A Statistics and UMLS-based Tool for Assisted Semantic Annotation of Brazilian Clinical Documents

Lucas E. S. Oliveira*, Caroline P. Gebeluca, Adalniza M. P. Silva, Claudia M. C. Moro
Health Technology Program
Pontifícia Universidade Católica do Paraná
Curitiba, PR, Brasil
e-mail: lucas.oliveira@pucpr.br

Sadid A. Hasan, Oladimeji Farri
AI Lab
Philips Research North America
Cambridge, MA, USA

***Abstract*—Natural Language Processing and Machine Learning techniques can be used to automatically identify, extract and manipulate textual clinical data. Many of these methods are strongly dependent on annotated corpora that are very difficult to find in the clinical domain, especially for the Brazilian Portuguese language. The annotation task is expensive and time-consuming; hence, it is important to provide intelligent computational tools to facilitate this kind of work. In this paper, we propose a collaborative annotation tool that assists the user by proposing the UMLS semantic types of the clinical concepts based on the previous annotation statistics and UMLS terminology access via REST API. Our evaluation was focused on the amount of effort saved by the annotation tool, reliability of the preliminary annotations and efficacy of the annotation assistant.**

Keywords:Natural Language Processing; Data Curation; Unified Medical Language System.

I. INTRODUCTION AND BACKGROUND

Most biomedical data are stored in Electronic Health Records (EHR) as unstructured texts that are essential to obtain certain types of clinical information [1]. Biomedical Information Extraction (BioIE) applications aim to automatically find semantic structures like clinical concepts and its relations from biomedical texts and can be applied to various tasks and relevant domains such as Clinical Decision Support (CDS), integrative biology, pharmacovigilance, etc. [2].

Natural Language Processing (NLP) techniques are widely used in BioIE applications where knowledge-, linguistics- and statistics-driven approaches can be utilized to extract meaningful information from texts. Supervised Machine Learning (ML) methods are also commonly used in statistical NLP tasks [3].

Although statistics-focused NLP techniques have achieved good results, they require laborious and repetitive manual annotation of texts to train a prediction model. Basically, manual annotation is a step of mapping text to knowledge domain structures, and can be both linguistic (syntactic) and semantic [3].

When it comes to semantic annotation the goal is to disambiguate biomedical information in unstructured data and connect its meaning to a well-described concept/relation in a terminological resource (i.e., UMLS [4] and SNOMED-CT [5]) [6-7]. Such annotated data is useful to train efficient models for biomedical NLP and information retrieval tasks. Clinical decision support tools can utilize these models to understand the correct diagnosis from an underlying clinical scenario for an effective treatment plan by administering appropriate tests and procedures; and ultimately, optimize the patient outcome across the care continuum [8].

It is difficult to find de-identified and semantically annotated corpora in the biomedical domain readily available to use, especially for the Brazilian Portuguese language (pt-br). Hence, it is essential to provide an environment that can facilitate such annotation of biomedical corpora. However, fully manual annotation is almost impossible due to the large amount of text that needs to be analyzed, and completely automatic annotation methods do not always bring acceptable results [9].

There exist some automatic systems that perform semantic annotation with good precision, e.g. MetaMap [10] and cTakes [11], but none of those are available for pt-br texts. Moreover, there are some annotation tools available for pt-br texts, like corte-e-costura [12], an annotation system that applies rules automatically to reduce the annotation effort in journalistic texts.

Lingren et al. [13] performed some experiments on the impact of pre-annotating English-written clinical notes on speed and potential bias of annotation. Besides that, they cited others' research who did similar evaluations. The conclusion was that an annotator with pre-annotated text takes less time to annotate the text and the use of such functionality does not introduce bias to the annotations.

The main objective of our work is to develop a collaborative tool to semi-automatically annotate pt-br biomedical texts assisted by terminological sources and previous annotations. Furthermore, we assess the amount of effort saved, the efficacy of the tool's annotation assistant, and measure preliminary annotation reliability.

II. MATERIALS AND METHODS

In this section, we describe the development of the present work, including the data used, an outline of the developed tool and the conducted experiments.

