Fusion d'ontologies biomédicales par des modèles siamois et validation par modèles de langue

S Menad, S Abdeddaïm, L F Soualmia

Fusion d'ontologies biomédicales par des modèles siamois et validation par modèles de langue

S. Menad[1], S. Abdeddaïm[1], LF. Soualmia[1]

[1] Univ. Rouen Normandie, Normandie Univ, LITIS UR 4108, FR-76000 Rouen, France

{safaa.menad1,said.abdeddaim,soualfat}@univ-rouen.fr

Résumé

Dans cette étude, nous proposons de nous appuyer sur des modèles siamois afin d'intégrer dans une même ressource sémantique les ontologies les plus pertinentes dans le domaine de la santé. Un premier axe concerne les maladies, symptômes, médicaments et événements indésirables. Nos modèles neuronaux siamois sont entraînés sur des données biomédicales et génèrent de nouvelles relations sémantiques entre concepts. Nous avons exploité des ressources du domaine et un grand modèle de langue comme moyen de validation de ces nouvelles relations. Les résultats obtenus permettent d'envisager des expérimentations à plus large échelle avec d'autres ontologies du domaine.

Mots-clés

Fusion d'ontologies, Ontologies Biomédicales, Grands Modèles de Langue, Modèles Neuronaux Siamois.

Abstract

In this study, we propose to use Siamese models to integrate the most relevant ontologies in the biomedical field into a single semantic resource. The first focus is on diseases, symptoms, drugs, and adverse events. Our Siamese neural models are pre-trained on biomedical data and allow to generate new semantic relations between concepts. Domain knowledge resources and a large language model are used as a means of validating these new relationships. The results obtained allow us to envisage larger-scale experimentation with other ontologies of the domain.

Keywords

Ontology Merging, Biomedical Ontologies, Large Language Models, Siamese Neural Models.

1 Introduction

Les ontologies jouent un rôle essentiel dans la représentation, l'organisation et la compréhension des connaissances, notamment dans le domaine biomédical. Elles sont utilisées dans un grand nombre de tâches telles que la recherche d'informations, la normalisation et l'intégration de données hétérogènes. À mesure que le volume des données biomédicales augmente, les exploiter efficacement et les analyser à des fins de recherche devient de plus en plus difficile. Malgré la disponibilité d'ontologies biomédicales, elles peinent souvent à couvrir tous les concepts et relations pertinentes. Afin de combler le manque de ressource unifiée, nous proposons une méthode d'intégration d'ontologies biomédicales en utilisant l'approche sémantique SiMHOMer (Siamese Models for Health Ontologies Merging) que nous avons développée [20]. Un premier focus de l'étude s'articule autour des maladies, symptômes, médicaments et événements indésirables. La méthode s'appuie sur des modèles neuronaux siamois que nous avons spécifiquement entraînés sur des données biomédicales. L'objectif est d'identifier des relations significatives entre différents concepts et d'établir de nouvelles relations sémantiques, permettant d'obtenir une nouvelle ressource sémantique. Afin de vérifier la validité des relations générées par SiMHOMer, nous nous appuyons sur des ressources existantes, comme les relations issues du Metathesaurus de l'UMLS (Unified Medical Language System [1]) et son Semantic Network. Afin de compléter cette validation pour des relations qui n'existeraient pas dans l'UMLS, nous proposons d'exploiter un grand modèle de langue.

Les principales contributions de cette étude peuvent être résumées comme suit : i) Nous décrivons le modèle neuronal siamois que nous avons proposé dans [21] qui a montré sa performance sur d'autres tâches par rapport à d'autres modèles.. Ce modèle est entraîné sur des données biomédicales et permet de détecter les similarités sémantiques entre les concepts. Nous avions utilisé notre modèle pour intégrer l'ontologie des maladies et l'ontologie des médicaments dans une première étude [20] afin de permettre la proposition d'un médicament potentiel pour une maladie donnée ; ii) Nous développons notre approche pour générer de nouvelles relations entre d'autres ontologies (maladies et symptômes) et sources de données (OpenFDA pour les effets indésirables liés aux médicaments), iii) Enfin, nous décrivons comment nous validons les relations proposées, d'abord par l'utilisation du Metathesaurus de l'UMLS et son Semantic Network, puis par un grand modèle de langue (LLM).

2 État de l'Art

Plusieurs méthodes ont été proposées afin d'enrichir et d'intégrer des ontologies dans une même ressource. Les

1. https://www.nlm.nih.gov/research/umls/index.html

approches consistent à identifier des potentielles relations entre concepts, et vont de méthodes classiques (distance entre chaînes de caractères) [1, 6, 11] à des méthodes plus sophistiquées reposant sur l'apprentissage automatique. [4] exploitent les capacités des transformeurs pour la résolution de la tâche de correspondance d'entités, démontrant une amélioration significative par rapport aux approches classiques en apprentissage profond. Dans une démarche similaire, le système de mise en correspondance d'entités DITTO [15] propose une architecture complète, incluant des techniques de blocage et d'augmentation de données, s'appuyant sur des modèles basés sur les transformeurs. Les applications des transformeurs pour la tâche de fusion d'ontologies sont moins fréquemment utilisées que pour la tâche de mise en correspondance d'entités. [13] ont montré qu'en rajoutant un composant transformeurs dans le framework MELT [12] pour aligner deux ontologies permet d'obtenir de meilleurs résultats.

Avec l'avancée continue des techniques d'apprentissage automatique, particulièrement dans le domaine du traitement du langage naturel (TALN), les LLMs ont aussi émergé comme des alternatives majeures. Ces modèles sophistiqués, construits sur des architectures d'apprentissage profond et entraînés sur de vastes corpus de données textuelles, ont révolutionné divers domaines d'étude. Par exemple, dans certaines études des réseaux neuronaux permettent de compléter les graphes de connaissances [5]. Les LLMs conçus à des fins générales, tels que BERT [7] et GPT [24], ont fait l'objet de recherches approfondies en raison de leur efficacité dans diverses tâches liées au langage. Dans le domaine biomédical, ils ont été appliqués pour aligner des concepts sources locaux avec des terminologies cliniques standard, telles que SNOMED-CT [16] et LOINC [28]. Cependant, ces travaux étaient limités à la relation de subsomption (is-a). Dans une recherche antérieure, nous avions proposé d'utiliser nos modèles neuronaux siamois pour fusionner l'ontologie des maladies et l'ontologie des médicaments [20, 21]. Dans ce travail, nous proposons d'étendre notre approche sur d'autres ontologies du domaine de santé.

