A PRE-ANNOTATION TOOL FOR EVENT EXTRACTION IN CARDIOLOGY AMBULATORY TEXTS IN BRAZILIAN PORTUGUESE

Yohan Bonescki Gumiel[1], Lucas Emanuel Silva e Oliveira[1], Carolina de Oliveira Montenegro[2], Carolina Dorigon Bantle[2], Luisa Superti Dal Magro[2], Eros Maranezzi Fadel[2], Caroline Pilatti Gebeluca[2], Lilian Mie Mukai Cintho[1], Claudia Moro[1], Deborah Ribeiro Carvalho[1]

[1]Graduate Program in Health Technology, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil
[2]School of Medicine, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil

E-mails: yohan.gumiel@gmail.com, kunkaweb@gmail.com, caarol_montenegro@hotmail.com, carolbantle@gmail.com, luisasdalmagro@gmail.com, eros-26@hotmail.com, carolgebeluca21@gmail.com, miemukai@hotmail.com, claudia.moro@gmail.com, drdrcarvalho@gmail.com

Abstract: Clinical texts offer medical information regarding the patients that are not present elsewhere, so creating tools that automatically extract this information can provide better and more personalized patient care. Creating these tools demands advanced machine learning techniques that require annotated data provided by burdensome annotation processes. Thus, we proposed a dictionary-based pre-annotation tool to diminish the burden of manually annotating all mentions over the texts. We developed a pre-annotation tool to help in our event annotation for cardiology ambulatory texts. The pre-annotation tool was based on a dictionary created during the annotator's training phase and four rounds of the annotation process. We annotated 126 texts with three annotators from the medicine course. We evaluated the pre-annotation performance based on the inter-annotator agreement, the annotation time, the annotation speed, and the pre-annotation coverage (amount of correct pre-annotations that were present in the gold standard). We concluded that our dictionary's refinement was beneficial to our pre-annotation; it raised the pre-annotation coverage while not reducing the inter-annotator agreement. We noticed that our annotation time decreased over the rounds, which is expected due to the annotators getting used to the annotation guideline and annotation tool over time.

Keywords: Natural language processing, clinical texts, pre-annotation, annotation.

1. INTRODUCTION

Noncommunicable diseases (NCDs) are the primary cause of death and are responsible for more than 70% of the deaths globally (World Health Organization, 2020). Among the NCDs, cardiovascular diseases are the most common cause of death in Brazil . Hence, developing tools that can support cardiologists is essential.

Chronic diseases have a longitudinal nature, providing extensive and continuous patient data represented over electronic health records (EHRs) (Sheikhalishahi et al., 2019). Amid the EHRs, clinical texts represent the unstructured data, which can not be understood by machines without processing (Jensen; Jensen; Brunak, 2012; Jiang et al., 2017). Natural language processing (NLP) methods transform this unstructured data type into a structured format that can be used to develop tools.

One of the NLP fields that can be beneficial to cardiologists is temporal relation extraction due to the possibility of inferring order among relevant medical events and creating a patient's clinical timeline. The first step for temporal relation extraction is to create a framework that automatically extracts medical events, requiring machine learning (ML) methods that need annotated data, i.e., positive examples of event annotations over the texts. These ML methods, especially deep learning-based methods, need an abundant amount of annotated texts. The annotation process is burdensome; besides the high number of annotations, the annotators need to the trained to ensure the annotation quality. This process extends for several months, especially in complex annotation projects.

We annotated ambulatory texts in Brazillian Portuguese from the cardiology department, and we proposed a pre-annotation tool to relieve to annotation burden. Pre-annotation tools are essential in an ambulatory scenario due to specific recurrent mentions (e.g., patient symptoms, medications, and medical tests). For example, patients with hypertension have several routine tests that need to be checked by the physician every consult. These routine tests are defined by the guidelines for the management of arterial hypertension. Similarly, patients generally use several medications that are mentioned in a list format along with their dosage. Hence, pre-annotations are quite effective for these types of events.

