Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230268

Ontology-Based Semantic Annotation of French Psychiatric Clinical Documents

Ons AOUINA[a,1], Jacques HILBEY[a] and Jean CHARLET[a,b]
[a]*Sorbonne Université, Sorbonne Paris Nord, INSERM, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé - LIMICS, Paris, France*
[b]*Assistance Publique-Hôpitaux de Paris, Paris, France*

Abstract. Building a timeline of psychiatric patient profiles can answer many valuable questions, such as how important medical events affect the progression of psychosis in patients. However, the majority of text information extraction and semantic annotation tools, as well as domain ontologies, are only available in English and cannot be easily extended to other languages, due to fundamental linguistic differences. In this paper, we describe a semantic annotation system based on an ontology developed in the PsyCARE framework. Our system is being manually evaluated by two annotators on 50 patient discharge summaries, showing promising results.

Keywords. Semantic Annotation, GATE, Ontology, Psychiatry, NLP

1. Introduction

Schizophrenia and chronic psychosis are among the most debilitating disorders in adolescents and young adults and are associated with cognitive impairment, poorer occupational success, and poor quality of life. Studies have shown that the longer the duration of untreated psychosis, the worse the outcome of intervention, the worse the recovery and general functioning, and the greater the long-term social impairment [1]. This issue is addressed by the RHU PsyCARE[2] project, which aims to improve early detection and intervention in psychoses.

In this context, the analysis of Patient Discharge Summaries PDSs can give us the opportunity to study many valuable questions such as how important medical events affect the progression of psychosis in patients. These summaries provide information about the patient's history (e.g., associated with the onset of symptoms or the start of treatment). However, extracting such information to trace the history of psychosis and develop the timeline is a complex matter that requires carefully annotated corpora. Being able to automatically extract this information would improve medical care and support clinical research.

In this project, we propose a method of semantic annotation of PDSs based on an ontology developed in the framework of PsyCARE. Our approach will not only extract medical entities from a text but also transform them into structured and formalized knowledge. Ontologies allow the design of semantic data indexes that leverage medical

[1] Corresponding Author: Ons AOUINA; E-mail: ons.aouina@etu.sorbonne-universite.fr.
[2] https://psy-care.fr/

knowledge to improve information retrieval and search [2]. This proposal is a first brick in a large project that aims to build a complete timeline of a patient's psychosis based on his medical records.

2. Material and Methods

2.1. Psychiatry Dataset

The clinical documents used in this work are derived from the French PsyCARE project. It is a compilation of about 8000 anonymized Patient Discharge Summaries covering a period of ten years, which represents a volume of about 3,500,000 words. These summaries come from the Groupe Hospitalier Universitaire Psychiatrie et Neurosciences de Paris, the largest psychiatric hospital in Paris. They are semi-standardized, in Word format and have been strictly anonymized beforehand, by deleting all names, dates, places, etc. In addition, the diagnosis is indicated at the end of each document and is referenced to the ICD-10[3].

These records are written in French and describe the patient's history and social context, medications, details of hospital admission, and current and previous psychiatric diagnoses.

2.2. Domain Ontology description

The ontology developed in the framework of PsyCARE both to integrate the data and to allow their semantic annotation represents domains such as psychiatric clinical aspects, drugs with their ATC code, imaging, biology, etc. In our context we are interested in the psychiatric clinical branch which describes the signs, symptoms, disorders related to psychiatry as well as the medication branch.

Based on this ontology, we reconstructed an annotation scheme to which we added a branch describing the structure of the PDSs[4]. This allows us to relate concepts to their context of occurrence. For example, drugs appearing in the Disease History section do not share the same context as those appearing in the Discharge Treatment section.

The version we present here is not yet complete. As we have already mentioned, our final goal is to model clinical events in psychiatry. Hence the need to include a temporal representation of medical knowledge [3].

2.3. Semantic Annotation

Several tools have been developed in the field of natural language processing and semantic annotation that can be used for French texts. Among them are GATE [4], ECMT (*http://ecmt.chu-rouen.fr*) and SIFR [2] annotator. In our proposed approach, we use GATE which provides different components for semantic information extraction. We used components previously developed by the OnBaSsam [5] project and subsequently improved to meet the needs of psychiatry-related texts. Our pipeline is composed of

[3] A medical classification list by the WHO https://icd.who.int/browse10/2019/en

[4] This branch of the ontology is available at the following address https://github.com/AouinaOns/PartOfDocuments_ontology

several Processing Resources (PRs) that run sequentially on a given document, as shown in Figure 1.

