Clinical Concept Extraction with Contextual Word Embedding

Henghui Zhu
Boston University
henghuiz@bu.edu

Ioannis Ch. Paschalidis
Boston University
yannisp@bu.edu

Amir Tahmasebi
Philips Research North America
Amir.Tahmasebi@philips.com

Abstract

Automatic extraction of clinical concepts is an essential step for turning the unstructured data within a clinical note into structured and actionable information. In this work, we propose a clinical concept extraction model for automatic annotation of clinical problems, treatments and tests in clinical notes utilizing domain-specific contextual word embedding. A contextual word embedding model is first trained on a corpus with a mixture of clinical reports and relevant Wikipedia pages in the clinical domain. Next, a bidirectional LSTM-CRF model is trained for clinical concept extraction using the contextual word embedding model. We tested our proposed model on the I2B2 2010 challenge dataset. Our proposed model achieved the best performance among reported baseline models and outperformed the state-of-the-art models by 3.4% in terms of F1-score. [1]

1 Introduction

Automatic clinical concept extraction is a crucial step in transforming the abandon of unstructured patient clinical data into a set of actionable information. A major challenge towards developing clinical Named Entity Recognition (NER) tools is access to a corpus of labeled data. The 2010i2b2/VA challenge [1] released such corpus of annotated clinical notes to facilitate the development of clinical concept extraction systems. Since the release of the 2010 i2b2/VA dataset, there have been numerous efforts on developing NER tools reported in the literature. In earlier works, Conditional Random Fields (CRF) with token-level features were proposed for concept extraction [2, 3, 4]. Nevertheless, such approaches require manual feature engineering, which makes its application limited. Recently, with the advancement of deep learning models and its usage for natural language processing, Recurrent Neural Network (RNN) models, such as Long Short-Term Memory (LSTM), have been widely used for deriving contextual features for CRF model training and have demonstrated promising performance for clinical concept extraction tasks [5, 6, 7, 8]. Although RNN models provide a good set of features for NER, they depend heavily on the word embedding models [9, 10], which are not good at dealing with the complex characteristics of word use under different linguistic contexts.

Recently, Peters et al. proposed a contextual word-embedding model (ELMo), which claims to improve the performance of various NLP tasks such as sentiment analysis, question answering and sequence labeling [11]. By including the features of a language model, the word representations of the ELMo contain richer information compared to a standard word embedding such as skip-gram [9] or Glove [10]. Although the ELMo model is shown to have a good performance in some NER

[1] Codes are available at https://github.com/noc-lab/clinical_concept_extraction

Machine Learning for Health (ML4H) Workshop at NeurIPS 2018.

tasks such as the CoNLL 2003 NER task, it is trained on a corpus in a general domain [12] and as a result, does not demonstrate a desired performance for a clinical concept extraction task. This could be due to the fact that clinical text is structured differently compared to text in a general domain and therefore, the language model trained on a general domain corpus fails to generalize well on it.

In this work, we train an ELMo model in a corpus with a mixture of clinical reports and relevant Wikipedia pages in the clinical domain. Next, a bidirectional LSTM-CRF model is applied to identify the clinical concepts. This model is trained and tested on the 2010 i2b2/VA challenge [1]. Our proposed model achieves the best performance among others the state-of-the-art models by 3.4% in terms of F1-score.

2 Dataset

In this work, we used the data provided by the 2010 i2b2/VA challenge [1] for training a clinical concept extraction system. Due to the restrictions introduced by Institutional Review Board (IRB), only a portion of data from the original dataset is available. The released dataset consists of clinical summaries from three different medical sites: Partners Healthcare, Beth Israel Deaconess Medical Center, and the University of Pittsburgh Medical Center. There are three clinical concepts annotated in this corpus: problems, tests, and treatments. There are 170 summaries for training and 256 for test. The dataset statistics are shown in Table 1.

