CliNER 2.0: Accessible and Accurate Clinical Concept Extraction

Willie Boag
MIT CSAIL
Cambridge, MA
wboag@mit.edu

Elena Sergeeva
MIT CSAIL
Cambridge, MA
elenaser@mit.edu

Saurabh Kulshreshtha
UMass Lowell
Lowell, MA
skul@cs.uml.edu

Peter Szolovits
MIT CSAIL
Cambridge, MA
psz@mit.edu

Anna Rumshisky
UMass Lowell
Lowell, MA
arum@cs.uml.edu

Tristan Naumann
MIT CSAIL
Cambridge, MA
tjn@mit.edu

Abstract

Clinical notes often describe important aspects of a patient's stay and are therefore critical to medical research. Clinical concept extraction (CCE) of named entities — such as problems, tests, and treatments — aids in forming an understanding of notes and provides a foundation for many downstream clinical decision-making tasks. Historically, this task has been posed as a standard named entity recognition (NER) sequence tagging problem, and solved with feature-based methods using hand-engineered domain knowledge. Recent advances, however, have demonstrated the efficacy of LSTM-based models for NER tasks, including CCE. This work presents *CliNER 2.0*, a simple-to-install, open-source tool for extracting concepts from clinical text. CliNER 2.0 uses a word- and character- level LSTM model, and achieves state-of-the-art performance. For ease of use, the tool also includes pre-trained models available for public use.

1 Introduction

Although there is a trend toward digitizing patient records in an increasingly structured manner, much information is still hidden in unstructured narrative text. In their primary role, electronic health record (EHR) notes facilitate patient care by recording communication among care staff. These clinical notes capture patient data that provide insight into a patient's status and courses of care, such as patient history, recommended treatments, records of meetings, and more. Often, this granularity of data does not appear, or does not appear in equivalent detail, in a structured form elsewhere in the EHR. It is no surprise then, that leveraging clinical notes is critical to the successful analysis of EHR data — an important secondary use.

Clinical concept extraction (CCE) improves our understanding of notes and our ability to analyze them; identifying for example, *problems* and symptoms a patient has exhibited, *tests* that have been run, and *treatments* that have been administered. CCE, much like standard named entity recognition (NER), has traditionally been posed as a sequence tagging task, handled by feature-based methods using hand-engineered domain knowledge. In this formulation, tokens (i.e., pre-processed words) are predicted to be inside, outside, or beginning (IOB) a concept span; thus, allowing the identification concepts that span multiple, contiguous tokens. Subsequently, each identified span is predicted to be one of the specified concepts, as shown in Figure 1.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

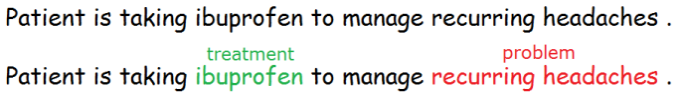

Figure 1: Identifying concept spans in clinical text. Concept spans can be any number of contiguous tokens; in this case there are spans of length one and two, respectively.

Community efforts to advance concept extraction have led to numerous shared task competitions. Notably, in 2010 Informatics for Integrating Biology & the Bedside (i2b2) and the U.S. Department of Veterans Affairs (VA) partnered to hold the Workshop on Natural Language Processing Challenges for Clinical Records [Uzuner et al., 2011]. This workshop included a task for concept extraction from clinical discharge summaries, the objective of which was to identify contiguous spans of tokens as in narrative text and classify their concept type.

The 2010 i2b2/VA workshop made an important step forward in releasing the data required for building tools to identify clinically important concepts. However, despite the existence of tasks like this, the competition format has not incentivized open and easy-to-use tools for researchers who require simple solutions. Most submissions are not released as open systems, thus limiting the availability of concept extraction tools. To address these concerns, the CliNER project was developed as an open-source clone of a top-performing i2b2/VA 2010 challenge system Boag et al. [2015].

This work presents *CliNER 2.0*, a simple-to-install, open-source tool for extracting concepts from clinical text, which includes pre-trained models for additional ease of use. [1] Much like general domain natural language processing (NLP), clinical NLP has also been shown to benefit from deep learning models that can better learn complex patterns from data. Recently, Dernoncourt et al. [2016] proposed a word- and character-level LSTM model for the de-identification task that outperformed all existing baselines. In *CliNER 2.0*, we adopt this approach for concept extraction, integrating the same state-of-the-art deep learning NER architecture described above into the tool.

2 Related Work

The 2010 i2b2/VA Workshop on NLP Challenges for Clinical Records [Uzuner et al., 2011] promoted the development of 22 systems towards the task of concept extraction from discharge summaries. The winning system achieved an exact F measure of 0.852 by using a discriminative semi-Markov HMM, trained using passive-aggressive online updates [deBruijn et al., 2011]. Many other top performing methods used a Conditional Random Field (CRF) to model the sequence learning problem [Roberts and Harabagiu, 2011].

