq i} \\ \cap_{j \neq i}(T_j \text{ does not assign } B)) \end{aligned} \tag{6}$$

$$\begin{aligned} P(I{-}e_i) = P\,((T_i \text{ assigns } I) \cap_{j \neq i} \\ \cap_{j \neq i}(T_j \text{ does not assign } I)) \end{aligned} \tag{7}$$

$$P(O) = P\,(\cap_i (T_i \text{ assigns } O)) \tag{8}$$

Given the independence between teachers and the mutual exclusivity characterizing each teacher distribution, we can then compute the probabilities of the aggregated distribution $\mathcal{A}$ as follows:

$$\mathcal{A}^j(y_j = B\text{-}e_i|\mathbf{x}; \theta_T^1, \ldots, \theta_T^{|\mathbb{E}|}) = p_B^i \prod_{j \neq i} \left(p_I^j + p_O^j\right) \tag{9}$$

$$\mathcal{A}^j(y_j = I\text{-}e_i|\mathbf{x}; \theta_T^1, \ldots, \theta_T^{|\mathbb{E}|}) = p_I^i \prod_{j \neq i} \left(p_B^j + p_O^j\right) \tag{10}$$

$$\mathcal{A}^j(y_j = O|\mathbf{x}; \theta_T^1, \ldots, \theta_T^{|\mathbb{E}|}) = \prod_i p_O^i \tag{11}$$

Given a sentence token $x_j \in \mathbf{x}, j \in \{1, \ldots, \mathrm{H}\}$, the output of this phase is the distribution of $\mathcal{Y}'$ labels:

$$\mathcal{A}^j\{y_j = k|\mathbf{x}; \theta_T^1, \ldots, \theta_T^{|\mathbb{E}|}\}, k \in \mathcal{Y}' \tag{12}$$

Running Example: Let $x_j \in \mathbf{x}$ be the input token and $T_{disease}^j = \{0.8, 0.1, 0.1\}$, $T_{gene}^j = \{0.05, 0.05, 0.9\}$, $T_{drug}^j = \{0.1, 0.1, 0.8\}$ its associated teacher predictions for labels B, I and O. Resulting from the aggregation of distributions we will have:

$$\begin{aligned} \mathcal{A}^j &= \{0.8 \cdot (0.05 + 0.9) \cdot (0.1 + 0.8), \\ &\quad 0.1 \cdot (0.05 + 0.9) \cdot (0.1 + 0.8), \ldots, 0.1 \cdot 0.9 \cdot 0.8\} \\ &= \{0.68, 0.09, 0.01, 0.04, 0.02, 0.09, 0.07\}, \end{aligned}$$

where results for labels $\{B\text{-}disease, I\text{-}disease, B\text{-}gene, I\text{-}gene, B\text{-}drug, I\text{-}drug, O\}$ are reported in order.

4) Student Training: Let us represent the student model with its parameters θ_S and its output distribution $S\{y_t = k|\mathbf{x}; \theta_\mathbf{S}\}$, $k \in \mathcal{Y}'$. The fine-tuning procedure aims to minimize a loss function composed by two terms: the former measuring the distance of the student distribution from its teachers distribution, the latter representing its error on the ground-truth. Formally, we can define our loss as shown below:

$$\begin{aligned} \mathcal{L}(\mathcal{D}_S; \theta_S, \theta_T^1, \ldots, \theta_T^{|\mathbb{E}|}) = \lambda \mathcal{L}_{KD}(\mathcal{D}_S; \theta_S, \theta_T^1, \ldots, \theta_T^{|\mathbb{E}|}) + \\ + (1 - \lambda)\mathcal{L}_{GT}(\mathcal{D}_S; \theta_S), \end{aligned} \tag{13}$$

where $\mathcal{L}_{KD}$ and $\mathcal{L}_{GT}$ are the knowledge distillation and ground-truth loss, respectively, while λ is an hyperparameter controlling their weight on the overall loss $\mathcal{L}$.

Despite the Kullback-Leibler divergence being suitable for this knowledge-distillation task, similarly to [39] and in compliance with [34] which proves that minimizing the Kullback-Leibler divergence is equal to minimize the cross-entropy error between two distributions, it is sufficient to train the student model to minimize the following loss function:

$$\begin{aligned} &\mathcal{L}_{KD}(D_S; \theta_S, \theta_\mathbf{T}) = \\ &- \sum_{(x,y) \in D_S} \sum_{t=1}^{H} \sum_{k \in \mathcal{Y}'} T\{y_j = k|\mathbf{x}; \theta_T^1, \ldots, \theta_T^{|\mathbb{E}|}\} \cdot \\ &\cdot logS\{y_t = k|\mathbf{x}; \theta_\mathbf{S}\}, \end{aligned} \tag{14}$$

where $\mathcal{H}$ is the sequence length and $S\{\cdot\}$ denotes the student distribution.

The ground-truth-based loss function is:

$$\begin{aligned} &\mathcal{L}_{GT}(D_S; \theta_S) = \\ &- \sum_{(x,y) \in D_S} \sum_{t=1}^{H} \mathbb{1}\{y_t = k\} logS\{y_t = k|\mathbf{x}; \theta_\mathbf{S}\}, \end{aligned} \tag{15}$$

where the indicator $\mathbb{1}\{\cdot\}$ represents the one-hot label annotated in the ground truth.