3 Approche Proposée

3.1 Modèles Siamois

Les transformeurs représentent une architecture révolutionnaire en apprentissage profond, particulièrement éminente en traitement du langage naturel. Contrairement aux modèles séquentiels traditionnels, tels que les réseaux neuronaux récurrents (RNN), les transformeurs s'appuient sur des mécanismes d'auto-attention, ce qui leur permet de capturer efficacement les dépendances globales dans les séquences d'entrée. Ce mécanisme permet à chaque mot dans une séquence d'observer tous les autres mots, permettant une meilleure compréhension contextuelle sans être limité par un traitement séquentiel. Les transformeurs ont considérablement avancé diverses tâches de NLP, y compris la traduction automatique, la génération de texte, etc.

Les transformeurs traditionnels utilisent généralement une architecture de type cross-encoder, nécessitant la combinaison de deux phrases en une seule entrée pour prédire la variable cible. Cependant, cette approche devient peu pratique lorsqu'il s'agit de traiter de nombreuses comparaisons par paires.

Les Sentence-transformeurs [26] comblent cette limitation en introduisant une approche qui génère des plongements (embeddings) pour les phrases d'entrée. Ces embeddings encapsulent des informations sémantiques sur les phrases, garantissant que deux textes ayant des significations similaires sont positionnés à proximité dans l'espace d'embedding. La méthode implique d'entraîner simultanément deux modèles de transformeurs, en utilisant une architecture de réseau siamois, permettant l'extraction de représentations de phrases significatives favorables à l'évaluation de similarité sémantique, et facilitant des tâches de NLP. Pour chaque entrée, le modèle produit un vecteur de taille fixe (u et v). La fonction objectif est choisie de telle sorte que l'angle entre les deux vecteurs u et v soit plus petit lorsque les entrées sont similaires. La fonction objectif utilise le cosinus de l'angle :

$$cos(u, v) = \frac{u.v}{||u||||v||} \quad (1)$$

Si $cos(u, v) = 1$, les phrases sont similaires et si $cos(u, v) = 0$, les phrases n'ont aucun lien sémantique.

3.2 Modèles Proposés

Les transformeurs siamois fonctionnent bien dans le domaine général, mais pas dans les domaines de spécialité, comme le domaine biomédical. Nous avons donc besoin de modèles entraînés sur des données biomédicales.

Nous avons proposé un nouveau modèle siamois BioSTransformers [21] pré-entraîné sur le corpus d'articles scientifiques en anglais PubMed. Les transformeurs siamois ont été initialement conçus pour transformer des phrases (de taille similaire) en vecteurs. Dans notre approche, nous proposons de transformer les termes du thésaurus MeSH (Medical Subject Headings), les titres et les résumés des articles PubMed dans le même espace vectoriel en entraînant un modèle de transformeur siamois sur ces données. Nous voulons assurer un espace de correspondance entre le texte court et le texte long dans ce même vecteur. Le modèle est entraîné avec des paires d'entrées (titre, terme MeSH) et (résumé, terme MeSH). Sur la base de ces données, nous avons construit notre modèle basé sur un transformeur pré-entraîné sur des données biomédicales.

Dans cette étude, nous utilisons le modèle spécifique entraîné basé sur le Bio_ClinicalBERT [2] pour la tâche de fusion d'ontologies. La construction de BioSTransformers s'est inspirée du modèle Sentence-BERT [26] en remplaçant BERT par d'autres transformeurs. Nous avons utilisé des transformeurs qui ont été entraînés sur des données biomédicales (bio-transformers) pour créer des transformeurs siamois en ajoutant une couche de pooling et en modifiant la fonction objectif. La couche de pooling calcule le vecteur moyen des vecteurs de sortie du transformeur (embeddings de tokens). Les deux textes d'entrée passent successivement à travers le transformeur, produisant deux vecteurs u et v à

la sortie de la couche de pooling, qui sont ensuite utilisés par la fonction objectif.

Dans un Sentence-transformeur, les données supervisées sont représentées par des triplés (*phrase 1, phrase 2, score de similarité*) où le *score de similarité* est calculé entre les deux phrases *phrase 1* et *phrase 2*. Dans notre cas, comme il n'y a pas de score, ni pour les résumés, ni pour les titres et les termes MeSH correspondants, nous avons considéré :

— qu'un résumé, un titre et les termes MeSH associés au même article (identifié par un PMID) sont similaires, et le score est égal à 1 ;

— qu'un résumé (ou un titre) avec des termes MeSH non associés au même article ne sont pas similaires, et le score est égal à 0.

Nous avons utilisé une fonction objectif d'apprentissage contrastif auto-supervisé basée sur la fonction de perte Multiple Negative Ranking Loss (MNRL) dans le paquetage Sentence-Transformers [2]. La fonction MNRL nécessite des couples positifs en entrée (le titre ou le résumé et un terme MeSH associés au même article dans notre cas). Pour un couple positif (titre_i ou résumé_i, MeSH_i) la fonction MNRL considère que chaque couple (titre_i or résumé_i, MeSH_j) avec $i \neq j$ dans le même batch est négatif. Comme un article peut être associé à plusieurs termes MeSH, nous nous assurons que dans le batch de génération qu'un titre (ou résumé) associé à un terme MeSH dans un article PubMed ne soit jamais considéré comme négatif.

4 Fusion d'ontologies

4.1 Éléments Clés

Nous nous sommes inspirés des définitions [23, 8, 22] et les avons adaptées ci-dessous pour notre contexte de fusion d'ontologies biomédicales.

