SemOntoMap : une méthode hybride pour l'annotation sémantique de textes cliniques en psychiatrie

O. Aouina[1], J. Hilbey[1,2], J. Charlet[1,2]

[1] Sorbonne Université, Inserm, Université Sorbonne Paris-Nord, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, Paris, France
[2] Assistance Publique-Hôpitaux de Paris, Paris, France

ons.aouina@etu.sorbonne-universite.fr

Résumé

Les descriptions en texte libre contenues dans les dossiers patients informatisés (DPI) revêtent un intérêt significatif pour la recherche clinique et l'optimisation des soins. Toutefois, la capacité des ordinateurs à interpréter directement ce texte libre est limitée, réduisant ainsi sa valeur potentielle. Bien que l'annotation sémantique offre une solution pour rendre le texte libre des DPI interprétable par les machines, elle rencontre des obstacles majeurs lorsqu'elle est appliquée aux ontologies de domaine spécifiques, particulièrement en français. Ces difficultés sont encore plus marquées dans le domaine psychiatrique où l'on cherche non seulement à extraire les concepts du domaine mais à les normaliser et à extraire les relations de textes décrivant longuement l'histoire d'une maladie d'un patient et ses ascendants. Face à ces enjeux, nous proposons un système fondé sur des techniques d'apprentissage non supervisé pour extraire les entités et leurs interrelations en utilisant une ontologie de domaine. Ce système est évalué dans le cadre du projet PsyCARE sur un échantillon de 60 comptes rendus analysés par deux évaluateurs.

Mots-clés

annotation sémantique, ontologie, plongement de l'ontologie, apprentissage automatique non supervisé, TALN, BERT, Word2Vec

Abstract

The free-text descriptions contained in Electronic Health Records (EHR) hold significant interest for clinical research and the optimization of care. However, computers' ability to directly interpret this free text is limited, thereby reducing its potential value. While semantic annotation offers a solution to make the free text of EHRs machine-interpretable, it faces major obstacles when applied to specific domain ontologies, particularly in French. These difficulties are even more pronounced in the psychiatric field, where there is an attempt to extract domain concepts relations from texts that extensively describe a patient's disease history and their ancestors. Faced with these challenges, we propose a system based on unsupervised learning techniques to extract entities and their interrelations using a domain ontology. This system is evaluated within the framework of the PsyCARE project on a sample of 60 reports analyzed by two evaluators.

Keywords

Semantic annotation, ontology embedding, unsupervised machine learning, Ontology, NLP, BERT, Word2Vec

1 Introduction

Dans le domaine de la recherche biomédicale et des soins aux patients, les documents techniques, et plus spécifiquement les textes biomédicaux, sont cruciaux. Ces documents, sont indispensables pour faire avancer les pratiques cliniques et stimuler l'innovation en santé. La complexité et la richesse de ces textes ont mené à l'utilisation de diverses ontologies biomédicales, telles que l'Ontologie des Gènes (GO) et la Nomenclature Systématisée de Médecine – Termes Cliniques (SNOMED CT), marquant des efforts significatifs pour structurer cette information vitale et améliorer son accessibilité [34]. L'annotation sémantique, qui associe le texte à des balises significatives issues de ces ontologies, joue un rôle clé dans de nombreuses applications, renforçant l'interopérabilité et l'efficacité de la récupération d'informations [27].

L'effort d'annotation sémantique varie du manuel au totalement automatisé, exploitant les avancées en traitement automatique du langage naturel (TALN) [25]. Notre étude se concentre sur l'annotation automatique de textes cliniques, tâche rendue complexe par la nature du langage médical. Une attention particulière est portée aux sections narratives des dossiers cliniques, surtout dans les résumés de sortie en psychiatrie (dans notre cas des comptes rendus d'hospitalisation ou CRH), qui recèlent des informations sur les événements cliniquement significatifs affectant la trajectoire médicale du patient. Ces informations incluent les antécédents familiaux, l'historique de la maladie, les traitements prescrits, ainsi que les relations temporelles entre ces événements. Des questions comme « comment la maladie a évolué chez le patient ? » ne peuvent être interprétées et on ne peut y répondre que si l'on prend en compte le contexte complet des antécédents du patient et les relations temporelles entre les différents concepts repérés. Ce problème est

abordé en psychiatrie par le projet RHU PsyCARE [1] qui vise à améliorer l'intervention précoce dans la psychose en fournissant des outils pour faciliter l'accès aux soins et offrir des programmes de traitement personnalisés.
Notre travail vise donc à annoter sémantiquement des CRH, en capturant les segments textuels qui correspondent à des ontologies ou des terminologies standardisées mais aussi en déchiffrant les modalités, les relations temporelles et les informations détaillées sur les antécédents et l'évolution de la psychose. Dans cet article, nous proposons une méthode d'annotation sémantique des CRH fondée d'abord sur une ontologie développée dans le cadre de PsyCARE. Cette ontologie est combinée avec des modèles de langue et des algorithmes d'apprentissage pour construire un modèle formel précis du texte [15].

2 Contexte

Les CRH sont rédigés par des professionnels de la santé divers et aux styles d'écriture variés. Le traitement efficace de ces documents est essentiel pour ces derniers qui doivent parcourir d'importants volumes de dossiers médicaux électroniques pour dégager les informations clés. La normalisation des entités nommées joue un rôle crucial dans la réduction de l'ambiguïté des comptes rendus cliniques. Néanmoins, ces étapes peuvent aboutir à l'extraction de concepts redondants ou de faible valeur informative. Face à ce défi, le développement des méthodes d'extraction d'information (EI) non supervisées devient une évidence. Ces algorithmes permettent d'identifier des informations significatives sans dépendre des corpus préalablement annotés, offrant ainsi une réponse efficace aux contraintes des approches traditionnelles [22].
Parallèlement, la normalisation des entités, également connue sous les termes de désambiguïsation ou de liaison d'entités, joue un rôle crucial dans l'extraction d'informations. Cette démarche consiste à associer les mentions d'entités présentes dans le texte avec des catégories ou des concepts issus d'un vocabulaire de référence [28] ou d'une ontologie spécifique, ce qui permet d'uniformiser la représentation de ces mentions. Pour améliorer l'efficacité de la normalisation des entités, certaines recherches proposent d'intégrer des données concernant la structure des graphes de connaissances [24], tandis que d'autres études mettent en avant les bénéfices de combiner les plongements de mots et d'entités afin de créer des connexions significatives entre les entités. Ces approches visent à renforcer les performances de cette tâche en exploitant les relations sémantiques profondes [24]. Dans ce contexte, l'*embedding* (ou plongement) d'ontologies représente un domaine de recherche prometteur. P. Devkota *et al.* [9] montre que l'intégration d'informations issues du plongement d'ontologies peut significativement affiner la détection des concepts ontologiques dans la littérature scientifique, renforçant ainsi la concordance sémantique entre les informations textuelles et les structures ontologiques [3].
Dans cette section, nous explorons les techniques d'EI qui, dans notre contexte, concernent l'extraction de syntagmes, y compris les syntagmes nominaux (SN) et les syntagmes verbaux (SV) ainsi que le plongement d'ontologies pour structurer et enrichir les connaissances extraites.

