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

doi:10.3233/SHTI240789

SemOntoMap: A Hybrid Approach for Semantic Annotation of Clinical Texts

Ons AOUINA[a,1], Jacques Hilbey [a,b] and Jean CHARLET [a,b]
[a]*Sorbonne Université, Inserm, Université Sorbonne Paris-Nord, LIMICS, Paris, France*
[b]*Assistance Publique-Hôpitaux de Paris, Paris, France*
ORCiD ID: Ons Aouina https://orcid.org/0000-0002-6455-0314

Abstract. This study addresses the challenge of leveraging free-text descriptions in Electronic Health Records (EHR) for clinical research and healthcare improvement. Despite the potential of this data, its direct interpretation by computers is limited. Semantic annotation emerges as a method to make EHR free text machine-interpretable but struggles with specific domain ontologies and faces heightened difficulties in psychiatry. To tackle these challenges, this study proposes a system based on unsupervised learning techniques to extract entities and their relationships, aligning them with a domain ontology. The effectiveness of this system has been validated within PsyCARE project by analyzing 60 patient discharge summaries.

Keywords. Semantic annotation, ontology embedding, unsupervised NLP.

1. Introduction

Biomedical texts are pivotal for enhancing clinical practices and innovation in patient care, with semantic annotation playing a crucial role by linking texts to meaningful ontological tags, thus improving interoperability and retrieval efficiency. The effort of semantic annotation is increasingly automated, leveraging advancements in natural language processing (NLP) [1]. Our study focuses on the automatic annotation of clinical texts, a task made complex by the nature of medical language that, which requires named entity recognition (NER) and disambiguation and relation extraction (RE). Special attention is paid to the narrative sections of clinical records, especially in psychiatric discharge summaries (PDSs), which contain information about clinically significant events affecting the patient's medical trajectory. This information includes family history, disease history, prescribed treatments, test results, and the temporal relationships between these events. Questions like "*How has the disease progressed in the patient?*" can only be interpreted and answered if the complete context of the patient's history and the temporal relationships between the identified concepts are considered. This problem is addressed in psychiatry by the RHU PsyCARE project, which aims to improve early intervention in psychosis by providing tools to facilitate access to care and offer personalized treatment programs. Therefore, our work aims to capture PDSs content with standard ontologies to decipher modalities, such as temporal relationships, and detailed

[1] Corresponding Author: Ons Aouina; E-mail: ons.aouina@etu.sorbonne-universite.fr.

information on the history and progression of psychosis. In this paper, we propose a method for the semantic annotation of PDSs by combining a domain ontology with language models and unsupervised learning algorithms to construct an accurate model of the text [2].

2. Material and Methods

This section presents SemOntoMap, a system architecture for enriching psychiatric PDSs with semantic annotations and based on unsupervised methods. The process unfolds in three main phases as shown in Figure 1. Methodologies for text processing (1), section identification (a), temporal entities extraction (b), context extraction (i) and gazetteer creation (d) based on the ontology are further detailed in [3].

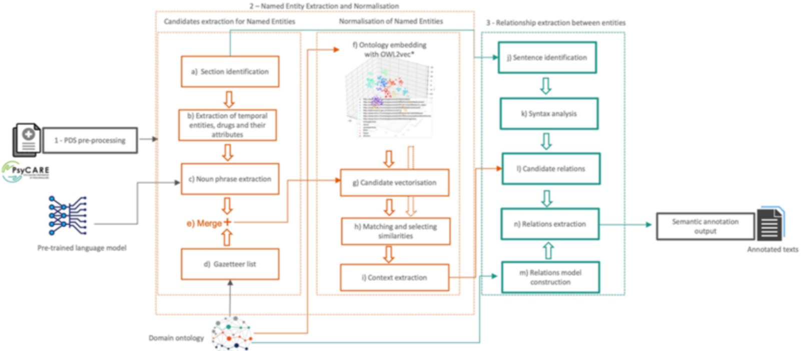
2 – Named Entity Extraction and Normalisation
Candidates extraction for Named Entities
Normalisation of Named Entities
3 - Relationship extraction between entities
1 - PDS pre-processing
PsyCARE
Pre-trained language model
a) Section identification
b) Extraction of temporal entities, drugs and their attributes
c) Noun phrase extraction
e) Merge
d) Gazetteer list
f) Ontology embedding with OWL2vec*
g) Candidate vectorisation
h) Matching and selecting similarities
i) Context extraction
j) Sentence identification
k) Syntax analysis
l) Candidate relations
n) Relations extraction
m) Relations model construction
Semantic annotation output
Annotated texts
Domain ontology

