Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210195

DeIDNER Corpus: Annotation of Clinical Discharge Summary Notes for Named Entity Recognition Using BRAT Tool

Mahanazuddin SYED[a,1], Shaymaa AL-SHUKRI[a], Shorabuddin SYED[a] , Kevin SEXTON[a], Melody L. GREER[a], Meredith ZOZUS[b], Sudeepa BHATTACHARYYA[c], and Fred PRIOR[a]

[a] *Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA*

[b] *Department of Population Health Sciences, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA*

[c] *Department of Biological Sciences and Arkansas Biosciences Institute, Arkansas State University, Jonesboro, AR, USA*

Abstract. Named Entity Recognition (NER) aims to identify and classify entities into predefined categories is a critical pre-processing task in Natural Language Processing (NLP) pipeline. Readily available off-the-shelf NER algorithms or programs are trained on a general corpus and often need to be retrained when applied on a different domain. The end model's performance depends on the quality of named entities generated by these NER models used in the NLP task. To improve NER model accuracy, researchers build domain-specific corpora for both model training and evaluation. However, in the clinical domain, there is a dearth of training data because of privacy reasons, forcing many studies to use NER models that are trained in the non-clinical domain to generate NER feature-set. Thus, influencing the performance of the downstream NLP tasks like information extraction and de-identification. In this paper, our objective is to create a high quality annotated clinical corpus for training NER models that can be easily generalizable and can be used in a downstream de-identification task to generate named entities feature-set.

Keywords. Named Entity Recognition, Annotation, De-identification, Clinical Corpus, Natural Language Processing.

1. Introduction

Named entity recognition (NER) has become a critical pre-processing task in Natural Language Processing (NLP) pipeline to identify the entities such as person name, organization, and other temporal expressions from the unstructured documents [1]. NER is usually accomplished by either rule-based or Machine Learning (ML) algorithms. The majority of NER ML algorithms available off-the-shelf are trained with general corpora and are used to generate named entities for classification tasks such as sentiment analysis, information extraction, and de-identification [2]. However, such models' performance

[1] Corresponding Author, Mahanazuddin Syed, Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA; E-mail: msyed2@uams.edu.

degrades when applied to cross-domain settings and often needs 1) to be retrained on a domain-specific corpus and 2) embedding domain specific vocabulary to improve the accuracy in the downstream tasks [3].

One such downstream task is the de-identification of clinical notes using ML techniques. The NER feature-set generated by the aforementioned ML methods is used as input to the de-identification algorithms [4]. The majority of existing de-identification ML algorithms rely on off-the-shelf NER models for generating NER input feature-sets, thereby influencing the final end model's performance, i.e., de-identification [4; 5]. This problem can be circumvented by using a clinical corpus for training the NER models, however, there is a dearth of such training data due to privacy concerns and data protection. Recently, interest is increasing in creating workshops like Informatics for Integrating Biology and the Bedside (I2B2) to make clinical corpus available, but the information is de-identified with realistic surrogates in place of real identifiers, which may hinder the generalizability and quality of the resulting NER feature-set [6]. To address these problems, we have created a corpus to train the NER models, which will help generate accurate, generalizable named entities specifically for a downstream de-identification task. The guidelines and non-PHI code are available from corresponding author upon reasonable request.

2. Methods

2.1. Dataset and Sampling

This study was conducted with University of Arkansas for Medical Sciences (UAMS) institutional review board approval (IRB #228649). The initial study corpus consisted of 500 clinical discharge summary notes (n=385, confidence interval=95%) collected from UAMS medical record system of the patients hospitalized during May 2014 to December 2019.

2.2. Annotation Guidelines and Tools

The annotation guidelines were initially drafted from CoNLL-2003 and the I2B2 2014 challenge workshop that was revised through a rigorous and iterative process. Experienced annotators methodically updated the guidelines at the end of every iteration. The final schema contained eight entities, as shown in Table 1. The annotators' task was to use these guidelines to annotate the documents using an open-source annotation tool called BRAT [7].

2.3. Annotation Process and Workflow

As shown in Figure 1, we divided the process into two different stages: 1) preparation stage, where we identified annotators, tools, and guidelines before starting the actual annotation process and 2) annotation stage, where we annotated the documents and finalized the standard gold corpus. To annotate the documents, we chose a double annotation method to avoid errors made by a single annotator and improve the annotation quality. In addition, we have used pre-annotation for one of the annotator, and the other did not. Pre-annotation was done using Stanford NER.