TABLE III
DATASETS AND PERFORMANCE OF TEACHERS USED IN THE EXPERIMENTS

				Teacher		
Dataset	**Size**	**Entity type**	**# Mentions**	**Precision**	**Recall**	**F1**
NCBI	793 abstracts	Disease	6, 881	87.31	89.58	88.43
BC5CDR	1500 articles	Chemical	15, 935	94.38	94.19	94.28
BC2GM	20, 000 sentences	Gene	24, 583	85.24	86.17	85.70

IV. EXPERIMENTS

In this section, we report our empirical evaluation of *TaughtNet*. In the first place, we train three students for three different biomedical entity types: diseases, chemical compounds and genetic information. Thereafter, we train several student architectures with different size and parameters, and report our results in the *Results* subsection. Specifically, we report: (1) a comparison of our best student with several state-of-the-art baselines; (2) results of different students with different architectures and size; (3) a comparison of all the students in terms of their level of agreement on predictions; (4) an error analysis w.r.t different error types; and (5) an explainability experiment which investigates how the inner workings change from the teachers to the student.

A. Datasets and Teachers

We evaluate the performance of our approach with three benchmark datasets, each of which has been constructed from PubMed abstract: NCBI-Disease [41], BC5CDR [42], BC2GM [43]. All the datasets — with their training, development and test splits — have been downloaded from: https://github.com/dmis-lab/biobert. We encoded word labels by using the IOB2 notation format [44].

For each one of the datasets, we trained our teachers by fine-tuning for 30 epochs a RoBERTa-large architecture which had been pre-trained on PubMed and PMC and MIMIC-III with a BPE Vocab learnt from PubMed [12].

A summary of the datasets, in terms of size and entity-type, and of the teachers, in terms of their precision, recall and F1 scores, is provided in Table III.

B. Evaluation Details

For all the datasets, we used the same dataset splits as BioBERT [8], which are based on earlier publications for a fair evaluation. In particular, training/development/test splits of NCBI-disease and BC5CDR corpora are the same as their original version, while the training set of BC2GM has been modified because the original corpus does not provide a development set. Thus, 2,500 sentences are split off from the training data to generate the development set.

C. Metrics

Quality: For the evaluation of the quality of the named entity recognition approaches, we used the *Precision*, *Recall* and *F1* metrics computed with the `seqeval` Python framework. In simple terms, *Precision* is the percentage of entities which are correctly found by the system, while *Recall* is the percentage of entities of the test set which are found by the system. A system with a low *Precision* is not able to differentiate between entity types, while a low *Recall* indicates the inability to recognize entities.

To measure the degree of agreement among different models, we used the *Cohen's Kappa* metric which can be computed as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{16}$$

where p_o is the relative observed agreement among predictions, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category.

Memory occupation and inference time: The efficiency of models has been evaluated based on their *size* (in terms of MB of memory occupied) and the *samples-per-second* (SPS) required during the training and inference phases. A model with too many parameters is difficult to deploy on hardware systems with strict memory constraints, while a slow model is difficult to integrate in complex systems where the NER engine is just a step in a pipeline. Our experiments have been performed on a Oracle Cloud Infrastructure (OCI) with an Intel(R) Xeon(R) Platinum 8167 M CPU @ 2.00 GHz (12 cores) and a NVIDIA Tesla V100 SXM2 GPU.

D. Settings and Hyperparameters

We developed our framework on top of the HuggingFace *transformers* library [45]. We experimented with several model architectures and weights with varying size. Specifically, we used `RoBERTa-large-PM-M3-Voc` and `RoBERTa-base-PM-M3-Voc-train-longer` from Lewis et al. [12], and `huawei-noah/TinyBERT_General_4L_312D` and `distilroberta-base` from the huggingface model hub.

To fine-tune our models, we used a Adam optimizer with an initial learning rate of 5e-5 and $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\mathrm{e}{-8}$. The batch size was set to 8 and the maximum sequence length to 128.

Although the results in terms of quality are always satisfactory from the first epochs, we usually found the highest performance after 20 epochs. This result is consistent with findings from Lee et al. [8].

As concerns the loss function designed to train the student model, we used the `KLDivLoss` and `NLLLoss` PyTorch implementations for the knowledge distillation $\mathcal{L}_{KD}$ and ground-truth $\mathcal{L}_{GT}$ loss components.

E. Results

1) Comparison With Baselines: We compare the quality of our best student with several baselines, described as follows:

- *Merged:* the simplest way to train a multi-label NER model from single-entity datasets is to merge them in one aggregated dataset to be used for training and testing. We fine-tuned until convergence the same RoBERTa-large model architecture used for teachers on such dataset.

TABLE IV
COMPARATIVE EXPERIMENTS

Dataset	Metric	Merged	MTM-CW	CollaboNet	MT-BioNER	TaughtNet
NCBI	Precision	83.88	85.86	85.48	86.73	**88.51**
	Recall	85.10	86.42	87.27	89.70	**89.90**
	F1	84.49	86.14	86.36	88.10	**89.20**
BC5CDR	Precision	94.08	89.10	94.26	88.46	**94.51**
	Recall	83.87	88.47	92.38	90.52	**93.40**
	F1	88.69	88.78	93.31	89.50	**93.95**
BC2GM	Precision	83.29	82.10	80.49	82.01	**84.90**
	Recall	78.94	79.42	78.99	84.04	**83.45**
	F1	81.06	80.74	79.73	83.01	**84.84**

Best scores are reported in bold.