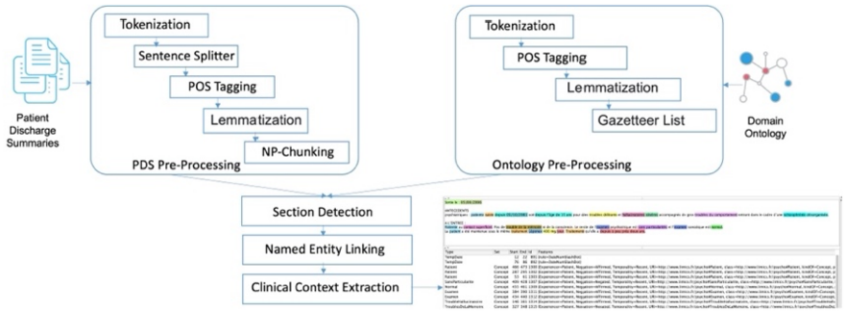

Figure 1. Illustration of the ontology-based semantic annotation process.

Task 1 - Pre-processing: First, Word tokenization is applied on PDSs, then sentence splitting, followed by Part Of Speech (POS) tagging and lemmatization, and finally, Noun Phrases extraction or NP-chunking.

For this purpose, we used Gate Corpus Pipeline, a Gate component for corpus processing, to which we added the following PRs: a French tokenizer for tokenization and the french TreeTagger[5] for annotating texts with information about parts of speech and lemmas. It has been successfully used to tag French texts. Finally, to retrieve noun phrases from PDSs, OpenNLP[6] is used, and adapted to the French language. This step allows us to build a list of noun phrases from the output of the TreeTagger. In our solution, we assume that entities such as signs and symptoms, diseases, disorders, and clinical events are noun phrases [5]. Regarding the ontology processing, we follow the same steps for the building of the gazetteer list. This list consists of the ontology concepts and their labels, pre-processed according to the steps presented in figure 1, as well as the URIs.

Task 2 - Section detection: The discourse structure of a document can be very useful for improving information retrieval tools. The identification of sections, e.g., Current Illness Story, Family Story, is crucial in our context because it is the major key to identifying the temporal context of narrative passages. As mentioned earlier, the PDSs are organized according to taxonomy (Cf. sec. 2.2). This structure is not always followed, some sections may be missing, merged or in a different order, and the same section may have several titles. For example, for the section "*family history*" we may find "*fam history*", "*psychiatric family history*", etc.

Therefore, we apply JAPE [7] rules and fuzzy string matching on the terms of the previously constructed section name gazetteer to identify section boundaries. It uses the concept of the Levenshtein distance which is a metric representing the number of character changes between words. This rule identifies the beginning of a section by the appearance of a term available in the terminology and the end before the beginning of the next term.

Task 3 - Named Entity Extraction and Linking: For this task, we classify the NP-chunk candidates obtained in the pre-processing phase to ontology concepts by assigning

[5] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

[6] https://github.com/GateNLP/gateplugin-OpenNLP

them a URI. For this task, we develop a rule-based system combined with a fuzzy matching string classification. We define the annotation acceptance threshold as being proportional to the length of the dictionary term.
For extracting temporal entities, we used the TIMEX[7] GATE plugin to annotate the documents with TIMEX3 tags using the SUTime (Stanford Temporal Tagger) library. This allowed us to extract dates, duration e.g. "*since about 1 month*", and frequencies e.g. "*2 times a day*".

Task 4 - Clinical Context Extraction: In addition to extracting the named entities themselves, it is necessary to identify the context in which they appear in the text. Documents are pre-annotated and mapped to ontology concepts (signs, diseases, treatments, symptoms, temporal expressions, etc.) using the process described above. Then, the French FastContext [7] algorithm is applied, to identify the context of the clinical conditions annotated in a sentence. It considers three contexts: negation, hypothesis, and determination of the subject, whether the patient, patient's relative or healthcare professional.

3. Results

Table 1. Quantitative results of named entity extraction evaluations by the 2 annotators.