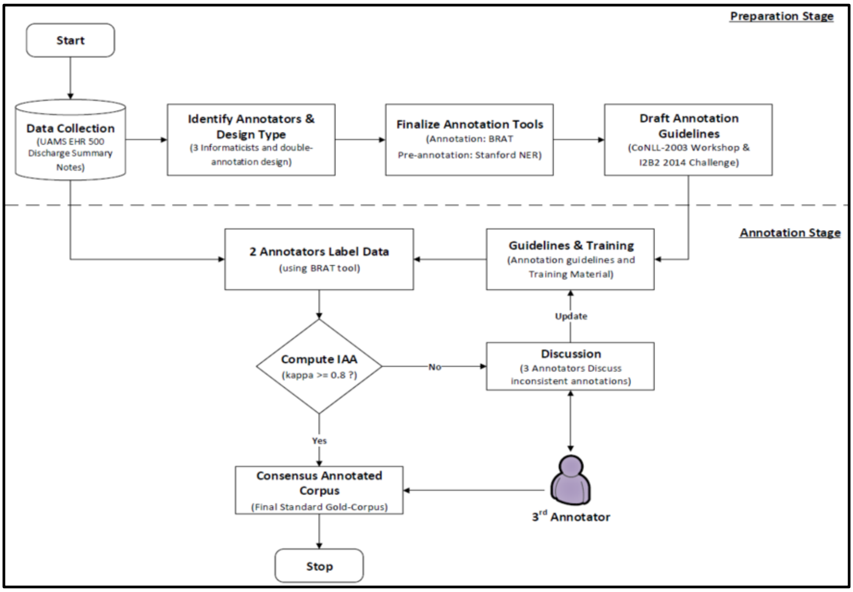
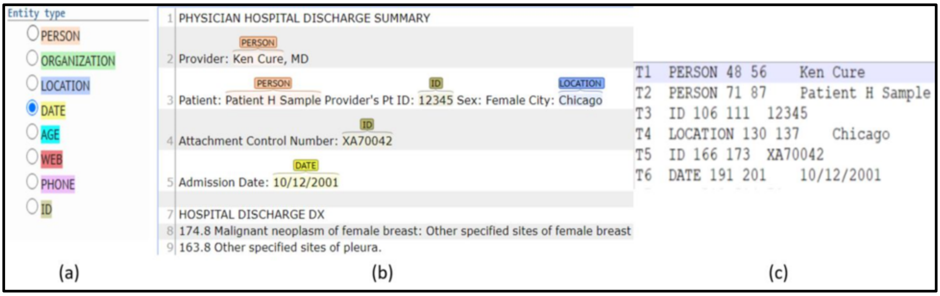

Preparation Stage
Start
Data Collection
(UAMS EHR 500 Discharge Summary Notes)
Identify Annotators & Design Type
(3 Informaticists and double-annotation design)
Finalize Annotation Tools
(Annotation: BRAT
Pre-annotation: Stanford NER)
Draft Annotation Guidelines
(CoNLL-2003 Workshop & I2B2 2014 Challenge)
Annotation Stage
2 Annotators Label Data
(using BRAT tool)
Guidelines & Training
(Annotation guidelines and Training Material)
Update
Compute IAA
(kappa >= 0.8 ?)
No
Discussion
(3 Annotators Discuss inconsistent annotations)
Yes
Consensus Annotated Corpus
(Final Standard Gold-Corpus)
3rd Annotator
Stop

Figure 1. A workflow depicting the annotation process, which was divided into preparation and annotation stage. The preparation stage finalizes the pre-requisites, and the annotation stage performs annotation, computes inter-annotator agreement, and finalize the corpus based on discussion and consensus.

Annotation was performed in five iterations that covered 100 clinical notes; a sample file is shown in Figure 2. At the end of each iteration, we measured the Inter Annotator Agreement (IAA) metric defined in equation 1 and the differences in the analysis was performed and discussed with a third annotator (MS) for consensus. If IAA was not acceptable, that batch was re-annotated with updated guidelines. Guideline updates most frequently consisted of simple clarification and the addition of supporting examples. If IAA was acceptable, the batch was queued for consensus annotation by the third annotator (MS) to finalize the annotations. The consensus annotation was based on a simple match mechanism while preserving the concepts identified by either of the annotators.



1 PHYSICIAN HOSPITAL DISCHARGE SUMMARY
2 Provider: Ken Cure, MD
3 Patient: Patient H Sample Provider's Pt ID: 12345 Sex: Female City: Chicago
4 Attachment Control Number: XA70042
5 Admission Date: 10/12/2001
7 HOSPITAL DISCHARGE DX
8 174.8 Malignant neoplasm of female breast: Other specified sites of female breast
9 163.8 Other specified sites of pleura.

```
T1  PERSON 48 56    Ken Cure
T2  PERSON 71 87    Patient H Sample
T3  ID 106 111   12345
T4  LOCATION 130 137    Chicago
T5  ID 166 173  XA70042
T6  DATE 191 201    10/12/2001
```

Figure 2. Example file depicting annotations using BRAT tool. (a) Set of pre-defined entities, (b) Sample BRAT annotation document, and (c) Final metadata file with annotations that will be used for training.