1. https://psy-care.fr/

2.1 Extraction de syntagmes

Pour l'extraction de syntagmes, nous adoptons l'hypothèse selon laquelle les syntagmes correspondent à une liste de N-grammes, soit des séquences de n mots manifestant une structuration grammaticale particulière. Cette tâche consiste à déterminer un ensemble de séquences de mots qui encapsulent les thèmes centraux ou les idées présentées dans un document, offrant un aperçu de son contenu le plus critique. Ces algorithmes sont classés en méthodes supervisées [32] et non supervisées [31]. Compte tenu de la polyvalence et de l'applicabilité générale des méthodes non supervisées, se concentrant sur les attributs inhérents du texte pour l'extraction de syntagmes, notre proposition se concentre sur l'extraction non supervisée. Trois méthodes principales se distinguent dans ce domaine :

Méthodes fondées sur les Graphes. Ces méthodes convertissent le document en un graphe et classent les phrases candidates dans le graphe [21, 31] . Les nœuds correspondent à des éléments textuels tels que les mots ou les phrases, et les arêtes reflètent les liens entre eux, par exemple la co-occurrence ou la similarité sémantique. Cette approche permet d'évaluer l'importance des phrases candidates en fonction de leur position et de leurs connexions au sein du graphe. En exploitant les relations contextuelles entre les éléments textuels, ces méthodes se distinguent par leur capacité à identifier avec précision les syntagmes clés qui sont directement liés aux thèmes centraux du document. Ainsi, elles améliorent la pertinence et l'efficacité des systèmes de récupération d'information en facilitant l'identification de syntagmes essentiels qui récapitulent de manière efficace le contenu central du texte.

Méthodes Statistiques. Ces méthodes, telles que YAKE [5], sont fondées sur TF-IDF (Fréquence du Terme - Inverse de la Fréquence des Documents), TextRank [21] ou SingleRank [37]. Ces méthodes analysent les propriétés de distribution des mots et des phrases dans un texte par rapport au corpus de travail pour identifier des phrases clés et mettre en évidence des termes qui sont spécifiques et informatifs du contenu du document. L'importance de combiner des analyses statistiques avec des informations contextuelles est spécialement mise en avant par YAKE qui se distingue par son utilisation de métriques statistiques avancées pour saisir le contexte et la dispersion des termes à travers le document. Mais bien que ces méthodes soient efficaces d'un point de vue computationnel et simples à mettre en œuvre, elles ne capturent pas toujours la richesse sémantique du texte, limitant potentiellement leur efficacité dans certains contextes de récupération d'informations.

Méthodes fondées sur l'apprentissage profond. Ces méthodes exploitent les réseaux neuronaux pour apprendre des représentations du texte qui capturent des relations et des motifs sémantiques. Des approches non supervisées

d'apprentissage profond telles que les auto-encodeurs ou les modèles fondés sur les transformateurs comme BERT [10], peuvent modéliser implicitement l'importance des syntagmes. Parmi celles-ci, KeyBERT [12] tire parti des modèles tels BERT, pour identifier de manière efficace les syntagmes clés dans les textes. KeyBERT combine la capacité des transformateurs à comprendre le contexte profond du texte avec une approche ciblée pour l'extraction des phrases clés, permettant ainsi une identification précise et contextuellement riche des informations clés contenues dans les documents. PatternRank [26] s'appuie sur des modèles de langage et des parties de discours (PoS) pré-entraînés pour l'extraction non supervisée de phrases-clés à partir de documents uniques. Cet algorithme représente l'état de l'art dans l'extraction de phrases clés, grâce à son intégration de modèles de partie du discours pour la sélection des phrases candidates, permettant ainsi son adaptation à divers domaines. Cette approche permet une granularité et une précision accrues dans l'extraction des phrases clés, en se basant sur des critères syntaxiques spécifiques pour identifier les éléments les plus informatifs du texte. Dans notre approche, nous avons adapté PatternRank pour améliorer sa capacité à capturer des syntagmes nominaux et verbaux, en exploitant le *part of speech*, augmentant la précision de notre méthode.

2.2 Plongement des ontologies

Les modèles de plongement de graphe de connaissances (Knowledge Graph Embedding ou KGE) sont utilisés pour la transformation des vastes réseaux complexes d'entités et de relations au sein d'un graphe de connaissances en des espaces vectoriels de faible dimension, ainsi gérables [8]. L'essence du KGE réside dans sa capacité à transformer des informations complexes et de haute dimension d'un graphe de connaissances – comprenant diverses entités et les relations à facettes multiples entre elles – en une forme à la fois efficace sur le plan du calcul et sémantiquement riche.

Plusieurs modèles pour KGE, tels que DistMult [39] et RotatE [33], ont été proposés pour relever ces défis, montrant de bons résultats sur des ensembles de données de graphes de connaissances à usage général comme FB15K-237 [35]. Cependant, leur efficacité dans des domaines spécialisés, tels que la médecine, peut ne pas être aussi satisfaisante en raison de difficultés liées à la représentation et au raisonnement autour des entités et relations médicales [11]. Les méthodes existantes ne capturent pas adéquatement les relations complexes, les structures hiérarchiques, et l'hétérogénéité des entités médicales, ni n'abordent les problèmes de données bruyantes, incomplètes et la haute dimensionnalité souvent rencontrés dans les graphes de connaissances médicales.

Dans ce contexte, le plongement d'ontologies se présente comme une approche prometteuse, complétant le KGE. Axée sur la modélisation des relations directes entre entités, le plongement d'ontologies utilise la richesse sémantique et la structure logique des ontologies [19] . Cette méthode permet de capturer non seulement les relations entre entités mais aussi les concepts abstraits, les hiérarchies de classes et les axiomes qui structurent les connaissances dans un domaine spécifique.

L'intégration des techniques de prédiction par apprentissage automatique et d'analyse statistique des ontologies gagne en popularité et des méthodes pour plonger la sémantique des ontologies OWL commencent à émerger dans la littérature. Contrairement aux graphes de connaissances, les ontologies OWL ne se limitent pas à une structure graphique mais incorporent également des constructeurs logiques, et les entités sont souvent enrichies d'informations lexicales détaillées, spécifiées via $rdfs{:}label$, $rdfs{:}comment$ et de nombreuses autres propriétés d'annotation personnalisées ou intégrées. Dans cette approche, le but du plongement d'ontologie OWL est de représenter chaque entité nommée OWL (classe, instance ou propriété) par un vecteur, de manière à conserver dans l'espace vectoriel les relations inter-entités indiquées par les informations mentionnées ci-dessus et à maximiser la performance des tâches en aval où les vecteurs d'entrée peuvent être considérés comme des caractéristiques apprises.