In the years following the shared task workshop, the dataset proved very useful as a research benchmark. Numerous systems and methods that have been developed can be compared against one another using this dataset. Early successful attempts utilized the strengths of workshop participants (sequential models, such as a CRF) and added generalized word representations using distributional semantics [Fu and Ananiadou, 2014, Jonnalagadda et al., 2012, Wu et al., 2015]. Since then, deep learning and recurrent neural networks have increased in popularity and easiness-to-implement, leading to a many LSTM-based approaches to clinical concept extraction [Chalapathy et al., 2016, Unanue et al., 2017].

The most widely-used clinical NLP tool, cTAKES [Savova et al., 2010], relies nearly exclusively on UMLS-based dictionary lookups [Bodenreider, 2004]. In doing so, cTAKES achieves high recall (at the cost of low precision) by identifying all phrases that have any potential to be a relevant concept. While this property may be desirable for search-related tasks, it's lack of relevance to many downstream clinical decision-making tasks has been noted as the reason for the development of additional tools [Divita et al., 2014, Kang et al., 2017, Soysal et al., 2017]. This limitation, in addition to a desire for out-of-the-box usability motivated the creation of the original CliNER [Boag et al., 2015]

[1] CliNER 2.0 can be downloaded from `https://github.com/text-machine-lab/CliNER`.

3 Data

We use data from the i2b2/VA 2010 challenge. [2] This dataset contains 16,107 sentences, which are 6–9 words long on average, from patient discharge summaries. There are 169 summaries made available for training and 255 summaries available for testing. Note the training data are smaller than the testing data, a result of nearly a third of the training data being revoked following the challenge.

4 Methods

CliNER 2.0 has two options for building machine learning models: 1) traditional CRF-based learning with domain expert features, or 2) deep learning with a state-of-the-art neural architecture. These options afford flexibility with respect to desired time and space constraints; notably, the CRF model is smaller than a large hierarchical LSTM. Both models employ a word-level prediction using the IOB format. For three concept types — problem, test, and treatment — this results in a 7-way tag prediction for each token.

4.1 CRF with UMLS Features

The CRF-based option heavily employs feature extraction using both linguistic features (e.g., ngrams and wordshapes) and domain knowledge (e.g., UMLS Metathesaurus). POS tagging was performed with the general-domain nltk pos_tagger [Bird, 2002]. Table 1 shows a full list of features that are extracted for each token. These features are extracted for each individual token, except for *prev1-all-feats* and *next1-all-feats*, which include all word-level features of the previous and next tokens, respectively. These features are then fed into a wrapper for CRFsuite, a fast CRF implementation [Okazaki, 2007].

Table 1: Features for the CRF.

word unigram	last-2 characters	word shape	part-of-speech
regexes of units	length	umls-cui	umls-lui
umls-rel	umls-sty	umls-tty	umls-abr
prev3-unigrams	next3-unigrams	*prev1-all-feats*	*next1-all-feats*

4.2 Hierarchical LSTM

The hierarchical LSTM option employs both word- and character- level bidirectional LSTMs (w+c Bi-LSTM). For each word w_t in a sentence, we consider the sequence of characters that comprise that word: $w_t^1, w_t^2, w_t^3, ...$ The embeddings for this sequence of characters w_t^i are fed into the Bi-LSTM$_t$ corresponding to the t^{th} word, and concatenated to the final forward and backward hidden states to create a character-level representation of the word. All character-level Bi-LSTMs share the same weights. Finally, the character-level representation is concatenated with its standard word embedding (pre-trained GloVe vectors available at `http://neuroner.com/data/word_vectors/glove.6B.100d.zip`) to form a rich word- and character- level representation of token t. From there, this word representation is fed in to a standard (Bi-)LSTM-CRF framework. Figure 2 depicts the w+c LSTM-CRF model.

5 Results

Table 2 displays the results from recent work on the 2010 i2b2/VA concept extraction task. Notably excluded are the results of the top-performing systems from the original task. These systems are removed because their reported performance was obtained using now-revoked training data and the systems are not available to train again using the limited subset. Consequently, their reported performances from the 2010 competition are not comparable to recent work.

The remaining top-performing systems are are all deep neural models using LSTMs and neural embeddings. Their relative performances vary and no single, clear winner stands out. While the

[2] i2b2/VA 2010 challenge data are available at `https://www.i2b2.org/NLP/DataSets/Main.php`.

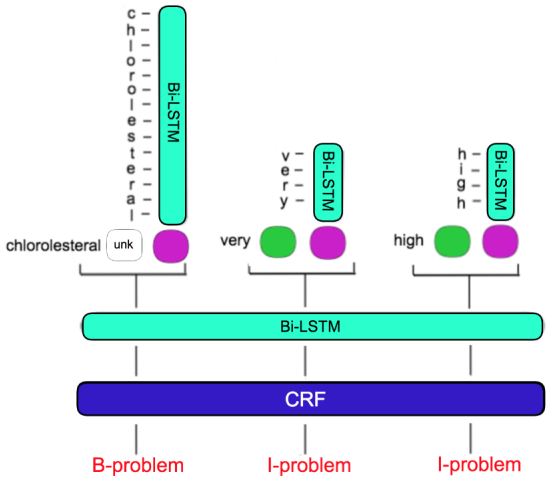

Figure 2: Character-level LSTMs that feed into word-level LSTMs. Unlike purely word-level models, this approach is sensitive to misspellings such as "cholorolesteral."