A. Data

For annotation purposes, we used a de-identified database of nursing notes from the Nephrology department of a University Hospital. These nursing notes are free-texts, but with semi-structured aspects and including many kinds of abbreviations, as shown in Figure 1.

```
RIM UNICO POR DOAÇÃO RENAL PARA O FILHO
INTOLERANCIA A GLICOSE
S: SEM QUEIXAS
O: PA 140/80  140/85   P 78   AFEBRIL   PESO 65
CORADA E HIDRATADA
CPP LIVRES
PC BCRNF
ABD RHA+, FLACIDO, INDOLOR, CIC SUPRAPUBICA A ESQUERDA
MMII SEM EDEMA
```

Figure 1. An example of a nursing note with its abbreviations marked. Abbreviations meanings: S=Subjective, O=Objective, PA= Blood Pressure, P: Weight, CPP= Lung pleural fields, PC=Request for consultation, etc.

For experimentations (described in the Experiments section) in the first part (E1), we selected 3 random sets with 10 notes each (named S1, S2 and S3). And for the second part (E2), 2 random sets with 15 notes were selected (S4 and S5). The inclusion criterion was to have a minimum of 40 words in the text.

The current online version of UMLS has a mixture of European (pt-pt) and Brazilian Portuguese (pt-br) terms. Hence, we used the UMLS 2013AA version for our experiments, which has been recently adapted to pt-br by Oliveira et al. [14].

B. Tool Description

One of the objectives of the tool is to enable medical experts without high computer familiarity and skills to easily work with the annotation process. So, the tool mainly involves allowing for text selection with the mouse to annotate terms (Figure 2) and simple “Yes or No” (“Sim”,”Não”) question-answering (Figure 3).

All the UMLS semantic types were defined as the possible tags to be annotated, as UMLS “integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services” [4]. Additionally, we cannot choose other terminologies, like SNOMED-CT alone, because they are not translated to pt-br.

In addition to the UMLS types, we also included two other annotation types: “Negation” and “Abbreviation” as this information is very common and can be very useful for a biomedical annotated corpus in its application to downstream NLP tasks.

It is worth mentioning that is possible to label a concept with more than one semantic type. For instance the text “Patient presented with NVD”. The word NVD can be annotated as “Sign or Symptom” and “Abbreviation” (Nausea-Vomiting-Diarrhea).

1) Technical Aspects

To ease future collaborative annotation work and be a platform-independent system, we developed the annotation tool as a Web Application, using Java (JSP and Servlets), MySQL database and heavy JavaScript client to decentralize processing and avoid server overload during the annotation process.

To access the terminology data, we used the UMLS Terminology Services (UTS) REST API Service [15].

2) Data Workflow

The data workflow runs through six main modules: text importation, text review, text assignment, text annotation, annotation adjudication and annotation exportation.

- Text importation: functionality to import texts from “.xls” and “.csv” files into the system.
- Text review: users with special permissions can read the imported texts, remove identifiable information that the de-identification algorithm might have mistakenly failed to anonymize.
- Text assignment: define which texts will be assigned to each annotator.
- Text annotation: the main functionality, where the annotator can label the assigned texts with semantic types, supported by the annotation assistant that will suggest possible semantic types to terms in the text (more details in the annotation assistant section). Each concept can be annotated with one or multiple semantic types.
- Annotation adjudication: where one or more adjudicators can resolve the annotators’ divergences about the same texts and set a gold standard.
- Annotation exportation: to export the annotated information as JSON or XML file.

3) Annotation Assistant

In order to reduce the time and effort of the annotation work, a component called annotation assistant was featured in the text annotation module. With the assistant the annotator can “curate” annotation suggestions rather than manually annotate all texts from scratch. This is different than pre-annotation that existing approaches [13, 16] use, because the assistant does not label the texts in advance. Rather only asks the annotator if certain annotation is preferred or not.

The annotation suggestions are both statistics- and terminology-based. The statistics-based approach searches for terms in the text and queries previous annotations to retrieve annotation statistics for each term. All the previous annotations made by the annotator are used to suggest a semantic type (i.e., if “Myocardial Infarction” was already labeled as “Disease or Syndrome” before, it will suggest the same semantic type to the annotator for this term. If

