Clinical Named Entity Recognition using Contextualized Token Representations

Yichao Zhou, Chelsea Ju, J. Harry Caufield, Kevin Shih, Calvin Chen, Yizhou Sun, Kai-Wei Chang,
Peipei Ping, and Wei Wang
University of California, Los Angeles
{yz,chelseaju,lspss89221,calvinyhchen,yzsun,kwchang,weiwang}@cs.ucla.edu
{jcaufield,pping}@mednet.ucla.edu

Abstract

The clinical named entity recognition (CNER) task seeks to locate and classify clinical terminologies into predefined categories, such as diagnostic procedure, disease disorder, severity, medication, medication dosage, and sign symptom. CNER facilitates the study of side-effect on medications including identification of novel phenomena and human-focused information extraction. Existing approaches in extracting the entities of interests focus on using static word embeddings to represent each word. However, one word can have different interpretations that depend on the context of the sentences. Evidently, static word embeddings are insufficient to integrate the diverse interpretation of a word. To overcome this challenge, the technique of contextualized word embedding has been introduced to better capture the semantic meaning of each word based on its context. Two of these language models, ELMo and Flair, have been widely used in the field of Natural Language Processing to generate the contextualized word embeddings on domain-generic documents. However, these embeddings are usually too general to capture the proximity among vocabularies of specific domains. To facilitate various downstream applications using clinical case reports (CCRs), we pre-train two deep contextualized language models, Clinical Embeddings from Language Model (C-ELMo) and Clinical Contextual String Embeddings (C-Flair) using the clinical-related corpus from the PubMed Central. Explicit experiments show that our models gain dramatic improvements compared to both static word embeddings and domain-generic language models. The pre-trained embeddings of these two models will be available soon.

Keywords natural language processing, clinical named entity recognition, clinical case report, contextualized token embedding, deep language model

1 Introduction

Clinical case reports (CCRs) are written descriptions of the unique aspects of a particular clinical case (Cabán-Martinez and García-Beltrán, 2012). They are intended to serve as educational aids to science and medicine, as they play an essential role in sharing clinical experiences about atypical disease phenotypes and new therapies (Caufield et al., 2018). Unlike other types of clinical documents (e.g., electronic medical records, or EMRs), CCRs generally describe single clinical narratives at a time: these are stories of diseases as they were observed and treated, written in language requiring domain familiarity but otherwise generally interpretable. Conveniently, accessing and reading any of the more than 2 million CCRs in publication does not involve the privacy responsibilities required by EMRs and other protected health information. CCRs therefore serve as rich, plentiful examples of clinical language.

Clinical named entity recognition (CNER) is an important text mining task in the domain of biomedical natural language processing. It aims to identify clinical entities and events from the case reports. For example, in the sentence "CT of the maxillofacial area showed no facial bone fracture." "CT of the maxillofacial area" is a "diagnostic procedure" and "facial bone fracture" belongs to the "disease and disorder" category. As with documents describing experimental procedures and results—often the focus of general biomedical annotated corpora such as PubTator (Wei et al., 2013)—CCRs include a large variety of entity types and potential orders of events (Caufield et al., 2018). Methods to better enable biomedical and clinical NLP at scale, across numerous entity types, and with generalizable approaches across topics are necessary, as single-task or single-entity type methods provide insufficient detail for comprehensive CNER. Fine-grained CNER supports development of precision medicine's hope to leverage advanced computer technologies to deeply digitize, curate and understand medical records and case reports (Bates et al., 2014, Rajkomar et al., 2018).

Biomedical NER (BioNER), of which CNER is a subtask, has been a focus of intense, groundbreaking research for decades but has recently undergone a methodological shift. Its foundational methods are largely rule-based (e.g., Text Detective (Tamames, 2005)), dictionary-based (e.g., BioThesaurus (Liu et al., 2006) or MetaMap (Aronson, 2001)), and basic statistical approaches (e.g., the C-value / NC-value method (Frantzi et al., 2000)). Source entities for NER are sourced from extensive knowledgebases such as UMLS (Bodenreider, 2004) and UniProtKB (The UniProt Consortium, 2017). Readily applicable model-based BioNER methods, including those built upon non-contextualized word embeddings such as Word2Vec and GloVe (Mikolov et al., 2013, Pennington et al., 2014) now promise to more fully address the challenges particular to the biomedical domain: concepts may have numerous names, abbreviated forms, modifiers, and variants. Furthermore, biomedical and clinical text assumes readers have extensive domain knowledge. Its documents follow no single structure across sources or topics, rendering their content difficult to predict.