We calculated agreement between the two annotators to continually address subjectivity in judging things that are not observed with the senses. In our process, we used Cohen's kappa statistical measure described in literature [8; 9] for every entity.

$$k = (p_o - p_c) / (1 - p_c) \quad (1)$$

Here p_o is the observed agreement between two annotators, and p_c is the probability of expected agreement by chance representing agreements and disagreements. The annotation is considered as perfect when the k value is >= 0.8 based on the interpretation of kappa value on the Landis and Koch scale [10].

3. Results

The annotator 1, without pre-annotation, has taken 30 hours to finish annotation of 5603 entities, whereas annotator 2 with pre-annotation has taken 34.5 hours to finish annotation of 5776 entities as shown in Table 1. The final gold corpus consisted of 5,852 entities. With both the annotators, DATE entity had the maximum number of entries, followed by PERSON and AGE. The agreement metric IAA between the two annotators was averaged for 5 iterations and resulted in greater than 80% for every entity type.

We have noticed increased IAA with iterative method, with IAA values in every round is higher than the previous. We believe that discussions, meetings, and updated documentation or guidelines during the annotation process helped achieve better IAA metrics.

Table 1. Details of entities, time taken to annotate, and Inter Annotator Agreement (IAA) measured.

Entity Type	Total Entities identified by Annotator 1 (n=5603, time=30 hours) (no pre-annotation set)	Total Entities identified by Annotator 2 (n=5776, time=34.5 hours) (pre-annotation set)	Average Inter Annotator Agreement (IAA) after 5 iterations
DATE	3493	3541	0.94
PERSON	1403	1481	0.91
AGE	253	256	0.94
ID	169	171	0.98
ORGANIZATION	114	134	0.87
LOCATION	113	130	0.86
PHONE	55	59	0.93
WEB	3	4	0.84

4. Discussions

To address the scarcity of high-quality clinical corpora for training NER models for a de-identification task, we have built a foundational corpus using a double annotation strategy and guidelines adopted from the literature. We believe this work is foundational to future NER model training on clinical data and will be fundamental to future NLP projects for clinical data. In addition, the algorithms that treat de-identification as named entity tasks can utilize the corpus in training/fine-tuning the models [11]. This study demonstrates the amount and importance of accurate named entities for training NER models in the clinical domain and using such models in de-identification tasks. With stricter federal laws on privacy in US and European countries, to complete NLP work with clinical data, one must build a highly accurate de-identification algorithm to avoid a protected health information (PHI) violation.

In lieu with other studies, we did not find pre-annotation useful; the process took longer than without pre-annotation because of correcting the existing annotations.

However, pre-annotation did help to identify a few additional entities that were missed by the other annotator. There are some inherent limitations in our work; for instance, the study corpus consists of 500 discharge summary notes and we have annotated only 100 documents. The goal is to use the strategies and guidelines developed herein to annotate the remainder of the documents using clinician annotator. The other limitation is, we have a limited number of instances of some entities, e.g., WEB and PHONE, and we may have to evaluate the results on the full corpus. Finally, our future work is to develop and train an NER model using above corpus and analyze the effectiveness by experiment.

5. Conclusion

Depending on the task and domain, the corpus used for training can significantly impact NLP algorithms. This project built a high-quality annotated corpus using discharge summary notes for NER models training in the clinical domain. Our work methodology involving training, refining guidelines, and discussions in iterations ensured high-quality annotations that we have quantified using IAA metric.

Acknowledgment: This study was supported in part by the Translational Research Institute (TRI), grant UL1 TR003107 received from the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH). The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] Giorgi JM, Bader GD. Transfer learning for biomedical named entity recognition with neural networks, Bioinformatics 2018;34:4087-94.

[2] Augenstein I, Derczynski L, Bontcheva K. Generalisation in named entity recognition: A quantitative analysis, Computer Speech & Language 2017;44:61-83.

[3] Tai W et al. exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources, Proceedings of EMNLP; 2020 Available at: https://www.aclweb.org/anthology/2020.emnlp-main. p. 1433-39.

[4] Dehghan A et al. Combining knowledge- and data-driven methods for de-identification of clinical narratives, J Biomed Inform 2015;58:S53-S59.

[5] Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. J Biomed Inform. 2017 Nov;75S:S34-S42.

[6] Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform. 2015 Dec;58(Suppl):S20-9.

[7] Stenetorp P et al. BRAT: a web-based tool for NLP-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics;2012; Avignon, France: p. 102–7.

[8] Cohen J. A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement 1960;20:37-46.

[9] Carletta J. Assessing agreement on classification tasks: the kappa statistic, Comput. Linguist. 1996; 22:249–54.

[10] Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data, Biometrics 1977;33:159-174.

[11] Catelli R et al. A Novel COVID-19 Data Set and an Effective Deep Learning Approach for the De-Identification of Italian Medical Records, IEEE Access 2021;9:19097-19110.