Journal of the American Medical Informatics Association, 27(12), 2020, 1935–1942
doi: 10.1093/jamia/ocaa189
Advance Access Publication Date: 29 October 2020
Research and Applications

Research and Applications

Clinical concept extraction using transformers

Xi Yang,[1,2] Jiang Bian,[1,2] William R. Hogan,[1] and Yonghui Wu[1,2]

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA and [2]Cancer Informatics and eHealth core, University of Florida Health Cancer Center, Gainesville, Florida, USA

Corresponding Author: Yonghui Wu, PhD, Clinical and Translational Research Building, 2004 Mowry Road, PO Box 100177 Gainesville, FL, USA (yonghui.wu@ufl.edu)

Received 7 May 2020; Revised 8 July 2020; Editorial Decision 18 July 2020; Accepted 25 July 2020

ABSTRACT

Objective: The goal of this study is to explore transformer-based models (eg, Bidirectional Encoder Representations from Transformers [BERT]) for clinical concept extraction and develop an open-source package with pretrained clinical models to facilitate concept extraction and other downstream natural language processing (NLP) tasks in the medical domain.

Methods: We systematically explored 4 widely used transformer-based architectures, including BERT, RoBERTa, ALBERT, and ELECTRA, for extracting various types of clinical concepts using 3 public datasets from the 2010 and 2012 i2b2 challenges and the 2018 n2c2 challenge. We examined general transformer models pretrained using general English corpora as well as clinical transformer models pretrained using a clinical corpus and compared them with a long short-term memory conditional random fields (LSTM-CRFs) mode as a baseline. Furthermore, we integrated the 4 clinical transformer-based models into an open-source package.

Results and Conclusion: The RoBERTa-MIMIC model achieved state-of-the-art performance on 3 public clinical concept extraction datasets with F1-scores of 0.8994, 0.8053, and 0.8907, respectively. Compared to the baseline LSTM-CRFs model, RoBERTa-MIMIC remarkably improved the F1-score by approximately 4% and 6% on the 2010 and 2012 i2b2 datasets. This study demonstrated the efficiency of transformer-based models for clinical concept extraction. Our methods and systems can be applied to other clinical tasks. The clinical transformer package with 4 pretrained clinical models is publicly available at https://github.com/uf-hobi-informatics-lab/ClinicalTransformerNER. We believe this package will improve current practice on clinical concept extraction and other tasks in the medical domain.

Key words: named entity recognition, transformer models, deep learning, natural language processing

INTRODUCTION

Electronic health records (EHRs), which contain both structured, coded data and unstructured clinical text, have now been widely used for research and various clinical applications. A critical challenge of using EHRs is to unlock patient information from the unstructured clinical text.[1] Much of the critical information of patients, such as family history, drug adverse events, and social, behavioral, and environmental determinants of health, is often only well-documented in narrative clinical text.[2–5] Therefore, researchers have invested significant effort into developing natural language processing (NLP) methods and tools to extract important clinical concepts from narrative clinical text.[6] Various NLP architectures—including rule-based, machine learning-based, and hybrid models—have been developed and studied to enhance the accuracy of clinical concept extraction.[7] With the emergence of deep learning models, research on clinical concept extraction has shifted from traditional machine learning models that rely heavily on semantic and lexical

features manually crafted by domain experts to deep learning models that can automatically learn feature representations (eg, word embeddings) from large volumes of unlabeled clinical text.[8–10] Recently, studies have reported that a new deep learning-based architecture, named "transformers," achieved state-of-the-art performance for a number of benchmark tasks[11–16] in the general English domain. Although several studies have examined transformer-based models for clinical individually,[17–21] there is no study that has systematically explored and compared their performance in the biomedical domain. In addition, there is a lack of package with pretrained clinical transformers that could facilitate researchers and other users adopting these state-of-the-art NLP models in various downstream clinical NLP tasks.

The goal of this study is to explore transformer-based models for clinical concept extraction and develop a software package with pretrained clinical models to facilitate clinical concept extraction and other downstream clinical NLP tasks. Here, we systematically explored 4 widely used transformer models (or encoders)—including bidirectional encoder representations from transformers (BERT),[12] RoBERTa,[15] ALBERT,[14] and ELECTRA[22]—to extract various types of clinical concepts and benchmarked them against 3 public datasets developed by the 2010 i2b2 challenge,[23] the 2012 i2b2 challenge,[24] and the 2018 n2c2 challenge.[25] We evaluated and compared the 4 models for clinical concept extraction with standard precision, recall, and F1-score metrics calculated using official evaluation scripts from the i2b2 and n2c2 challenges. In addition, we integrated the 4 transformers with pretrained clinical models into an open-source software package (available at https://github.com/uf-hobi-informatics-lab/ClinicalTransformerNER) to facilitate clinical concept extraction and other downstream NLP tasks. To the best of our knowledge, this is the first comprehensive study to systematically explore the 4 widely used transformer-based models for clinical concept extraction.