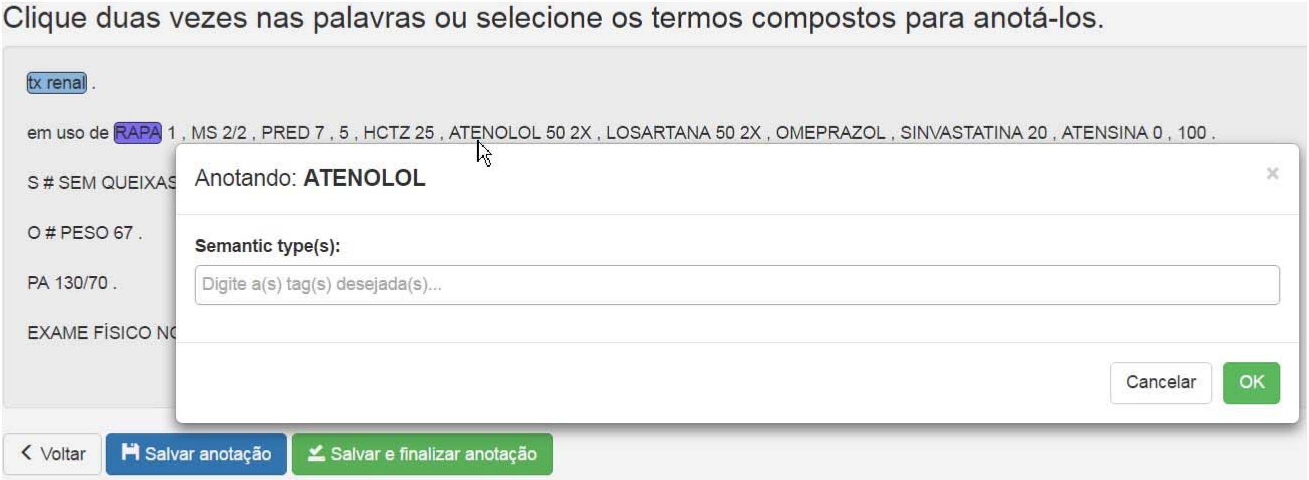

Figure 2. A view of the annotation screen. When a term is selected or double-clicked in the text, a dialog box to select the related semantic type is opened. In this example the user selected the word "ATENOLOL" and it is possible to select one or more semantic types related to this word.

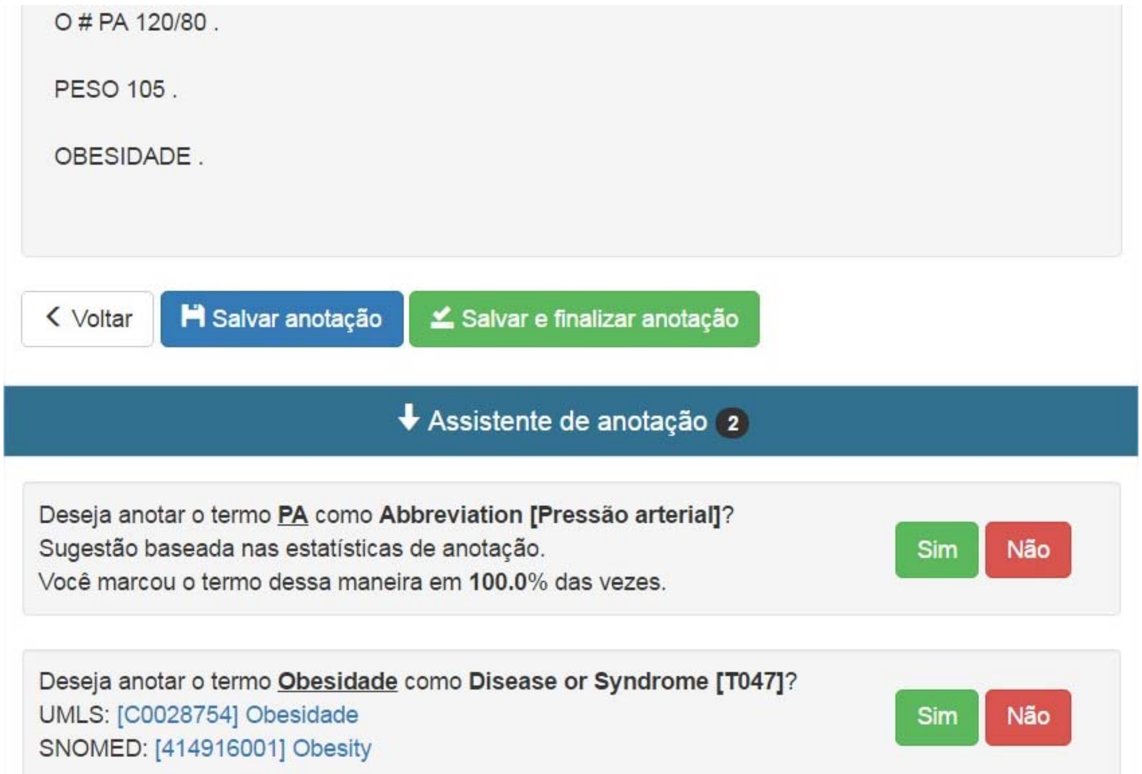

Figure 3. Annotation assistant ("Assistente de anotação" bottom bar) suggesting semantic types. The first suggestion is based on statistics of previous annotations ("PA" it is the "Blood pressure" abbreviation) where the annotators annotated "PA" as "Abbreviation" 100% of the occurrences. And the second is based on UMLS terminology, showing the UMLS CUI of the term and the related term in SNOMED-CT.