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

2.1. Data

Psychiatry Dataset: For our work, we have at our disposal 8,000 PDSs from a ten-year span, collected from the University Hospital Group for Psychiatry and Neuroscience in Paris. These documents have been pseudo-anonymized to remove personal identifiers and they include a diagnosis per the ICD-10. Written in French, the PDSs give a comprehensive view of each patient's medical history, social background, medication, hospital admission details, and psychiatric diagnoses. For evaluation, 60 reports were randomly selected based on ICD-10 codes to represent a diverse array of psychiatric conditions.
Domain Ontology description: The ontology developed for the PsyCARE project [4], plays a crucial role beyond just NLP; it serves as a comprehensive model for interoperability within the project. The ontology for NLP (subsequently called OntoPSY) consists of branches of interest such as psychiatric clinical aspects (signs, symptoms, psychiatric disorders), medications identified by their ATC code, elements related to imaging, and a temporal dimension to adequately represent medical knowledge. A branch of the ontology dedicated to the structure of PDSs is added to link concepts to their context of appearance in the document [5]. In addition, it details the relationships between different concepts, thereby enriching our understanding of interactions and connections within clinical data. Finally, it includes 8,114 entities and 41,506 axioms, which enabled us to create an annotation schema as a basis for semantic annotation [3].

2.2. Named Entity Recognition and Normalization

Extracting Noun Phrases (NPs) from French psychiatric narratives is pivotal to our methodology. These NPs are instrumental in capturing essential entities that reflect the intricacies of psychiatric assessments. Our approach extends beyond simple identification; we aim to understand the clinical significance of these terms. This aspect is highlighted by Liu et al. as crucial for enhancing the accessibility of clinical documents [5]. After the initial extraction, we normalize these entities. This process, also referred to as disambiguation or entity linking, associates textual mentions with standardized categories in an ontology to achieve consistency across documents. To augment this phase, we incorporate data from knowledge graph structures and leverage both word and entity embeddings. These methods facilitate the establishment of meaningful connections between entities [6]. Moreover, ontology embedding is indispensable for infusing ontology-driven insights into our process. This significantly enhances the identification of ontological concepts within academic texts, ensuring that the extracted knowledge is semantically aligned with the relevant ontology [6]. This section delves into the tasks of named entity recognition and normalization as follows:
Task (c+e) - Entities Candidate Extraction: We perform unsupervised extraction of NPs (c), adapting PatternRank [7] to the narrative complexity of PDSs. This algorithm represents the state of the art in key-phrase extraction, thanks to its integration of part-of-speech models for the selection of candidate phrases, thus allowing its adaptation to various domains. This approach involves text segmentation, syntactic tagging (POS), and selection of phrases that meet specific criteria. In our experiments, we used Sentence-CamemBERT model[2]and, the custom pattern *<(NOUN.|JJ|ADV)+> <NOUN.>* to identify entities like clinical signs and diseases. The final extracted list is merged with the output of our pipeline described in [3] (e), which is based on fuzzy matching.
Task (f) - Ontology Embedding: OntoPSY's semantic embedding is accomplished via the OWL2Vec* [8] tool. This algorithm effectively condenses the semantic and structural information of the ontology into a compact vector space, facilitating the use of this data by machine learning algorithms for downstream tasks. Configured to leverage a Word2Vec model trained on a french medical corpus, the model undergoes fine-tuning, intricately aligned with ontology peculiarities (https://github.com/Kureman/NegBiRNNs) .
Task (g+h) - Matching and Selecting Similarities: Each extracted NP is transformed into a vector using the output from OWL2Vec*, allowing its representation in the same vector space as the ontology's axioms. We then rank the ten ontological concepts closest to each NP vector by cosine semantic similarity. A re-ranking module, based on syntactic analysis, refines this list to identify the concept that aligns most accurately with the NP [9]. Thus, by analyzing the grammatical structure to pinpoint identify the main entity. Finally, we select the top-ranked concept, i.e. the first, as the most appropriate match. More details about the parameters and outcomes of the OWL2Vec* training are made available through a GitHub[3] repository.

2.3. Unsupervised Relation Extraction

At this stage of the annotation process, the extracted information is linked to the concepts of OntoPSY. Figure 1 describes the architecture of the RE process within an

[2] https://huggingface.co/dangvantuan/sentence-camembert-large

[3] https://github.com/AouinaOns/Semantic-Annotation

unsupervised framework. Initial relation selection is guided by the ontology's structure, restricting possible relations to those supported by ontological knowledge.

This approach unfolds through four main phases: Phase 1 involves selecting relations initially guided by the ontology's structure and relationships. This approach restricts the set of possible relations to those supported by ontological knowledge.

In Phase 2, a dependency parser based on Spacy's transition model4 is used for syntactic analysis. A dependency tree is generated for each sentence, by assigning a syntactic function to each word. Phase 3 focuses on identifying relations by analyzing syntactic paths within the word dependency tree. This analysis reveals syntactic connections, validating relations that link hospitalization to conditions like “has as a motive”. Phase 4 applies specific rules centered on words located before temporal entities to determine the precise nature of the temporal relation. The methodology is focused on the extraction of 14 distinct relationships, organized into five categories: temporal sequences, causative motives, participation links, qualification details, and dosage instructions. By navigating the challenge of identifying patterns, syntactic dependencies, and semantic clues without labeled data for each entity pair, this unsupervised approach significantly leverages the ontology's structure and advanced parsing techniques to elucidate complex relationships within the text.