EL Embedding [18] et Quantum Embedding [16] sont deux algorithmes de plongement d'ontologie OWL. Ils élaborent des fonctions de score et des fonctions de perte spécifiques pour les axiomes logiques issus respectivement d'EL++ et d'ALC, en transformant les relations logiques en relations géométriques. Cela encode la sémantique des constructeurs logiques mais néglige la sémantique supplémentaire apportée par les informations lexicales de l'ontologie. De plus, bien que la structure graphique soit explorée en considérant les axiomes de sous-classement et d'appartenance à une classe, l'exploration reste incomplète car elle se limite uniquement aux arêtes $rdfs{:}subClassOf$ et $rdf{:}type$ et ignore les arêtes impliquant d'autres relations.

Onto2Vec [29] et OPA2Vec [30] sont deux algorithmes de plongement d'ontologie utilisant le paradigme du plongement de mots, fondés sur l'architecture skip-gram ou CBOW. Onto2Vec utilise les axiomes d'une ontologie comme corpus pour l'entraînement, tandis qu'OPA2Vec enrichit le corpus d'Onto2Vec avec les informations lexicales fournies par, par exemple, $rdfs{:}comment$. Les deux méthodes adoptent la fermeture déductive d'une ontologie avec un raisonnement par inférence. Les deux méthodes traitent chaque axiome comme une phrase, ce qui signifie qu'elles ne peuvent pas explorer la corrélation entre les axiomes. Cela rend difficile l'exploration complète du graphe et de la relation logique entre les axiomes, et peut également conduire à un problème de pénurie de corpus pour les ontologies de petite à moyenne échelle.

OWL2Vec* [6] propose une solution aux limitations des approches précédentes en enrichissant leur corpus d'axiomes avec des données générées par des parcours sur des graphes RDF issus de la transformation des ontologies OWL. Cette approche prend en compte à la fois le graphe et les constructeurs logiques de l'ontologie. En outre, OWL2Vec* maximise l'exploitation des informations lexicales en créant des plongements non seulement pour les entités de l'ontologie mais également pour les termes lexicaux. Ainsi, OWL2Vec* condense efficacement les informations séman-

tiques et structurelles d'une ontologie dans un espace vectoriel compact, facilitant l'utilisation de ces données par des algorithmes d'apprentissage automatique pour des tâches en aval.

Le cadre d'OWL2Vec* est structuré autour de deux étapes clés comme illustré dans la figure 1 : (i) l'extraction d'un corpus à partir de l'ontologie, et (ii) l'entraînement d'un modèle de plongement de mots avec ce corpus. Ce corpus se compose de trois documents distincts : un document de structure, un document lexical, et un document combiné. Les deux premiers documents sont conçus pour explorer la structure de l'ontologie, ses constructeurs logiques et ses informations lexicales, permettant ainsi l'activation du raisonnement par inférence. Le troisième document vise à maintenir la corrélation entre les entités (IRIs) et leurs étiquettes lexicales (mots), en utilisant le premier document comme base tout en intégrant les informations lexicales disponibles de l'ontologie.

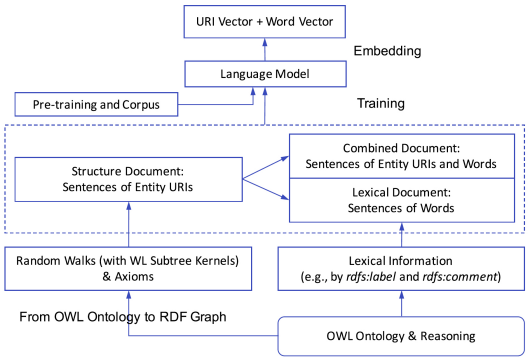

FIGURE 1 – Le contexte général d'OWL2Vec* Source : [6]

3 Système d'annotation

Dans cette section, nous présentons l'architecture du système, SemOntoMap, conçu pour enrichir les CRH de psychiatrie avec des annotations sémantiques. Notre approche s'appuie sur un corpus de textes de psychiatrie non annotés et une ontologie dédiée à ce domaine, visant à structurer ces documents.

Comme le montre la figure 2, la tâche d'annotation sémantique se déroule en trois grandes étapes : le prétraitement des textes, l'identification et la normalisation des entités nommées et l'extraction des relations entre ces entités.

3.1 Jeu de données en psychiatrie

Les documents cliniques exploités dans cette étude sont une compilation de près de 8000 CRH s'étendant sur une période de dix ans, totalisant environ 3,5 millions de mots. Ces CRH ont été collectés au sein du Groupe Hospitalier Universitaire Psychiatrie et Neurosciences de Paris. Ils sont semi-standardisés, en format Word et ont été pseudo-anonymisés au préalable, en remplaçant tous les noms, dates, lieux, etc. Chaque document se conclut par un diagnostic formulé selon la Classification Internationale des Maladies, version 10 (CIM-10 [2]). Rédigés en français, ces comptes rendus fournissent un aperçu détaillé de l'historique et du contexte social des patients, des prescriptions médicamenteuses, des circonstances d'admission à l'hôpital ainsi que les diagnostics psychiatriques actuels et antérieurs. Pour les besoins de l'annotation sémantique, une sélection de 30 comptes rendus a été annotée d'une façon aléatoire, en se basant sur le code CIM-10. Cette méthode de sélection vise à garantir une diversité dans les cas cliniques étudiés, couvrant un large éventail de diagnostics psychiatriques.

3.2 Ontologie de domaine

L'ontologie utilisée dans le processus d'annotation, appelée par la suite OntoPSY est une version fusionnée des modules ontoDOPSY, ontoMEDPSY, ontoDOME et ontoPOF de l'ontologie développée dans le cadre de PsyCARE [3] pour l'intégration des données et leur annotation sémantique. Ces modules contiennent les branches d'intérêt tels que les aspects cliniques psychiatriques (signes, symptômes, troubles psychiatriques), les médicaments identifiés par leur code ATC, des éléments relatifs à l'imagerie ainsi qu'une dimension temporelle pour représenter les connaissances médicales de manière adéquate. À partir de cette base, un schéma d'annotation est construit. Une branche de l'ontologie dédiée à la structure des CRH est ajoutée pour lier les concepts à leur contexte d'apparition dans le document, c'est-à-dire la section dans laquelle ils sont repérés [13] (« Histoire de la maladie », « Traitement de sortie », etc.) . En plus de décrire les aspects cliniques et les entités médicales, l'ontologie détaille les relations entre les différents concepts, enrichissant ainsi notre compréhension des interactions et des liens au sein des données cliniques.

