HHS Public Access

Author manuscript

Proceedings (IEEE Int Conf Bioinformatics Biomed). Author manuscript; available in PMC 2022 April 22.

Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2019 November ; 2019: 1548–1550. doi:10.1109/bibm47256.2019.8983406.

Clinical named-entity recognition: A short comparison

Juan Antonio Lossio-Ventura[α], **Sebastien Boussard**[β], **Juandiego Morzan**[θ], **Tina Hernandez-Boussard**[α]

[α]Department of Medicine, Biomedical Informatics, Stanford University, USA

[β]College of Engineering, Boston University, USA

[θ]School of Engineering, Universidad del Pacífico, Lima, Peru

Abstract

The adoption of electronic health records has increased the volume of clinical data, which has opened an opportunity for healthcare research. There are several biomedical annotation systems that have been used to facilitate the analysis of clinical data. However, there is a lack of clinical annotation comparisons to select the most suitable tool for a specific clinical task. In this work, we used clinical notes from the MIMIC-III database and evaluated three annotation systems to identify four types of entities: (1) procedure, (2) disorder, (3) drug, and (4) anatomy. Our preliminary results demonstrate that BioPortal performs well when extracting disorder and drug. This can provide clinical researchers with real-clinical insights into patient's health patterns and it may allow to create a first version of an annotated dataset.

Index Terms—

clinical research; electronic health records; named-entity recognition; natural language processing

I. Introduction

Healthcare organizations and goverment agencies have successfully adopted electronic health records (EHRs), which has created an explosion of clinical data available for research [1]. EHRs are mostly composed of unstructured free text, this free text often records critical information. The conversion of EHRs to structured data is labor-intensive, expensive, and can not successfully capture all relevant information.

The use of natural language processing (NLP) techniques allows to decrease the time and the human intervention needed to obtain critical information from free text which has positively been impacting biomedical and clinical research [2], [3]. On the other hand, clinical research has become difficult to measure due to the absence of available gold standard datasets, which are mainly manually created. To ensure a high quality of gold standard datasets, an inter-annotator agreement is calculated between the experts that are requested to independently annotate the data. This makes the creation of these datasets a

jlossio@stanford.edu .

costly process. Biomedical named-entity recognition (BioNER) is a very important task in NLP that aims to automatically recognize and classify biomedical entities; such as genes, disease names, medication names, procedures, proteins, chemicals, among others; from biomedical text. Different BioNER systems have been proposed to extract clinical concepts from text, including MedLEE [4], MetaMap [5], MetaMap Lite [6], KnowledgeMap [7], Apache cTAKES [8], HiTEX [9], MedTagger [10], CLAMP [11], QuickUMLS [12]. There is a wide range of BioNER systems in the literature. Overviews of these existing bioNER applications have been provided in [3], [13], [14]. However, there is a lack of comparison studies providing information to choose the most suitable tool for a specific entity extraction task in the clinical domain, for instance the extraction of treatments, disorders, drugs, procedures, anatomy. The ultimate goal of our project is to demonstrate the feasibility to create a clinical gold standard dataset using an ensemble of the more appropriate annotation systems, and to recommend the most suitable tool for the annotation of a specific clinical entity type (i.e., categories). In sum, the aim of this study is to compare three state-of-the-art annotation systems used successfully in the clinical and biomedical domain. We selected four categories for the evaluation: (1) drug (2) disorder, (3) procedure, and (4) anatomy.

II. Methods

A. Data

We used the publicly available Medical Information Mart for Intensive Care (MIMIC-III) database [15], [16]. For this study, we compiled a subset of 1,000 clinical notes from the table “NOTEEVENTS” which contains deidentified notes, including nursing and physician notes, ECG reports, imaging reports, and discharge summaries. The single rule to preprocess the extracted clinical notes was to delete null entries.

B. Annotation systems

We implemented the following three applications:

1. **BioPortal**[1]: is a web portal that provides access to a library of more than 1,100 biomedical ontologies and terminologies. BioPortal enables community participation in the evaluation and evolution of ontology content by providing features to add mappings between terms, to add comments linked to specific ontology terms and to provide ontology reviews [17], [18]. The open biomedical annotator provides REST Web service for the annotation task.
2. **CLAMP**[2]: is a Java-based clinical language annotation, modeling, and processing toolkit. CLAMP provides different state-of-the-art NLP modules such as entity recognition, entity linking, normalization. It also provides an integrated development environment with user-friendly graphic user interfaces to allow users to quickly build customized NLP pipelines for individual applications [11].
3. **ScispaCy**[3]: is a specialized Python NLP library for processing biomedical, scientific, and clinical texts which leverages the spaCy library[4], used and

[1]BioPortal annotator: https://bioportal.bioontology.org/annotator
[2]CLAMP: https://clamp.uth.edu/

evaluated on several NLP tasks such as part-of-speech tagging, dependency parsing, named entity recognition, and sentence segmentation [19].