BACKGROUND

Clinical concept extraction is a fundamental task to support downstream clinical applications such as computable phenotyping, clinical decision support, and question-answering.[1] Many clinical NLP systems have been developed to extract various clinical concepts from clinical narratives, such as MedLEE,[26,27] MetaMap,[28] cTAKES,[29] and CLAMP.[30] The clinical NLP community has organized a series of open challenges with a focus on clinical concept extraction, including Informatics for Integrating Biology & the Bedside (i2b2),[23,31,32] National NLP Clinical Challenges (n2c2),[25] SemEval,[33–35] and ShARe/CLEF[36,37] in the past decade. Researchers have explored rule-based, machine learning-based methods, and hybrid approaches. Rule-based methods (eg, MedLEE, MetaMap, cTAKES) heavily depend on domain experts to identify patterns and design rules manually to capture clinical concepts based on medical dictionaries. The machine learning-based methods (eg, CLAMP) typically approach clinical concept extraction as a named entity recognition (NER) task—to identify the boundaries (ie, the start and end positions in the document) of concepts and classify the semantic categories of the identified concepts (eg, disease, medication). Most state-of-the-art NLP methods for clinical concept extraction are based on supervised machine learning models. Examples include the winning systems in the 2010 and 2012 i2b2 challenges, the 2014 and 2015 SemEval challenges, and the 2018 n2c2 challenge, which are all built using various supervised machine learning models.[32,38–41] Both traditional machine learning models (eg, conditional random fields [CRFs], structured support vector machines [SSVMs]) and deep learning models (eg, convolutional neural networks [CNN], recurrent neural networks [RNN]) have been explored.

Early studies of clinical concept extraction mainly focused on traditional machine learning algorithms (eg, SVMs, CRFs, and SSVMs) and various linguistic features (eg, part of speech, dependency parsing, and n-grams). CRFs is 1 of the commonly used NER methods as it achieved the best performance in several clinical NLP challenges (eg, the 2010 and 2012 i2b2 challenge). Then, researchers explored several unsupervised machine learning algorithms (eg, the Brown clustering algorithm[42] and the random indexing algorithm[43,44]) to generate clusters of words with similar meanings as novel unsupervised features. The experimental results from Tang et al[45] showed that the unsupervised semantic clusters could improve the performance of NER in addition to the features manually identified by domain experts. We also have examined word embeddings as features in a traditional CRFs model and demonstrated improved performance.[46]

Subsequently, with the development of deep learning models, the focus of concept extraction research shifted to algorithms for automated word-representation learning.[47] Several deep learning methods such as CNNs and RNNs demonstrated better performance than traditional machine learning methods. Compared with traditional machine learning methods, deep learning methods usually apply word embedding algorithms (eg, word2vec,[48] GloVe,[49] and FastText[50]) to learn vector representation of words (ie, word embeddings) to achieve better performance and avoid time-consuming, manual feature identification by domain experts. The RNN-based neural NER model implemented using bidirectional long-short term memory (LSTM) with a CRFs layer (LSTM-CRFs), first proposed by Lample et al,[51] achieved state-of-the-art performance in several clinical NER challenges (eg, i2b2, n2c2, SemEval, and ShARe/CLEF). Based on the LSTM-CRFs architecture, we also explored algorithms to combine factual medical knowledge embeddings with word embeddings to better handle medical terminologies not commonly used in general domain corpora.[52]

Inspired by the idea of representing words as vectors,[48,53] NLP researchers continued to explore new representations with rich architectures using technologies such as the attention mechanism and position embeddings. In 2017, Vaswani et al first used the word "transformer" to name a novel language model architecture that was solely constructed from self-attention blocks.[54] Typically, the training of transformer-based models (eg, BERT) consists of a pretraining stage and a fine-tuning stage. Pretraining is a procedure for optimizing the transformer-based models using large volumes of unlabeled text data by language-modeling methods (eg, statistical language modeling[55] and masked language modeling[56]) which are independent of any specific downstream NLP tasks—that is why it is called "pretraining." Fine-tuning is a procedure to further optimize the pretrained transformer-based models towards a specific NLP task (eg, clinical concept extraction) using annotated corpora. The transformer-based models only need to be pretrained once and then they can be applied to various downstream NLP tasks through fine-tuning. Several transformer-based models with different architectures and pretraining strategies have been proposed in the past 2 years including BERT, ALBERT, RoBERTa, and ELECTRA. These models demonstrated improvements on most NLP benchmarking tasks and outperformed the LSTM-CRFs model as the new state-of-the-art. Compared with CNNs and RNNs based on word-level embeddings, transformer-based models further break down words into sublevel tokens (as shown in Figure 1) to learn fine-grained