Table 1: Number of reports, sentences, tokens and named entities in training and test corpus.

corpus	reports	sentences	tokens	problem	treatment	test
training	170	16,414	149,541	7,073	4,844	4,606
test	256	27,763	267,249	12,592	9,344	9,225

3 Methods

The proposed model consists of an ELMo model trained using a corpus in a clinical domain and a bidirectional LSTM-CRF model for the clinical context extraction.

3.1 ELMo Training

ELMo [11] is a recently proposed model that extends a word embedding model with features produced by a bidirectional language model. It has been shown that the utilization of ELMo for different NLP tasks results in improved performance compared to other types of word embedding models such as skip-gram and Glove[11, 13, 14]. Although [11] released several pretrained ELMo models, all of them are trained on a general text corpus [12]. We believe the clinical reports are structured differently compared to such general language corpora. Therefore, it is necessary to train an ELMo model specifically for the clinical domain in order to achieve the desired performance for clinical NLP tasks including concept extraction. To do so, the following corpora were considered for training the ELMo model.

- Wikipedia pages with titles that are items (medical concepts) in a standard clinical ontology, known as SNOMED CT [15]. The following sections were excluded: 'See also', 'References', 'Further reading' and 'External links'. Furthermore, if a term has more than one Wikipedia page, we exclude all these pages in the corpus for avoiding introducing ambiguity.
- The discharge summaries and radiology reports from MIMIC-III dataset [16]. Phrase normalization is performed on the deidentified entities in the dataset, e.g., converting all names to 'John Does', removing the prefixes and suffixes for deidentified data and numbers. The reason for the normalization is to transform the deidentified reports to be more similar to actual reports.

This paper uses our in-house built sentence segmentation tool[2] to detect sentence boundary. NLTK [17] is used for tokenization. Statistics of the training corpus is shown in Table 2. For

[2] `https://github.com/noc-lab/simple_sentence_segment`

Table 2: The corpus statistics for training the ELMo model.

	number of sentences	average sentence length
Wikipedia pages	3,687,501	20.84
discharge summaries from MIMIC-III	10,457,035	9.61
radiology reports from MIMIC-III	12,786,115	9.62

training the ELMo model, we use the default hyperparameter settings shown in [11]. In detail, a character-based Convolutional Neural Network (char-CNN) embedding layer [18] is used with character embeddings dimension 16, filter widths $[1, 2, 3, 4, 5, 6, 7]$ and number of filters $[32, 32, 64, 128, 256, 512, 2014]$. Next, a two-layer bidirectional LSTM (bi-LSTM) with 4,096 hidden units in each layer is considered. After each char-CNN embedding and LSTM layer, the output is projected to 512 dimensions and a high-way connection [19] is applied. In addition, we define the vocabulary in the language model as the tokens that appear not less than 5 times in the corpus.

We randomly split the whole corpus into a training corpus (90%) and a testing corpus (10%). We train an ELMo model using the training corpus for 10 epochs. The average perplexity in the testing corpus is 17.80. On the other hand, the ELMo model[3] trained in a corpus of a general domain only achieves a perplexity of 628.26. Despite the fact that comparing these two perplexities is not fair due to the different vocabularies in the two language models, such large gap between them suggests that training a specific ELMo model for the clinical domain is necessary.

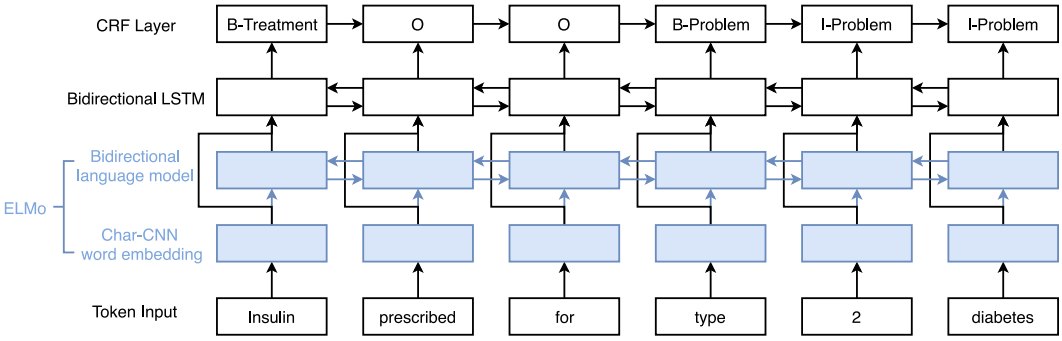

Figure 1: The proposed bidirectional LSTM model diagram for medical concept extraction. The char-based CNN word embedding and the bidirectional language model layers (ELMo), shown in the blue color, are pretrained with corpus from the clinical domain and hold fixed during training the NER model.