"cortex" was labeled with 2 semantic types before, it will make 2 suggestions: "Body Part, Organ, or Organ Component" marked in 80% of occurrences and "Body Location or Region" in 20% of cases).

The suggestions originated from the terminology-based approach come from terms in text that are found in pt-br translated part of UMLS like shown in Figure 3. If annotators answer yes to a suggestion, immediately the term is labeled with the same semantic type. The overview of the assistant is presented in Figure 4.

With some exceptions, the algorithm iterates through the words and searches them in the pt-br UMLS to suggest annotations. For instance, the text span "Exame fisico normal" indicates a compound term corresponding to

"[C0855737] Physical examination normal" in UMLS. In this case, we do not want that a suggestion of "Exame" that corresponds to "[C1261322] Evaluation procedure" to appear. So, before searching the single word "Exame" in UMLS, a recursive method looks for all combinations of multiple sequential words (i.e., "Exame fisico", "Exame fisico normal", etc.). If an occurrence is found, the algorithm does not suggest for the single words inside the compound term, but only for the compound terms found.

Another issue is when a term appears in its plural form, but the corresponding UMLS term is defined in its singular form. For example, the text span "Paciente com múltiplas paradas cardíacas" (Patient with multiple cardiac arrest) indicates the concept "[C0018790] Cardiac Arrest". In pt-br, the singular form of this term is "Parada cardíaca" and the plural form is "Paradas cardíacas". As we are using an exact match in our search, it is needed to normalize plural words to singular to be able to find those cases. We used the CoGrOO POS Tagger [17] and a simple set of replacing rules to generate the singular form of a term, and consequently, find the correct matching term in UMLS. The CoGrOO algorithm was used to parse sentences and tokens as well.

Whenever the annotation of a text is finished by all assigned annotators, the adjudication is enabled by the tool.

The adjudication interface (presented in Figure 5) is very similar to the annotation assistant, where the user simply answers if s/he wants to annotate the divergent annotations or not. The adjudicator cannot edit the concepts for which the annotators agreed on the semantic types.

C. *Experiments*

Our experiments are outlined in two main parts. In Experiment 1 (E1), we seek to assess if the tool reduces the annotation effort and improve the understanding of the annotation process, aiming to define the annotation guidelines. Two annotators (A1 and A2), performed a manual annotation (without the tool) of S1 set of notes together, labeling the semantic types of the terms. Both annotators have experience and practice in elaborating clinical narratives. After this, the annotators and a part of the research team discussed errors, discordances, difficulties and how certain terms need to be annotated. All these observations were summarized into an Annotation Guideline, together with UMLS semantic types' meaning translation (in pt-br) and some annotation examples.

Furthermore, in E1, A1 and A2 performed an individual manual and semi-automatic (using the tool) annotation of S2. And then, the S3 set was annotated individually by A1 and A2 only using the tool. All the E1 steps are shown in Figure 6.

In Experiment 2 (E2) the focus was to measure the annotation reliability and the efficacy of the annotation assistant by comparing annotations using and not using the assistant. By this time, the annotators had much more familiarity with the tool and the nursing notes. Additionally, the annotation tool's statistics-based suggestions already had some previous annotated text to work on.

In E2, two annotators conducted the annotation of S4 and S5, where S4 was annotated using the annotation assistant and S5 not using it. Both sets were adjudicated by a medical expert who had experience in other biomedical annotation tasks.