Pre-annotations tools can be based on dictionaries or machine learning. Depending on the annotation purpose, terminologies such as the Unified Medical Language System (UMLS) (Bodenreider, 2004) can be used as a dictionary lookup. Some examples of studies about pre-annotations by dictionaries in the clinical domain are Oliveira et al. (2017), Névéol, Dogan and Lu (2011), Hamon et al. (2017), and Lingren et al. (2012). Oliveira et al. (2017) created an annotation assistant for clinical texts based on statistics and the UMLS. Névéol, Dogan, and Lu (2011) and Hamon et al. (2017) focused on biomedical texts in English, the first aiming to identify named entities and the second to identify food-drug interactions. Lingren et al. (2012) focused on pre-annotating clinical notes and clinical trial announcements for named entities and terminology coding. Some examples of machine learning-based pre-annotation tools are South et al. (2014), Hernandez et al. (2014) and Grouin, and Névéol (2014). South et al. (2014) and Grouin and Névéol (2014) focused on pre-annotating for the text de-identification task, an essential task to ensure patient privacy. Hernandez et al. (2014) identified drug-drug interactions, comparing the annotation performance of experts and non-experts.

Machine learning-based approaches rely on models trained on different corpora or training specific models for the corpus, which need previously annotated data. There was no suitable trained model for our annotation, and there was not enough previously annotated data. Thus, we used a dictionary-based approach to pre-annotate events in Brazilian Portuguese ambulatory texts. We created our dictionary based on correct annotations from the annotator's training and four rounds of the annotation process.

Our objective was to diminish the annotation effort while maintaining the annotation quality. We evaluated our pre-annotation tool and the impact of refining the pre-annotation dictionary along with the annotation process.

2. MATERIAS AND METHODS

Our event definitions were based on the i2b2 2012 (Sun; Rumshisky; Uzuner, 2013) and THYME (Styler et al., 2014) annotation guidelines, with adaptations to fulfill the cardiology and ambulatory text characteristics. Events were defined as relevant mentions over the patient's timeline, the same criteria used for i2b2 2012 and THYME annotations. Our event categories were based on the i2b2 2012 guideline. In our definitions: (I) problems were mentions that differed from normal conditions (e.g., diseases); (II) treatments were mentions that referred to procedures and interventions used to treat problems (e.g., medications); (III) tests were mentions that aimed to detect and evaluate problems (e.g., laboratory exams); (IV) evidences were mentions that alluded to words which connected the fount of information to the information (e.g., words such as "deny"); (V) clinical departments were mentions that referred to departments, places, health professionals; (VI) occurrences mentions could be related to several types of information (e.g., medication change mentions such as "keep" and "reduced"; encounter-related mentions, such as "return" and "consult"), as the occurrence category coverage was extensive.

We selected three students from the medicine course to annotate the texts. First, we started the guideline refinement and annotator training cycle. This cyclic process is shown in Figure 1, in the guideline refinement section. We provided a small set of documents to the annotators in the training cycle and verified their concordance with the inter-annotator agreement (IAA) every cycle. In every cycle, we had meetings to discuss the annotations discordances and refine the guidelines. We continued this process until consistent IAA values were obtained from all the annotators. When this condition was satisfied, we started the annotation process of our 126 texts (Figure 1 – annotation process section). The documents were double annotated and adjudicated by a Ph.D. student with a background in Biomedical Informatics.