3.2 Bidirectional LSTM CRF model for NER

We use a bidirectional LSTM-CRF [20] model for the NER task. The architecture of the proposed model is shown in Figure 1. The input is a list of tokens. Contextual word embeddings are generated as a learnable aggregation of char-CNN word-embedding layer and two bi-LSTM layers for the language model. Suppose $\mathbf{x}_k$ is the context-independent token representation for the kth token in the sentence produced by the character CNN layer and denote $\mathbf{h}_k^0 = [\mathbf{x}_k, \mathbf{x}_k]$ as the token layer. Also denote $\mathbf{h}_k^1 = [\overleftarrow{\mathbf{h}_k^1}, \overrightarrow{\mathbf{h}_k^1}]$ and $\mathbf{h}_k^2 = [\overleftarrow{\mathbf{h}_k^2}, \overrightarrow{\mathbf{h}_k^2}]$ the two bidirectional LSTM layers in ELMo with forward and backward language models. Following [11], we use

$$\mathbf{y}_k = \gamma \sum_{i=0}^{2} s_i \mathbf{h}_k^i$$

as the features for the NER model, where γ is a scale factor and s_i's are softmax-normalized weights. During training the NER model, the parameters of the ELMo model is fixed while γ and s_i's are learnable parameters. Next, a two layer bidirectional LSTM is applied with $\mathbf{y}_k$'s as input. Finally,

[3]The 'Original' model in section 'Pre-trained ELMo Models' at `https://allennlp.org/elmo`

a linear-chain CRF layer [20] is applied for predicting the label of each token. In this paper, the BIO-tagging format is used, as shown in Figure 1.

4 Results

A two-layer bidirectional LSTM is used for NER task, each of which consists of 256 hidden states. For regularization, dropout [21] is applied to the LSTM layer with a rate of 0.5. We train the model with the Adam optimizer [22] using a learning rate of 0.001, a batch size of 32, and 200 epochs.

Table 3: Performance comparison between proposed models and the state-of-the-art models on the 2010 i2b2/VA dataset.

Methods	Precision	Recall	F1
Distributional semantics CRF [3] *	85.60	82.00	83.70
Hidden semi-Markov model [2] *	86.88	83.64	85.23
Truecasing CRFSuite [4]	80.83	71.47	75.86
CliNER [5]	79.5	81.2	80.0
Binarized neural embedding CRF [23]	85.10	80.60	82.80
Glove-BiLSTM-CRF [6]	84.36	83.41	83.88
CliNER 2.0 [7]	84.0	83.6	83.8
Att-BiLSTM-CRF + Transfer [8]	86.27	85.15	85.71
ELMo(General) + BiLSTM-CRF (Single) **	83.26 ± 0.25	81.84 ± 0.22	82.54 ± 0.14
ELMo(Clinical) + BiLSTM-CRF (Single) **	87.44 ± 0.27	86.25 ± 0.26	86.84 ± 0.16
ELMo(Clinical) + BiLSTM-CRF (Ensemble)	**89.34**	**87.87**	**88.60**

* These models were trained using the original larger labeled dataset of the 2010 i2b2/VA challenge.
** Performances are reported as mean $\pm$ standard deviation across 10 runs with different random seeds.