These models neither avoid time-consuming feature engineering, nor make full use of semantic and syntactic information from each token's context. Context can thoroughly change an individual word's meaning, e.g., an "infarction" in the heart is a heart attack but the same event in the brain constitutes a stroke. Context is crucial for understanding abbreviations as well: "MR" may represent the medical imaging technique *magnetic resonance*, the heart condition *mitral regurgitation*, the concept of a *medical record*, or simply the honorific *Mister*. Non-contextualized word embeddings exacerbate the challenge of understanding distinct biomedical meanings as they contain only one representation per word. The most frequent semantic meaning within the training corpus becomes the standard representation.

Inspired by the recent development of contextualized token representations (Akbik et al., 2018, Devlin et al., 2018, Peters et al., 2018) supporting identification of how the meaning of words changes based on surrounding context, we refresh the technology of CNER to better extract clinical entities from unstructured clinical text. The deep contextualized token representations are pre-trained with a large corpus using a language model (LM) objective. ELMo (Peters et al., 2018) takes word tokens as input and pre-trains them with a bidirectional language model (biLM). Flair (Akbik et al., 2018) proposes a pre-trained character-level language model by passing sentences as sequences of characters into a bidirectional LSTM to generate word-level embeddings. BERT (Devlin et al., 2018) is built with bidirectional multi-layered Transformer encoders on top of the WordPiece embeddings, position embeddings, and segment embeddings. In this paper, we address the CNER task with contextualized embeddings (i.e., starting with ELMo and Flair), then and compare structural differences in the resulting models. Following recent work demonstrating impressive performance and accuracy of pre-training word representations with domain-specific documents (Sheikhshabbafghi et al., 2018), we collected domain-specific documents all related to CCRs, roughly a thousandth of PMC documents, and pre-trained two deep language models, C-ELMo and C-Flair. In this paper, we focus on the CNER task and evaluate the two language models across three datasets. Our two pre-trained language models can support applications beyond CNER, such as clinical relation extraction or question answering.

Our contributions are as follows:

- To the best of our knowledge, we are the first to build a framework for solving clinical natural language processing tasks using deep contextualized token representations.
- We pre-train two contextualized language models, C-ELMo and C-Flair for public use. We evaluate our models on three CNER benchmark datasets, MACCROBAT2018, i2b2-2010, NCBI-disease, and achieve dramatic improvements of 10.31%, 7.50%, and 6.94%, respectively.
- We show that pre-training a language model with a light domain-specific corpus can result in better performance in the downstream CNER application, compared with domain-generic embeddings.

In Section 3.1, we introduce the ELMo and Flair language models. In Section 3.2 we propose the CNER model.Section 4 describes our pre-training corpus and the datasets for evaluating our CNER tasks. We show experimental results and detail a brief case analysis.

2 Related work

2.1 Clinical named entity recognition

Clinical named entity recognition (CNER) is a fundamental technique to acquire knowledge from descriptions of clinical events and disease presentations from a wide variety of document types, published case reports and sets of electronic medical records. CNER has drawn broad attention, but heavy feature engineering is intentional for traditional CNER methods (Aronson, 2001, De Bruijn et al., 2011, Demner-Fushman et al., 2017, Savova et al., 2010). In recent years, deep learning methods have achieved significant success in CNER tasks. Zhang et al. (2018) leveraged transfer learning to use existing knowledege. Wang et al. (2018) applied another semi-supervised learning method, multi-task learning, to obtain useful information from other datasets. Xu et al. (2018) improved the performance of CNER by using a global attention mechanism. A residue dilated convolution network helped fast and accurate recognition on Chinese clinical corpus (Qiu et al., 2018). However, these deep learning methods all depend on the token representations that are not contextualized. The failure to track different semantic and syntactic meanings of each token leads to sub-optimal learning and modeling on named entity recognition. In this work, inspired by the recent development of contextualized token representations, we explore the ensemble of contextualized language models and simple deep learning methods for CNER.