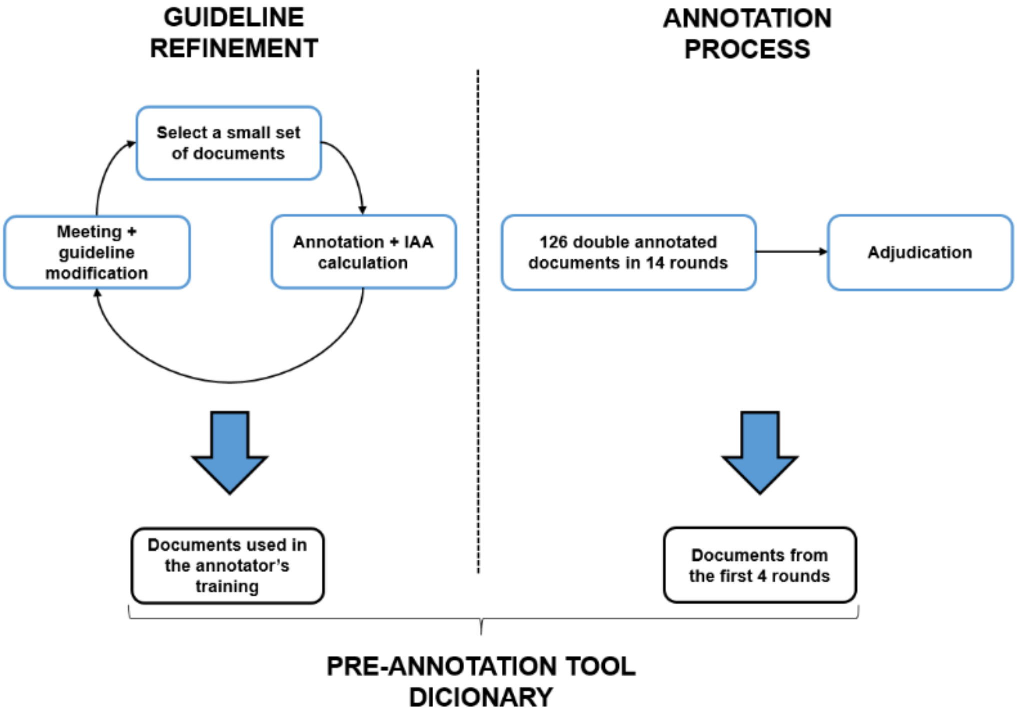

GUIDELINE REFINEMENT
ANNOTATION PROCESS
Select a small set of documents
Annotation + IAA calculation
Meeting + guideline modification
126 double annotated documents in 14 rounds
Adjudication
Documents used in the annotator's training
Documents from the first 4 rounds
PRE-ANNOTATION TOOL DICIONARY

Figure 1: Annotation process and pre-annotation tool development

Our pre-annotation tool was based on a dictionary created over the annotator's training rounds and four rounds from the annotation process (Figure 1 – pre-annotation tool section). We selected correct annotations from these rounds and added them to the dictionary. We added regular expressions to add medication dosage into the pre-annotations, such as the correct annotation of the dosage in "enalapril 10 mg". We also added specific rules to pre-annotate the longer spans, as in our annotation scheme, events do not overlap their spans. For instance, correctly pre-annotating "lower extremity edema" and not restricting to "edema" when both events are in the pre-annotation dictionary.

To measure the performance of our pre-annotation tool, we used the following criteria: (i) IAA; (ii) annotation time per document; (iii) speed (event annotations per minute); (iv) coverage (amount of pre-annotations found in the gold standard). These criteria were selected based on the studies of Lingren et al. (2014), Fort and Sagot (2010), Grouin and Névéol (2014), Oliveira et al. (2017), South et al. (2014), Névéol, Dogan and Lu (2011).

We used the F1-score between two annotators for the IAA, used in both THYME and i2b2 2012 evaluations. We considered only exact matches, where agreements occurred when both annotators annotated the same tokens for the entity. For instance, if one annotator annotated "dyspnea on exertion" and another "dyspnea", that would be a disagreement; to achieve an agreement, both would have to annotate "dyspnea on exertion".

