Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240546

Accelerating Clinical Text Annotation in Underrepresented Languages: A Case Study on Text De-Identification

He A. XU[a,1], Valentin LOFTSSON[a], Bogdan KULYNYCH[a], Bayrem KAABACHI[a] and Jean Louis RAISARO[a]
[a] *Biomedical Data Science Center, Lausanne University Hospital (CHUV) and Lausanne University, Lausanne, Switzerland*
ORCiD ID: He A. Xu https://orcid.org/0000-0003-0248-8604, Bogdan Kulynych https://orcid.org/0000-0001-5923-3931, Bayrem Kaabachi https://orcid.org/0009-0002-7534-8493, Jean Louis Raisaro https://orcid.org/0000-0003-2052-6133

Abstract. Clinical notes contain valuable information for research and monitoring quality of care. Named Entity Recognition (NER) is the process for identifying relevant pieces of information such as diagnoses, treatments, side effects, etc., and bring them to a more structured form. Although recent advancements in deep learning have facilitated automated recognition, particularly in English, NER can still be challenging due to limited specialized training data. This exacerbated in hospital settings where annotations are costly to obtain without appropriate incentives and often dependent on local specificities. In this work, we study whether this annotation process can be effectively accelerated by combining two practical strategies. First, we convert usually passive annotation tasks into a proactive contest to motivate human annotators in performing a task often considered tedious and time-consuming. Second, we provide pre-annotations for the participants to evaluate how *recall* and *precision* of the pre-annotations can boost or deteriorate annotation performance. We applied both strategies to a text de-identification task on French clinical notes and discharge summaries at a large Swiss university hospital. Our results show that proactive contest and average quality pre-annotations can significantly speed up annotation time and increase annotation quality, enabling us to develop a text de-identification model for French clinical notes with high performance (F1 score 0.94).

Keywords. NLP, Name Entity Recognition, Clinical Notes, De-identification

1. Introduction

One of the most common types of information stored in electronic health record systems is free-text clinical notes. Although the primary use of clinical notes is to facilitate healthcare and billing, they contain valuable information that can be leveraged for medical research and improving care practices [1]. Named entity recognition (NER) is a common building block for extracting relevant information from a clinical text [2]. Vocabulary systems such as unified medical language systems (UMLS) were developed

[1] Corresponding Author: He A. Xu, Rue du Bugnon 21, 1011 Lausanne, Switzerland; E-mail: he.xu@chuv.ch.

to facilitate information extraction from clinical text. Extracting named entities by matching words against a vocabulary, however, suffers from low recall [3]. In the recent years, deep-learning based language models for NER have seen significant progress, enabling effective automated information extraction from clinical notes. This progress, however, has been most prominent for the English language. Deep-learning based NER in other languages is still challenging due to scarce availability of open annotated training datasets specialized to clinical settings [5]. Moreover, even if more specialized training datasets for clinical NER were available, the performance of trained NER models could deteriorate in a local context of a specific deployment scenario [1,4]. These two problems can be solved by creating context-specific (e.g., hospital-specific) NER datasets.

As a result, healthcare research based on clinical notes in underrepresented languages requires a time-consuming annotation process for developing useful NER models. As the annotation is done on sensitive patient data, it must be performed by hospital practitioners, and is therefore costly. In this work, we ask a question: *Can we accelerate the annotation process with strategies that are simple and affordable to implement in a hospital setting?*

First, inspired by the 'agile annotation' methodology [7], we propose to convert the passive annotation tasks into a proactive annotation contest with frequent feedback. For this, we set up a gamified procedure by assigning the annotators to different teams to compete for a final reward, aiming to maintain their motivation throughout the process. Second, we provided pre-annotations, hypothesizing that supplying the annotators with partial pre-annotations obtained via a simple rule-based algorithm would increase the annotation efficiency. To evaluate the hypothesis, we design two controlled experiments which aim to answer whether (1) *low-recall*, and (2) *low-precision* pre-annotations affect the annotation performance.

As a case study, we apply both strategies as part of the development of a clinical text de-identification model—one of the important use cases of medical NER—for French language at the Lausanne University Hospital. Text de-identification is a standard method to lower privacy risks when sharing clinical text with external researchers.