2.2 Contextualized token representations

Deep contextualized word representation models complex characteristics of word use and how these uses vary across linguistic contexts (Akbik et al., 2018, Devlin et al., 2018, Peters et al., 2018). As a result, the representation of each token is a function of the entire input sentence, which is different from the traditional word type embeddings. Peters et al. (2018) leveraged a two-layer bidirectional language model (biLMs) with character convolutions to construct this function. Devlin et al. (2018) followed the idea of self-attention mechanism (Vaswani et al., 2017) and pre-trained a deep bidirectional Transformer by jointly conditioning on both left and right context. Akbik et al. (2018) developed contextual string embeddings by leveraging the internal states of a trained character language model. So far, these deep language models has brought massive improvement in different NLP applications including question answering, relation extraction, and sentiment classification.

Some researchers have applied contextualized embeddings to the biomedical domain. Lee et al. (2019) pre-trained a BioBERT with the settings of base BERT (Devlin et al., 2018) using billions of tokens from PubMed abstracts and PMC full text articles. Peng et al. (2019) pre-trained also a BERT using the complete PubMed abstract and MIMIC III corpus, tested with ten datasets of five tasks. These two works improved the performance of several representative biomedical text mining tasks; however, it required a large number of computational resources and inevitably a long time to train the language model. Inspired by (Sheikhshabbafghi et al., 2018), we pre-trained two light-loaded language models with a much smaller domain-specific clinical dataset selected from the PMC corpus.

3 Method

In this section, we firstly introduce the architectures of both word-level and character-level language models in Section 3.1. Then we explain our CNER model in Section 3.2.

3.1 Contextualized Embeddings

3.1.1 ELMo

ELMo is a language model that produces contextualized embeddings for words. It is pre-trained with a two-layered bidirectional language model (biLM) with character convolutions on a large corpus. The left lower part in Figure 1 is the high level architecture of ELMo, where R(·) means the representation of a word.

ELMo takes a sequence of words $(w_1, w_2, ..., w_N)$ as input and generates context-independent token representations using a character-level CNN. Then ELMo feeds the sequence of tokens $(t_1, t_2, ..., t_N)$ into the biLM which is a bidirectional Recurrent Neural Network (RNN). The forward-LM computes the probability of each sequence by:

$$p(t_1, t_2, ..., t_N) = \prod_{k=1}^{N} p(t_k | t_1, t_2, ..., t_{k-1}) \tag{1}$$

Thus at each position k, the RNN layer outputs a hidden representation h_k for predicting the token t_{k+1}. The backward-LM has the same structure as the forward-LM, except the input is the reverse sequence. Then, we jointly maximize the log-likelihood of both directions:

$$\begin{aligned}\sum_{k=1}^{N} (\log p(t_k | t_1, t_2, ..., t_{k-1}; \theta_x, \theta_f, \theta_s) \\ + \log p(t_k | t_{k+1}, t_{k+2}, ..., t_N; \theta_x, \theta_b, \theta_s))\end{aligned} \tag{2}$$

where θ_x is the token representation and θ_s is the Softmax layer for both the forward and backward LM's, and θ_f and θ_b denotes the parameters of RNNs in two directions.

3.1.2 Flair

Flair is a character-level word representation model that also uses RNN as the language modeling structure. Different from ELMo, Flair treats the text as a sequence of characters.

The goal of most language models is to estimate a good distribution $p(t_0, t_2, ..., t_T)$ where $t_0, t_1, ..., t_n$ is a sequence of words. Instead of computing the distribution of words, Flair aims to estimate the probability $p(x_0, x_1, ...x_T)$, where $x_0, x_1, ..., x_T$ is a sequence of characters. The joint distribution over the entire sentence can then be represented as follows:

$$p(x_0, x_1, ..., x_T) = \prod_{t=0}^{T} p(x_t | x_1, x_2, ..., x_{t-1}) \tag{3}$$

where $p(x_t | x_0, ..., x_{t-1})$ is approximated by the network output h_t from one RNN layer.

$$p(x_t | x_0, ..., x_{t-1}) = \prod_{t=0}^{T} p(x_t | h_t; \theta) \tag{4}$$

h_t is the hidden state that records the entire history of the sequence, which is computed recursively with a memory cell. θ denotes all the parameters of the RNN model. On top of the hidden layer, there is a fully-connected softmax layer, so the likelihood of a character is defined as:

$$p(x_t | h_t; W) = \text{softmax}(W \cdot h_t + b) \tag{5}$$

where W and b are the weights and biases.