3.3 Prétraitement des données textuelles

Ce processus implique le traitement du format du document et l'extraction de segments textuels pertinents à partir du document source, tout en écartant les balises et les éléments non pertinents. À ce stade, une analyse TALN de base est réalisée, incluant la tokenisation, la normalisation, et le marquage morphosyntaxique (*part-of-speech tagging*, POS). Le produit de cette phase est un texte brut enrichi de certaines annotations. Les algorithmes employés lors de cette étape ont été décrits dans un article antérieur [1].

3.4 Reconnaissance d'entités et normalisation

Dans cette section, nous détaillons les différentes étapes consacrées à l'extraction des candidats pour la reconnaissance des entités nommées (REN) ainsi qu'à leur normalisation en concordance avec les concepts de l'ontologie.

3.4.1 Extraction de syntagmes

L'importance des SN dans l'analyse des textes médicaux et psychiatriques est soulignée par les travaux de chercheurs comme Liu et al. [14] qui mettent en évidence

2. https://icd.who.int/browse10/2019/en
3. Cette ontologie sert à plus de processus que le seul TALN ; elle sert en particulier de modèle d'interopérabilité général pour le projet [13].

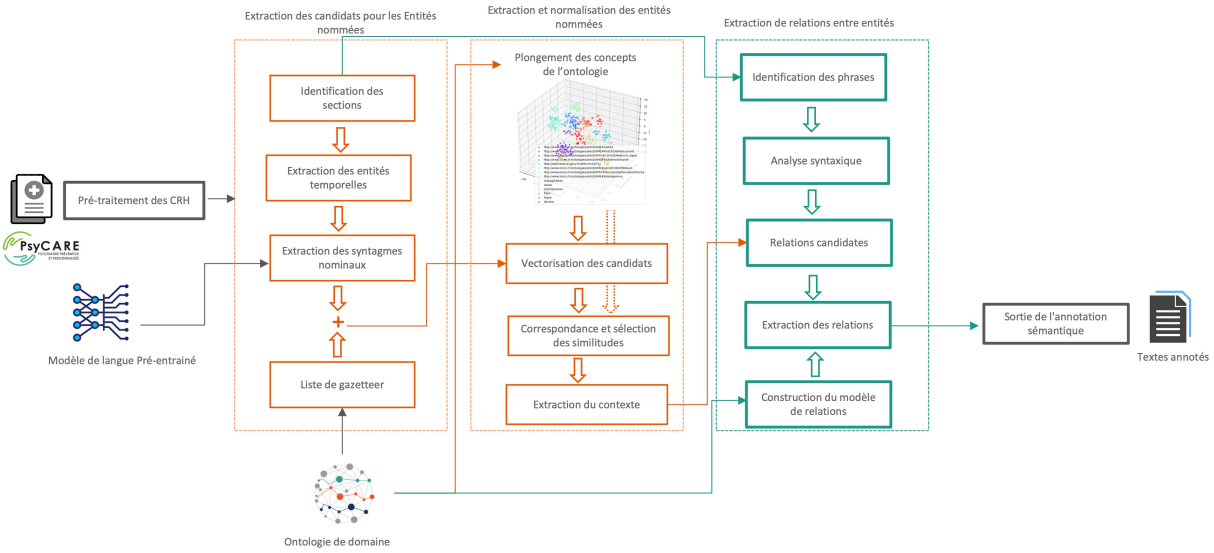

FIGURE 2 – Architecture du système d'annotation sémantique proposé.

la valeur de l'identification précise des termes médicaux pour améliorer l'accès à l'information dans les documents cliniques. Comme mentionné précédement, nous modifions PatternRank (Cf. sec. 2.1) pour l'adapter à la complexité narrative de ces documents. Les étapes sont détaillées dans la figure 3. Cette approche implique la segmentation du texte, le marquage syntaxique (POS), et la sélection des syntagmes qui répondent à des critères spécifiques (e.g., <(NN.|JJ|ADV)+><NN.> pour identifier des séquences commençant par un nom, adjectif, ou adverbe suivis de noms). Par la suite, les similarités cosinus entre la représentation du document et les représentations des syntagmes candidats sont calculées et ces derniers sont classées en ordre décroissant en fonction des scores trouvés. Les candidats sont à la fin transformés en vecteurs et classés par similarité cosinus avec le document pour extraire les termes les plus significatifs.

Dans nos expériences, nous utilisons le modèle de langage pré-entraîné Sentence-CamemBERT-Large développé par La Javaness [4]. Il s'agit d'un modèle SBERT qui a démontré sa capacité à produire de bonnes représentations textuelles pour des tâches de similarité sémantique.

3.4.2 Extraction d'information temporelles, medicaments et dosage

L'extraction des informations temporelles, médicamenteuses et de dosages est essentielle pour le suivi clinique et l'évaluation des traitements des patients. De nombreux travaux se sont concentrés sur ces extractions, notamment en ce qui concerne la temporalité et dans le domaine plus large du TALN médical [4]. Nous avons adopté la solution proposée par Aumiller Dennis [2] pour la reconnaissance et la normalisation des expressions temporelles, en combinant les capacités des librairies HeidelTime et SUTime pour l'identification complète des expressions temporelles dans les textes, car elles couvrent dates, heures, fréquences, et durées, et permettent l'ajustement de la date de référence pour une interprétation contextuelle. HeidelTime est utilisée pour son efficacité dans l'extraction temporelle de narrations non cliniques, adapté ici aux contextes cliniques. SUTime, en complément, offre la flexibilité d'une date de référence, utile pour notre analyse documentaire. Nous intégrons également le Temporal Tagger Service pour une détection précise de ces informations temporelles. Pour l'extraction des mentions de médicaments dans les comptes rendus hospitaliers en psychiatrie, EDS-NLP [36], développé par l'AP-HP, est employé pour sa spécialisation dans le traitement des données de santé en français. Enfin, le *GATE Tagger* [5] est utilisé pour identifier dosages et unités, facilitant l'interprétation des prescriptions.