- *MTM-CW:* multi-task model built upon a single-task BiLSTM-CRF model with an additional context-dependent BiLSTM layer to model character sequences.
- *CollaboNET:* aggregates the results of *collaborator* single-task models, and uses them as an additional input to the target multi-task model.
- *MT-BioNER:* multi-task transformer-based neural architecture, where different models for different datasets share some layers to build a "shared" knowledge across tasks.

Table IV reports results over the three benchmark datasets in terms of *Precision*, *Recall* and *F1* scores. Thanks to the utilization of high-performing teachers, our *student* model achieves the best results for each of the datasets. Interestingly, performance obtained for the *NCBI* dataset surpasses the related teacher thanks to the indirect positive effect of the (1) data augmentation obtained by merging all the dataset and the (2) joint training based on both the ground-truth and teacher predictions. A comparative discussion with baselines is provided in Section IV-F.

2) Smaller and Smaller Students: Thanks to its knowledge distillation based architecture, one of the advantages of using *TaughtNet* is its straightforward way to train multi-task small models by leveraging the knowledge of large and high-performing teachers. In our experiments, we compare results of the student architectures described as follows:

- *Large:* same as teachers,' i.e. RoBERTa-large architecture pre-trained on PubMed and PMC and MIMIC-III with a BPE Vocab learnt from PubMed.
- *Base:* RoBERTa-base architecture pre-trained on PubMed and PMC and MIMIC-III with a BPE Vocab learnt from PubMed with an additional 50 K steps.
- *Distil:* distilled version of BERT base introduced by Sanh et al. [27]. It has 40% less parameters and runs 60% faster than BERT-base.
- *Tiny:* distilled version of BERT-base introduced by Jiao et al. [31], 7.5x smaller and 9.4x faster on inference than BERT-base.

TABLE V
PERFORMANCE OF VARIOUS STUDENT ARCHITECTURES WITH DECREASING SIZE

	SPS scores			*F1 scores*		
Model	train	inference	Size (MB)	NCBI	BC5CDR	BC2GM
Large	38	125	1,416.3	89.20	93.95	84.84
Base	111	324	495.5	87.62	93.63	84.25
Distil	202	566	265.5	80.71	85.24	77.45
Tiny	475	914	57	77.80	81.42	71.94

SPS stands for samples-per-second.

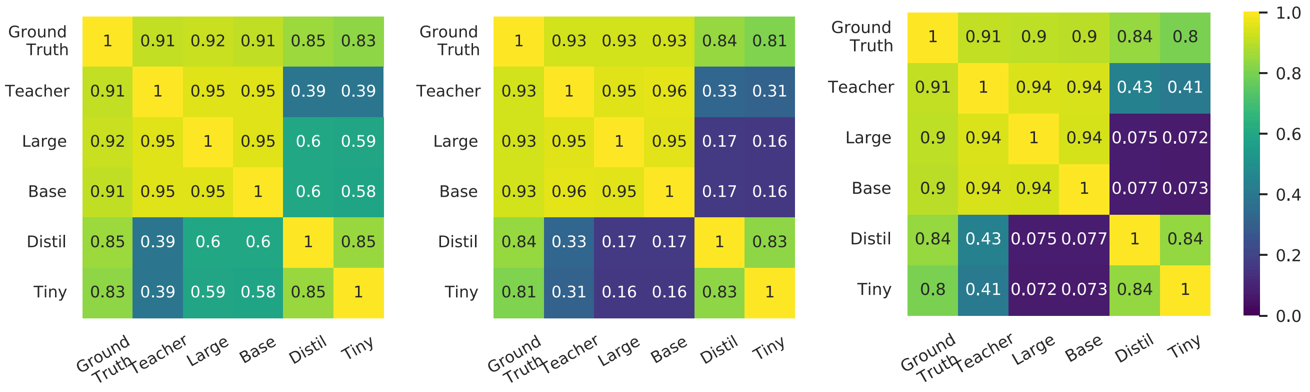
Results are reported in Table V in terms of model size, samples-per-second (SPS) processed during the training and inference phase, and F1 scores over the three benchmark datasets. Interestingly, the *Base* architecture achieves F1 scores closely resembling its *Large* counterpart, probably resulting in the best choice in the trade-off between quality of predictions and model size/inference time. The distilled architectures (*Distil* and *Tiny*) result in lower F1 scores, but their considerable improvement in memory occupation and processing time could make them a suitable choice in limited-resource scenarios. In the experiments that follow, we delve into the differences between these students and their corresponding teachers.

3) Levels of Agreement (Cohen's Kappa): We computed the Cohen's Kappa metric to measure the degree of agreement among models and the ground-truth.[5] Heatmaps in Fig. 2 show agreements over the three benchmark datasets among the ground truth, the teacher, and the size-decreasing student architectures. Despite the disagreement between distilled models and their teacher — which highlights a limitation in distilling their knowledge, which will be explored in future work — results show an overall agreement between teachers and their students and among student architectures.