Besides, Flair also has a backward RNN layer. Flair extracts the token embeddings from the hidden states of both the forward and backward models. Given a word that starts at index t_s and ends at t_e in a sequence of characters, the embeddings of this word are defined as a concatenation of the hidden states from both forward and backward models:

$$r^{Flair} := h^f_{t_e+1} \oplus h^b_{t_s-1} \tag{6}$$

where h^f denotes the hidden states from the forward model and h^b are the hidden states from the backward model. The details are illustrated in the left upper part in Figure 1.

3.2 CNER Model

We used a well-established BiLSTM-CRF sequence tagging model (Habibi et al., 2017, Huang et al., 2015, Wang et al., 2018) to address the downstream sequence labeling tasks.

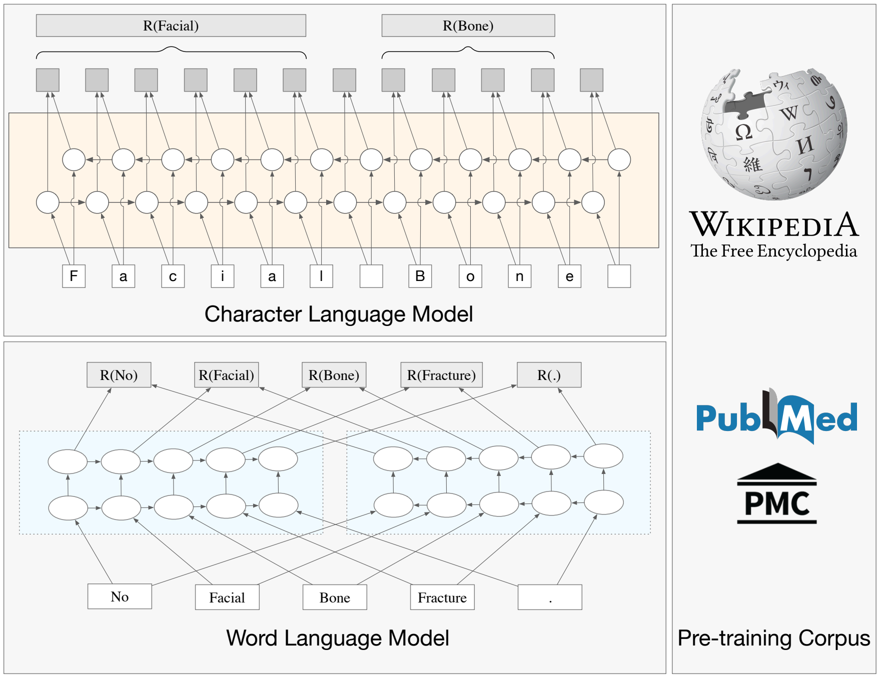

Figure 1. Character and Word Language models

First, it passes sentences to a user-defined token embedding model, which converts a sequence of tokens into word embeddings: $r_0, r_1, r_2, ..., r_n$. We may concatenate embedding vectors from different sources to form a new word vector. For example, the concatenated embeddings of GloVe and Flair is represented as:

$$r_i = r_i^{GloVe} \oplus r_i^{Flair} \tag{7}$$

Then, the concatenated embeddings are passed to the BiLSTM-CRF sequence labeling model to extract the entity types.

4 Experiments

In this section, we first introduce three benchmark datasets, MACCROBAT2018, i2b2-2010 and NCBI-disease. Then, we explain the corpus we used to pre-train the language models. The performance comparison among a set of baseline models and our methods is discussed in Section 4.4.

4.1 Datasets

4.1.1 MACCROBAT2018

Caufield et al. (2018) developed a standardized metadata template and identified text corresponding to medical concepts within 3,100 curated CCRs spanning 15 disease groups and more than 750 reports of rare diseases. MACCROBAT2018 is a subset of the case reports which were annotated by clinical experts. In total, there are 200 annotated case reports and 3,652 sentences containing 24 different entity/event types. We randomly selected 10% case reports as development set and 10% as test set. The remaining documents are used to train the CNER model. Detailed description is shown in Table 1, 2, and 3.