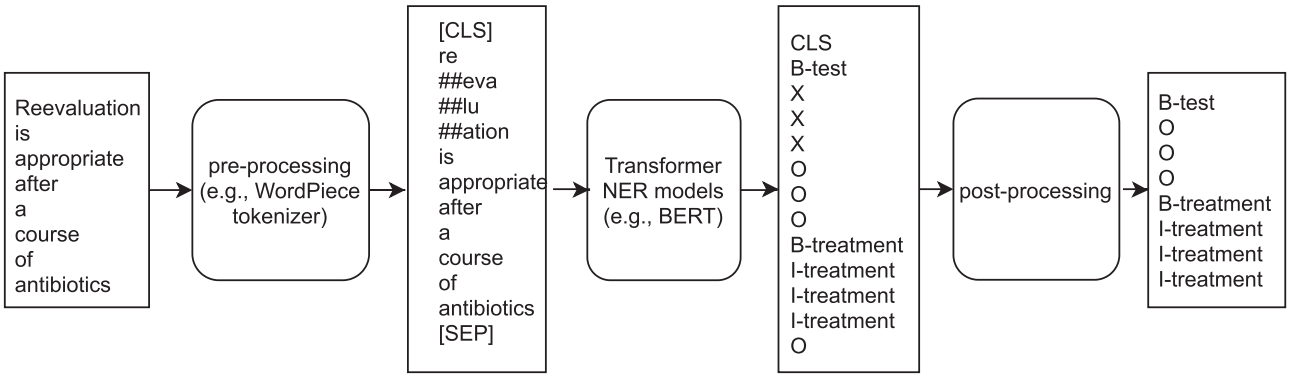

Figure 1. An overview of the workflow for a transformer-based NER pipeline using BERT as an example.

structured representations of words or subwords. Several studies[12,57] from the general domain have reported that transformers achieved better performance for NER, outperforming the LSTM-CRFs model even without a traditional CRFs layer. In the clinical domain, Alsentzer et al[21] and Si et al[18] explored the BERT model for clinical concept extraction.

However, there is no study that has systematically explored transformers in the medical domain even though studies from the general English domain have reported several transformers outperformed BERT on many benchmarks (eg, GLUE,[58] MultiNLI,[59] and SQuAD[60]) There is also a lack of publicly available software packages to facilitate researchers adopting transformers in clinical NLP tasks. In this study, we systematically examined 4 widely used transformer models for clinical concept extraction and developed an open-source clinical transformer package.

MATERIALS AND METHODS

Dataset

This study used 3 public clinical concept extraction datasets developed by the 2010 i2b2 challenge, 2012 i2b2 challenge, and 2018 n2c2 challenge, respectively. Table 1 shows the descriptive statistics of these 3 datasets. In the 2010 i2b2 dataset, there are 3 types of clinical concepts including PROBLEM, TREATMENT, and TEST. The 2012 i2b2 challenge has 6 types of clinical concepts including PROBLEM, TREATMENT, TEST, CLINICAL_DEPT, EVIDENTIAL, and OCCURRENCE. The 2018 n2c2 challenge focused on concepts related to drug adverse events, including drugs, drug associated attributes (ie, Dosage, Strength, Form, Frequency, Duration), drug related reasons, and drug-induced adverse events (ADE).

Table 1. Distribution of notes and clinical concepts in the 3 datasets

Challenge dataset	Subset	Number of notes	Number of clinical concepts
2010 i2b2	Training	349	27 837
	Test	477	45 009
2012 i2b2	Training	190	16 468
	Test	120	13 594
2018 n2c2	Training	303	50 951
	Test	202	32 918

Transformer models

This study systematically explored 4 widely used transformer-based models including BERT, ALBERT, RoBERTa, and ELECTRA.

BERT: a bidirectional transformer-based encoder model pretrained using masked language modeling and optimized using next sentence prediction. The base model architecture has 12 transformer blocks with a hidden size of 768 and 8 attention heads. The total number of parameters is ~110 million.

ALBERT: a "lite" version of BERT. ALBERT simplified the architecture of BERT to reduce the total number of parameters to optimize large-scale configurations and memory efficiency. More specifically, the ALBERT model reduced the token-embedding layer size from 768 to 128. Parameters are shared across all layers. ALBERT is pretrained using masked language modeling but optimized using sentence-order prediction instead of next sentence prediction. Therefore, the ALBERT is significantly smaller than BERT. The base model of ALBERT has 12 transformer blocks with an embedding size of 128 and the hidden size of 768 with 8 attention heads. The total number of parameters is ~12 million, which is only approximately one-tenth the number of parameters in BERT.

RoBERTa: a transformer-based model with the same architecture as BERT but pretrained using a dynamic masked language modeling and optimized using different strategies (eg, removing the next sentence prediction).

ELECTRA: another transformer-based model with the same architecture as BERT but pretrained using a novel strategy called replaced token detection. The central idea of pretraining ELECTRA is detecting replaced tokens in the input sentences. Specifically, ELECTRA consists of 2 transformer models with 1 as a generator, to create tokens to replace some of the original tokens, and the other as a discriminator, to predict whether input tokens are original or replaced by the generator.

Workflow of a transformer-based NER pipeline

Similar to other machine learning-based NER methods, transformer-based models use NER tags such as the B-I-O tags to label the tokens, where "B" indicates the first token of a concept, "I" indicates tokens inside of a concept, and "O" indicates tokens that do not belong to any concepts. Thus, the clinical concept extraction task can be formulated as a classification problem—classify a predefined NER tag for each token. Different from the previous deep learning models (eg, LSTM-CRFs) using word-level embeddings, transformer-based models further break down words into pieces of subtokens (pieces of tokens that are frequently used to form words). A special tag, "X," is introduced to indicate the subtokens. Figure 1 shows an overview of the workflow for transformer-based NER methods.