4) Error Analysis: We further explored the differences among models based on the number of correctly-retrieved entity mentions (CORRECT), new predictions deriving from the application of the framework (NEW) and their errors, which can be divided into five categories described as follows:

- *Complete False Positive (CPF):* the model recognizes an entity which was not annotated as a named entity.
- *Complete False Negative (CFN:* the model does not recognize an entity which was annotated as a named entity.

[5] When computing the Cohen's Kappa between a teacher and a student, we assign the *outside* (O) label to predictions of entity types which are different from the type in which the teacher is specialized.



(a) NCBI (disease) (b) BC5CDR (chem) (c) BC2GM (gene)

Fig. 2. Heatmaps of Cohen's Kappas among models and ground truth.

TABLE VI
NUMBER OF CORRECT PREDICTIONS, NEW PREDICTIONS (FOR ENTITY TYPES NOT ANNOTATED IN THE GROUND TRUTH) AND DIFFERENT TYPES OF ERROR

Dataset	Model	CORRECT	NEW	CFP	CFN	WLRS	WLOS	RLOS
NCBI	Large	785	803	43	24	9	24	113
	Base	785	801	51	19	8	18	125
	Distil	742	172	61	126	1	7	79
	Tiny	720	185	83	158	3	5	69
BC5CDR	Large	4753	4612	190	189	58	93	279
	Base	4750	4654	194	168	50	97	307
	Distil	4187	307	143	995	48	42	100
	Tiny	4032	357	339	1120	48	41	131
BC2GM	Large	5330	3426	257	213	48	37	608
	Base	5306	3453	309	202	56	32	640
	Distil	4797	625	468	699	18	31	685
	Tiny	4393	764	617	987	35	48	767

- *Wrong Label Right Span (WLRS):* the model correctly recognizes the presence of an annotated named entity, but assigns the wrong label.
- *Wrong Label Overlapping Span (WLOS):* the model recognizes the presence of an annotated named entity, but assigns the wrong label and the span is wrong.
- *Right Label Overlapping Span (RLOS):* the model recognizes the presence of an annotated named entity, but the span is wrong.

It can be seen from the data in Table VI that students trained with *TaughtNet* allow us to retrieve a considerable number of novel entity mentions which were not annotated in the ground-truth, thanks to the knowledge of the teachers employed. Concordant with the above-reported experiments, *Large* and *Base* students are able to detect a significantly higher number of new entity mentions w.r.t. distilled architectures. The highest limitation of distilled architectures w.r.t. to their "larger" counterparts is in the number of CFN errors, i.e. they are not able to identify mentions which are actually annotated.

The majority of errors fall in the RLOS category, meaning that models are able to identify an entity mention, but the range detected is not the same as the ground truth. However, previous works have shown that this type of errors are often a result of the subjectivity and inconsistency of span annotations [46], [47]. Some examples are shown in Table VII. It is important to note that many of the errors are due to the ability of our model to recognize multiple entity types: for example, the two words *gene* mention "estrogen receptor" (see WRLS, 2nd example) are assigned by our model to two different entity types ("estrogen" as a *chemical* compound, "receptor" as a *gene*).

5) Explainability: We apply *Integrated Gradients* [48] to assign an importance score to each input token by approximating the integral of gradients of the output w.r.t the inputs.[6] To investigate how the inner workings of the models change from Teachers to Student, we report in Fig. 3 the explanations from the three *large* Teachers and the resulting Student to the sentence: *"Subchronic inhibition of nitric-oxide synthesis modifies haloperidol-induced catalepsy and the number of NADPH-diaphorase neurons in mice"*, which contains at least one mention per entity type. Interestingly, despite our experiment being carried out with just the aim to prove the effortlessly interpretability of our method — which does not modify the architecture of the Student model and thus can leverage off-the-shelf methods to explain its predictions —, we also observed that the explanations provided by the Student are better targeted (i.e. lower number of influential tokens) and understandable.

F. Discussion

In our experiments, we have studied in-depth the effects of learning from several single-task transformer-based teachers and contrasted TaughtNet with strong baselines from the current literature.