2. Method

Designing the Annotation Task. We designed an annotation task to identify protected health information (PHI) in clinical documents. The PHIs were defined based on the HIPAA privacy rule and the Swiss Federal Data Protection Act. The defined PHIs contain 9 categories (*hospital unit, contact information, demographics, location, time, patient's ID, person's name, organization name, other personal information*), with each category containing several sub-categories. In total, we defined 26 named entities to be annotated. We collected 3010 clinical documents from the hospital's research data warehouse, including discharge letters, consultation letters, transfer letters, and lab reports. The text was split into chunks (each chunk ranging between one and several sentences, with 20–500 tokens in total) and then fed into the Prodigy annotation tool [8] for manual annotation. We refer to the annotation of a single chunk of text as a *task.*

Developing a Rule-Based Algorithm. To test whether we can easily accelerate manual annotation, we built a rule-based model to identify the defined categories of PHIs as follows. For categories such as 'hospital unit', we matched against a list of unit names at

Table 1. The assignment of tasks and performance with different relative recall and precision of pre-annotations. Recall of 100% indicates we provide the pre-annotations using all predictions from the rule-based model. Precision of 100% indicates we do not alter the rule-based pre-annotations, and Precision 40% indicates we leave 40% of the pre-annotations unperturbed. *** means p-value < 10^{-4}, * means p-value between 0.01 and 0.05 when comparing experimental and control arms separately within each group.

CONTEST	GRP.	ARM	RECALL	PRECISION	DURATION	SEG. ERR.	CAT. ERR.
1 (*VARYING* RECALL)	A	exp.	*100%*	100%	**6.12 ± 15.00*****	**0.03 ± 0.17*****	**0.03 ± 0.17*****
		ctrl.	0%	—	9.08 ± 18.60	0.13 ± 0.33	0.13 ± 0.33
	B	exp.	*50%*	100%	**6.01 ± 16.88*****	**0.06 ± 0.23***	**0.05 ± 0.23***
		ctrl.	0%	—	9.16 ± 19.50	0.08 ± 0.28	0.08 ± 0.27
2 (*VARYING* PRECISION)	C	exp.	100%	*40%*	8.09 ± 19.39***	0.03 ± 0.19	0.04 ± 0.20
		ctrl.	0%	—	**5.95 ± 13.80*****	0.03 ± 0.17	0.03 ± 0.17
	D	exp.	100%	*20%*	7.63 ± 15.68	0.10 ± 0.30	0.07 ± 0.26
		ctrl.	0%	—	**6.57 ± 14.95*****	**0.02 ± 0.15*****	**0.02 ± 0.14*****

the hospital. For other categories such as contact information and telephone numbers, we manually implemented matching patterns based on regular expressions. Although we did not know the performance of the algorithm prior to the completion of the annotation phase, the algorithm obtained an average 77% F1 score across all sub-categories (recall 0.86, precision 0.75), showing modest performance.

Study Design. We recruited 32 annotators from the hospital IT and research staff who had the right to access the patient records. We organized two contests with final rewards aiming to motivate annotators to annotate as many tasks as possible. Each contest included 16 participants and lasted four weeks. For each contest, we divided the participants into two groups. Participants in each group were randomly assigned to either an experimental arm or control arm. Within a group, annotators in both arms received the same tasks in the same order. The tasks across groups and contests were different. We provided pre-annotations to all experimental arms using the predictions by the rule-based model described above.

The first contest aimed to test how *low-recall pre-annotations*, i.e., fewer entities are pre-annotated, affect annotation efficiency. To evaluate this, we used two experimental arms in two groups with varying proportions of provided pre-annotations (see Table 1). The second contest aimed to test how *low-precision* pre-annotations, i.e., incorrect pre-annotations, affect annotation efficiency. To simulate low precision, we perturbed the pre-annotations by either changing the entity category or its boundaries. For the experimental arms in both groups, we used different proportions of perturbations. The participants were given a detailed description of the 26 sub-categories to be annotated. Following the 'agile annotation' methodology [7], they were given interactive feedback on their performance, and we continuously updated annotation guidelines that were available to all participants. Moreover, unlike prior work [10], to maintain participants' motivation, we encouraged the participants to compete on the quantity and quality of their annotations. We created a daily updated leaderboard to show the performance of each participant and group. We awarded the best performing participant and group at the end of each contest. Finally, using the annotated corpus, we fine-tuned a standard French Transformer-based NLP model CamemBERT [11].

3. Results

In Contest 1, participants finished a total of 3515 tasks, spending on average 6.5 hours over the duration of the contest (13.9 minutes per day). In contest 2, participants finished 8274 tasks, spending on average 14.75 hours (31.6 minutes per day). The time participants spent on the tasks is shorter than traditional annotation settings with pre-annotations [12]. We measured the Inter-Annotator Agreement using the Gamma metric [13], which considers both the categorization and segmentation agreement. Annotators achieved agreement with gamma value 0.85 in contest 1 and 0.91 in contest 2, both of which are higher than in prior annotation studies in a similar context [10].

We evaluate the effects of pre-annotations by measuring the (1) *time* of annotation of a single task, (2) *segmentation error*, i.e., error in defining an annotation's boundary, and (3) *categorization error*, i.e. error in identifying a sub-category.