Preprocessing and postprocessing for transformer models

Transformer-based models used various word segmentation algorithms[61] to break words into subtokens to alleviate out-of-vocabulary issue and learn contextual representations at the subtoken level. For example, BERT and ELECTRA used the WordPiece,[62] RoBERTa adopted the Byte Pair Encoding,[63] and ALBERT employed the SentencePiece.[64] To integrate the 4 transformer-based models into a unified package, we developed a preprocessing module to dynamically select a word segmentation algorithm. The preprocessing module also aligns the word-level NER tags to the subtoken NER tags. More specifically, it will: 1) assign the original word-level tag if the subtoken is the first token split from a word; 2) assign a special label "X" to other subtokens after the first 1. After preprocessing, the system applies a transformer to predict subtoken level NER tags. Then, a postprocessing module was developed to decode the token-level predictions back to word-level NER tags.

Pretraining and fine-tuning of transformer models

Pretraining

For each of the 4 transformer-based models, we explored general models pretrained using general English corpora and clinical models pretrained with clinical corpora, as previous studies[17,18] showed that pretraining on a clinical corpus improved the performance of clinical concept extraction. For the general models, we adopted the existing benchmark transformer-based models pretrained using general English domain corpora including bert-base-uncased (BERT-general), roberta-base (RoBERTa-general), albert-base-v2 (ALBERT-general), and electra-base (ELECTRA-general). Based on the general models, we further pretrained them using clinical notes from the Medical Information Mart for Intensive Care III (MIMIC-III) database.[65] We denote the clinical transformer-based models pretrained using MIMIC-III corpus as BERT-MIMIC, ALBERT-MIMIC, RoBERTa-MIMIC, and the ELECTRA-MIMIC, respectively.

Fine-tuning

Based on the transformer models pretrained using the MIMIC data (an unsupervised training procedure that doesn't require labels), we further added a linear classification layer to predict NER tags using clinical concepts labeled in the training corpora. At this stage, both the parameters of transformer models and the parameters of the classification layer will be optimized for clinical concept extraction.

Experiments and evaluation

Experiment set up

We developed a clinical concept extraction package on top of existing transformer architectures implemented in the Transformers library developed by the HuggingFace team[11] using PyTorch.[66] The default parameters were used to pretrain the transformer models using the MIMIC-III corpus. To measure the perplexity scores, we held out 5% of the MIMIC-III corpus as an evaluation set. The detailed pretraining hyperparameters are available in Supplementary Material Table S1. To fine-tune transformer models for clinical concept extraction, we split 10% of annotated notes from the training dataset as a validation set and used the rest of the notes as a (short) training set. The best model was selected according to the validation performance measured by strict F1-scores on the validation set. We adopted an early stop strategy to stop the training when there were no improvements observed in 5 consecutive epochs. We conducted all experiments using 2 Nvidia P6000 GPUs. An LSTM-CRFs model developed in our previous study[52] was used as the baseline.

Evaluation metrics

Following the standard evaluation of clinical concept extraction, we compared the performance of transformer-based NER models using the strict microaveraged precision, recall, and F1-score aggregated from all entity categories. The official evaluation scripts provided by the 2010 i2b2 and the 2018 n2c2 challenges were used to calculate the scores.

RESULTS

Table 2 compares the 4 transformer models for clinical concept extraction on 3 public clinical NER datasets developed by the 2010 i2b2 challenge, 2012 i2b2 challenge, and 2018 n2c2 challenge. For each transformer model, we compared a general model pretrained using general English domain corpora and a clinical model pretrained using the clinical notes from the MIMIC III database. Among all models, the RoBERTa-MIMIC achieved the best performance on the 3 datasets with F1-scores of 0.8994, 0.8053, and 0.8907, respectively. Compared to the baseline LSTM-CRFs model, the RoBERTa-MIMIC significantly improved the F1-scores by ~4% and 6% on the 2010 and 2012 i2b2 datasets. Compared with the general transformer models, the clinical transformer models improved the performance for clinical concept extraction. For example, on the 2010 i2b2 dataset, the RoBERTa-MIMIC outperformed the RoBERTa-general by ~2% in terms of F1-score (0.8994 vs 0.8822). The BERT and ALBERT-based models achieved similar performances on the 3 datasets, indicating that the lightweight BERT model, ALBERT, works as well as the full BERT model. Since ELECTRA shared the same model architecture as the BERT, the ELECTRA models achieved similar performances as the BERT models on all 3 datasets. Notably, the RoBERTa-general model matched or exceeded the performance of our original LSTM-CRFs and the RoBERTa-MIMIC model consistently outperformed it.

DISCUSSION

Accurate identification of clinical concepts from clinical narratives is a fundamental task to enable many downstream research and clinical applications leveraging patient information documented in unstructured clinical text. This study systematically explored 4 widely used transformer architectures for clinical concept extraction and compared them with the LSTM-CRFs model as a strong baseline (ie, the previous state-of-the-art model). Our evaluation using 3 public clinical NER datasets showed that 2 transformer models, RoBERTa and BERT, outperformed the previous widely used LSTM-CRFs deep learning model for clinical concept extraction. Among the 4 transformer models, RoBERTa achieved the best performance on all 3 public datasets. Compared with the baseline performance of LSTM-CRFs, RoBERTa remarkably improved the F1-score by approximately 4% and 6% on the 2010 and 2012 i2b2 datasets. The ALBERT and ELECTRA based models achieved performances comparable to the BERT based models on all 3 clinical NER datasets.