4.1.2 i2b2-2010

The i2b2/VA 2010 Workshop (Uzuner et al., 2011) on NLP challenges for clinical records presented three tasks: a concept extraction task, an assertion classification task, and a relation classification task. The i2b2-2010 dataset provides “layered” linguistic annotation over a set of clinical notes. In this study, we focus on the first task: given the plain text, we extract the clinical entities. The dataset contains three entity types which are “test”, “problem”, “treatment”. We followed Uzuner et al. (2011) to split the dataset into train/development/test sets.

4.1.3 NCBI-disease

The NCBI-disease (Doğan et al., 2014) dataset is fully annotated at the mention and concept level to serve as a research resource for the biomedical natural language processing community. The dataset contains 793 PubMed abstracts with 6,892 disease mentions which leads to 790 unique disease concepts. Therefore, the dataset only has one types which is “disease”.

4.2 Pre-training Corpus

To pre-train the two language models, we obtained articles through the PubMed Central (PMC) FTP server[1], and in total picked 47,990 documents that are related to clinical case

[1] ftp://ftp.ncbi.nlm.nih.gov/pub/pmc

reports. We indexed these documents with some keyword including “case report” and “clinical report”. This corpus contains 0.1 billion words which is around 1/10 of the corpus used for the domain-generic ELMo (Peters et al., 2018) and Flair (Akbik et al., 2018). We will release our pre-trained language models soon.

4.3 Pre-trained Language Model

We proposed C-ELMo and C-Flair, which are respectively a pre-trained ELMo and a pre-trained Flair with the domain-specific corpus. To fairly compare the two models, we do not initialize C-ELMo and C-Flair with any pre-trained ELMo and Flair, and pre-train them on the same clinical case report corpus described in Section 4.2. Moreover, we tried to set both models’ parameter sizes to a similar scale. Since Flair’s parameter size is 20M when it performs at its best (hidden size of 2048), we chose the medium size ELMo model correspondingly, which has 25M parameters according to AllenNLP (Peters et al., 2018). All models were pre-trained on one NVIDIA Tesla V100 (16GB), with each requiring roughly one week to complete.

For C-Flair, we followed the default settings of Flair, a hidden size of 2048, a sequence length of 250, and a mini-batch size of 100. The initial learning rate is 20, and the annealing factor is 4.

For C-ELMo, we chose the medium-size model among all configurations, which has a hidden size of 2048 and projection dimension of 256. For the convolutional neural network token embeddings, the maximum length of a word is 50 and the embedding dimension is 16.

Table 1. Number of sentences in each CNER dataset

Dataset Name	Train	Dev	Test
MACCROBAT2018	2,894	380	351
i2b2-2010	14,683	1,632	27,626
NCBI-disease	5,423	922	939

Table 2. Number of tokens in each CNER dataset

Dataset Name	Train	Dev	Test
MACCROBAT2018	64,879	862	7,955
i2b2-2010	134,586	14,954	267,250
NCBI-disease	135,701	23,969	24,497

Table 3. Number of entity types in each CNER dataset

Dataset Name	# of Entity Types
MACCROBAT2018	24
i2b2-2010	3
NCBI-disease	1

4.4 Results

To fairly compare the performance of each model, we pre-trained C-Flair and C-ELMo on the same subset of PubMed Central (PMC) documents. We then applied the BiLSTM-CRF model (Huang et al., 2015) to evaluate the downstream sequence labeling tasks. The results of our experiments are shown in Table 4. Note that “Embeddings” in Table 4 denotes the stacking embeddings which can be the concatenation of different word embedding vectors. We used the pre-trained GloVe embeddings of 100 dimensions [2]. The Flair embeddings are pre-trained with a 1-billion word corpus (Chelba et al., 2013). ELMo denotes the pre-trained medium-size ELMo on the same 1-billion word corpus and ELMoPubMed denotes the pre-trained ELMo model with the full PubMed and PMC corpus [3]. We used the micro F1-score as the evaluation metric.