Contest 1 (Table 1) showed that providing low-recall pre-annotations reduces the average time for annotation (*Group A: $t=6.57$, $p<10^{-5}$; Group B: $t=7.89$, $p<10^{-5}$*). Pre-annotations also helped reduce the segmentation error (*Group A: $t=6.33$, $p<10^{-5}$; Group B: $t=2.43$, $p = 0.015$*) and categorization error (*Group A: $t=6.19$, $p<10^{-5}$; Group B: $t=2.34$, $p = 0.019$*), thus improving the annotation efficiency. Contest 2 showed that providing low-precision pre-annotations hurt the average time for annotation *(Group C: $t=-5.94$, $p<0.0001$; Group D: $t=-4.86$, $p<0.0001$)*, with very low-precision annotations significantly increasing the segmentation error *(Group D: $t=-15.87$, $p < 0.0001$)* and categorization error *(Group D: $t=-11.74$, $p<0.0001$)*.

Using the annotated corpus we managed to obtain high-quality performance with a CamemBERT [11] model, achieving 0.94 average F1 score across the NER categories.

4. Discussion

We proposed strategies for facilitating NLP annotation tasks designed to be useful for small resource-constrained research groups, although they are also applicable to larger scale crowdsourcing tasks. An alternative approach to our strategies could be to use large language models (LLMs) in a few-shot setting [14], as this would not require large-scale dataset annotation. At the moment of writing, however, the NER performance of openly available LLMs in a few-shot settings is still lacking [15].

5. Conclusions

In this paper, we proposed two strategies for facilitating the collection of annotations for NLP tasks. The first is to turn the passive annotation task into a proactive contest. Conventionally, researchers employ two annotators to annotate a certain amount of text, with or without pre-annotations [9,12]. In our case study, the daily feedback and the final rewards kept participants motivated throughout the annotation process. Our results showed that the annotation contest encouraged annotators to perform tasks faster and with higher quality compared to traditional settings. The second strategy is to provide pre-annotations to the tasks using a rule-based model. Our results showed that with 50%

of pre-annotations provided using a medium-quality rule-based model (F1 score of 0.77), the speed and quality of the final annotations were significantly improved. However, if the precision of the pre-annotation is too low, the annotation performance deteriorated.

References

[1] Boag W, Doss D, Naumann T, Szolovits P. What's in a note? unpacking predictive value in clinical note representations. AMIA Summits Transl Sci Proc. 2018;2018:26.

[2] Ahmed A, Abbasi A, Eickhoff C. Benchmarking modern named entity recognition techniques for free-text health record deidentification. AMIA Summits Transl Sci Proc. 2021;2021:102.

[3] Wang Y, Patrick J. Cascading Classifiers for Named Entity Recognition in Clinical Notes. In: Proceedings of the Workshop on Biomedical Information Extraction. Association for Computational Linguistics; 2009. p. 42–9.

[4] Ahmed T, Aziz MMA, Mohammed N. De-identification of electronic health record using neural network. Sci Rep. 2020;10(1):18600.

[5] Tchouka Y, Couchot JF, Coulmeau M, Laiymani D, Selles P, Rahmani A, et al. De-identification of french unstructured clinical notes for machine learning tasks. ArXiv Prepr ArXiv220909631. 2022;

[6] Garfinkel S, others. De-identification of Personal Information:. US Department of Commerce, National Institute of Standards and Technology; 2015.

[7] Voormann H, Gut U. Agile corpus creation. 2008;

[8] Montani I, Honnibal M. Prodigy: A new annotation tool for radically efficient machine teaching. Artif Intell Appear. 2018;

[9] Grouin C, Névéol A. De-identification of clinical notes in French: towards a protocol for reference corpus development. J Biomed Inform. 2014 Aug 1;50:151–61.

[10] Bourdois L, Avalos M, Chenais G, Thiessard F, Revel P, Gil-Jardiné C, et al. De-identification of Emergency Medical Records in French: Survey and Comparison of State-of-the-Art Automated Systems. Fla Artif Intell Res Soc. 2021 May 11;34(1).

[11] Martin L, Muller B, Suárez PJO, Dupont Y, Romary L, de la Clergerie ÉV, et al. CamemBERT: a Tasty French Language Model. Proc 58th Annu Meet Assoc Comput Linguist. 2020;7203–19.

[12] Lingren T, Deleger L, Molnar K, Zhai H, Meinzen-Derr J, Kaiser M, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. J Am Med Inform Assoc. 2014;21(3):406–13.

[13] Mathet Y, Widlöcher A, Métivier JP. The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. Comput Linguist. 2015 Sep;41(3):437–79.

[14] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.

[15] Liu Z, Yu X, Zhang L, Wu Z, Cao C, Dai H, et al. Deid-GPT: Zero-shot medical text de-identification by GPT-4. ArXiv Prepr ArXiv230311032. 2023;