C. Data analysis

We evaluated 1,000 clinical notes in terms of the following aspects: A) a general overview of all annotated entities for the four clinical categories; B) an overview of the entities extracted per category; and C) implementation requirements.

III. Experimental results

A. General overview

To evaluate the annotation of all entities of all categories with the three systems, we first evaluated the number of entities extracted by each application as shown in Table I. Then, we studied the co-identification (i.e., overlapping) of different entities by the three applications over the dataset. Fig. 1 illustrates the distribution of distinct entities extracted by the three different applications, as well as the percentage of entities extracted by more than two applications. As shown in Fig. 1, 10% of distinct entities were extracted by the three applications for the four categories.

B. Extraction of distinct entities per category

We also evaluated the extraction of the four categories: drug, disorder, procedure, and anatomy. Fig. 2 illustrates the distribution of distinct entities extracted per category with the three applications and all entities extracted in common between the three systems.

C. Implementation requirements

We also mention some details needed to set up the three applications. A registration is needed to use BioPortal. It allows 15 queries per second per IP address. The limit of characters per query is 7,250. CLAMP and ScispaCy need an UMLS account. In general, ScispaCy requires less time for execution followed by CLAMP.

IV. Discussion

As shown in Table I, BioPortal extracted the highest number of entities (190,898) from the 1,000 clinical records. CLAMP got the lowest number of entities (55,258). However, CLAMP extracted more distinct entities (6,291) than BioPortal and ScispaCy. Since CLAMP is based on machine learning algorithms, it seeks to extract only relevant entities from clinical records. Also, 10% of distinct entities were extracted by the three applications. Moreover, CLAMP performed better when extracting distinct procedure entities (2,023), see Fig. 2. BioPortal performed better when extracting distinct disorder (2,050) and drug (719) entities. ScispaCy obtained better results over the distinct anatomy entities only (19). We also investigated that ScispaCy performs better with more training.

[3]ScispaCy: https://github.com/allenai/scispacy
[4]spaCy: https://spacy.io/

V. Conclusion

Our preliminary comparison of annotation systems consisted of three annotation tools over a sample of 1,000 clinical notes from MIMIC-III database. We extracted four entity categories. In general, BioPortal performed better than CLAMP and ScispaCy. In the near future, we will investigate the feasibility of creating a first version of an annotated dataset using the common entities extracted by two or more applications. Finally, further exploration of the annotation tasks is planned with: (a) additional state-of-the-art annotation applications, such as QuickUMLS, Apache cTAKES, and MetaMap; (b) an annotated dataset; and (c) a bigger sample from the MIMIC-III database.

Acknowledgment

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA183962. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1]. Evans R, “Electronic health records: then, now, and in the future,” Yearbook of medical informatics, vol. 25, no. S 01, pp. S48–S61, 2016.

[2]. Yim W.-w., Yetisgen M, Harris WP, and Kwan SW, “Natural language processing in oncology: a review,” JAMA oncology, vol. 2, no. 6, pp. 797–804, 2016. [PubMed: 27124593]

[3]. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S et al. , “Clinical information extraction applications: a literature review,” Journal of biomedical informatics, vol. 77, pp. 34–49, 2018. [PubMed: 29162496]

[4]. Friedman C, Alderson PO, Austin JH, Cimino JJ, and Johnson SB, “A general natural-language text processor for clinical radiology,” Journal of the American Medical Informatics Association, vol. 1, no. 2, pp. 161–174, 1994. [PubMed: 7719797]

[5]. Aronson AR and Lang F-M, “An overview of metamap: historical perspective and recent advances,” Journal of the American Medical Informatics Association, vol. 17, no. 3, pp. 229–236, 2010. [PubMed: 20442139]

[6]. Demner-Fushman D, Rogers WJ, and Aronson AR, “Metamap lite: an evaluation of a new java implementation of metamap,” Journal of the American Medical Informatics Association, vol. 24, no. 4, pp. 841–844, 2017. [PubMed: 28130331]

[7]. Denny JC, Irani PR, Wehbe FH, Smithers JD, and Spickard A III, “The knowledgemap project: development of a concept-based medical school curriculum database,” in AMIA Annual Symposium Proceedings, vol. 2003. American Medical Informatics Association, 2003, p. 195.