4.4.1 Domain-specific v.s. Domain-generic corpus

From Table 4, we can observe that the models pre-trained on the selected case report corpus outperformed all the other language models pre-trained on the domain-generic corpus. The concatenated embedding of GloVe and C-ELMo performs the best on MACCROBAT2018 and NCBI-disease datasets, while GloVe plus C-Flair achieved the best performance on i2b2-2010. We can conclude that pre-training the language models with a small domain-specific corpus can be more efficient and effective for improving the performance of some downstream tasks. The domain-specific knowledge can alter the distribution and the proximity among words, thus contributing a better understanding of the relationship between word and entity types in our task.

4.4.2 Contextualized v.s. Non-contextualized embeddings

We also used the static word embeddings, GloVe itself, to represent the tokens in the sequence labeling task. The results in Table 4 show that the stacking contextualized embeddings dramatically boosted the F1-score on three different datasets by 10.31%, 7.50%, and 6.94%. It proves that the deep language models absorb more intensive semantic and syntactic knowledge from the contexts. We noticed that the F1-score of Flair on MACCROBAT2018 dataset was surprisingly low. It showed that the performance of a purely character-level language model may be not as robust as the word-level models.

4.4.3 Compared with other baseline models

Ma and Hovy (2016) proposed a bi-directional LSTM-CNNs-CRF model to make use of both word- and character-level representations. Wang et al. (2018) leveraged multi-task learning and attention mechanisms to improve the performance of

[2] https://nlp.stanford.edu/projects/glove/

[3] https://allennlp.org/elmo/

Table 4. The comparison of F1-scores (%) on three datasets among different types of embeddings

Embeddings	MACCROBAT2018	i2b2-2010	NCBI-disease
GloVe	59.63	81.35	82.18
ELMo	61.69	84.61	84.50
Flair	57.25	81.65	84.23
GloVe+ELMo	63.09	84.82	85.37
GloVe+Flair	62.63	81.21	85.58
GloVe+ELMoPubMed	64.56	86.50	87.04
GloVe+C-ELMo	**65.75**	87.29	**87.88**
GloVe+C-Flair	64.18	**87.45**	86.60

Table 5. The performance of three baseline methods and our best model on three datasets. Our models only leverage a simple LSTM-CRF sequence labeling module with the pre-trained contextualized embeddings.

Models	MACCROBAT2018	i2b2-2010	NCBI-disease
Our best model	**65.75**	**87.45**	87.88
Ma and Hovy (2016)	60.13	81.41	82.62
Wang et al. (2018)	63.10	84.97	86.14
Lee et al. (2019)	64.38	86.46	**89.36**

biomedical sequence labeling task. Compared with these two state-of-the-art models, as shown in Table 5, our methods perform consistently better. We suppose that with the help of pre-trained contextualized embeddings, even a light-loaded downstream model can achieve extraordinary performances.

The BioBERT proposed in (Lee et al., 2019) was pre-trained using a language model with around 110M parameters and using a large number of computational resources (8 NVIDIA V100 32GB GPUs). However, this contextualized language model only gets better performance in the simplest dataset (NCBI-disease) with only one entity type. On MACCROBAT2018 and i2b2-2010, we improved the performance by 2.13% and 1.15%. This shows that good experimental results can be achieved by making rational use of limited resources.

Table 6. The comparison of F1-scores (%) between C-ELMo and C-Flair on different entity types of MACCROBAT2018

Entity	GloVe+C-ELMo	GloVe+C-Flair
Biological Structure	63.94	**64.88**
Detailed Description	**45.81**	40.00
Diagnostic Procedure	**74.93**	74.71
Disease Disorder	**50.84**	50.83
Dosage	77.42	**80.00**
Lab Value	**74.48**	72.31
Medication	**76.34**	72.13
Non-biological Location	**80.77**	76.00
Severity	**72.41**	61.81
Sign Symptom	**62.27**	60.64

4.5 Case Study and Analysis

We analyze the C-Flair and C-ELMo on specific categories for the MACCROBAT2018 dataset. We look into the F1-scores of 10 different entity types. All these types appear more than 50 times in the dataset.

From Table 6, we can see that the character-level language model C-Flair shows an advantage in the type "Dosage". We find that this entity type has a number of entities that do not appear in the word-level vocabulary, such as "60 mg/m2", "0.5 mg", and "3g/d". On the other hand, C-ELMo has a better performance in the type "Severity", which contains words like "extensive", "complete", "significant", and "evident". C-ELMo also extensively outperforms C-Flair in "Detailed Description". The representations of tokens rely more on the word-level context in these types. Therefore, C-ELMo shows better power of capturing the relationship between the word-level contextual features with the entity types.

