Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Annotation d'interactions hôte-microbiote dans des articles scientifiques par similarité sémantique avec une ontologie

Oumaima El Khettari[1] Solen Quiniou[1], Samuel Chaffron[1]
(1) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000
oumaima.el-khettari@ls2n.fr, solen.quiniou@ls2n.fr,
samuel.chaffron@ls2n.fr

RÉSUMÉ

Nous nous intéressons à l'extraction de relations, dans des articles scientifiques, portant sur le microbiome humain. Afin de construire un corpus annoté, nous avons évalué l'utilisation de l'ontologie OHMI pour détecter les relations présentes dans les phrases des articles scientifiques, en calculant la similarité sémantique entre les relations définies dans l'ontologie et les phrases des articles. Le modèle BERT et trois variantes biomédicales sont utilisés pour obtenir les représentations des relations et des phrases. Ces modèles sont comparés sur un corpus construit à partir d'articles scientifiques complets issus de la plateforme ISTEX, dont une petite sous-partie a été annotée manuellement.

ABSTRACT

Annotating host-microbiota interactions in scientific articles using semantic similarity to an ontology

We are interested in the task of relation extraction in scientific articles on the subdomain of the human microbiome. In order to build an annotated corpus, we investigated the use of the OHMI ontology to detect the relations appearing in scientific article sentences by computing the semantic similarity between the relations defined in the ontology and the sentences. The BERT model and three biomedical variations are used to compute the representations of both relations and sentences. These models are compared on a corpus built from full-text scientific articles from the ISTEX platform, with a small subpart being manually annotated.

MOTS-CLÉS : Extraction de relations, ontologie, similarité sémantique, BERT, ISTEX.

KEYWORDS: Relation extraction, Ontology, Semantic similarity, BERT, ISTEX.

1 Introduction

Ces dernières années, le nombre de publications dans le domaine biomédical a énormément augmenté. L'extraction d'information devient une tâche indispensable pour la veille scientifique et l'intégration des connaissances dans ce domaine. Nous nous intéressons plus particulièrement à l'extraction de relations, dans les articles scientifiques, pour l'étude du microbiome humain. Cette tâche est généralement considérée comme une tâche de classification (Zhou *et al.*, 2014) sur un corpus annoté. Néanmoins, il n'existe aucun corpus annoté pour la tâche d'extraction de relations sur le microbiome humain et ses interactions. Il existe cependant des ontologies, dans le domaine biomédical et dans certains de ses sous-domaines : c'est le cas pour le microbiome humain, pour lequel il existe l'ontologie OHMI : Ontology of Host-Microbiome Interactions (He *et al.*, 2019).

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 21–26.

Dans la suite, nous nous appuyons sur l'hypothèse distributionnelle (Harris, 1954), affirmant que la similarité des contextes dans lesquels apparaissent deux mots permet de mesurer leur proximité sémantique. Cela nous permet de calculer la similarité sémantique entre une phrase d'un article scientifique et les relations d'une ontologie, afin d'étudier la correspondance de ce score avec la présence ou l'absence d'une relation dans la phrase. Nous nous appuyons également sur l'utilisation de modèles adaptés de type BERT, pour obtenir les représentations sémantiques des relations de l'ontologie et des phrases des articles, ces modèles ayant permis d'obtenir d'importants progrès dans les domaines de spécialité (Pankaj & Gautam, 2022).

2 Ressources : ontologie, corpus et modèles

Ontologie des interactions hôte-microbiome (OHMI) Afin de définir les noms des interactions explicitement liées à l'étude du microbiome humain, nous nous appuyons sur l'ontologie des interactions hôte-microbiome OHMI (Ontology of Host-Microbiome) (He *et al.*, 2019). Celle-ci se définit comme une ontologie communautaire des interactions de l'être humain avec les éléments de son microbiome, le microbiote. OHMI contient un ensemble d'informations incluant le microbiote, avec une taxonomie microbienne des espèces hôtes, les entités anatomiques des hôtes, et les interactions hôte-microbiote dans différentes conditions. Dans la suite de notre étude, nous utilisons les 129 relations présentes dans l'ontologie. Ces relations sont plus ou moins longues et plus ou moins précises, comme l'illustrent les 2 exemples suivants : *'negatively regulated by'* et *'microbe susceptibly depleted in host with disease'*. Malgré le fait que certaines relations extraites aient des similitudes sémantiques et syntaxiques, elles ont un niveau de précision différent. Par conséquent, chaque relation est exprimée de manière unique. Pour illustrer cela, nous pouvons citer les deux relations *'causally upstream of'*, *'causally upstream of or within, negative effect'*. Lors de l'annotation, il est demandé d'annoter avec la relation la plus précise. Il est important de préciser que ce cas particulier des relations proches représente 19 relations, ce qui correspond à approximativement 15% de la liste des relations.