[8]. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, and Chute CG, “Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications,” Journal of the American Medical Informatics Association, vol. 17, no. 5, pp. 507–513, 2010. [PubMed: 20819853]

[9]. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, and Lazarus R, “Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system,” BMC medical informatics and decision making, vol. 6, no. 1, p. 30, 2006. [PubMed: 16872495]

[10]. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, Ravikumar K, Wu ST, Kullo IJ, and Chute CG, “An information extraction framework for cohort identification using electronic health records,” AMIA Summits on Translational Science Proceedings, vol. 2013, p. 149, 2013.

[11]. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, and Xu H, “Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines,” Journal of the American Medical Informatics Association, vol. 25, no. 3, pp. 331–336, 2017.

[12]. Soldaini L and Goharian N, “Quickumls: a fast, unsupervised approach for medical concept extraction,” in Medical Information Retrieval (MedIR) Workshop, ser. in SIGIR ‘16, 2016.

[13]. Doan S, Conway M, Phuong TM, and Ohno-Machado L, “Natural language processing in biomedicine: a unified system architecture overview,” in Clinical Bioinformatics. Springer, 2014, pp. 275–294.

[14]. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M, and Botsis T, “Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review,” Journal of biomedical informatics, vol. 73, pp. 14–29, 2017. [PubMed: 28729030]

[15]. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG, “Mimic-iii, a freely accessible critical care database,” Scientific data, vol. 3, p. 160035, 2016.

[16]. Johnson AE, Stone DJ, Celi LA, and Pollard TJ, “The mimic code repository: enabling reproducibility in critical care research,” JAMIA, vol. 25, no. 1, pp. 32–39, 2017.

[17]. Jonquet C, Shah NH, and Musen MA, “The open biomedical annotator,” Summit on translational bioinformatics, vol. 2009, p. 56, 2009. [PubMed: 21347171]

[18]. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, and Musen MA, “Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications,” Nucleic acids research, vol. 39, no. suppl 2, pp. W541–W545, 2011. [PubMed: 21672956]

[19]. Neumann M, King D, Beltagy I, and Ammar W, “Scispacy: Fast and robust models for biomedical natural language processing,” 2019.