		Quantity	Precision	Recall	F1
Sign Or Symptom		1747	0.9544	0.9503	0.9524
Disease		150	0.9826	0.7635	0.8593
Trouble		459	0.9894	0.9493	0.9690
Clinical Event.		1459	0.9744	0.8140	0.8870
Personal Situation		188	0.9895	0.8468	0.9126
Drug	Drug Name	1034	0.8200	0.9805	0.8822
	Drug Dose	650	0.9848	0.9610	0.9727
Temporal Inf.	Date	840	0.9942	0.9709	0.9824
	Duration	529	0.9574	0.9777	0.9674
	Frequency	212	0.9459	0.8373	0.8883

The evaluation of our approach consists in manually analyzing the named entity extraction phase. Our evaluation corpus consists in 50 PDSs randomly extracted from the dataset (4100 sentences, 4213 non-unique ontology concepts annotated). Two persons evaluate the annotations and their context including negation, hypothetical, temporality, and experiencer. We have, therefore, grouped them into 10 unique higher-level concept ontologies to facilitate manual evaluation.

The precision, recall, and F-measure are presented in Table 1. An inter-annotator agreement has been calculated (0.88). The results were promising with an overall precision value of 0.9674, a recall of 0.9780, and a F1 of 0.9727.

4. Discussion and Conclusion

The objective of this work is to reconstruct structured patient data from PDSs to complete

[7] https://github.com/pkourdis/gateplugin-SUTime

the patient data in the PsyCARE project. In this paper we presented a first step which consists in semantically annotating the French psychiatry PDSs. To this end, from an unstructured set, we were able to perform a semantic annotation using Gate plugins and algorithms that we modified to fit PDSs structure.

A first evaluation of the semantic annotations shows that we are able to correctly identify the concepts of the ontology with GATE and modified algorithms to take into account French texts. These good results can be explained in particular by taking into account the structure of the PDSs, which could prove to be a weakness for other types of psychiatric documents.

The next steps in our work are to integrate into the ontology the clinical events and non-psychiatric diseases most frequently present in the PDSs. Then, it will be necessary to identify the relationships between the concepts. Given the state of the art, this will be done with machine learning and deep learning algorithms.

Acknowledgment

This work has been supported by the French government's “Investissements d'Avenir” program, which is managed by the Agence Nationale de la Recherche (ANR), under the reference PsyCARE ANR-18-RHUS-0014.

References

[1] Souaiby L, Gaillard R, Krebs MO. Durée de psychose non traitée: état des lieux et analyse critique [Duration of untreated psychosis: A state-of-the-art review and critical analysis]. 2016 Aug;42(4):361-6.
[2] Tchechmedjiev A, Abdaoui A, Emonet V, Zevio S, Jonquet C. SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes. BMC Bioinformatics. 2018.
[3] Hilbey J, Aimé X, Charlet J. Temporal Medical Knowledge Representation Using Ontologies. Challenges of Trustable AI and Added-Value on Health: Proceedings of MIE 2022. 2022 May 1;294:337.
[4] Cunningham H, Maynard D, Bontcheva K, Tablan V, Aswani N et al. 2011. Text processing with GATE (Version 6). Sheffield: GATE. doi: https://doi.org/10.1016/j.procs.2014.05.444.
[5] Cardoso S, Meneton P, Aimé X, Meininger V, Grabli D, et al. Use of a modular ontology and a semantic annotation tool to describe the care pathway of patients with amyotrophic lateral sclerosis in a coordination network. PLOS ONE 16(1): e0244604. doi: https://doi.org/10.1371/journal.pone.0244604
[6] Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. Journal of biomedical informatics. 2013 Dec 1;46(6):1088-98.
[7] Yacoubi Ayadi N, Charrad M, Vidal ME, Ben Ahmed M, Amdouni S. A Semantic Framework for Web service Annotation, Matching and Classification in Bioinformatics. Information interaction intelligence. 2011 Nov 1;11(2).
[8] Mirzapour M, Abdaoui A, Tchechmedjiev A, Digan W, Bringay S, Jonquet C. French FastContext: A publicly accessible system for detecting negation, temporality and experiencer in French clinical notes. Journal of Biomedical Informatics. 2021 May 1;117:103733..