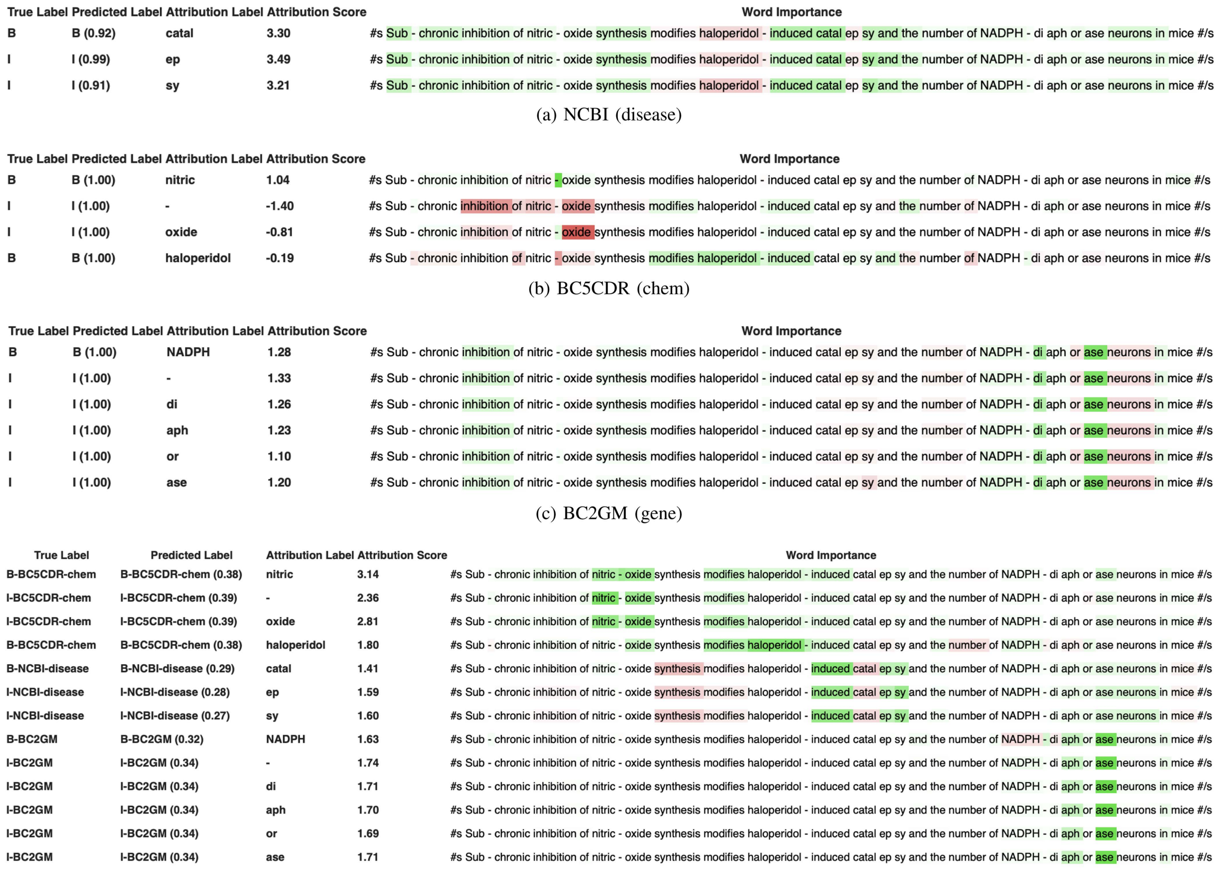
In Table VIII we show a methodological comparison of state-of-the-art methods accompanied by averaged precision, recall and F1 scores on the benchmarking datasets used in this work. We can observe that multi-task methods that are based on high-performing pre-trained transformer models consistently outperform CollaboNet in most situations, despite the fact that

[6] Integrated Gradients for Transformers interpretability, code: https://github.com/cdpierse/transformers-interpret

TABLE VII
SAMPLES ANNOTATED BY THE LARGE MODEL WHICH CONTAIN ERRORS IN DISEASE, CHEMICAL AND/OR GENE MENTION PREDICTIONS. ERRORS ARE HIGHLIGHTED WITH A HIGHER LEVEL OF TRANSPARENCY (DISEASE, CHEMICAL, GENE). ONE INPUT SAMPLE IS RANDOMLY SELECTED FOR EACH DATASET AND ERROR