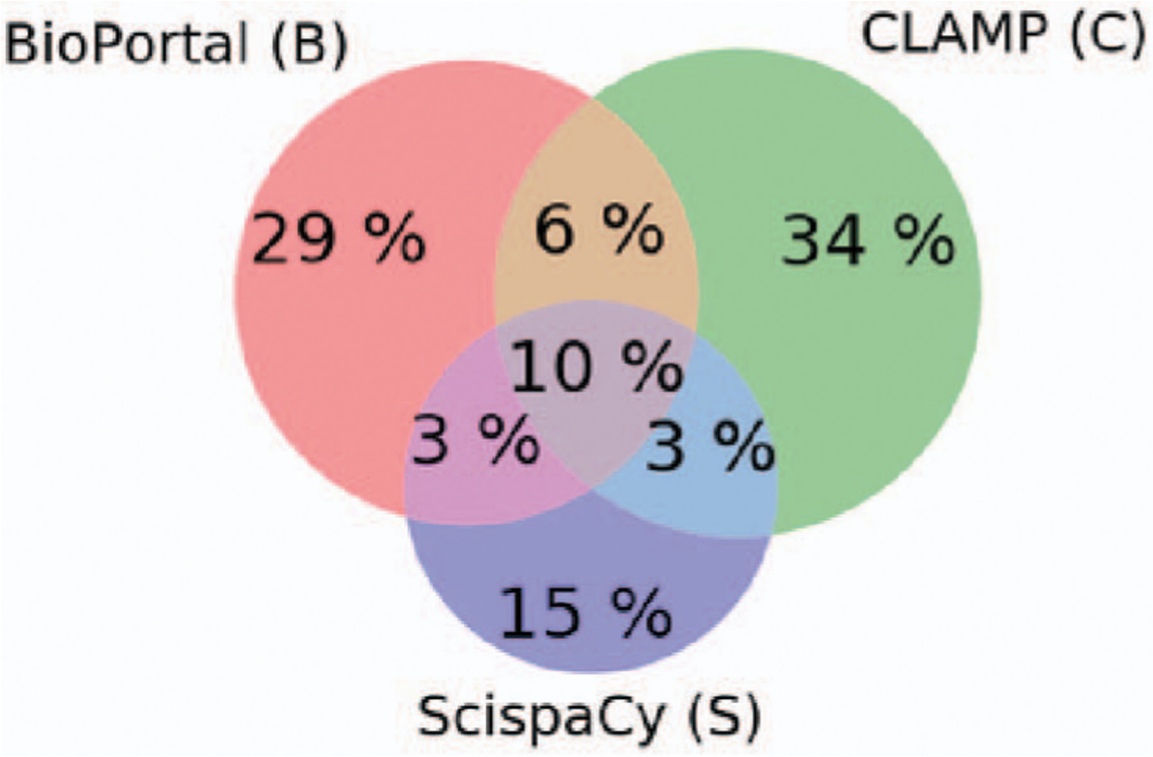

Fig. 1.
Distribution of the 11,777 distinct entities annotated with BioPortal, CLAMP, and ScispaCy.

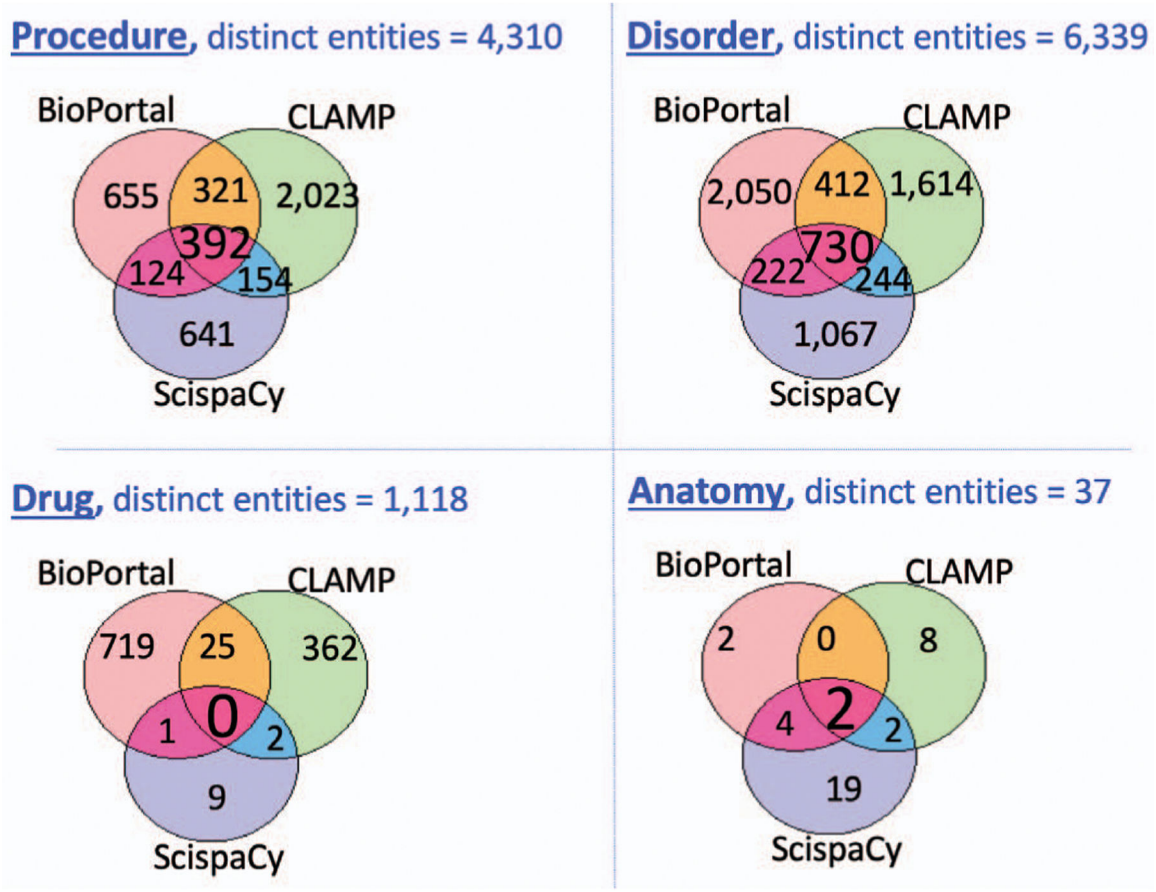

Fig. 2.
Distribution of distinct entities extracted per category with BioPortal, CLAMP, and ScispaCy.

TABLE I

Number of entities extracted by the three applications.

	Total entities	Distinct entities
BioPortal (B)	190,898	5,632
CLAMP (C)	55,258	6,291
ScispaCy (S)	157,597	3,613
B ∪ **C** ∪ **S**	403,753	11,777