We noticed in Table 6, "Disease Disorder" achieved around 50% F1-score with both models. Though they performed well on NCBI-disease dataset, it is hard for them to correctly recognize complex phrase-level disease entities on MACCROBAT2018, such as "Scheuer stage 3", and "feeding difficulties".

5 Conclusion

In our study, we showed that contextual embeddings show a sizable advantage against non-contextual embeddings for clinical NER. In addition, pre-training a language model with a domain-specific corpus results in better performance in the downstream CNER task, compared to the off-the-shelf corpus. We also developed a comparatively fair comparison between C-ELMo and C-Flair. We found that the two language models demonstrate variability in labeling different entity types in our datasets, presumably due to their separate focuses on word-level and character-level contextual features.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 1638–1649.

A. R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* (2001), 17–21. http://view.ncbi.nlm.nih.gov/pubmed/11825149

David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs* 33, 7 (2014), 1123–1131.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32, 90001 (2004), D267–70. https://doi.org/10.1093/nar/gkh061

Alberto J Cabán-Martinez and Wilfredo F García-Beltrán. 2012. Advancing medicine one research note at a time: the educational value in clinical case reports. *BMC Research Notes* 5, 1 (2012), 293. https://doi.org/10.1186/1756-0500-5-293

J Harry Caufield, Yijiang Zhou, Anders O Garlid, Shaun P Setty, David A Liem, Quan Cao, Jessica M Lee, Sanjana Murali, Sarah Spendlove, Wei Wang, et al. 2018. A reference set of curated biomedical data and metadata from clinical case reports. *Scientific data* 5 (2018), 180258.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005* (2013).

Berry De Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association* 18, 5 (2011), 557–562.

Dina Demner-Fushman, Willie J Rogers, and Alan R Aronson. 2017. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *Journal of the American Medical Informatics Association* 24, 4 (2017), 841–844.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018).

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI Disease Corpus. *J. of Biomedical Informatics* 47, C (Feb. 2014).

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries* 3, 2 (2000), 115–130. https://doi.org/10.1007/s007999900023

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* 33, 14 (2017), i37–i48.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR* abs/1508.01991 (2015).

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *CoRR* abs/1901.08746 (2019).

H. Liu, Z.-Z. Hu, M. Torii, C. Wu, and C. Friedman. 2006. Quantitative Assessment of Dictionary-based Protein Named Entity Tagging. *Journal of the American Medical Informatics Association* 13, 5 (2006), 497–507. https://doi.org/10.1197/jamia.M2085

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *CoRR* abs/1603.01354 (2016).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). 3111–3119.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *arXiv preprint arXiv:1906.05474* (2019).

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR* abs/1802.05365 (2018).

Jiahui Qiu, Qi Wang, Yangming Zhou, Tong Ruan, and Ju Gao. 2018. Fast and Accurate Recognition of Chinese Clinical Named Entities with Residual Dilated Convolutions. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 935–942.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* 1, 1 (2018), 18.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. 17 (2010).

Golnar Sheikhshabbafghi, Inanc Birol, and Anoop Sarkar. 2018. In-domain Context-aware Token Embeddings Improve Biomedical Named Entity Recognition. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*.

Javier Tamames. 2005. Text Detective: a rule-based system for gene annotation in biomedical texts. *BMC Bioinformatics* 6, Suppl 1 (2005), S10. https://doi.org/10.1186/1471-2105-6-S1-S10

The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 45, D1 (2017), D158–D169. https://doi.org/10.1093/nar/gkw1099

Ö. Uzuner, B.R. South, S. Shen, and S.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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

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. *CoRR* abs/1801.09851 (2018). arXiv:1801.09851 http://arxiv.org/abs/1801.09851

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research* 41, W1 (2013), W518–W522. https://doi.org/10.1093/nar/gkt441

Guohai Xu, Chengyu Wang, and Xiaofeng He. 2018. Improving clinical named entity recognition with global neural attention. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 264–279.

Edmond Zhang, Quentin Thurier, and Luke Boyle. 2018. Improving Clinical Named-Entity Recognition with Transfer Learning. *Studies in health technology and informatics* 252 (2018), 182–187.