Our study demonstrated that it is necessary to pretrain transformer-based models using clinical text when applying them in the medical domain. The 4 clinical transformer models outperformed their corresponding general models for clinical concept extraction. For example, the RoBERTa-MIMIC model outperformed

Table 2. The strict level performances on the 2010 i2b2, 2012 i2b2 and 2018 n2c2 test set

Model	2010 i2b2 test set			2012 i2b2 test set			2018 n2c2 test set		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
LSTM-CRFs[a]	0.8737	0.8509	0.8621	0.7742	0.7215	0.7469	**0.9016**	0.8593	0.8849
BERT-general	0.8636	0.8752	0.8694	0.7353	0.7748	0.7546	0.8887	0.8728	0.8807
BERT-MIMIC	0.8847	0.8965	0.8905	0.7682	0.8184	0.7925	0.8835	0.8871	0.8853
RoBERTa-general	0.8777	0.8867	0.8822	0.7649	0.8007	0.7824	0.8821	0.8804	0.8812
RoBERTa-MIMIC	**0.8963**	**0.9026**	**0.8994**	**0.7930**	0.8181	**0.8053**	0.8927	0.8888	**0.8907**
ALBERT-general	0.8619	0.8758	0.8688	0.7561	0.7766	0.7662	0.8772	0.8766	0.8769
ALBERT-MIMIC	0.8937	0.8893	0.8915	0.7836	**0.8210**	0.8019	0.8776	**0.8909**	0.8842
ELECTRA-general	0.8689	0.8781	0.8735	0.7512	0.7983	0.7740	0.8797	0.8805	0.8801
ELECTRA-MIMIC	0.8801	0.8955	0.8877	0.7851	0.8146	0.7996	0.8814	0.8857	0.8836

[a]The LSTM-CRFs results for 2010 i2b2 and 2018 n2c2 were originally reported in our previous works.[52,67] Best precision, recall, and F1-score are highlighted in bold.

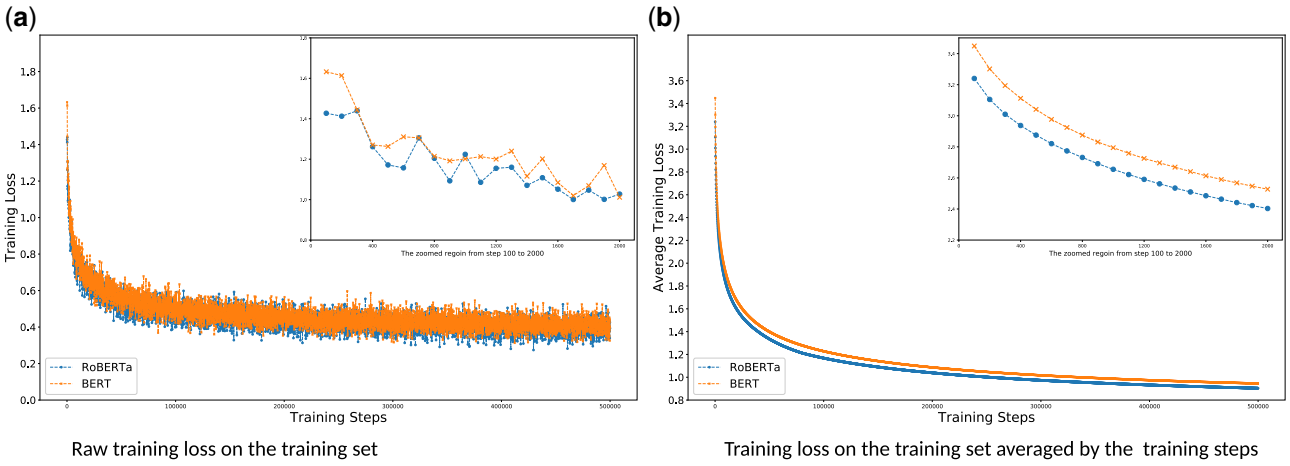

Figure 2. Comparison of pretraining loss for BERT and RoBERTa.

the RoBERTa-general by approximately 1.7%, 2%, and 1% on the 3 datasets, respectively (Table 2). The same trend was observed for other transformer models. The general RoBERTa transformer nevertheless performed comparably to LSTM-CRFs. We further examined the training loss and perplexity scores for the top 2 transformer models (ie, BERT and RoBERTa), using 5% notes held out from the MIMIC-III corpus. Supplementary Material Table S10 compares the detailed perplexity scores and associated F1-scores of the 2 models during the pretraining. Without pretraining using clinical corpora, the perplexities for the general (ie, pretrained using general English corpora only) BERT and RoBERTa were 41.6099 and 32.5643, respectively. Both of their perplexities decreased along with the pretraining using clinical notes from MIMIC-III. After pretraining for 10 epochs, the perplexities for BERT and RoBERTa dropped to 2.1818 and 2.0823, respectively. Figure 2 compares the pretraining loss between BERT and RoBERTa.