3.5 Correspondance entre les informations extraites et les concepts de l'ontologie

Plongement de OntoPSY. Afin de produire le plongement sémantique de l'ontologie OntoPSY, nous avons mis en œuvre l'outil OWL2Vec* (voir Section 2.2). Cet outil a été configuré pour se servir d'un modèle Word2Vec préalablement entraîné sur un corpus diversifié, comprenant des articles de Wikipédia en français, des textes biomédicaux, ainsi que des corpus spécialisés [17]. Le modèle a ensuite été finement ajusté pour s'aligner avec les spécificités de l'ontologie, dont le prétraitement a été détaillé dans une publication antérieure [1]. Le réglage fin du modèle avec le corpus de l'ontologie a été réalisé à travers des marches aléatoires d'une profondeur de trois, permettant une exploration approfondie de la granularité de l'ontologie. Réalisée sur une série de 100 itérations, ce réglage a utilisé la même stratégie de marche aléatoire pour garantir une compréhen-

4. https://huggingface.co/dangvantuan/sentence-camembert-large

5. https://github.com/GateNLP/gateplugin-Tagger_Measurements

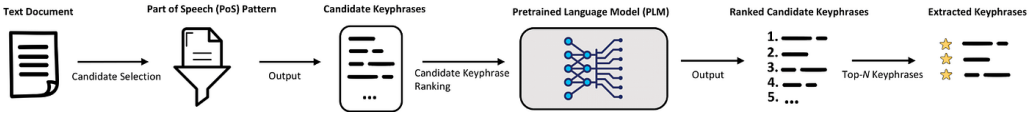

FIGURE 3 – Schéma du processus d'extraction non supervisée des syntagmes en utilisant PatternRank. Source : [26].

sion complète et adéquate de l'ontologie.

Vectorisation des candidats. Chaque syntagme extrait est transformé en un vecteur en utilisant la sortie de OWL2vec*. Cette étape de vectorisation permet leur représentation dans le même espace vectoriel que les axiomes de l'ontologie, facilitant ainsi les comparaisons sémantiques directes entre ces syntagmes et les concepts de l'ontologie.

Correspondance et sélection des similarités. Pour chaque vecteur de syntagmes, nous déterminons les dix concepts ontologiques les plus proches, classés selon leurs scores de similarité sémantique. Pour affiner cette sélection initiale et identifier avec précision le concept adéquat parmi les premiers candidats, nous employons un module de reclassement décrit plus en détail dans cet article [16], lequel se fonde sur une analyse syntaxique poussée. Ce module, exploite l'analyse syntaxique pour distinguer le concept le plus pertinent parmi les options pré-sélectionnées, en se fondant sur les scores de similarité issus de notre modèle de plongement. Le cœur de notre innovation réside dans l'utilisation de l'analyseur syntaxique de SpaCy [6], un outil conçu pour isoler le mot ou le syntagme le plus significatif au sein d'une phrase. En analysant la structure grammaticale du syntagme, le module de reclassement peut identifier avec précision l'entité principale, permettant ainsi une correspondance plus exacte entre le syntagme analysé et le concept de l'ontologie pertinent.

3.6 Extraction non supervisée de relations

Dans le contexte de l'analyse des CRH, l'extraction de relations (ER) constitue une étape cruciale du TALN. Cette tâche vise à identifier et à définir les liens sémantiques existant entre les entités nommées détectées dans le texte. Il existe deux principales techniques d'extraction de relations entre entités : les méthodes fondées sur des règles de modèles (*template rule-based*) et les méthodes fondées sur des vecteurs propres (*eigenvector-based*).

Dans la méthode fondée sur des règles, les caractéristiques linguistiques des relations entre entités sont d'abord organisées par des linguistes. Ensuite, les règles sont compilées [23], enfin, les relations entre entités sont extraites à travers une correspondance fondée sur ces règles.

Les méthodes fondées sur des vecteurs propres peuvent être divisées en deux types : l'apprentissage automatique traditionnel et l'apprentissage profond [38]. Pour répondre à nos besoins, nous utilisons une combinaison des méthodes décrites ci-dessus. Nous combinons l'analyse syntaxique des dépendances et la struture de l'ontologie pour identifier et classer les relations entre les entités.

6. https://spacy.io/models/fr

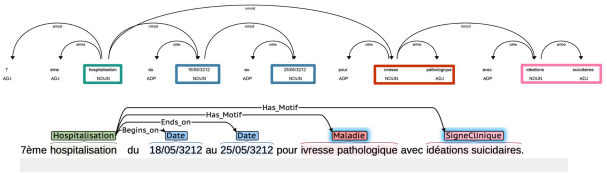

FIGURE 4 – Analyse Syntaxique et Structuration de Texte : Panneau Supérieur - Arbre d'Analyse Syntaxique avec Positions de Mots et Types de Dépendances ; Panneau Inférieur - Regroupement en SN et Identification de hospitalisation/date et hospitalisation/Maladie/SigneClinique.

3.6.1 Définition de la Tâche

À ce stade du processus d'annotation, les informations extraites sont liées aux concepts de l'ontologie OntoPSY. La figure 2 décrit l'architecture du processus d'extraction de relations (représenté par le rectangle vert en pointillés). Notre approche se décompose en quatre phases principales que nous détaillons dans cette section. Nous ciblons spécifiquement l'extraction de 14 relations que nous regroupons en cinq catégories : a) les relations temporelles, qui articulent un ordre ou une séquence d'événements ou d'épisodes de soin ; ensuite, b) les associations « a pour motif » qui relient des événements ou des épisodes de soin à leurs causes sous-jacentes, telles que des maladies ou des symptômes ; puis, c) la relation « participe », qui lie des individus ou des médicaments à l'épisode de soin auquel ils sont associés ; quatrièmement, d) les relations de qualification offrent des précisions sur les entités, en se référant à des attributs comme le niveau scolaire ou la texture pour les individus, ou pour caractériser des soins et événements. Finalement e) les relations de dosage qui concernent la connexion entre des médicaments ou substances chimiques et leurs dosages, modes d'administration, et fréquences d'administration. Cette approche est illustrée dans la figure 4 où nous présentons un exemple de ces relations.

3.6.2 Solution proposée

Dans un cadre non supervisé, le principal défi est l'absence d'échantillons étiquetés indiquant la relation spécifique r pour chaque paire d'entités (e_h, e_t) dans la phrase avec e_h est l'entité tête et e_t et l'entité queue. Par conséquent, l'ensemble des relations $\mathcal{R}$ (ensemble de r_i) est explicitement défini grace aux relations de l'ontologie et la tâche repose sur l'identification de motifs, de dépendances syntaxiques et d'indices sémantiques au sein de la phrase x pour inférer la relation potentielle r.