Table 2: Precision, recall, and F1 of selected concept extraction models.

	Exact Class Match		
	Precision	Recall	F1
Truecasing CRFSuite [Fu and Ananiadou, 2014]	0.808	0.715	0.759
Binarized Neural Embedding CRF [Wu et al., 2015]	**0.851**	0.806	0.828
LSTM-CRF: GloVe Chalapathy et al. [2016]	0.844	0.834	**0.839**
CliNER 2.0: feats+CRF	0.835	0.758	0.795
CliNER 2.0: w+c LSTM-CRF: GloVe	0.840	**0.836**	0.838

Binarized Neural Embedding CRF achieves the best precision, the other LSTM-CRF models (some of which also use character-based LSTMs) all independently achieve the best recall and F1 scores.

The nearly identical performance of three systems — and the limited gains of additional character LSTMs, pre-trained word embeddings, and other enhancements — suggest that the community might be reaching the limits of this task. Performance on the (rather small) i2b2 dataset has effectively plateaued, with little-to-no recent improvements in performance.

Notably, both the shallow CRF and deep w+c LSTM-CRF models available in CliNER 2.0 match, or exceed, the top performing systems in their respective machine learning paradigm. This puts CliNER 2.0 performance among the highest possible using either of the methods.

6 Conclusion

We present *CliNER 2.0*, an updated, open-source clinical named entity recognition tool that matches state-of-the-art performance, and achieves the highest reported recall among systems trained on the 2010 i2b2 data. While other tools like cTakes often identify a large number possible concepts for a given span, it can be overwhelming when many are not relevant. *CliNER 2.0*, on the other hand, has a much less intrusive number of false positives, and focuses specifically on the identification of relevant concepts: problems, tests, and treatments.

Further, *CliNER 2.0* is easy to install. Pre-trained models are available for public use, which allow the tool to run out-of-the-box. To our knowledge, this represents the first open-source and pre-trained deep learning model available for state-of-the-art concept extraction. In addition, the tool has an optional flag that can disable the deep network; instead, backing off to a simple CRF model with UMLS-derived features. This option could be useful in resource constrained settings.

Acknowledgments

This research was funded in part by the Intel Science and Technology Center for Big Data, a Philips-MIT research agreement, the National Science Foundation Graduate Research Fellowship Program grant No. 1122374, and grants from the National Institutes of Health (NIH): National Library of Medicine (NLM) Biomedical Informatics Research Training grant 2T15 LM007092-22.

References

S. Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

W. Boag, K. Wacome, T. Naumann, and A. Rumshisky. CliNER: A lightweight tool for clinical named entity recognition. In *AMIA Joint Summits on Clinical Research Informatics*, 2015.

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

R. Chalapathy, E. Z. Borzeshi, and M. Piccardi. Bidirectional LSTM-CRF for clinical concept extraction. *arXiv preprint arXiv:1611.08373*, 2016.

B. deBruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. 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*, 18:557–562, 2011.

F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits. De-identification of patient notes with recurrent neural networks. 2016.

G. Divita, Q. T. Zeng, A. V. Gundlapalli, S. Duvall, J. Nebeker, and M. H. Samore. Sophia: a expedient UMLS concept extraction annotator. In *AMIA Annual Symposium Proceedings*, volume 2014, page 467. American Medical Informatics Association, 2014.

X. Fu and S. Ananiadou. Improving the extraction of clinical concepts from clinical records. *Proceedings of BioTxtM14*, 2014.

S. Jonnalagadda, T. Cohen, S. Wu, and G. Gonzalez. Enhancing clinical concept extraction with distributional semantics. *Journal of biomedical informatics*, 45(1):129–140, 2012.

T. Kang, S. Zhang, Y. Tang, G. W. Hruby, A. Rusanov, N. Elhadad, and C. Weng. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, page ocx019, 2017.

N. Okazaki. Crfsuite: a fast implementation of conditional random fields (CRFs), 2007. URL http://www.chokkan.org/software/crfsuite/.

K. Roberts and S. Harabagiu. A flexible framework for deriving assertions from electronic medical records. *J Am Med Inform Assoc*, 18:568–573, 2011.

G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, and H. Xu. Clamp – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, page ocx132, 2017. doi: 10.1093/jamia/ocx132. URL +http://dx.doi.org/10.1093/jamia/ocx132.

I. J. Unanue, E. Z. Borzeshi, and M. Piccardi. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *CoRR*, abs/1706.09569, 2017. URL http://arxiv.org/abs/1706.09569.

O. Uzuner, B. South, S. Shen, and S. DuVal. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. volume 18, page 552–6, 2011.

Y. Wu, J. Xu, M. Jiang, Y. Zhang, and H. Xu. A study of neural word embeddings for named entity recognition in clinical text. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2015:1326—1333, 2015. ISSN 1942-597X. URL http://europepmc.org/articles/PMC4765694.