We also compared the precision and recall of the RoBERTa-MIMIC with the baseline LSTM-CRFs model and found that the improvement on F1-score mainly from the recall. For example, the recalls of RoBERTa-MIMIC are 2.95% (0.8888 vs 0.8593), 9.66% (0.8181 vs 0.7215), and 2.95% (0.8888 vs 0.8593) higher than the recalls of the LSTM-CRFs on the 3 datasets, respectively, demonstrating the efficiency of transformer models for clinical concept extraction.

The performance of the best model, RoBERTa-MIMIC, is comparable with state-of-the-art models reported in the medical domain. Si et al reported that a BERT-large model achieved F1-score of 0.9025 on the 2010 i2b2 dataset.[18] Although our RoBERTa-MIMIC model (here we adopted the RoBERTa base model) is only ~1/3 of the BERT-large model, our performances are comparable to the BERT-large model with the advantage of a short training time. We compared our RoBERTa-MIMIC model with the BERT-large model reported by Si et al[18] on the concept level using the 2010 i2b2 dataset. Our RoBERTa-MIMIC model achieved better F1-scores for PROBLEM and TEST (0.9084 vs 0.8926 on PROBLEM; 0.8955 vs 0.8880 on TEST) compared to the BERT-large model. For TREATMENT, the F1-score achieved by our RoBERTa-MIMIC was only 0.0008 lower than the BERT-large model (0.8906 vs 0.8914). Supplementary Table S9 compares the training time of the 4 transformer models examined in this study with the BERT-large model previously reported. The average training time per epoch was 922 seconds for our RoBERTa-MIMIC, which is about 1/3 of the BERT-large model of 2,804 seconds. The Alibaba team reported a hybrid system that achieved the best strict F1-score of 0.8956 in the 2018 n2c2 challenge focusing on extracting adverse drug events (ADE).[25] Here, the single RoBERTa-MIMIC model achieved comparable performance (F1 score of 0.8907) on the 2018 n2c2 dataset. A key difficulty in the 2018 n2c2 challenge is that the performance of

extracting ADEs is relatively low compared with other tasks in the challenge across all participants. Our RoBERTa-MIMIC model achieved a better lenient F1-score (see Supplementary Material Table S8) than the best system in the 2018 n2c2 challenge (0.5936 vs 0.5731) for extraction of ADEs.

Clinical concept extraction is 1 of the well-established clinical NLP tasks that have been extensively explored in recent decades. Thanks to the open challenges organized by the NLP research community, we can see how breakthroughs in general NLP algorithms can improve the accuracy of clinical concept extraction, thus reducing the noise passed to the downstream components. Compared with the best system reported in the 2010 i2b2 challenge (deBruijn et al[38]) the proposed RoBERTa-MIMIC model improved the strict F1-score from 0.8520 to 0.8994.

This study has limitations. We mainly focused on clinical concept extraction, which is a word-level NLP task. Recent studies from the general English domain have explored BERT for sentence-level NLP tasks and reported promising results for semantic textual similarity,[68] clinical records classification,[69] relation extraction,[70,71] and question-answering.[72] Further studies should examine the transformer-based models for sentence-level and document-level NLP tasks.

CONCLUSION

In this study, we systematically evaluated 4 transformer models using 3 public clinical datasets and developed an open-source clinical transformer package for clinical concept extraction. Our study demonstrated the efficiency of transformers for the extraction of various types of clinical concepts from clinical narratives. Our methods and systems can be applied to other clinical NLP tasks through fine-tuning. The transformer package with 4 pretrained clinical transformer models is publicly available at https://github.com/uf-hobi-informatics-lab/ClinicalTransformerNER. We believe this package will improve current practice on clinical concept extraction and benefit other related clinical NLP tasks.

FUNDING

This study was partially supported by a Patient-Centered Outcomes Research Institute Award (ME-2018C3-14754), a grant from the National Cancer Institute, 1R01CA246418 R01, a grant from the National Institute on Aging, NIA R21AG062884, the University of Florida Informatics Institute Junior SEED Program (00129436), and the Cancer Informatics and eHealth core jointly supported by the University of Florida Health Cancer Center and the University of Florida Clinical and Translational Science Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding institutions.

AUTHOR CONTRIBUTIONS

XY, JB, and YW were responsible for the overall design, development, and evaluation of this study. XY, JB, and YW wrote the initial drafts and revisions of the manuscript. WRH also contributed to writing and editing of this manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

DATA AVAILABILITY

The software package is publicly available under the MIT license at https://github.com/uf-hobi-informatics-lab/ClinicalTransformerNER.