Définition d'ontologie : une ontologie O_i est un ensemble de vocabulaire défini au moyen de taxonomies pour décrire un domaine d'intérêt donné. Ce vocabulaire est considéré comme un ensemble d'éléments $e_i =< C_i, R_i, I_i >$; avec C_i étant l'ensemble des classes, R_i agrégeant les relations pour relier les classes, et I_i rassemblant l'ensemble des instances pour interpréter les classes et les relier avec R_i. Une ontologie O_i est également enrichie sémantiquement avec X_i pour définir des axiomes qui formalisent les classes en utilisant des langages logiques tels que les logiques de description.

Alignement d'ontologies : un alignement décrit l'ensembles des correspondances entre deux ontologies. Formellement, étant donné deux ontologies O_1 et O_2, nous limitons la définition d'un alignement A à un ensemble de triplés. Chaque triplé est spécifié par la terminologie de la relation binaire $r(e_1, e_2)$; où r représente la relation entre les deux éléments $e_1 \in O_1$ et $e_2 \in O_2$. En conséquence, l'alignement est le processus de recherche de ces ensembles de correspondances. Un score de confiance c peut également être ajouté au triplé de correspondance pour mesurer la similarité entre e_1 et e_2 (par exemple, la valeur de $c \in$ [0,1]).

Processus d'alignement : il est défini comme une fonction d'alignement ayant plusieurs paramètres calculant la similarité entre les entités. $F_m(O_1, O_2, A_j, P_c, B)$ est une fonction d'alignement avec P_c comme paramètre qui contient la valeur de confiance de similarité et B l'ensemble des ressources externes utilisées pour identifier un possible alignement A_j entre l'élément e_1 et e_2.

Fusion d'ontologies : suivant le travail présenté dans [22], nous définissons la fusion d'ontologies comme l'enrichissement sémantique d'une ontologie cible O_1 en utilisant des éléments d'une ontologie source O_2. Le résultat obtenu est une nouvelle ontologie O_3 grâce à l'alignement $A =< r_j, e_{1,j}, e_{2,j}, c_j >$

4.2 Portée des travaux

Dans la section suivante, nous décrivons les ontologies de santé que nous avons utilisées pour cette étude.

4.2.1 Ontologie des maladies

L'ontologie des maladies humaines (DOID) [17] décrit les maladies et le vocabulaire médical grâce à l'alignement de plusieurs ressources externes. Elle a été initialement construite en utilisant la Classification Internationale des Maladies (CIM-9) comme vocabulaire fondamental. Les premières versions ont été largement réorganisées par processus, système affecté et causes (troubles génétiques, maladies infectieuses, troubles métaboliques). Les révisions ultérieures ont été améliorées avec la réorganisation de DOID basée sur les concepts des maladies de l'UMLS en conjonction avec les mappings de concepts de termes vers l'ontologie SNOMED-CT (Systematized Nomenclature of Medicine- Clinical terms) et la classification CIM-9 [17]. Ces mappings se basent sur les identifiants uniques de concepts (CUI) de l'UMLS de chaque terme de maladie. Son développement a été motivé par la nécessité de représenter les connaissances avec une richesse sémantique qui permet de lier des données biomédicales à des gènes et des maladies. DOID permet ainsi d'identifier, d'intégrer et de relier des concepts de maladies synonymes qui sont inclus dans le MeSH, la SNOMED-CT, OMIM (Online Mendelian Inheritance in Man qui relie maladies et gène), et la CIM-9. Les mappings de vocabulaire sont mis à jour deux fois par an à partir d'une extraction des CUI de termes du fichier de mapping de vocabulaire MRCONSO.RRF de l'ULMS. Elle contient 13 910 concepts.

4.2.2 Ontologie des médicaments

L'ontologie des médicaments (DRON) [10] est un dictionnaire d'entités moléculaires décrivant des composants chimiques. Les entités moléculaires en question sont soit des produits naturels, soit des produits synthétiques. En plus des entités moléculaires, ChEBI (Chemical Entities of Biological Interest) contient des groupes (parties des entités moléculaires) et des classes d'entités. Ce dictionnaire comprend donc une classification ontologique, dans laquelle les relations entre les entités moléculaires ou les classes d'entités et leurs parents et/ou enfants sont spécifiées. Elle

2. https://www.sbert.net/docs/package_reference/losses.html\#multiplenegativesrankingloss

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/DOID_0112374">
    <rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/DOID_0050557"/>
    <obo:IAO_0000115>A congenital muscular dystrophy characterized by muscular dystrophy resulting from defective
    glycosylation of dystroglycan.</obo:IAO_0000115>
    <oboInOwl:hasDbXref>ICD10CM:G71.2</oboInOwl:hasDbXref>
    <oboInOwl:hasDbXref>ORDO:370953</oboInOwl:hasDbXref>
    <oboInOwl:hasExactSynonym>CMD due to dystroglycanopathy</oboInOwl:hasExactSynonym>
    <oboInOwl:hasExactSynonym>MDDG</oboInOwl:hasExactSynonym>
    <oboInOwl:hasExactSynonym>congenital muscular dystrophy due to dystroglycanopathy</oboInOwl:hasExactSynonym>
    <oboInOwl:hasOBONamespace>disease_ontology</oboInOwl:hasOBONamespace>
    <oboInOwl:id>DOID:0112374</oboInOwl:id>
    <oboInOwl:inSubset rdf:resource="http://purl.obolibrary.org/obo/doid#DO_rare_slim"/>
    <rdfs:label>muscular dystrophy-dystroglycanopathy</rdfs:label>
</owl:Class>
```

FIGURE 1 – Exemple d'une classe de l'ontologie de maladies DOID.

contient 8 282 concepts.

4.2.3 Ontologie des symptômes

L'ontologie des symptômes [3] (SYMP) a été développée comme une ontologie normalisée pour les symptômes des maladies humaines, à l'École de médecine de l'université du Maryland, à l'Institut des sciences du génome. Elle contient des symptômes, avec leur définition, libellés et synonymes. Elle est composée de 1 013 concepts.

4.2.4 Événements indésirables dûs aux médicaments

À notre connaissance, il n'existe aucune ontologie intégrant à la fois les médicaments et leurs effets indésirables. Par conséquent, nous avons choisi d'utiliser la base de données OpenFDA pour combler cette lacune. L'OpenFDA [4] [14] est une initiative de la Food and Drug Administration (FDA) des États-Unis qui offre un accès public à des ensembles de données et à des API liées aux produits réglementés par la FDA. Elle vise à promouvoir la transparence des données, à faciliter la recherche et l'analyse, à surveiller la sécurité des produits et à encourager le développement d'applications. Elle propose notamment de nombreuses API : événements indésirables liés aux médicaments, événements indésirables liés aux dispositifs médicaux, étiquettes de médicaments, classifications de dispositifs, rappels de produits et rapports d'application de la loi. En interrogeant ces APIs, il est possible d'accéder aux informations sur les événements indésirables liés aux médicaments. Pour répondre à nos questions de recherche, nous avons utilisé l'API des événements indésirables liés aux médicaments à partir de `https://open.fda.gov/apis/drug/event/`. Nous avons par exemple utilisé les requêtes suivantes :