Three different scenarios of the proposed ELMo-based model were considered and compared in this study: Two BiLSTM-CRF models were trained using 1) an ELMo model trained on a general domain corpus [11], referred to as "ELMo(General) + BiLSTM-CRF (Single)"; and 2) an ELMo model trained on a clinical corpus as described before, referred to as "ELMo(Clinical) + BiLSTM-CRF (Single)". The training was performed 10 times starting with 10 different random seeds and the mean and standard deviation of the performance metrics were reported. Besides, we also trained an ensemble model based on the most voted label by the 10 models for each token. The performance of our models and several previously published baseline models for the 2010 i2b2/VA challenge is reported in Table 3 in terms of precision, recall and F1-score for exact class spans using the definition given in [1]. As expected, it can be observed from the table that an ELMo model trained using domain-specific data results in significant improvement in performance. The BiLSTM-CRF model with ELMo trained on the clinical corpus outperforms other alternatives. Furthermore, it was observed that our best model yielded similar performance among three types of named entities (problem, treatment and test).

5 Conclusions

Contextual word embedding approach such as ELMo has exhibited a promising performance in many natural language processing tasks. In this paper, we trained a domain-specific ELMo model using a clinical domain-specific corpus and furthermore, utilized it for building a clinical concept extraction tool. We trained and tested the proposed model using the dataset provided by the 2010 i2b2/VA challenge. To the best of our knowledge, our model yields the best performance compared to the reported prior work by a significant margin of 3.4% in terms of F1-score.

From the results reported in this work, one can conclude that training a domain-specific language model is essential to achieve high performance for NER tasks. Nevertheless, access to a large size domain-specific corpus is a known challenge for training a language model. In this work, we demonstrated an effective yet simple approach to create such corpus for a clinical domain by filtering a general domain corpus such as wiki-pages using a domain-specific ontology such as SNOMED CT.

Acknowledgments

Research partially supported by the ONR under MURI N00014-16-1-2832, by the NSF under grants DMS-1664644, CNS-1645681, CCF-1527292, and IIS-1237022, by the NVIDIA Corporation with the donation of a Titan Xp GPU, and by the Center for Information and Systems Engineering.

References

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

[2] Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 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):557–562, 2011.

[3] Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. Enhancing clinical concept extraction with distributional semantics. *Journal of biomedical informatics*, 45(1):129–140, 2012.

[4] Xiao Fu and Sophia Ananiadou. Improving the extraction of clinical concepts from clinical records. *Proceedings of BioTxtM14*, pages 47–53, 2014.

[5] William Boag, Kevin Wacome, Tristan Naumann, and Anna Rumshisky. CliNER: A lightweight tool for clinical named entity recognition. *AMIA Joint Summits on Clinical Research Informatics (poster)*, 2015.

[6] Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. Bidirectional LSTM-CRF for clinical concept extraction. *arXiv preprint arXiv:1611.08373*, 2016.

[7] Willie Boag, Elena Sergeeva, Saurabh Kulshreshtha, Peter Szolovits, Anna Rumshisky, and Tristan Naumann. CliNER 2.0: Accessible and Accurate Clinical Concept Extraction. *arXiv preprint arXiv:1803.02245*, 2018.

[8] Guohai Xu, Chengyu Wang, and Xiaofeng He. Improving clinical named entity recognition with global neural attention. In Yi Cai, Yoshiharu Ishikawa, and Jianliang Xu, editors, *Web and Big Data*, pages 264–279, Cham, 2018. Springer International Publishing.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[10] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[11] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

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

[13] Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[14] Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. Jointly predicting predicates and arguments in neural semantic role labeling. *arXiv preprint arXiv:1805.04787*, 2018.

[15] Jeremy Rogers and Olivier Bodenreider. SNOMED CT: Browsing the Browsers. In *KR-MED*, pages 30–36, 2008.

[16] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[17] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[18] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. In *AAAI*, pages 2741–2749, 2016.

[19] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[20] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[21] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027, 2016.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Yonghui Wu, Jun Xu, Min Jiang, Yaoyun Zhang, and Hua Xu. A study of neural word embeddings for named entity recognition in clinical text. In *AMIA Annual Symposium Proceedings*, volume 2015, page 1326. American Medical Informatics Association, 2015.