ACKNOWLEDGMENTS

We would like to thank the i2b2 and n2c2 challenge organizers for providing the annotated corpus. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Wang Y, Wang L, Rastegar-Mojarad M, *et al*. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018; 77: 34–49.
2. Wang Y, Wang L, Rastegar-Mojarad M, *et al*. Systematic analysis of free-text family history in electronic health record. *AMIA Jt Summ Transl Sci Proc* 2017; 2017: 104–13.
3. Luo Y, Thompson WK, Herr TM, *et al*. Natural language processing for EHR-based pharmacovigilance: a structured review. *Drug Saf* 2017; 40 (11): 1075–89.
4. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1*. Washington (DC): National Academies Press (US); 2014.
5. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. Washington (DC): National Academies Press (US); 2015.
6. Fu S, Chen D, He H, *et al*. Development of clinical concept extraction applications: a methodology review. *arXiv: 191011377 [cs]* Published Online First: 2 March 2020. http://arxiv.org/abs/1910.11377.
7. Masanz J, Pakhomov SV, Xu H, *et al*. Open Source Clinical NLP–more than any single system. *AMIA Jt Summ Transl Sci Proc* 2014; 2014: 76–82.
8. Shickel B, Tighe PJ, Bihorac A, *et al*. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform* 2018; 22 (5): 1589–604.
9. Li J, Sun A, Han J, *et al*. A survey on deep learning for named entity recognition. *CoRR*: abs/1812.09449; 2018.
10. Wu S, Roberts K, Datta S, *et al*. Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc* 2020; 27 (3): 457–70.
11. Wolf T, Debut L, Sanh V, *et al*. HuggingFace's transformers: state-of-the-art natural language processing. *arXiv*: abs/1910.03771; 2019.
12. Devlin J, Chang M-W, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv: 181004805*; 2018.
13. Yang Z, Dai Z, Yang Y, *et al*. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:190608237 [cs]* Published Online First: 2 January 2020. http://arxiv.org/abs/1906.08237.
14. Lan Z-Z, Chen M, Goodman S, *et al*. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv*: **abs/1909.11942;** 2019.
15. Liu Y, Ott M, Goyal N, *et al*. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*: **abs/1907.11692;** 2019.
16. Raffel C, Shazeer N, Roberts A, *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv: 191010683 [cs, stat]*; 2019. http://arxiv.org/abs/1910.10683.

17. Lee J, Yoon W, Kim S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2019; doi: 10.1093/bioinformatics/btz682.
18. Si Y, Wang J, Xu H, *et al*. Enhancing clinical concept extraction with contextual embeddings. *J Am Med Inform Assoc* 2019; 26 (11): 1297–304. doi : 10.1093/jamia/ocz096.
19. Huang K, Singh A, Chen S, *et al*. Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation. *arXiv:191211975 [cs]*/Published Online First: 26 December 2019. http://arxiv.org/abs/1912.11975.
20. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: proceedings of the 18th BioNLP Workshop and Shared Task; 2019. doi: 10.18653/v1/w19-5006.
21. Alsentzer E, Murphy J, Boag W, *et al*. Publicly available clinical BERT embeddings. In: proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019: 72–8.
22. Clark K, Luong M-T, Le QV, *et al*. ELECTRA: pre-training text encoders as discriminators rather than generators. *arXiv: 200310555 [cs]*; 2020. http://arxiv.org/abs/2003.10555.
23. Uzuner Ö, South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011; 18 (5): 552–6.
24. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc* 2013; 20 (5): 806–13.
25. Henry S, Buchan K, Filannino M, *et al*. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc* 2020; 27 (1): 3–12.
26. Friedman C, Johnson SB, Forman B, *et al*. Architectural requirements for a multipurpose natural language processor in the clinical environment. In: *proceedings of the Annual Symposium on Computer Application in Medical Care;* 1995: 347–51.
27. Friedman C. A broad-coverage natural language processing system. *Proc AMIA Symp* 2000; 2000: 270–4.
28. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010; 17 (3): 229–36.
29. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
30. Soysal E, Wang J, Jiang M, *et al*. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018; 25 (3): 331–6.
31. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* 2010; 17 (5): 524–7.
32. Xu Y, Wang Y, Liu T, *et al*. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *J Am Med Inform Assoc* 2013; 20 (5): 849–58.
33. Pradhan S, Chapman W, Man S, *et al*. SemEval-2014 Task 7: Analysis of Clinical Text. In: *proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics; 2014: 54–62.
34. Elhadad N, Pradhan S, Gorman S, *et al*. SemEval-2015 Task 14: Analysis of Clinical Text. In: proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics. doi: 10.18653/v1/S15-2051. 2015: 303–10.
35. Bethard S, Savova G, Chen W-T, *et al*. SemEval-2016 Task 12: Clinical TempEval. In: proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Diego, California: Association for Computational Linguistics. doi: 10.18653/v1/S16-1165. 2016: 1052–62.
36. Suominen H, Salanterä S, Velupillai S, *et al*. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Forner P, Müller H, Paredes R, *et al*., eds. *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*. Berlin: Springer; 2013: 212–31.
37. Kelly L, Goeuriot L, Suominen H, *et al*. Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In: Kanoulas E, Lupu M, Clough P, *et al*., eds. *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*. Cham: Springer International Publishing; 2014: 172–191.
38. deBruijn B, Cherry C, Kiritchenko S, *et al*. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc* 2011; 18: 557–62. doi:10.1136/amiajnl-2011-000150.
39. Zhang Y, Wang J, Tang B, *et al*. UTH_CCB: a report for semeval 2014–task 7 analysis of clinical text. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*; 2014: 802–6.
40. Xu J, Zhang Y, Wang J, *et al*. UTH-CCB: The Participation of the SemEval 2015 Challenge—Task 14. In: proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015); 2015: 311–4.
41. Uzuner Ö, Stubbs A, Lenert L. Advancing the state of the art in automatic extraction of adverse drug events from narratives. *J Am Med Inform Assoc* 2020; 27 (1): 1–2.
42. Brown PF, deSouza PV, Mercer RL, *et al*. Class-based N-gram Models of Natural Language. *Comput Linguist* 1992; 18: 467–79.
43. Kanerva P, Kristoferson J, Holst A. Random Indexing of Text Samples for Latent Semantic Analysis. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*; 2000: 10–36.
44. Sahlgren M. An introduction to random indexing. In: Methods and Applications of Semantic Indexing Workshop at the TKE 2005. TermNet News, 87: 1–9.
45. Tang B, Cao H, Wu Y, *et al*. Clinical entity recognition using structural support vector machines with rich features. ACM *Sixth International Workshop on Data and Text Mining in Biomedical Informatics*, Maui, HI; 2012: 13–20.
46. Wu Y, Xu J, Jiang M, *et al*. A study of neural word embeddings for named entity recognition in clinical text. *AMIA Ann Symp Proc* 2015; 2015: 1326–33.
47. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 2013; 35 (8): 1798–828.
48. Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, Volume 2. USA: Curran Associates Inc.; 2013: 3111–9. http://dl.acm.org/citation.cfm? id=2999792.2999959.
49. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014. 1532–43.
50. Joulin A, Grave E, Bojanowski P, *et al*. FastText.zip: Compressing text classification models. *arXiv: 161203651*; 2016.
51. Lample G, Ballesteros M, Subramanian S, *et al*. Neural Architectures for Named Entity Recognition. *CoRR*: abs/1603.01360. http://arxiv.org/abs/1603.01360; 2016.
52. Wu Y, Yang X, Bian J, *et al*. Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition. *AMIA Annu Symp Proc* 2018; 2018: 1110–7.
53. Bengio Y, Schwenk H, Senécal J-S, *et al*. Neural probabilistic language models. In: Holmes DE, Jain LC, eds. *Innovations in Machine Learning: Theory and Applications*. Berlin: Springer; 2006: 137–86.
54. Vaswani A, Shazeer N, Parmar N, *et al*. Attention Is All You Need. *arXiv:170603762 [cs]* Published Online First: 5 December 2017. http://arxiv.org/abs/1706.03762.
55. Liu X, Croft WB. Statistical language modeling for information retrieval. *Ann Rev Info Sci Technol* 2006; 39 (1): 1–31.
56. Taylor WL. Cloze procedure: a new tool for measuring readability. *Journalism Quarterly* 1953; 30 (4): 415–33.
57. Yan H, Deng B, Li X, *et al*. TENER: adapting transformer encoder for named entity recognition. *arXiv: 191104474 [cs]*; 2019. http://arxiv.org/abs/1911.04474
58. Wang A, Singh A, Michael J, *et al*. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv:180407461 [cs]* Published Online First: 22 February 2019. http://arxiv.org/abs/1804.07461.