III. RESULTS

With E1 we wanted to test the hypothesis that an annotation tool can reduce annotation effort (time). Table 1 presents E1 annotation results in every combination of set of notes (except S1 that was made only to align the guidelines) and annotators (A1 and A2). Information relative to the time

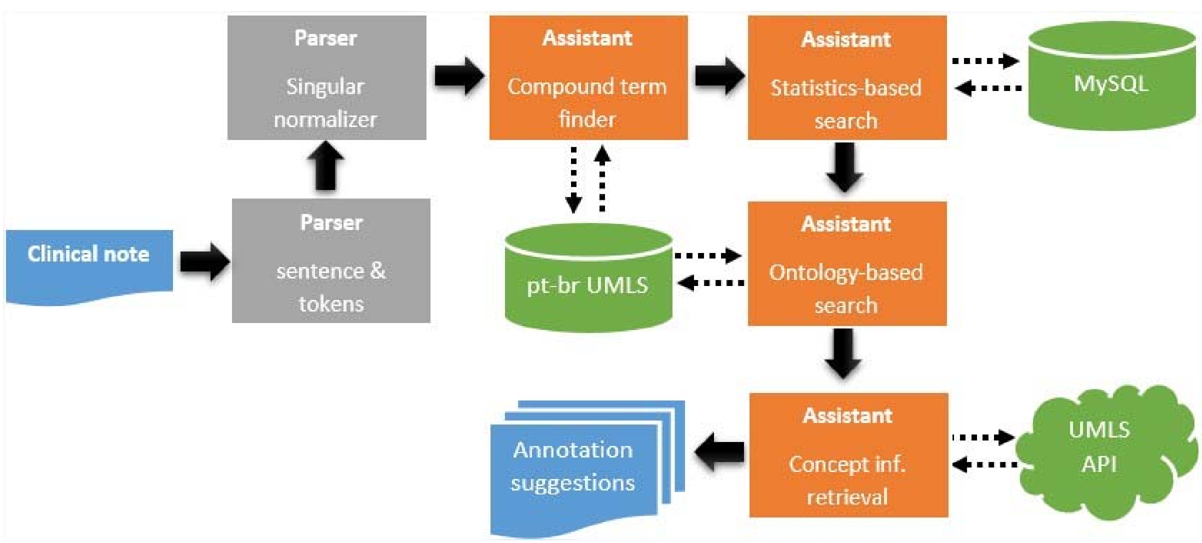

Figure 4. Annotation assistant overview. After the terms were found in the UMLS local database, the concept information retrieval method gets complete information in UMLS API (UTS) services