Phase 1 - Selection des relations candidates. La sélection des relations est initialement guidée par la structure et les

relations présentes dans l'ontologie. Par exemple, si entre les entités *Maladie* et *Signe Clinique* il n'existe pas de relation dans l'ontologie, aucune relation candidate n'est considérée entre ces entités dans notre système. Cette approche nous permet de restreindre l'ensemble des relations possibles à celles qui sont soutenues par des connaissances ontologiques, garantissant une cohérence initiale dans le processus de sélection. Dans une premiere étape, nous identifions les phrases contenant plusieurs entités nommées et établissons toutes les combinaisons possibles de relations entre elles. En tenant compte de l'ordre qui est déterminant dans notre analyse, nous considérons les entités et les relations potentielles illustrées dans notre corpus de données. Considérons la phrase type issue de notre corpus, illustrée dans la figure 4. Nous avons les entités *hospitalisation*, *ivresse pathologique*, *idéations suicidaires* ainsi que les dates spécifiques *18/05/3212* et *25/05/3212*.
Nous examinons alors les combinaisons suivantes :

- $r1(e_h, e_t)$ où e_h est *hospitalisation* et e_t est *ivresse pathologique*, la relation $r1$ pouvant être interprétée comme *a pour motif*;
- $r2(e_h, e_t)$ où e_h est *hospitalisation* et e_t est *idéations suicidaires*, la relation $r2$ étant également *a pour motif*;
- pour les dates qui peuvent être associées respectivement aux entités *hospitalisation*, *ivresse pathologique* et *idéations suicidaires* considérées comme e_h, plusieurs relations candidates sont envisageables en fonction des informations contextuelles et de l'ontologie :
  - $r3(e_h, e_t)$ pour *a pour date de début*,
  - $r4(e_h, e_t)$ pour *a pour date de fin*,
  - $r5(e_h, e_t)$ pour *a pour date*, utilisée si on ne fait pas la distinction entre la date de début et de fin,

  avec, dans les 3 cas, $e_h \in$ {*hospitalisation, ivresse pathologique, idéations suicidaires*}.

Phase 2 - Analyse Syntaxique. Nous utilisons, ensuite, un composant d'analyseur de dépendances fondé sur les transitions de Spacy. Ce dernier est fondé sur le modèle Transformer, notamment camembert-base [20] avec une précision de 0.95 [7]. Les phrases, une fois étiquetées avec des tags de parties du discours (POS), sont analysées par cet outil. Il génère alors un arbre de dépendances pour chaque phrase, illustré dans la figure 4 (panneau supérieur), et assigne une fonction syntaxique à chaque mot.

Phase 3 - Identification des relations. Cette phase exploite l'analyse des chemins syntaxiques au sein de l'arbre de dépendance par mot, traçant le parcours depuis un point de départ e_h, généralement l'effecteur, vers les entités cibles e_t. Cette analyse syntaxique révèle une absence de connexion syntaxique directe entre *ivresse pathologique* et *idéations suicidaires* et les dates, signifiant qu'aucun chemin de dépendance n'indique une relation temporelle explicite avec ces entités. Toutefois, partant de *hospitalisation*, il y a une dépendance syntaxique tant avec les entités temporelles qu'avec *Maladie* et *Signe Clinique*, indice d'une relation de modification ou de causalité. Par conséquent, nous validons les relations $r1$ et $r2$ qui relient *hospitalisation* à *ivresse pathologique* et *idéations suicidaires* via *a pour motif*, marquant un lien causal où l'hospitalisation résulte de ces conditions. Les relations $r3$, $r4$, et $r5$ associant *hospitalisation* aux dates spécifiques restent des candidats et sont sujettes à une analyse plus approfondie dans la phase suivante.

Phase 4 - Application des règles. Cette étape finale du processus d'extraction de relations tire parti de règles spécifiques centrées sur les mots situés avant les entités temporelles pour déterminer la nature précise de la relation temporelle. En s'appuyant sur des indicateurs lexicaux clairs, tels que des mots ou des expressions indiquant un commencement (« début », « commencement », « à partir du ») ou une conclusion (« fin », « jusqu'au », « termine le »), nous pouvons affiner notre compréhension des relations $r3$, $r4$, et $r5$ restantes entre *hospitalisation* et les instances temporelles. La présence de ces indicateurs dans le contexte immédiat avant une entité temporelle nous permet d'attribuer avec précision la relation la plus adéquate. Par exemple, si un indicateur de début ou de fin précède une entité temporelle associée à *hospitalisation*, la relation $r3$ ou $r4$ sont validées. Si le contexte ne spécifie pas clairement un début ou une fin, ou que les indicateurs sont ambigus ou absents, la relation générale $r5$ *a pour date* est considérée comme appropriée. En résultat, illustré dans la figure 4 (panneau Inférieur), les relations finales sélectionnées pour *hospitalisation* sont directement influencées par ces indicateurs lexicaux, renforçant l'exactitude sémantique et contextuelle de notre modèle relationnel.

4 Analyse et résultats

4.1 Analyse des performances du système

Notre approche a été évaluée de manière distincte sur trois composantes clés : l'extraction des entités nommées, la normalisation des entités, et l'extraction des relations. Pour chaque composante, nous avons réalisé une analyse manuelle approfondie en utilisant un schéma d'annotation dédié conçu pour mesurer précisément les performances de notre pipeline. L'outil d'annotation BRAT a été utilisé pour sa facilité d'usage et sa capacité à répondre à nos critères spécifiques. Dans le cadre de l'optimisation de l'évaluation manuelle, nous avons classifié les entités extraites en 17 concepts uniques de haut niveau de l'ontologie OntoPSY. Les annotations dans BRAT incluaient les URI correspondant à chaque concept de l'ontologie, donnant ainsi aux annotateurs la possibilité de corriger les annotations au besoin. Pour les relations, nous avons retenu 14 types de relation. Il faut noter la différence d'approche d'évaluation entre les entités et les relations : les entités correspondent pour une grande majorité à des concepts médicaux et sociétaux : ils peuvent être appréhendés par les évaluateurs qui ont 17 concepts de haut niveau à leur disposition et peuvent préciser les concepts repérés à l'envi en balayant l'ontologie. les relations décrites dans l'ontologie sont très précises en raison du rôle tenu par icelles – modèle de données de

7. https://spacy.io/models/fr#fr_dep_news_trf

la plateforme gérant les données cliniques – dans le projet PsyCARE. Les relations telles qu'elles sont appréhendées par les experts sont plus proches de dépendances syntaxiques visibles dans les phrases : c'est pour cela qu'on en a retenu 14 (synthétisées en 5 types, Cf.3.6.1) et que nous sollicitons les experts dessus sans leur demander d'approfondir les URI.

Notre corpus d'évaluation est composé de 60 CRH extraits aléatoirement de l'ensemble de données (5120 phrases, 10013 concepts ontologiques non uniques annotés). Deux personnes ont évalué l'annotation sémantique et leur contexte, notamment le repérage de la négation, l'hypothétique, la temporalité et la personne impliquée (p. ex. le patient vs un membre de sa famille). Dans les résultats, pour les tâches REN et ER, nous indiquons les scores de précision, de rappel et de F1.