3. Results

The evaluation of our approach consists in manually analyzing the NER and normalization task as well as the RE task. The inter-annotator agreement, assessed through contributions from two annotators on all tasks, showed a significant consistency level of 0.79. Scores of NER and normalization as well as RE are detailed in Table 1, demonstrated an overall precision of 0.95, a recall of 0.91 and an F1 score of 0.92.

Table 1. Quantitative results of NER and RE tasks evaluations by the 2 annotators.

	Category	Quant.	Precision	Recall	F1
Named Entities	Episodes [Disease, Clinical and life events]	2429	0.9840	0.9680	0.9530
	Substance [Name, Dose, Drug Form]	1656	0.8827	0.9646	0.9142
	Temporal Inf. [Date, Duration, Time, Freq.]	2028	0.9625	0.9179	0.9638
	Clinical Statements [Clinical signs and Exam’s]	3987	0.9850	0.9640	0.9570
	Other entities [Individual, Qualifier, Body Part]	1162	0.9893	0.8284	0.9040
Relations	Temporal has	860	0.9130	0.8077	0.8571
	has for cause	1050	0.9750	0.9070	0.9398
	Participate	286	0.9831	0.5800	0.7296
	Qualifies	756	0.9211	0.7368	0.8188
	Medication Dosage	521	0.9046	0.8333	0.8678

The analysis highlights also the system's effective entity normalization, achieving 84.8% accuracy in correctly identifying the most relevant URI for an entity on the first attempt (Hits@1), and 90.4% accuracy within the top five suggestions (Hits@5).

As for the Mean Reciprocal Rank MRR, which provides an overview of system performance by considering the rank of the correct the score obtained is 85%. Further details of the system's performance and the entities and relationships extracted from the PDS are available on GitHub.

[4] https://spacy.io/models/fr a CamemBERT-based transformer pipeline with a precision of 0.95

4. Discussion and Conclusions

This study demonstrates the effectiveness of a systematic approach to analyzing entities and relationships within complex psychiatric texts. Using unsupervised learning methods and OntoPSY, a specialized ontology for psychiatry, it retrieves and normalizes biomedical entities and discerns relationships between them. Firstly, OntoPSY's extensive psychiatric vocabulary enhances the quality of ontological embeddings and normalization processes. Additionally, the integration of specific structuring in the relation extraction process improves precision but presents challenges in generalizability. Our method can be adapted to different domains and languages; however, it is crucial to consider the language model and domain ontology used, as they are closely related and can have a significant impact on the precision and accuracy of the NER and normalization phase. We tested this method on French nephrology texts using a nephrology ontology, and the results are promising. The potential of unsupervised learning and domain-specific ontology integration to enhance learning precision and reduce manual annotation dependency is also explored. The system's accuracy heavily relies on precise initial phrase extraction, as early errors can affect subsequent analyses. Despite satisfactory performance, future research will focus on improving various components. Efforts will involve a comparative analysis of non-contextual (OWL2Vec*) and contextual ontology embeddings for semantic annotation. This will entail leveraging already annotated PDSs for weakly supervised learning and testing the effectiveness of the approach with publicly available annotated French medical data.

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] Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A. Text processing through Web services: calling Whatizit. Bioinformatics. 2008 Jan 15;24(2):296-8. doi: 10.1093/bioinformatics/btm557.

[2] Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet. 2006 Feb;7(2):119-29. doi: 10.1038/nrg1768. PMID: 16418747.

[3] Aouina O, Hilbey J, Charlet J. Ontology-Based Semantic Annotation of French Psychiatric Clinical Documents. Stud Health Technol Inform. 2023 May 18;302. doi: 10.3233/SHTI230268.

[4] Hilbey J, Aimé X, Charlet J. Temporal Medical Knowledge Representation Using Ontologies. Stud Health Technol Inform. 2022 May 25;294:337-341. doi: 10.3233/SHTI220470.

[5] Hur A, Janjua N, Ahmed M. A survey on state-of-the-art techniques for knowledge graphs construction and challenges ahead. 2021 IEEE Fourth International AIKE ; doi:10.1109/aike52691.2021.00021

[6] Devkota P, Mohanty SD, Manda P. A gated recurrent unit based architecture for recognizing ontology concepts from biological literature. BioData Mining. 2022 Sept 28. doi:10.1186/s13040-022-00310-0

[7] Schopf T, Klimek S, Matthes F. Patternrank: Leveraging pretrained language models and part of speech for unsupervised keyphrase extraction. In Proc. of the 14th Int. Joint Conf. on Knowl Discov and KM. 2022; doi:10.5220/0011546600003335

[8] Chen J, Hu P, Jimenez-Ruiz E, Holter OM, Antonyrajah D, Horrocks I. Owl2vec*: Embedding of owl ontologies. Machine Learning. 2021 Jun 16; doi:10.1007/s10994-021-05997-6

[9] Karadeniz İ, Özgür A. Linking entities through an ontology using word embeddings and syntactic re-ranking. BMC Bioinformatics. 2019 Mar 27;20(1). doi:10.1186/s12859-019-2678-8 1.