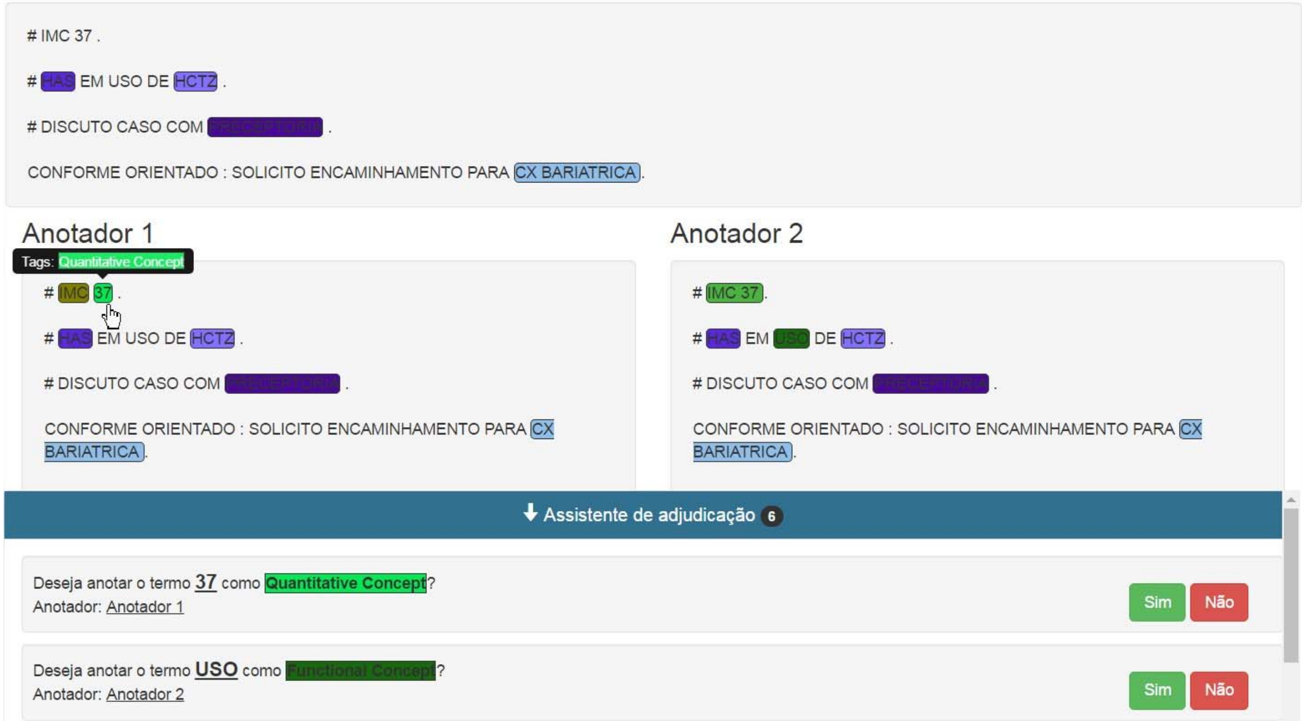

Figure 5. Adjudication interface. Inside the top frame we show the concepts agreed by the 2 annotators. The original annotations made by Annotator 1 are inside the left frame and by the Annotator 2 inside the right frame. Each semantic type has its own color (i.e., Quantitative concept is represented by a specific tonality of green). The adjudicator can curate the divergent annotations presented in the bottom frame

spent in all annotations (in minutes), number of annotations made and annotation speed (annotations per minute) are shown.

The annotation of S2 using the tool compared to manual annotation showed a time reduction of 44 minutes (53%) with A1 and 32 minutes (45.7%) with A2. The annotation speed increased at the same level for both annotators.

The annotation speed of S3 remained close to S2 even without previous manual annotation of these notes, different from S2 where the annotators had previous contact with the set of notes before annotating them with the tool.

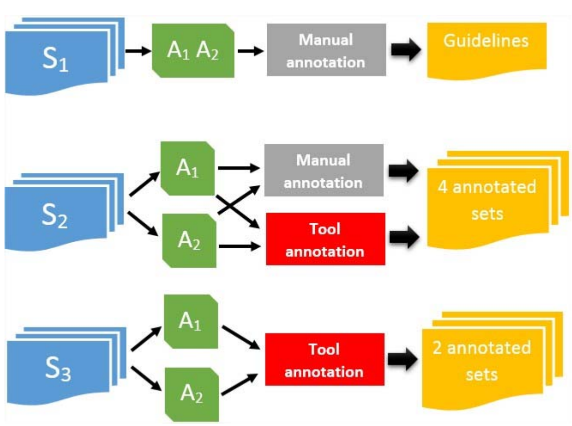

Figure 6. Overview of Experiment 1 annotations

Table I. Experiment 1 results: Time (minutes), Speed (annotations per minute) and quantity of annotations

	Time spent	***Annotation quantity***	***Annotation speed***
S_2 | A_1 | Manual	83	98	1.18
S_2 | A_1 | Tool	39	178	4.56
S_2 | A_2 | Manual	70	107	1.53
S_2 | A_2 | Tool	38	154	4.05
S_3 | A_1 | Tool	46	214	4.65
S_3 | A_2 | Tool	40	159	3.98

Table 2 shows that A1 accepted 77 suggestions based on previously annotated data and 83 on UMLS lookup that represents 40.8% of all annotations made by A1. For A2 the percentage is similar: 40.3%.

After the adjudication (curation) of E2 annotations (S4 + S5), the generated ground-truth was used to measure the performance of annotators A1 and A2. A1 scored 0.81 for F-Measure while A2 scored 0.63.

As expected, the most used semantic type was the “Abbreviation” with 24.1% of all annotations made in E2. Table 3 shows the Observed agreement (Ao) and the proportion of annotations of each one of the most used semantic types.

Due to the low agreement in 3 important semantic types, it is unlikely that E2 annotations are reliable, even with the adjudicator "corrections".

Table II. Experiment 1 results: Annotation suggestions based on statistics and terminology

	Terminology-based	*Statistics-based*
S_2 | A_1	28	21
S_3 | A_1	55	56
S_2 | A_2	29	18
S_3 | A_2	38	41

After E2 finished the team reunited to discuss the results. The "Abbreviation" result can be explained because A2 it less experienced than A1, and did not know the meaning of part of the abbreviations, and even knowing that some word was an abbreviation, A1 did not label it for not knowing the expanded form.