Dans le processus de normalisation des entités, où il est possible d'attribuer à chaque entité détectée plusieurs URIs candidats issus de l'ontologie, l'utilisation de métriques fondées sur le classement s'avère essentielle pour évaluer avec précision les correspondances établies. Ainsi, nous mettons en œuvre des mesures largement utilisées dans ce domaine : Hits@1, Hits@5, ainsi que le rang réciproque moyen MRR. Hits@1 et Hits@5 évaluent le rappel en mesurant la présence des correspondances correctes parmi les 1 et 5 premiers résultats proposés par notre système de normalisation. Le MRR, quant à lui, offre une perspective quant à la qualité du classement en calculant la moyenne des inverses des positions attribuées aux correspondances correctes. Hits@1, en particulier, permet de déterminer dans quelle mesure l'URI le mieux classé par notre système coïncide avec une correspondance vérifiée. Le MRR complète cette analyse en appréciant de manière globale l'exactitude du classement des URIs candidats, grâce à l'agrégation des positions relatives des correspondances avérées.

4.2 Résultat

Nous avons évalué l'accord inter-annotateurs à travers les contributions des deux annotateurs sur l'ensemble des tâches. Cet accord s'est avéré être de 0,79, indiquant une cohérence significative dans les annotations fournies. En conséquence, nous avons procédé à la consolidation de tous les comptes rendus annotés. La REN et l'extraction du contexte des concepts ont démontré une précision globale de 0.9610, un rappel de 0.9248 et un score F1 de 0.9425. Les résultats détaillés sont présentés dans le Tableau 1.

Dans le cadre de l'évaluation de la ER, nous avons distingué 14 types de relations différents. Ces derniers ont été synthétisés dans le Tableau 2, représentant un ensemble de 3473 annotations. Les performances globales atteintes pour cette tâche sont résumées par une précision de 0.92, un rappel de 0.81, et un score F1 de 0.86.

Dans le cadre de notre analyse des performances de normalisation des entités, les résultats obtenus pour les métriques clés sont particulièrement révélateurs. Pour Hits@1, nous avons atteint un taux de 84.8%. Cette performance souligne l'efficacité du système à déterminer l'URI le plus pertinent pour chaque entité dès la première proposition.

TABLE 1 – Résultats quantitatifs de l'évaluation de la reconnaissance d'entités nommées.

Entité nommée		Total	Precision	Recall	F1
Age		164	0.977	0.904	0.939
Substance	Name	1034	0.8200	0.9805	0.8822
	Dosage	608	0.99	0.94	0.97
	DrugForm	14	0.857	0.857	0.857
Temporal Inf.	Date	1208	0.9942	0.9709	0.9824
	Duration	221	0.7481	0.9619	0.8417
	Frequency	511	0.9954	0.7688	0.9819
	Time	88	0.8750	0.9459	0.9091
EpisodeDeSoin		400	0.975	0.9485	0.8273
EvenementVecu		343	0.9589	0.9333	0.9459
ExamenClinique		308	0.9799	0.8811	0.8875
Hospitalisation		520	0.9954	0.9688	0.9819
Individu		540	0.991	0.7774	0.8747
Maladie		1166	0.9894	0.9852	0.9843
PartieDuCorps		22	0.98	0.6667	0.8000
Qualifier		600	0.9882	0.8802	0.9342
SigneClinique		1250	0.9870	0.9775	0.9823
AnnotationsToAdd		721	0.9256	0.8077	0.8626

TABLE 2 – Résultats quantitatifs des évaluations de l'extraction de relations.

Relation	Total	Precision	Rappel	F1
Relations Temporelles	860	0.9130	0.8077	0.8571
A pour motif	1050	0.9750	0.9070	0.9398
Participe	286	0.9831	0.5800	0.7296
Qualifie	756	0.9211	0.7368	0.8188
Relations Medicaments dosage	521	0.9046	0.8333	0.8678

En élargissant notre évaluation aux cinq premières propositions avec Hits@5, le taux s'améliore pour atteindre 90.4%, démontrant ainsi la capacité du système à inclure la correspondance exacte parmi les choix les plus privilégiés. Cette métrique confirme que même si la correspondance idéale n'est pas toujours première, elle figure presque toujours parmi les premières propositions.

Quant au MRR, qui offre une vue d'ensemble de la performance du système en prenant en compte le rang de la bonne réponse, le score obtenu est de 85%.

5 Discussion

Les résultats témoignent des performances obtenues au sein de notre étude. Notre démarche méthodique, adaptée tant à l'analyse des entités qu'à celle des relations, a été fructueuse et a mis en lumière les défis spécifiques liés au traitement de textes complexes, en particulier ceux du domaine de la psychiatrie.

Ces résultats encourageants s'expliquent principalement par deux facteurs. Premièrement, la richesse du vocabulaire de l'ontologie, notamment dans les domaines des signes, symptômes psychiatriques, des maladies, et des événements vécus, ce qui contribue directement à la qualité du plongements ontologique ainsi qu'à celle de la normalisation. En outre, nous avons réalisé des expérimentations sur la qualité du plongement générés par OWL2Vec*, qui ont

révélé notre aptitude à distinguer efficacement les classes de premier niveau de l'ontologie OntoPSY, cette analyse est disponible sur un notebook Jupyter GitHub [8].

Le second facteur déterminant est l'intégration de la structuration spécifique des CRH dans le processus d'extraction des relations, notamment à travers l'ajout de règles de filtrage. Cette adaptation améliore considérablement la précision de notre système, bien que cela puisse représenter un défi pour la généralisation de cette approche à d'autres types de documents.

De plus, il est essentiel de souligner que la performance globale de notre système est étroitement liée à l'efficacité du modèle d'extraction des syntagmes. Les imperfections inhérentes à ce processus ne se limitent pas à leur occurrence initiale ; elles se propagent à travers le système, impactant chaque étape subséquente de l'analyse. Cette interdépendance souligne la nécessité d'une extraction précise des syntagmes dès les premiers stades, étant donné que toute erreur générée peut être amplifiée et influencer l'ensemble des résultats obtenus. Cette limitation nécessite une étude d'ablation pour comprendre l'impact de chaque étape sur les résultats finaux du système d'annotation. En outre, l'analyse de dépendance influence également la précision et le rappel des tâches d'extraction de relations. Cette interdépendance met en exergue l'importance vitale d'une extraction précise des groupes nominaux dès les premiers instants, puisque les erreurs initiales peuvent être exacerbées, affectant de manière significative la qualité totale des résultats. Face à cette contrainte, il s'avère indispensable de recourir à une méthode d'ablation pour identifier avec exactitude l'impact de chaque élément sur la performance globale du système d'annotation.

L'utilisation de méthodes d'apprentissage non supervisé, intégrant une ontologie spécifique au domaine pour affiner la précision de l'apprentissage, pourrait diminuer le besoin d'annotations manuelles. La performance globale de l'annotation représente la somme des performances des différents composants. Suite à plusieurs améliorations apportées à chaque élément, comme l'intégration de règles spécifiques ou l'emploi de plongements pré-entraînés sur un corpus médical, nous avons atteint un niveau de performance jugé satisfaisant. Néanmoins, des opportunités d'amélioration de la performance de chaque composant du système proposé subsistent et feront l'objet de recherches approfondies dans nos travaux futurs.