Error	Student annotation	Ground Truth
CFP	Other complement components were normal during remission of lupus, but C1, C4, C2, and C3 levels fell during exacerbations.	Other complement components were normal during remission of lupus, but C1, C4, C2, and C3 levels fell during exacerbations.
	The encoded protein contains an amino terminal PDZ domain, followed by a predicted coiled-coil region, a PEST domain, and a carboxy-terminal SAM domain.	The encoded protein contains an amino terminal PDZ domain, followed by a predicted coiled-coil region, a PEST domain, and a carboxy-terminal SAM domain.
	We conclude that CNA and INA demonstrated similar profiles with regard to safety, morbidity, and mortality.	We conclude that CNA and INA demonstrated similar profiles with regard to safety, morbidity, and mortality.
CFN	If untreated, hemochromatosis can cause serious illness and early death, but the disease is still substantially under-diagnosed.	If untreated, hemochromatosis can cause serious illness and early death, but the disease is still substantially under-diagnosed.
	ORF3 encodes a putative periplasmic c-type cytochrome with a molecular mass of 94,000 Da and contains seven c-heme-binding motifs but shows no sequence homology to occ or ORF1.	ORF3 encodes a putative periplasmic c-type cytochrome with a molecular mass of 94,000 Da and contains seven c-heme-binding motifs but shows no sequence homology to occ or ORF1.
	BS pool size was decreased by 27% but total BS synthesis was not affected by EE in intact rats.	BS pool size was decreased by 27% but total BS synthesis was not affected by EE in intact rats.
WLRS	HFE is an MHC-related protein that is mutated in the iron-overload disease hereditary hemochromatosis.	HFE is an MHC-related protein that is mutated in the iron-overload disease hereditary hemochromatosis.
	Effects of long-term use of raloxifene, a selective estrogen receptor modulator, on thyroid function test profiles.	Effects of long-term use of raloxifene, a selective estrogen receptor modulator, on thyroid function test profiles.
	Nociceptin, also known as orphanin FQ, is an endogenous ligand for the orphan opioid receptor-like receptor 1 (ORL1) and involves in various functions in the central nervous system (CNS).	Nociceptin, also known as orphanin FQ, is an endogenous ligand for the orphan opioid receptor-like receptor 1 (ORL1) and involves in various functions in the central nervous system (CNS).
WLOS	Previous family studies suggested that these individuals may be compound heterozygotes for the common mutant TSD gene and a rare (allelic) mutant gene.	Previous family studies suggested that these individuals may be compound heterozygotes for the common mutant TSD gene and a rare (allelic) mutant gene.
	When expressed in Escherichia coli, SH-PTP2 displays tyrosine-specific phosphatase activity.	When expressed in Escherichia coli, SH-PTP2 displays tyrosine-specific phosphatase activity.
	Sub-chronic inhibition of nitric-oxide synthesis modifies haloperidol-induced catalepsy and the number of NADPH-diaphorase neurons in mice.	Sub-chronic inhibition of nitric-oxide synthesis modifies haloperidol-induced catalepsy and the number of NADPH-diaphorase neurons in mice.
RLOS	The evidence of a significant proportion of loss-of-function mutations and a complete absence of the normal copy of ATM in the majority of mutated tumours establishes somatic inactivation of this gene in the pathogenesis of sporadic T-PLL and suggests that ATM acts as a tumour suppressor.	The evidence of a significant proportion of loss-of-function mutations and a complete absence of the normal copy of ATM in the majority of mutated tumours establishes somatic inactivation of this gene in the pathogenesis of sporadic T-PLL and suggests that ATM acts as a tumour suppressor.
	A GT-rich sequence binding the transcription factor Sp1 is crucial for high expression of the human type VII collagen gene (COL7A1) in fibroblasts and keratinocytes.	A GT-rich sequence binding the transcription factor Sp1 is crucial for high expression of the human type VII collagen gene (COL7A1) in fibroblasts and keratinocytes.
	An increase in TDR by dl-sotalol facilitated transmural propagation of EADs that initiated multiple episodes of spontaneous TdP in 3 of 6 rabbit left ventricles.	An increase in TDR by dl-sotalol facilitated transmural propagation of EADs that initiated multiple episodes of spontaneous TdP in 3 of 6 rabbit left ventricles.

CollaboNet effectively addresses the low-precision problem of multi-task learning systems by defining a collaborative framework made of single-task BiLSTM-CRF models that also solves the *type conflict* problem, i.e. different models recognising the same mentions. In TaughtNet, we have leveraged the advantages of both multi-task learning and transformers by dealing with the same low-precision and type conflict problems as CollaboNet. The result is a single high-performing fine-tuned transformer model able to identify mentions of several entity types, which makes it (1) easy to lighten under constraining hardware and computing time requirements thanks to lighter students (e.g. DistilBERT, TinyBERT), and (2) easy to interpret by the use of



True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B	B (0.92)	catal	3.30	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (0.99)	ep	3.49	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (0.91)	sy	3.21	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s

(a) NCBI (disease)

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B	B (1.00)	nitric	1.04	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	-	-1.40	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	oxide	-0.81	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
B	B (1.00)	haloperidol	-0.19	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s

(b) BC5CDR (chem)

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B	B (1.00)	NADPH	1.28	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	-	1.33	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	di	1.26	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	aph	1.23	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	or	1.10	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I	I (1.00)	ase	1.20	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s

(c) BC2GM (gene)

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B-BC5CDR-chem	B-BC5CDR-chem (0.38)	nitric	3.14	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC5CDR-chem	I-BC5CDR-chem (0.39)	-	2.36	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC5CDR-chem	I-BC5CDR-chem (0.39)	oxide	2.81	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
B-BC5CDR-chem	B-BC5CDR-chem (0.38)	haloperidol	1.80	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
B-NCBI-disease	B-NCBI-disease (0.29)	catal	1.41	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-NCBI-disease	I-NCBI-disease (0.28)	ep	1.59	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-NCBI-disease	I-NCBI-disease (0.27)	sy	1.60	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
B-BC2GM	B-BC2GM (0.32)	NADPH	1.63	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC2GM	I-BC2GM (0.34)	-	1.74	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC2GM	I-BC2GM (0.34)	di	1.71	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC2GM	I-BC2GM (0.34)	aph	1.70	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC2GM	I-BC2GM (0.34)	or	1.69	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s
I-BC2GM	I-BC2GM (0.34)	ase	1.71	#s Sub - chronic inhibition of nitric - oxide synthesis modifies haloperidol - induced catal ep sy and the number of NADPH - di aph or ase neurons in mice #/s

(d) Student (disease, chem, gene)

Fig. 3. Visualization of attribution scores computed by applying Integrated Gradients to our Teachers (a)-(c) and Student (d). For each token, we show its true and predicted label, its attribution score and the original sentence where each token is highlighted based on its contribution to the prediction (green if positive, red otherwise).