```
https://api.fda.gov/drug/event.json?
    search=patient.drug.activesubstance.
    activesubstancename:"Collagen"&limit
    =1

https://api.fda.gov/drug/event.json?
    search=patient.drug.openfda.
    brand_name:"concerta"&limit=1
```

3. `https://www.ebi.ac.uk/ols4/ontologies/symp?viewMode=list`
4. `https://open.fda.gov/`

La première requête permet de rechercher un médicament en utilisant le nom de sa substance active, tandis que la seconde utilise son nom de marque. Étant donné que l'ontologie des médicaments est composée de classes contenant des noms de médicaments et d'entités chimiques, nous avons ainsi pu récupérer des médicaments en utilisant leur substance active principale. Par exemple, le terme *Collagène* concerne le champ ciblé que nous avons cherché à récupérer, qui est la substance active, et *Concerta* désigne le nom spécifique du médicament que nous recherchons.

4.3 SiMHOMer

Notre étude vise à fusionner des éléments des ontologies DOID, DRON et SYMP et à les aligner avec les données des événements indésirables de l'OpenFDA. Le résultat de ces alignements constitue une nouvelle ressource sémantique enrichie dans laquelle (i) chaque maladie est associée à un médicament potentiel et à un symptôme potentiel, et (ii) chaque médicament est lié à un ou plusieurs événements indésirables. Cette contribution s'inscrit dans le cadre du projet Predibioontlo (Predicting Clinical Diagnoses with Biomedical Ontologies and Language Models), dont l'objectif est d'utiliser le deep learning et les bases de connaissances biomédicales pour développer un outil visant à aider les médecins dans la prédiction des diagnostics et la proposition de médicaments pour leurs patients. Les phases listées dans [22] ont été adoptées pour le processus d'alignement.

La Figure 2 illustre l'approche globale permettant de générer les nouvelles relations entre les différentes ontologies. Nous commençons avec les trois ontologies distinctes que nous voulons fusionner, ainsi que la base de données OpenFDA contenant des informations sur les événements indésirables. La première étape (prétraitement) consiste à extraire les données textuelles des ontologies, y compris les définitions, les étiquettes et les synonymes. Dans la deuxième étape (fusion), ces données sont ensuite transformées en plongements (embeddings) en utilisant les modèles siamois. Ces embeddings permettent de calculer les similarités entre les différents éléments. En parallèle, un extracteur d'événements indésirables de médicaments est utilisé pour récupérer les événements indésirables associés aux médicaments en entrée dans DRON. Dans la dernière étape du processus, des correspondances sont établies entre

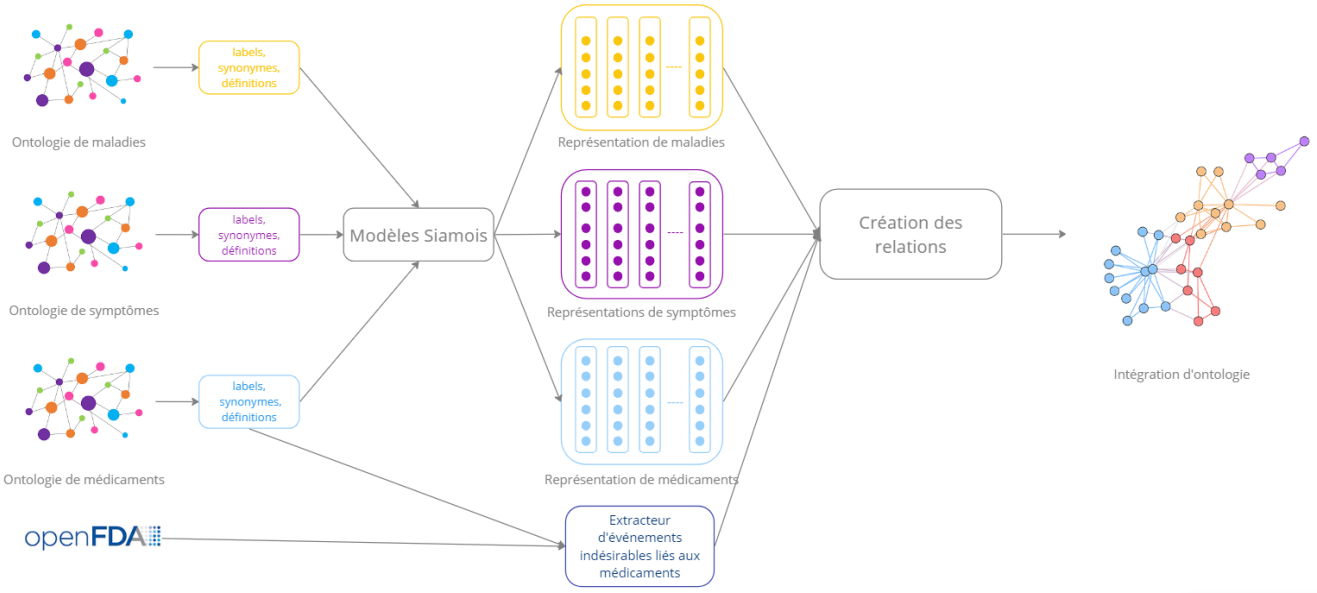

FIGURE 2 – Approche générale de la fusion des ontologies DRON, DOID, SYMP et alignement avec l'OpenFDA.

les relations suggérées par les modèles à travers les différentes ontologies et les données des événements indésirables de l'OpenFDA, créant ainsi une ontologie englobant ces diverses entités.

4.3.1 Prétraitement

Toutes les données textuelles ont été extraites des trois ontologies DOID, DRON et SYMP via la bibliothèque Owlready2 [5] [6]. Ces données sont liées :

- à la définition décrivant une maladie [7], les libellés et les synonymes ;
- aux métadonnées de ChEBI à partir desquelles l'ontologie DRON a été décrite. Ces métadonnées représentent des informations sur une maladie via une définition de propriété de données (métadonnées de ChEBI [8]), les libellés et les synonymes des médicaments ;
- à la définition décrivant un symptôme, son libellé et ses synonymes.

4.3.2 Fusion

Notre modèle est utilisé comme fonction de fusion, dans laquelle les bases de connaissances externes représentent les données sur lesquelles le modèle est entraîné : d'abord sur le corpus PubMed, puis sur la base de données clinique MIMIC III (une base de données contenant des dossiers patient électroniques).

Afin de trouver des relations entre maladies, symptômes et médicaments, nous avons procédé selon différentes manières. En effet, dans notre précédente étude [20], nous n'avions uniquement considéré que les noms des maladies de l'ontologie DOID ($rdfs : label$) et calculé les similarités entre ces éléments et les métadonnées de l'ontologie DRON ($obo : IAO_0000115$). Nous avons ensuite amélioré le processus en considérant deux approches différentes qui prennent en compte d'autres éléments constitutifs de DOID. En effet, le nom de la maladie n'étant pas suffisant, nous utilisons soit la concaténation des éléments de l'ontologie, soit nous ne considérons qu'un seul élément à la fois (celui représentant la similarité maximale).