6 Conclusion

L'objectif principal de ce travail est de reconstruire les données structurées des patients à partir de leurs CRH afin d'enrichir les données du projet PsyCARE. À cette fin, nous avons combiné l'utilisation de méthodes d'apprentissage non supervisées avec OntoPSY, une ontologie spécifique à la psychiatrie, pour récupérer et normaliser les entités biomédicales et identifier les relations entre ces paires d'entités dans le texte.

Initialement, pour l'extraction d'information, nous avons adapté l'algorithme PatternRank d'extraction de syntagmes clés. Nous avons ensuite exploité le plongement de l'ontologie dans un espace vectoriel avec OWL2Vec* pour associer ces informations aux concepts correspondants de l'ontologie. Enfin, en nous appuyant sur la structure de l'ontologie et l'analyse des dépendances syntaxiques, nous avons pu extraire les relations entre les entités.

Ce travail se distingue par l'exploitation des technologies du web sémantique combinées à l'apprentissage profond pour créer automatiquement des documents annotés dans le domaine de psychiatrie. Les performances de notre système sont prometteuses et ouvrent la voie à de nombreuses améliorations en termes de performances. Cette initiative a mis en lumière l'apport des plongements d'ontologie dans le contexte d'ontologies biomédicales variées et interconnectées, renforçant l'efficacité de l'annotation sémantique. Bien que cet article se concentre sur le domaine de la psychiatrie, des tests préliminaires dans le champ de la néphrologie avec une ontologie dédiée ont également révélé des perspectives encourageantes, bien que ces dernières ne soient pas l'objet principal de cette publication.

Les prochaines étapes de notre recherche incluent une analyse comparative entre notre méthode utilisant des plongements de mots non contextuels (word2vec) et les plongements sémantiques contextuels [7] pour l'annotation sémantique. Nous visons à améliorer le taux de rappel dans l'extraction de relations, en envisageant l'utilisation de notre base de données annotées et l'application d'apprentissage faiblement supervisé. Nous prévoyons également de tester l'efficacité de notre approche avec des données annotées en français disponibles publiquement.

8. https://github.com/AouinaOns/Semantic-Annotation

Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du Programme d'Investissements d'Avenir portant la référence PsyCARE ANR-18-RHUS- 0014.

Références

[1] Ons Aouina, Jacques Hilbey, and Jean Charlet. Ontology-Based Semantic Annotation of French Psychiatric Clinical Documents. *Studies in health technology and informatics*, 302 :793–797, May 2023.

[2] Dennis Aumiller et al. Online dateing : A web interface for temporal annotations. 07 2022.

[3] Antoine Bordes et al. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[4] Borui Cai et al. Temporal knowledge graph completion : A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI-2023. International Joint Conferences on Artificial Intelligence Organization, August 2023.

[5] Ricardo Campos et al. YAKE ! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509, 2020.

[6] Jiaoyan Chen et al. OWL2Vec* : Embedding of OWL Ontologies, January 2021. arXiv :2009.14654 [cs].

[7] Jiaoyan Chen et al. Contextual Semantic Embeddings for Ontology Subsumption Prediction, March 2023.

[8] Shivani Choudhary, Tarun Luthra, Ashima Mittal, and Rajat Singh. A Survey of Knowledge Graph Embedding and Their Applications, July 2021.

[9] Pratik Devkota et al. Using ontology embeddings with deep learning architectures to improve prediction of ontology concepts from literature. 2023.

[10] Jacob Devlin et al. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv :1810.04805 [cs].

[11] Aryo Pradipta Gema et al. Knowledge graph embeddings in the biomedical domain : Are they useful ? a look at link prediction, rule learning, and downstream polypharmacy tasks, 2023.

[12] Maarten Grootendorst. Keybert : Minimal keyword extraction with bert., 2020.

[13] Jacques Hilbey, Xavier Aimé, and Jean Charlet. *Temporal Medical Knowledge Representation Using Ontologies*. May 2022.

[14] Ali Hur et al. A Survey on State-of-the-art Techniques for Knowledge Graphs Construction and Challenges ahead, December 2021.

[15] Lars Juhl Jensen et al. Literature mining for the biologist : from information retrieval to biological discovery. *Nature Reviews Genetics*, 7(2), 2006.

[16] İlknur Karadeniz et al. Linking entities through an ontology using word embeddings and syntactic reranking | BMC Bioinformatics 2019 | Full Text.

[17] Dongkwan Kim et al. Supervised Graph Attention Network for Semi-Supervised Node Classification. 2019.

[18] Maxat Kulmanov et al. El embeddings : Geometric construction of models for the description logic el ++.

[19] Xuexiang Li et al. Efficient Medical Knowledge Graph Embedding : Leveraging Adaptive Hierarchical Transformers and Model Compression. 12, 2023.

[20] Louis Martin et al. Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, volume abs/1911.03894, 2019.

[21] Rada Mihalcea et al. TextRank : Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, September 2013.

[23] Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. A Novel Use of Statistical Parsing to Extract Information from Text. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000.

[24] Jose Moreno et al. Combining word and entity embeddings for entity linking. 2017.

[25] Dietrich Rebholz-Schuhmann et al. Text processing through Web services. *Bioinformatics*, 2008.

[26] Tim Schopf et al. PatternRank : Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction. In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 243–248, 2022.

[27] N. Shadbolt, T. Berners-Lee, and W. Hall. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3) :96–101, January 2006.

[28] Wei Shen et al. Entity linking with a knowledge base : Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27 :443–460, 2015.

[29] Fatima Zohra Smaili et al. Onto2Vec : joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics*, 2018.

[30] Fatima Zohra Smaili et al. OPA2Vec : combining formal and informal content of biomedical ontologies to improve similarity-based prediction. 35, 11 2018.

[31] Chengyu Sun et al. A Review of Unsupervised Keyphrase Extraction Methods Using Within-Collection Resources. *Symmetry*, 12(11) :1864, November 2020.

[32] Si Sun, Zhenghao Liu, Chenyan Xiong, et al. Capturing Global Informativeness in Open Domain Keyphrase Extraction, September 2021.

[33] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE : Knowledge Graph Embedding by Relational Rotation in Complex Space, February 2019.

[34] The OBI Consortium, Barry Smith, et al. The OBO Foundry : coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11) :1251–1255, November 2007.

[35] Kristina Toutanova et al. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, April 2015.

[36] Perceval Wajsburt et al. Eds-nlp : efficient information extraction from french clinical notes.

[37] Xiaojun Wan et al. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08. AAAI Press, 2008.

[38] Rui Xing et al. BioRel : towards large-scale biomedical relation extraction. *BMC Bioinformatics 2020*.

[39] Bishan Yang et al. Embedding entities and relations for learning and inference in knowledge bases, 2015.