TABLE VIII
OVERVIEW OF STATE-OF-THE-ART APPROACHES FOR MULTI-TASK BIONER. PRECISION, RECALL AND F1 SCORES SHOWN HERE HAVE BEEN AVERAGED ACROSS THE BENCHMARKING DATASETS USED IN THIS WORK

Method	Ref	Year	Model	Characteristics	Precision	Recall	F1
MTM-CW	[6]	2018	Multi-task BiLSTM-CRF model	**(+)** outperforms the previous state-of-the-art by leveraging multi-task learning; **(-)** type conflict problem	85.69	84.77	85.22
CollaboNet	[15]	2019	Collaborating BiLSTM-CRF models	**(+)** handles the low-precision problem of multi-task models; **(+)** no type conflict problem; **(-)** requires single-task models to be trained in advance and called for every inference	86.74	86.21	86.46
MT-BioNER	[14]	2020	Multi-task transformer model	**(+)** obtains state-of-the-art performance thanks to the use of a transformer-based architecture; **(-)** suffers from the low-precision problem of multi-task learning systems; **(-)** type conflict problem	85.73	88.08	86.87
TaughtNet	—	—	Transformer model	**(+)** combines the advantages of multi-task learning, transformer architectures and CollaboNet; **(+)** easy to lighten; **(+)** no low-precision problem; **(+)** no type conflict problem; **(+)** easy applicability of eXplainable AI techniques; **(-)** requires teachers to be trained in advance	**89.31**	**88.91**	**89.33**

Best results are reported in bold.

off-the-shelf explainability techniques, since we do not change any module in the architecture.

V. Conclusion & Future Work

The difficulty in finding a single dataset with all the entities required for a Biomedical Named Entity Recognition System (e.g. diseases, genes, species, drugs) has laid the foundations of this work. TaughtNet has the objective to integrate various publicly available single-task healthcare datasets in a single BERT architecture which can be used as a fast and highly performing BioNER engine in real applications, such as conversational agents or knowledge graph development.

Experimental results demonstrate that not only does TaughtNet surpass strong state-of-the-art baselines, but it also is a valuable option when constrained by strict computational and memory requirements thanks to its ability to train lightweight models that distill the knowledge from high-performing single-task teachers. Furthermore, we have shown the potential of TaughtNet to provide explainability, which is a valuable advantage, especially when dealing with healthcare data.

There is abundant room for further progress in exploring the use and application of knowledge distillation to bring the student performance as close as possible to that of teachers. As a future work, we would like to integrate more datasets and to extend the framework not only to other downstream tasks, but also to other application domains, since the technique is not dependent on the biomedical domain.

Acknowledgment

This work is carried out within the framework of the Knowledge graphs for next-generation health science applications project with Oracle America Inc. and with the support of the Oracle for Research program, within a research agreement between Oracle Research Lab and the Department of Electrical Engineering and Information Technologies at the University of Naples Federico II (DIETI). All opinions reflected in this paper are those of the authors and not necessarily those of the funding agency.

References

[1] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. Conf. North Amer. Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2016, pp. 260–270.

[2] S. Sahu and A. Anand, "Recurrent neural network models for disease name recognition using domain invariant features," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 2216–2225. [Online]. Available: https://aclanthology.org/P16-1209

[3] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1064–1074. [Online]. Available: https://aclanthology.org/P16-1101

[4] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 357–370, 2016.

[5] M. Habibi, L. Weber, M. L. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, pp. i37–i48, 2017.

[6] X. Wang et al., "Cross-typebiomedical named entity recognition with deep multi-task learning," *Bioinformatics*, vol. 35, no. 10, pp. 1745–1752, 2019.

[7] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. Conf. Empir. Methods Natural Lang. Process./Proc. Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3615–3620.

[8] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234–1240, 2020.

[9] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," 2019, *arXiv:1904.05342*.

[10] E. Alsentzer et al., "Publicly available clinical BERT embeddings," in *Proc. 2nd Clin. Natural Lang. Process. Workshop*, Minneapolis, MN, USA: Association for Computational Linguistics, 2019, pp. 72–78.

[11] S. Gururangan et al., "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, pp. 8342–8360.

[12] P. Lewis, M. Ott, J. Du, and V. Stoyanov, "Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art," in *Proc. 3rd Clin. Natural Lang. Process. Workshop*, 2020, pp. 146–157.

[13] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020,vol. 33, pp. 1877–1901.

[14] M. R. Khan, M. Ziyadi, and M. Abdelhady, "MT-BioNER: Multi-task learning for biomedical named entity recognition using deep bidirectional transformers," 2020, *arXiv:2001.08904*.

[15] W. Yoon, C. H. So, J. Lee, and J. Kang, "CollaboNet: Collaboration of deep neural networks for biomedical named entity recognition," *BMC Bioinf.*, vol. 20, no. S10, May 2019, Art. no. 249. [Online]. Available: http://dx.doi.org/10.1186/s12859-019-2813-6

[16] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in *Proc. 16th Conf. Comput. Linguistics*, 1996, pp. 466–471. [Online]. Available: https://doi.org/10.3115/992628.992709

[17] S. Zhao, T. Liu, S. Zhao, and F. Wang, "A neural multi-task learning framework to jointly model medical named entity recognition and normalization," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 817–824.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[19] M. E. Peters et al., "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technologies*, New Orleans, LA, USA: Association for Computational Linguistics, vol. 1, 2018, pp. 2227–2237.

