CliNER: A Lightweight Tool for Clinical Named Entity Recognition

William Boag[1*], Kevin Wacome[1*], Tristan Naumann, MS[2], Anna Rumshisky, PhD[1]
[1] University of Massachusetts Lowell, Lowell, MA
[2] Massachusetts Institute of Technology, Cambridge, MA

Summary

CliNER is a lightweight machine learning system for the extraction of named entities from clinical text. *CliNER* is open-source and built to be easy to set up. *CliNER* implements current state-of-the-art practices for named entity recognition. System performance on the 2010 i2b2 shared task data matches the top reported results.

Introduction

Named entity recognition (NER) is essentially a solved problem for information extraction in the general domain, with an abundance of both the systems and the annotated data. Clinical domain has been lagging behind, with many existing systems (including cTAKES, for example) relying nearly exclusively on UMLS-based dictionary lookup. Despite the recent advances in clinical concept extraction, spurred by such community-wide initiatives as the 2010 i2b2 shared task[1], no lightweight, easy to set up, open-source implementation is available to date. General domain systems cannot be adapted directly to clinical data, since clinical concepts require domain-specific features, including those using specialized knowledge sources. The point of the shared tasks has been generate publicly available annotated data and to identify best-performing methods, which has helped to push forward the state of the art in the clinical-domain information extraction. At the same time, the format of the workshops has not allowed or encouraged the participants to develop fully functioning, user-friendly systems. Many of the systems developed during the shared task challenges or published subsequently are never released for public use[2,3], and frequently are put together haphazardly from opportunistically developed code components. On the other hand, the available systems developed outside of the shared task paradigm tend to be heavy aggregations of multiple components that require extensive set up and configuration, which presents a significant barrier to initial use.

In response to these concerns, the authors developed *CliNER*, a lightweight open-source named entity recognition system which implements current state-of-the-art practices for this task in the clinical domain. *CliNER* provides for easy installation and use "out of the box". The system is free and open-source, and is available on GitHub (via http://cliner.org), open for improvement through community contributions.

Methods

CliNER is a two-pass supervised machine learning system. The first pass identifies concept boundaries using linear chain conditional random fields[4]. The second pass assigns clinical concept types to the phrases identified in the first pass, using support vector machines[5]. *CliNER* uses a collection of text features ranging from simple general-domain features such as word, stem, and part-of-speech n-grams to domain-specific features using GENIA, UMLS Metathesaurus and Semantic Network. Training data annotated with desired clinical concepts can be used to build a model which can then be used to recognize similar concepts in raw text. A built-in option for parameter tuning can be used to optimize system performance for the specific annotation scheme and data genre used. The system can also be used out of the box using existing annotated corpora, such as the 2010 i2b2 shared task data.

Results

Evaluation of the system was performed against the 2010 i2b2 shared task data which identified medical problems, treatments, and tests in discharge summaries. Table 1 shows per-category precision, recall, and F-measure under the exact evaluation as defined for the shared task[1]. Exact evaluation does not give credit for partial span overlaps between system output and gold standard. *CliNER* obtains the micro-average F-measure of 0.800, compared to the 0.821– 0.852 range for the F-measures achieved by the top three systems in the shared task.[6,7,8]

Discussion

System performance is comparable to the best submissions from the 2010 i2b2 challenge. This is an encouraging result, since a few of the custom features used by the top shared task systems have not been implemented yet (e.g. some of the top systems used rule-based features to augment the supervised learning). Note that since i2b2 systems were never publicly released, we were unable to perform a comparison in terms of ease of installation and use. We hope that releasing the *CliNER* system to the community at large will encourage additional improvements that can further narrow the gap between the performance of general-domain and clinical-domain named entity recognition.

[*] Equal contribution by the first two authors.

Table 1. Precision, recall, and F-measure for exact evaluation on the 2010 i2b2 shared task data.

	Precision	Recall	F-measure
Problem	0.710	0.858	0.777
Treatment	0.834	0.752	0.791
Test	0.840	0.825	0.833
Micro-average	0.795	0.812	0.800

References

[1] Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 2011;18(5):552-556.

[2] Rink B, Harabagiu S, Roberts K. 2011. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, *18*(5), 594-600.

[3] Gobbel, G. T., Reeves, R., Jayaramaraja, S., Giuse, D., Speroff, T., Brown, S. H., ... & Matheny, M. E. (2014). Development and evaluation of RapTAT: a machine learning system for concept mapping of phrases from medical narratives. *Journal of biomedical informatics*, 48, 54-65.

[4] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proc. 18th International Conf. on Machine Learning* 2001;18:282-289.

[5] Cortes C, Vapnik V. Support-vector networks. Machine learning 1995;20(3):273.

[6] de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. NRC at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. *2010 i2b2/VA Workshop on challenges in natural language processing for clinical data*. 2010.

[7] Jiang M, Chen Y, Liu M, et al. Hybrid approaches to concept extraction and assertion classification – Vanderbilt's systems for 2010 I2B2 NLP Challenge. *2010 i2b2/VA Workshop on challenges in natural language processing for clinical data.* 2010.

[8] Kang N, Barendse RJ, Afzal Z, et al. Erasmus MC approaches to the i2b2 Challenge. *2010 i2b2/VA Workshop on challenges in natural language processing for clinical data.* 2010.