Dans la concaténation de plusieurs éléments de l'ontologie DOID, ces éléments correspondent au nom de la maladie ($rdf : label$), à sa définition ($obo : IAO_0000115$) et à plusieurs noms de maladies synonymes ($oboInOwl : hasExactSynonym$). Nous appelons cette stratégie "multi-label". La concaténation est considérée comme une entrée pour notre modèle. Dans la seconde stratégie, nous ne considérons qu'un seul élément à la fois de DOID. Plus précisément, nous prenons en compte soit le nom de la maladie ($rdf : label$), soit la définition de la maladie ($obo : IAO_0000115$), soit un seul nom de maladie associé ($oboInOwl : hasExactSynonym$) dans chaque calcul de similarité. Ainsi, dans cette approche, pour chaque élément de DRON ou SYMP considéré par notre modèle, la correspondance est établie avec un élément de DOID, en choisissant le score de similarité maximum entre les métadonnées de DRON ou SYMP ($obo : IAO_0000115$), et l'une des métadonnées de DOID ($rdf : label$ ou $oboInOwl : hasExactSynonym$ ou $obo : IAO_0000115$).

4.3.3 Génération des Relations

Les alignements générés sont des correspondances entre un seul concept de DOID et un seul concept de DRON ou SYMP (alignement un à un). Une nouvelle relation est définie entre la maladie, le concept DRON et le concept SYMP. Cette nouvelle relation permet la génération d'une ontolo-

5. https://owlready2.readthedocs.io/en/v0.42/
6. https://github.com/arieme/OM_with_BioSTransformers
7. http://purl.obolibrary.org/obo/
8. http://purl.obolibrary.org/obo/$IAO_0000115$

gie plus complète (ontologie d'intégration) enrichie par les ontologies DRON, SYMP et DOID. Nous avons nommé ces nouvelles relations *has_drug* et *has_symptom*. Nous avons choisi un simple label qui décrit d'une manière très généraliste la relation. En raison des nuances et des questions très vaste qui existent dans le domaine de la médecine, par exemple : une médicament prescrit pour un certain type de patient qui a un certain âge ou sexe, etc.
Nous avons sélectionné la définition de la deuxième ontologie, SYMP, à des fins de comparaison, car elle contient un nombre significatif de synonymes nuls. Dans les cas où les définitions sont absentes, nous avons utilisé le libellé à la place.

4.3.4 Extraction des effets indésirables

Dans cette étape, nous avons extrait tous les effets indésirables liés à un médicament, soit en utilisant son nom de marque s'il est disponible, soit en utilisant sa substance active s'il n'existe pas dans la base de données. Ensuite, pour chaque maladie, nous avons créé la relation *has_side_effect*. Enfin, nous intégrons tout cela avec les relations *has_drug* et *has_symptom* dans un graphe que nous illustrons dans la figure 3.
La Figure 3 montre un exemple de relations générées entre les différentes ontologies. La maladie (en rouge) *Acquired angioedema* a comme symptôme (en marron) *anaphylaxis* et peut être traitée par le médicament (en vert) *icatibant* qui a plusieurs effets indésirables (en rose) comme *Laryngeal oedema* et *Stridor*.

5 Validation des Relations

Dans cette section, nous décrivons comment nous validons les alignements obtenus. À cette fin, nous nous appuyons sur l'utilisation de l'UMLS, le système de représentation des connaissances le plus utilisé dans le domaine biomédical. Nous utilisons également un grand modèle de langue.

5.1 UMLS

L'UMLS est une ressource exhaustive composée d'un Metathesaurus et d'un réseau sémantique (Semantic Network) développés par National Library of Medicine (NLM). Il sert de ressource unifiée pour les resources termino-ontologiques biomédicales, fournissant un moyen de relier divers vocabulaires et classifications biomédicales. Il comprend une vaste gamme de termes, concepts et relations provenant de sources diverses telles que la littérature médicale, les dossiers de santé électroniques, les terminologies cliniques et les ontologies.

5.1.1 Le Metathesaurus de l'UMLS

C'est le plus grand composant de l'UMLS. Il s'agit d'un vaste thésaurus biomédical organisé par concept, qui relie les noms similaires pour le même concept provenant de près de 200 vocabulaires différents. Le Metathesaurus identifie également les relations utiles entre les concepts et préserve les significations, les noms de concepts et les relations de chaque vocabulaire. Chaque terme dans le Metathesaurus est attribué un identifiant de concept unique (CUI), permettant l'interopérabilité et la liaison entre différents systèmes biomédicaux.
Nous avons identifié les relations issues du Metathesaurus les plus adaptées à notre cas d'étude, à savoir :