[20] Z. Yang et al., "Xlnet: Generalized autoregressive pretraining for language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[21] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, "Learning a health knowledge graph from electronic medical records," *Sci. Rep.*, vol. 7, 2017, Art. no. 5994.

[22] R. Caruana, "Multitask learning," in *Encyclopedia of Machine Learning and Data Mining*. Berlin, Germany: Springer, 1998.

[23] G. K. O. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen, "A neural network multi-task learning approach to biomedical named entity recognition," *BMC Bioinf.*, vol. 18, 2017, Art. no. 368.

[24] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[25] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, Austin, Texas, USA, 2016, pp. 1317–1327. [Online]. Available: https://doi.org/10.18653/v1/d16-1139

[26] M. Hu et al., "Attention-guided answer distillation for machine reading comprehension," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 2077–2086. [Online]. Available: https://doi.org/10.18653/v1/d18-1232

[27] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, A distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[28] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for BERT model compression," in *Proc. Conf. Empir. Methods Natural Lang. Process./Proc. Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 4323–4332.

[29] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: A compact task-agnostic BERT for resource-limited devices," in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, pp. 2158–2170.

[30] W. Wang, F. Wei, L. Dong, H.N.B. Yang, and M. Zhou, "MINILM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 5776–5788. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[31] X. Jiao et al., “TinyBERT: Distilling BERT for natural language understanding,” in *Proc. Findings Assoc. Comput. Linguistics: Empir. Methods Natural Lang. Process.*, 2020, pp. 4163–4174. [Online]. Available: https://doi.org/10.18653/v1/2020.findings-emnlp.372

[32] N. Reimers and I. Gurevych, “Making monolingual sentence embeddings multilingual using knowledge distillation,” in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 4512–4525. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.365

[33] Y.-C. Chen, Z. Gan, Y. Cheng, J. Liu, and J. J. Liu, “Distilling knowledge learned in BERT for text generation,” in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, pp. 7893–7905.

[34] Y. Chebotar and A. Waters, “Distilling knowledge from ensembles of neural networks for speech recognition,” in *Proc. Interspeech*, 2016, pp. 3439–3443.

[35] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, “Efficient knowledge distillation from an ensemble of teachers,” in *Proc. Interspeech*, 2017, pp. 3697–3701.

[36] U. Asif, J. Tang, and S. Harrer, “Ensemble knowledge distillation for learning improved and efficient networks,” in *Proc. 24th Eur. Conf. Artif. Intell.*, Santiago de Compostela, Spain, 2020, pp. 953–960. [Online]. Available: https://doi.org/10.3233/FAIA200188

[37] Y. Liu, W. Zhang, and J. Wang, “Adaptive multi-teacher multi-level knowledge distillation,” *Neurocomputing*, vol. 415, pp. 106–113, 2020.

[38] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[39] X. Tan, Y. Ren, D. He, T. Qin, Z. Zhao, and T. Liu, “Multilingual neural machine translation with knowledge distillation,” in *Proc. 7th Int. Conf. Learn. Representations*, New Orleans, LA, USA, 2019. [Online]. Available: https://openreview.net/forum?id=S1gUsoR9YX

[40] A. Ratnaparkhi and M. P. Marcus, “Maximum entropy models for natural language ambiguity resolution,” Ph.D. dissertation, Univ. Pennsylvania, USA, 1998.

[41] R. I. Dogan, R. Leaman, and Z. Lu, “NCBI disease corpus: A resource for disease name recognition and concept normalization,” *J. Biomed. Inform.*, vol. 47, pp. 1–10, 2014.

[42] J. Li et al., “Biocreative V CDR task corpus: A resource for chemical disease relation extraction,” *Database: J. Biol. Databases Curation*, vol. 2016, 2016, Art. no. baw068.

[43] L. Smith et al., “Overview of biocreative II gene mention recognition,” *Genome Biol.*, vol. 9, 2008, Art. no. S2.

[44] E. F. T. K. Sang and J. Veenstra, “Representing text chunks,” in *Proc. 9th Conf. Eur. Chapter Assoc. Comput. Linguistics*, University of Bergen, Bergen, Norway, 1999, pp. 173–179. [Online]. Available: https://aclanthology.org/E99-1023/

[45] T. Wolf et al., “Transformers: State-of-the-art natural language processing,” in *Proc. Conf. Empir. Methods Natural Lang. Process.: Syst. Demonstrations*, 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[46] R. T.-H. Tsai et al., “Various criteria in the evaluation of biomedical named entity recognition,” *BMC Bioinf.*, vol. 7, 2005, Art. no. 92.

[47] I. Nejadgholi, K. C. Fraser, and B. de Bruijn, “Extensive error analysis and a learning-based evaluation of medical entity recognition systems to approximate user experience,” in *Proc. 19th SIGBioMed Workshop Biomed. Lang. Process.*, 2020, pp. 177–186.

[48] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proc. 34th Int. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, pp. 3319–3328. [Online]. Available: http://proceedings.mlr.press/v70/sundararajan17a.html

Open Access funding provided by ‘Università degli Studi di Napoli “Federico II”’ within the CRUI CARE Agreement