Table III. Experiment 2 results: Agreement between annotators and percentage of use by semantic types in S4 and S5

	% of use	*Ao*
Abbreviation	24.2%	0.593
Finding	7.8%	0.106
Sign or Symptom	5.9%	0.166
Health Care Activity	5.3%	0.875
Patient or Disabled Group	3.6%	1.00
Spatial Concept	2.9%	1.00
Pharmacologic Substance	2.2%	0.875

The "Finding" and "Sign or Symptom" are related in UMLS semantic type's hierarchy tree, where "Sign or Symptom" is a specialization of "Finding". That's why they caused some misunderstandings to annotate, even with examples defined in Annotation Guidelines.

The same numbers presented in Table 3 were calculated separately for S4 and S5 with a small overall agreement improvement when using the annotation assistant (S4). Mainly because the Assistant proposed some underused semantic types (i.e., "Machine Activity" or "Professional Society" semantic types) that the annotators forgot in some cases, and just remembered to use because they had a suggestion from annotation Assistant.

IV. Discussion

The experiments with both annotators demonstrated a time savings when annotating texts using the tool. The results show that approximately 40% of all annotations in E1 were suggested by the annotation assistant. That is to say, the annotators were curators in almost half of the annotations, which took less time than annotating from scratch. On the other hand, E2 showed the effectiveness of the annotation assistant with regard to annotation speed although there was not much improvement as to the reliability of annotation data.

Another possible reason for the time savings is the availability of such a tool specifically designed for pt-br free text clinical note annotations. Moreover, showing the possible semantic tags and enabling functionalities like autocomplete, it is easier to perform the task than using a simple text editor.

An advantage of using a hybrid statistics- and terminology-based pre-annotation approach is that we do not have to deal with a limited machine pre-annotation assistance when annotating the first set of notes. Some statistics-based tools that do not have an initial training set, like RapTAT [18], face this problem.

With the statistics-based assistant, the hypothesis is that with more texts annotated, the quality of the suggestions improves, while this experiment was already using nursing notes from Nephrology that are more prone to have repetitive terms.

In Lingren's work [13], they mentioned that some terms "produce spurious annotations that cost time in removing" which is exactly what happened in South study [16], for example. This was not the case in our work since our tool does not pre-annotate the terms directly; rather it only provides suggestions to the annotators, so it is not necessary to spend additional time in removing erroneous pre-annotations.

Although the main objective of this work was to annotate only semantic types, the tool effortlessly mapped the concepts to UMLS terms, because every terminology-based suggestion that is accepted is automatically saved with the respective CUI (Concept Unique Id).

Instead of using more complex solutions for term extraction [19] to resolve the compound term issue, we opted for a simpler method to look up all the possible word combinations in a local database that demonstrated effective performance.

Some of the feedback given by the community of users of an already existing tool, the Semantator [9], were also implemented in our tool, including: annotators' information related to the annotations made, traceability of human and machine annotation, and the use of the same colors for each available semantic type.

As the availability of pt-br biomedical annotation tools and corpora are scarce, the presented tool can fill this gap for Health Informatics researchers in Brazil. However, E2 showed us that more work is needed in training annotators and defining guidelines especially for the most frequent semantic types.

A limitation of this study is that we did not evaluate the annotation process using multi-specialty documents to know how the annotation assistant would behave with an extra variety of vocabulary.

V. CONCLUSION AND FUTURE WORK

This study proposed and developed a semantic annotation tool for Brazilian Portuguese biomedical texts featuring semi-automatic (assisted) annotation based on a terminological source (UMLS) and statistics from previous annotations. The environment provides a comprehensive annotation workflow, from text importation and annotation to adjudication and exportation. We evaluated the amount of time saved using the tool compared to manual annotation and concluded that the pre-annotation can reduce the annotation effort, although there was a lack of improvement as to the reliability of annotation.

Several improvements are desirable for future work with regard to adding other sources for dictionary-based pre-annotation (like an abbreviation lexicon) and to develop a Machine Learning model to predict annotation for enhancing the capabilities of the annotation assistant. Further functionalities to enable crowdsourcing [20] would be useful to advance the Biomedical Information Extraction research for the pt-br language. We also intend to incorporate additional features for semantic annotation including: relationship mapping between annotated concepts and UMLS mapping through a search box to improve the user experience of the annotation assistant.

ACKNOWLEDGMENT

This work is supported by Philips Research North America and PUCPR.

REFERENCES

[1] P. Raghavan, J. L. Chen, E. Fosler-Lussier, and A. M. Lai, “How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?,” AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci., vol. 2014, pp. 218–23, 2014.

[2] F. Liu, J. Chen, A. Jagannatha, and H. Yu, “Learning for Biomedical Information Extraction: Methodological Review of Recent Advances,” 2016.