```
may_be_treated_by
treated_by
has_sign_or_symptom
sign_or_symptom_of
```

Parmi les relations que nous souhaitons valider, seuls certaines sont disponibles dans le Metathesaurus de l'UMLS. Par exemple, dans le cas de la relation *may_be_treated_by*, nous n'avons identifié que quelques relations correspondant aux maladies que nous recherchions. Les relations restantes concernant d'autres maladies ne sont pas incluses dans notre ontologie DOID ou qui existent dans notre ontologie mais pas dans UMLS. Cependant, ces 22 relations générées par notre approche sont incluses dans le Metathesarus, ce qui nous permet de les valider (Tableau 1).
En revanche, pour la relation *has_sign_or_symptom*, nous avons identifié plusieurs relations correspondant à celles que le modèle propose. Des exemples de ces relations trouvées dans le Metathesaurus sont présentés dans le Tableau 2. Il faut noter que le Metathesaurus ne fournit pas de relations spécifiques entre les maladies et les symptômes. Par exemple, pour *Irregular Heartbeat* (*Battement de cœur irrégulier*), il mentionne uniquement le symptôme *CIRC BLOOD*, qui est un terme trop générique. Les symptômes dans le Metathesaurus sont à un niveau élevé d'abstraction, même si nous sommes descendus au niveau maximum de l'arborescence au niveau des atomes, ce sont encore des termes généraux. Cependant, avec notre modèle, nous pouvons identifier le symptôme *Palpitation* qui a une relation avec *CIRC BLOOD* mais qui est plus spécifique.

5.1.2 Le Réseau Sémantique de l'UMLS

Le réseau sémantique (Semantique Network) se compose d'un ensemble de catégories, ou Types Sémantiques, qui fournissent une catégorisation cohérente de tous les concepts représentés dans le Metathesaurus de l'UMLS et d'un ensemble de relations sémantiques reliant les Types Sémantiques. Le réseau Sémantique contient 127 types sémantiques et 54 relations sémantiques. Pour valider nos relations, nous avons considéré les relations suivantes :

```
Pharmacologic Substance|treats|Sign or
    Symptom
Sign or Symptom|diagnoses|Pathologic
    Function
Virus|causes|Pathologic Function
Sign or Symptom|associated_with|Disease
     or Syndrome
Finding|associated_with|Disease or
    Syndrome
Finding|associated_with|Pathologic
    Function
```

Dans l'UMLS, les relations au sein du Metathesaurus sont idéalement alignées sur celles du Semantic Network. De

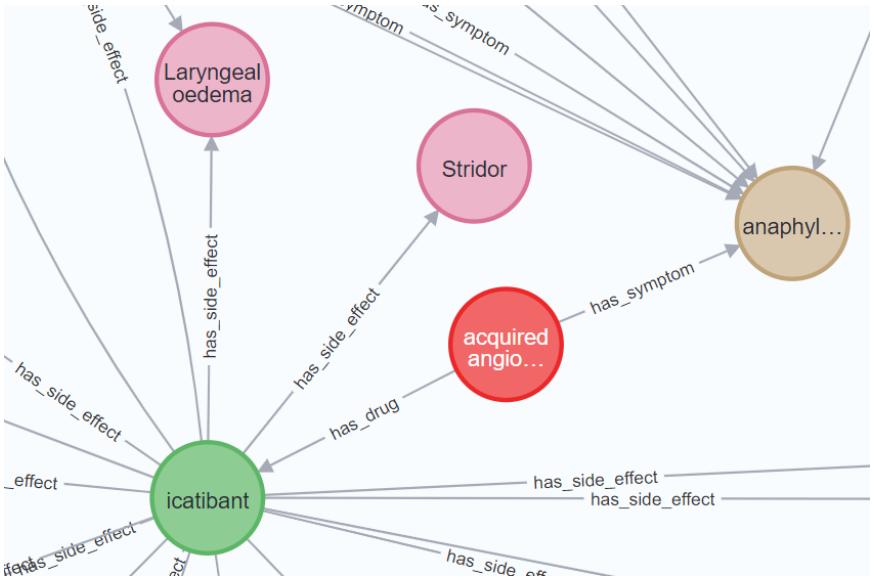

FIGURE 3 – Exemple de relations générées (has_side_effect ; has_drug ; has_symptom).

TABLE 1 – Exemples de la relation : *may_be_treated_by* proposées par notre modèle et retrouvées dans le Metathesaurus.

Maladie DOID	**CUI**	**Médicament DRON**	**CUI**
Obstructive sleep apnea	C0520679	Modafinil	C0066677
Enterocolitis, Pseudomembranous	C0014358	Fidaxomicin	C0065023
Pulmonary Fibrosis	C0034069	Pirfenidone	C0298067
Epilepsy	C0014544	Felbamate	C0060135
Ovarian neoplasm	C0919267	Niraparib	C2744440

nombreuses études ont été menées pour évaluer la pertinence de ces relations lorsqu'elles sont projetées vers le Metathesaurus [19, 29]. Dans notre étape de validation, nous supposons que des relations pourraient être déduites. En conséquence, le tableau 3 montre le nombre de relations trouvées dans le Réseau Sémantique qui correspondent à nos relations proposées, en fonction des relations sélectionnées. Une prochaine étape consistera à identifier toutes les autres relations potentielles pour une exploration plus approfondie.

5.2 Grands Modèles de Langage

Les progrès récents dans les grands modèles de langage (LLM) offrent de nouvelles opportunités pour le domaine biomédical. Ce sont des modèles puissants construits à l'aide de réseaux neuronaux contenant des milliards de paramètres. Ils sont entraînés sur de vastes volumes de texte généré par des humains [30] qui permettent ensuite de générer des textes, répondre à des questions, etc [9].

5.2.1 Le Modèle LlaMA

LlaMA [27] est un LLM publié par Meta en 2023. Plusieurs évaluations ont montré le potentiel de ces LLM dans le domaine biomédical [25]. Nous utilisons LlaMA 2 comme chatbot pour évaluer les relations que nous avons proposées. Bien que les LLM démontrent des capacités remarquables en zéro-shot, ils rencontrent des limitations lorsqu'ils sont confrontés à des tâches plus complexes. Pour y remédier, la sollicitation en quelques étapes peut faciliter l'apprentissage en contexte [3].

5.2.2 Les Prompts

Un prompt fait référence à l'entrée initiale fournie au modèle pour générer une sortie souhaitée. Il sert de repère ou d'instruction pour le modèle, le guidant pour produire du texte qui correspond à la tâche ou au contexte souhaité.

Diverses études ont examiné l'effet de l'ingénierie des prompts sur les performances des modèles de langage, ainsi que la variation des réponses générées en sortie pour une même tâche lors de l'utilisation de différents prompts. [18, 31] ont montré que la réponse générée par le modèle de langage présente une forte corrélation avec le prompt fourni en entrée.

Dans [9], les auteurs ont testé différents prompts pour évaluer les performances de ChatGPT dans sa réponse aux questions médicales et ont démontré que l'inclusion d'informations contextuelles supplémentaires a le potentiel d'influencer les réponses de ChatGPT.

Nous proposons d'utiliser différents prompts pour évaluer leur efficacité. Nous avons développé le prompt de manière itérative, en étudiant les résultats obtenus au fur et à mesure. Nous avons observé que le fait de fournir du contexte

TABLE 2 – Exemples de la relation : *has_sign_or_symptom* proposées par notre modèle et retrouvées dans le Metathesaurus.

Maladie DOID	**Symptôme SYMP**	**UMLS Disease**	**UMLS Symptom**
Obstructive sleep apnea	sleep apnea	Sleep apnea syndrome	Finding of sleep rest pattern
Esophageal varix	obsolete portal hepatitis	Varices	CIRC BLOOD
alcohol use disorder	concentration difficulty	Alcohol abuse	AOD use
spermatogenic failure 17	obsolete impotence	Infertility	Reproductive function
dilated cardiomyopathy 1X	muscle weakness	Muscle Weakness	Neuro-musculo-skeletal function
atrial fibrillation	palpitation	Irregular Heartbeat	CIRC BLOOD
Prinzmetal angina	tachycardia	Angina Pectoris	CIRC BLOOD
Perlman syndrome	renal involvement	Ascites	Digestion-hydration

TABLE 3 – Nombre de relations trouvées qui sont dans le Semantic Network.

Nom de relation	**Nombre**
Antibiotic | *treats* | Disease or Syndrome	191
Sign or Symptom | *associated_with* | Disease or Syndrome	907

Prompt 1 :