3. RESULTS

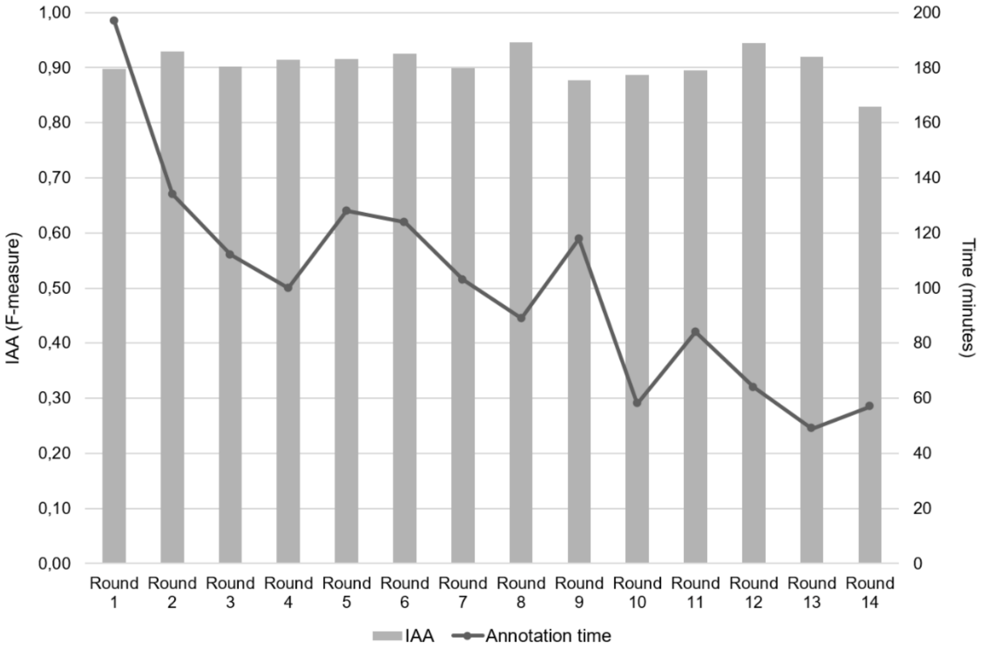
Our annotation process lasted for three months; we annotated 126 documents with 14 rounds of annotations and nine documents per round. Table 1 shows the annotation time, the annotation speed, the IAA values, and the pre-annotation coverage for each round; it also shows the average values. The annotation time decreases over the rounds, which is expected because the annotators got used to the annotation guideline and annotation tool over time. In the 1st round, the annotators took around 22 minutes per document, while they took around six minutes in the 14th round. However, on average, they took around 11 minutes per document.

Table 1: Annotation time and speed, IAA values and pre-annotation coverage per round and average values

Rounds	Annotation time (minutes)	Annotation speed (events per minute)	IAA	Pre-annotation coverage (percentage)
1	197	1.85	0.8983	54.12%
2	134	2.37	0.9288	43.08%
3	112	1.95	0.9019	49.54%
4	100	2.39	0.9145	58.58%
5	128	3.10	0.9161	72.04%
6	124	2.76	0.9258	62.57%
7	103	2.44	0.8988	65.34%
8	89	3.07	0.9460	69.23%
9	118	3.06	0.8766	62.60%
10	58	3.62	0.8873	62.38%
11	84	3.73	0.8949	63.26%
12	64	4.33	0.9449	67.87%
13	49	4.20	0.9193	64.56%
14	57	4.32	0.8295	58.94%
Average	**101**	**3.08**	**0.9066**	**61.17%**

In Table 1, it is noticeable that the annotation speed increased over the rounds. Further, there was an improvement in the annotation speed since the 5th round, in which our pre-annotation dictionary was completed. We noticed that the annotator's performance was stable over the rounds, except for the IAA value in the last round. We achieved an IAA value of 0.9066. We achieved an average coverage of 61.17% for our pre-annotation coverage, and our coverage raised considerably since the 5th round. In Figure 2, we show a plot between the IAA values and the annotation time per round. We noticed that the annotation time decreased over the rounds, but the IAA values remained stable.

Figure 2: IAA values and Annotation time for each round



4. DISCUSSION

We achieved positive results for the IAA (0,9066) in comparison with other event annotation projects. In the i2b2 2012 annotation, they achieved 0.83, and in Clinical TempEval 2016 annotation, related to the THYME project, they achieved 0.864. Both i2b2 2012 and Clinical TempEval 2016 corpora are references in annotating and extracting events and temporal relations in clinical texts. Thus, our annotation process was successful.

We did not have documents that were annotated without pre-annotations, so we could not directly evaluate the pre-annotation effect. However, we evaluated how much the dictionary refinement over the rounds (first four rounds) directly impacted the annotation results. The IAA values remained stable over the rounds, and the annotation time decreased, so refining the dictionary was effective. Our coverage increased since the 5th round, the round in which our dictionary was complete. Thus, a higher number of correct labeled events were provided to the annotators after the 4th round.

Our study concluded that pre-annotation were effective, corroborating with the results from the studies of Hernandez et al. (2014), Oliveira et al. (2017), Fort and Sagot (2010), Lingren et al. (2014), Névéol, Dogan and Lu (2011) and Grouin and Névéol (2014).

Clinical texts suffer from characteristics that negatively impact the development of pre-annotation tools. Aspects such as many orthographic errors, acronyms, and abbreviations directly impact the pre-annotation performance, especially in a scenario where acronyms and abbreviations can vary according to the institution. Dictionary-based pre-annotations are not robust to deal with these characteristics.