59. Williams A, Nangia N, Bowman S. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *arXiv:170405426 [cs]* Published Online First: 19 February 2018. http://arxiv.org/abs/1704.05426.
60. Rajpurkar P, Zhang J, Lopyrev K, *et al*. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv:160605250 [cs]* Published Online First: 10 October 2016. http://arxiv.org/abs/1606.05250.
61. Kudo T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. *arXiv:180410959 [cs]* Published Online First: 29 April 2018. http://arxiv.org/abs/1804.10959.
62. Wu Y, Schuster M, Chen Z, *et al*. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv*; abs/1609.08144; 2016.
63. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. *arXiv*; abs/1508.07909; 2015.
64. Kudo T, Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *arXiv:180806226 [cs]* Published Online First: 19 August 2018. http://arxiv.org/abs/1808.06226.
65. Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1): 160035.
66. Paszke A, Gross S, Massa F, *et al*. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, *et al*., eds. *Advances in Neural Information Processing Systems 32*. Curran Associates; 2019: 8024–35. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
67. Yang X, Bian J, Fang R, *et al*. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *J Am Med Inform Assoc* 2020; 27 (1): 65–72.
68. Xiong Y, Chen S, Qin H, *et al*. Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity. *BMC Med Inform Decis Mak* 2020; 20 (S1): 72.
69. Yao L, Jin Z, Mao C, *et al*. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. *J Am Med Inform Assoc* 2019; 26 (12): 1632–6.
70. Wei Q, Ji Z, Si Y, *et al*. Relation extraction from clinical narratives using pre-trained language models. *AMIA Annu Symp Proc* 2019; 2019: 1236–45.
71. Alimova I, Tutubalina E. Multiple features for clinical relation extraction: a machine learning approach. *J Biomed Inform* 2020; 103: 103382. doi : 10.1016/j.jbi.2020.103382.
72. Schmidt L, Weeds J, Higgins JPT. Data Mining in Clinical Trial Text: Transformers for Classification and Question Answering Tasks. *arXiv: 200111268 [cs]*; 2020. http://arxiv.org/abs/2001.11268.