```
Please say if there is a symptom-disease association for me:
When confirming, please return clear answers like 'true' or 'false'. Please provide
    the result in JSON format, with the following structure:
{'symptom': 'symptom name',
'disease': 'disease name',
'answer': 'true or false'}.

Text:
the disease {disease} defined by:
{Definiton}
the symptom {symptom} defined by:
{Definiton}
```

FIGURE 4 – Contenu du Prompt 1.

au modèle lui permettait de générer une réponse pertinente pour notre scénario. Le contenu du prompt en relation avec nos types de données et de relations est présenté dans Figure 4 et 5 :

5.3 Résultats

Nous avons testé une liste de relations validées via l'utilisation du Metatesaurus ou du Semantic Network de l'UMLS pour vérifier si le LLM renvoit le même résultat ou pas. Certaines relations non disponibles dans l'UMLS sans source ont également été considérées comme vraies.

La Table 4 montre que, parmi toutes les relations que nous lui avons fournies, le LLM a validé avec succès les relations présentes dans le Metathesaurus et le Semantic Network de l'UMLS, ainsi que des relations supplémentaires absentes de l'UMLS. Il n'a remis en question aucune des relations proposées.

6 Conclusion

Dans cette étude, nous avons présenté une approche novatrice visant à consolider plusieurs ontologies pertinentes pour le domaine de la santé dans un cadre d'intégration d'ontologies unifiée. En exploitant nos modèles entraînés sur des données biomédicales, nous avons proposé des relations potentielles entre des concepts provenant d'ontologies disparates.

Pour valider les relations proposées, nous les avons croisées avec celles qui existent dans le Metathesaurus de l'UMLS et son Semantic Network. Cependant, nous n'avons pas pu valider toutes nos relations avec cette ressource, qui s'est avérée insuffisante. Nous avons également utilisé les capacités des grands modèles de langue (LLM) pour compléter ce processus de validation. Cette validation initiale confirme l'exactitude de nos relations proposées par le LLM.

Nos premiers résultats sont très prometteurs, même si nous n'avons qu'une partie des relations proposées qui

Prompt 2 :

```
Please say if there is a drug-disease association for me:
When confirming, please return clear answers like 'true' or 'false'. Please provide
    the result in JSON format, with the following structure:
{'drug': 'drug name',
'disease': 'disease name',
'answer': 'true or false'}.