[3] C. Friedman, T. C. Rindflesch, and M. Corn, “Natural language processing : State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine q,” J. Biomed. Inform., vol. 46, no. 5, pp. 765–773, 2013.

[4] D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray, “The unified medical language system,” Methods Inf. Med., vol. 32, no. 4, pp. 281–291, 1993.

[5] R. A. Cote, S. Robboy. “Progress in medical information management. Systematized nomenclature of medicine (SNOMED)”. Union Med Can 1980; 109(9):1243–52.

[6] R. Berlanga, V. Nebot, and M. Pérez, “Tailored semantic annotation for semantic search,” J. Web Semant., vol. 30, pp. 69–81, 2015.

[7] T. Gschwandtner, K. Kaiser, P. Martini, and S. Miksch, “Easing semantically enriched information retrieval-An interactive semi-automatic annotation system for medical documents,” Int. J. Hum. Comput. Stud., vol. 68, no. 6, pp. 370–385, 2010.

[8] A. Prakash, S. Zhao, S. A. Hasan, V. Datla, K. Lee, A. Qadir, J. Liu, and O. Farri. “Condensed Memory Networks for Clinical Diagnostic Inferencing”. In Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI), pp. 3274-3280, 2017

[9] C. Tao, D. Song, D. Sharma, and C. G. Chute, “Semantator: Semantic annotator for converting biomedical text to linked data,” J. Biomed. Inform., vol. 46, no. 5, pp. 882–893, 2013.

[10] A. R. Aronson, F. M. Lang. “An overview of MetaMap: historical perspective and recent advances.” J Am Med Inform Assoc 2010;17(3):229–36.

[11] G. K. Savova, J. J. Masanz, P. V. Ogren, J Zheng, S. Sohn, K. C. Kipper-Schuler. “Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications”. J Am Med Inform Assoc 2010;17(5): 507–13.

[12] D. Santos and C. Mota, “Experiments in Human-computer Cooperation for the Semantic Annotation of Portuguese Corpora,” Proc. Seventh Conf. Int. Lang. Resour. Eval., pp. 1437–1444, 2010.

[13] T. Lingren, L. Deleger, K. Molnar, H. Zhai, J. Meinzen-Derr, M. Kaiser, L. Stoutenborough, Q. Li, and I. Solti, “Pre-annotating clinical notes and clinical trial announcements for gold standard corpus development: Evaluating the impact on annotation speed and potential bias,” Proc. - 2012 IEEE 2nd Conf. Healthc. Informatics, Imaging Syst. Biol. HISB 2012, p. 108, 2012.

[14] L. E. S. Oliveira, S. A. Hasan, O. Farri, C. M. C. M. Barra. “Translation of UMLS ontologies from European Portuguese to Brazilian Portuguese”. CBIS 2016 - XV Congresso Brasileiro de Informática em Saúde, 2016, Goiânia. CBIS 2016 - XV Congresso Brasileiro de Informática em Saúde, 2016.

[15] UMLS Terminology Services. https://uts.nlm.nih.gov/home.html (accessed 30 Aug 2017).

[16] B. R. South, D. Mowery, Y. Suo, J. Leng, Oscar Fernandez, S. M. Meystre, and W. W. Chapman, “Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text,” J. Biomed. Inform., vol. 50, pp. 162–172, 2014.

[17] J. Kinoshita, C. Eduardo, and D. De Menezes, “CoGrOO: a Brazilian-Portuguese Grammar Checker based on the CETENFOLHA Corpus,” October, pp. 2190–2193, 2003.

[18] G. T. Gobbel, J. Garvin, R. Reeves, R. M. Cronin, J. Heavirland, J. Williams, A. Weaver, S. Jayaramaraja, D. Giuse, T. Speroff, S. H. Brown, H. Xu, and M. E. Matheny, “Assisted annotation of medical free text using RapTAT,” J. Am. Med. Informatics Assoc., vol. 21, no. 5, pp. 833–841, 2014.

[19] M. da Silva Conrado, A. Felippo, T. Salgueiro Pardo, and S. Rezende, “A survey of automatic term extraction for Brazilian Portuguese,” J. Brazilian Comput. Soc., vol. 20, no. 1, p. 12, 2014.

[20] R. Khare, B. M. Good, R. Leaman, A. I. Su, and Z. Lu, “Crowdsourcing in biomedicine: Challenges and opportunities,” Brief. Bioinform., vol. 17, no. 1, pp. 23–32, 2016.