5. CONCLUSION

We created a pre-annotation tool based on a dictionary that was effective for our ambulatory clinical texts in Brazilian Portuguese. Refining the dictionary over the rounds improved the pre-annotation coverage and did not negatively impact the IAA values. Clinical text's characteristics (e.g., orthographic errors, acronyms, and abbreviations) directly influenced the dictionary-based pre-annotation tool's performance, lowering its effectiveness.

The pre-annotation tool was developed for cardiology texts, but it can be adapted to other medical specialties besides cardiology, enabling its usage in other annotation projects.

We plan to evaluate to annotate documents without the pre-annotation tool to have a more precise evaluation in our future work. Besides, we plan to combine dictionary-based pre-annotations with ML-based pre-annotations. Deep learning-based approaches can learn the event's context and be more robust to deal with clinical text writing characteristics.

ACKNOWLEDGEMENTS

This work was supported by the Brazilian Government Agency Coordination for the Improvement of Higher Education Personnel (CAPES).

REFERENCES

[1] BODENREIDER, O. The unified medical language system (UMLS): integrating biomedical terminology. **Nucleic acids research**, v. 32, p. 267-270, 2004.

[2] FORT, K.; SAGOT, B. Influence of pre-annotation on POS-tagged corpus development. In: **Proceedings of the ACL 2010 - LAW 2010: 4th Linguistic Annotation Workshop**. 2010. p. 56-63.

[3] GROUIN, C.; NÉVÉOL, A. De-identification of clinical notes in French: towards a protocol for reference corpus development. **Journal of biomedical informatics**, v. 50, p. 151-161, 2014.

[4] HAMON, T. et al. POMELO: Medline corpus with manually annotated food-drug interactions. In: **Proceedings of the Biomedical NLP Workshop associated with RANLP**. 2017. p. 73-80.

[5] HERNANDEZ, A. et al. Testing pre-annotation to help non-experts identify drug-drug interactions mentioned in drug product labeling. In: **Proceedings of the AAAI Conference on Human Computation and Crowdsourcing**. 2014. p. 14-15.

[6] JENSEN, P. B.; JENSEN, L. J.; BRUNAK, S. Mining electronic health records: towards better research applications and clinical care. **Nature Reviews Genetics**, v. 13, n. 6, p. 395-405, 2012.

[7] JIANG, F. et al. Artificial intelligence in healthcare: past, present and future. **Stroke and vascular neurology**, v. 2, n. 4, p. 230-243, 2017.

[8] LINGREN, T. 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. **Journal of the American Medical Informatics Association**, v. 21, n. 3, p. 406-413, 2014.

[9] LINGREN, T. et al. Pre-annotating clinical notes and clinical trial announcements for gold standard corpus development: Evaluating the impact on annotation speed and potential bias. In: **Proceedings of the 2012 IEEE 2nd Conference on Healthcare Informatics, Imaging and Systems Biology**. 2012. p. 108-108.

[10] NÉVÉOL, A.; DOĞAN, R. I.; LU, Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. **Journal of biomedical informatics**, v. 44, n. 2, p. 310-318, 2011.

[11] OLIVEIRA, L. E. S. et al. A statistics and UMLS-based tool for assisted semantic annotation of Brazilian clinical documents. In: **Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine**. 2017. p. 1072-1078.

[12] SHEIKHALISHAHI, S. et al. Natural language processing of clinical notes on chronic diseases: systematic review. **JMIR medical informatics**, v. 7, n. 2, p. e12239, 2019.

[13] SOUTH, B. R. et al. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. **Journal of biomedical informatics**, v. 50, p. 162-172, 2014.

[14] STYLER, W. F. et al. Temporal annotation in the clinical domain. **Transactions of the Association for Computational Linguistics**, v. 2, p. 143-154, 2014.

[15] SUN, Weiyi; RUMSHISKY, Anna; UZUNER, Ozlem. Evaluating temporal relations in clinical text: i2b2 2012 Challenge. **Journal of the American Medical Informatics Association**, v. 20, n. 5, p. 806-813, 2013.

[16] World Health Organization. Global status report on noncommunicable diseases. 2020.