Text:
the disease {disease} defined by:
{Definiton}
the symptom {drug} defined by:
{Definiton}
```

FIGURE 5 – Contenu du Prompt 2.

TABLE 4 – Relations testées avec le modèle LlaMA.

Maladie DOID	Symptôme ou Médicament	Relation	Source	Réponse LlaMA
angiosarcoma	spontaneous ecchymoses	_	_	true
Obstructive sleep apnea	sleep apnea	has_sign_or_symptom	Metathesaurus	true
obstructive sleep apnea	modafinil	may_be_treated_by	Metathesaurus	true
amodiaquine allergy	Exanthema	associated_with	Semantic Network	true
Zollinger-Ellison syndrome	Vomiting	associated_with	Semantic Network	true

est validée. Nos travaux futurs se concentreront sur un processus de validation plus approfondi, potentiellement en utilisant des ressources externes plus complètes telles que celles incluses dans le service de recherche d'ontologies (OLS) de l'Institut de bioinformatique européen de l'EMBL (`https://www.ebi.ac.uk/ols4`), ou encore par la sollicitation d'experts du domaine. Les nouvelles relations dérivées du cadre SiMHOMer offrent des avantages potentiels significatifs pour les spécialistes de la santé, promettant une intégration et une accessibilité des connaissances enrichies du domaine. Nous envisageons de comparer nos modèles, qui ont déjà démontré leur efficacité sur d'autres tâches, à d'autres modèles spécifiques à cette tâche (dans OAEI par exemple).

Références

[1] Ismail Akbari, Mohammad Fathian, and Kambiz Badie. An improved mlma+ and its application in ontology matching. In *2009 Innovative Technologies in Intelligent Systems and Industrial Applications*, pages 56–60. IEEE, 2009.

[2] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33 :1877–1901, 2020.

[4] Ursin Brunner and Kurt Stockinger. Entity matching with transformer architectures-a step forward in data integration. In *23rd International Conference on Extending Database Technology, Copenhagen, 30 March-2 April 2020*, pages 463–473. OpenProceedings, 2020.

[5] Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 904–915, 2022.

[6] William W Cohen, Pradeep Ravikumar, Stephen E Fienberg, et al. A comparison of string distance metrics for name-matching tasks. In *IIWeb*, volume 3, pages 73–78, 2003.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[8] Jérôme Euzenat, Pavel Shvaiko, et al. *Ontology matching*, volume 18. Springer, 2007.

[9] Zhenxiang Gao, Lingyao Li, Siyuan Ma, Qinyong Wang, Libby Hemphill, and Rong Xu. Examining the potential of chatgpt on biomedical information retrieval : Fact-checking drug-disease associations. *Annals of Biomedical Engineering*, pages 1–9, 2023.

[10] Josh Hanna, Eric Joseph, Mathias Brochhausen, and William Hogan. Building a drug ontology based on rxnorm and other sources. *Journal of biomedical semantics*, 4 :44, 12 2013.

[11] Wei He, Xiaoping Yang, and Dupei Huang. A hybrid approach for measuring semantic similarity between ontologies based on wordnet. In *International Conference on Knowledge Science, Engineering and Management*, pages 68–78. Springer, 2011.

[12] Sven Hertling, Jan Portisch, and Heiko Paulheim. Melt-matching evaluation toolkit. In *International conference on semantic systems*, pages 231–245. Springer International Publishing Cham, 2019.

[13] Sven Hertling, Jan Portisch, and Heiko Paulheim. Matching with transformers in melt. *arXiv preprint arXiv :2109.07401*, 2021.

[14] Taha A Kass-Hout, Zhiheng Xu, Matthew Mohebbi, Hans Nelsen, Adam Baker, Jonathan Levine, Elaine Johanson, and Roselie A Bright. Openfda : an innovative platform providing access to a wealth of fda's publicly available data. *Journal of the American Medical Informatics Association*, 23(3) :596–600, 2016.

[15] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, Jin Wang, Wataru Hirota, and Wang-Chiew Tan. Deep entity matching : Challenges and opportunities. *Journal of Data and Information Quality (JDIQ)*, 13(1) :1–17, 2021.

[16] Hao Liu, Yehoshua Perl, and James Geller. Concept placement using bert trained by transforming and summarizing biomedical ontology structure. *Journal of Biomedical Informatics*, 112 :103607, 2020.

[17] Schriml Lynn, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Kibbe. Disease ontology : A backbone for disease semantic integration. *Nucleic acids research*, 40 :D940–6, 11 2011.

[18] Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. Prompt engineering in large language models. In *International Conference on Data Intelligence and Cognitive Informatics*, pages 387–402. Springer, 2023.

[19] Alexa T McCray and Olivier Bodenreider. A conceptual framework for the biomedical domain. In *The semantics of relationships : an interdisciplinary perspective*, pages 181–198. Springer, 2002.

[20] Safaa Menad, Wissame Laddada, Saïd Abdeddaïm, and Lina F Soualmia. Biostransformers for biomedical ontologies alignment. In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2023, Volume 2 : KEOD*, pages 73–84, 2023.

[21] Safaa Menad, Wissame Laddada, Saïd Abdeddaïm, and Lina F Soualmia. New siamese neural networks for text classification and ontologies alignment. In *International Conference on Complex Computational Ecosystems*, pages 16–29. Springer, 2023.

[22] Inès Osman, Sadok Ben Yahia, and Gayo Diallo. Ontology integration : Approaches and challenging issues. *Information Fusion*, 71 :38–63, Jul 2021.

[23] Jan Portisch, Michael Hladik, and Heiko Paulheim. Background knowledge in ontology matching : A survey. *Semantic Web*, pages 1–55, 09 2022.

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8) :9, 2019.

[25] Arya Rao, Michael Pang, John Kim, Meghana Kamineni, Winston Lie, Anoop K Prasad, Adam Landman, Keith Dreyer, and Marc D Succi. Assessing the utility of chatgpt throughout the entire clinical workflow : development and usability study. *Journal of Medical Internet Research*, 25 :e48659, 2023.

[26] Nils Reimers and Iryna Gurevych. Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In *Proceedings of (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama : Open and efficient foundation language models, 2023.

[28] Tao Tu, Eric Loreaux, Emma Chesley, Adam D Lelkes, Paul Gamble, Mathias Bellaiche, Martin Seneviratne, and Ming-Jun Chen. Automated loinc standardization using pre-trained large language models. In *Machine Learning for Health*, pages 343–355. PMLR, 2022.

[29] Li Zhang, Michael Halper, Yehoshua Perl, James Geller, and James J Cimino. Relationship structures and semantic type assignments of the umls enriched semantic network. *Journal of the American Medical Informatics Association*, 12(6) :657–666, 2005.

[30] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv :2303.18223*, 2023.

[31] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers, 2023.