Corpus, pré-traitements et annotation La plateforme ISTEX (Cuxac & Thouvenin, 2017) permet d'accéder à plus de 25 millions de publications scientifiques. Nous l'utilisons pour construire un corpus de publications scientifiques, en anglais, portant sur le microbiome humain, à partir de la requête suivante : <"GUT MICROBIOTA" OR "GUT MICROBIOME" OR "INTESTINAL MICROBIOTA" OR "INTESTINAL MICROBIOME" AND LANGUAGE :ENG>. Nous obtenons, en résultat, un corpus de 8 657 publications scientifiques, dont les années de publication varient entre 2001 à 2019.

Le corpus obtenu est ensuite découpé en phrases, en utilisant la bibliothèque spaCy : on obtient ainsi 1 104 240 phrases. Afin de ne considérer que les phrases comportant potentiellement l'expression sémantique d'une relation sur le microbiome humain, nous partons de l'hypothèse que la présence d'une telle relation est positivement corrélée à la présence d'une ou plusieurs entités nommées liées au domaine biomédical. Nous faisons également l'hypothèse qu'une relation est contenue à l'intérieur d'une phrase, tout d'abord par simplicité, et aussi parce que c'est généralement le cas. Cela nous permet également de considérer les phrases, indépendamment les unes des autres, lors de l'annotation de la présence ou l'absence de relations, à l'intérieur de celles-ci.

La bibliothèque scispaCy (Neumann *et al.*, 2019) est utilisée pour identifier les entités nommées du domaine biomédical. Nous utilisons ainsi le modèle `en_ner_craft_md`, entraîné sur le corpus CRAFT (Bada *et al.*, 2012), et le modèle `en_ner_bc5cdr_md`, entraîné sur le corpus BC5CDR (Li

et al., 2016). Le premier modèle détecte les cellules (*Cell line*), les GGP (*Gene-or-Gene-Product*) et les taxons (*Taxon*) ainsi que les entités CHEBI présentes dans Chemical Entities of Biological Interest ontology (Degtyarenko *et al.*, 2007), les entités SO de Sequence Ontology (Eilbeck *et al.*, 2005) et les entités GO de Gene Ontology (Consortium, 2004). Quant au deuxième modèle, il détecte les noms de maladies (*Disease*) et les espèces chimiques (*Chemical*).

Afin d'étudier la distribution des scores de similarité sémantique, en fonction du nombre d'entités nommées biomédicales présentes dans les phrases, nous avons créé 4 sous-corpus, à partir de notre corpus initial de 1 104 240 phrases : le premier sous-corpus contient les 1 021 phrases avec une seule entité nommée, le deuxième contient les 762 phrases avec deux entités, le troisième contient les 454 phrases avec trois entités, et le dernier contient 618 phrases avec 4 entités nommées. Afin d'étudier l'utilisation du score de similarité sémantique pour la tâche d'extraction de relations, nous avons annoté une petite sous-partie de chacun des 4 sous-corpus : nous avons choisi aléatoirement 4 phrases, dans chaque sous-corpus, et les avons annotées avec la relation qui y était présente, en utilisant les 129 relations issues de l'ontologie OHMI. Cette tâche étant chronophage et le corpus actuellement construit étant de taille réduite, les premiers résultats obtenus sur celui-ci ne pourront donner que des indications sur l'utilisation du score de similarité sémantique pour la tâche d'extraction de relations.

Modèles de représentation du texte Compte tenu de la nature de la tâche de similarité sémantique, nous utilisons les modèles Sentence Transformers (Reimers & Gurevych, 2019) pour obtenir une représentation sémantique à la fois des relations extraites de l'ontologie et des phrases du corpus. Nous utilisons le modèle BERT (Devlin *et al.*, 2018) ainsi que trois variantes entraînées sur des données biomédicales, à savoir BioBERT (Lee *et al.*, 2019), SciBERT (Beltagy *et al.*, 2019) et PubMedBERT (Gu *et al.*, 2021). BioBERT est issu du pré-entraînement continu de BERT sur des résumés de PubMed. Quant à SciBERT, il est entièrement entraîné sur des articles complets de Semantic Scholar dont 82% s'inscrit dans le domaine biomédical. Enfin, PubMedBERT est entièrement entraîné sur des résumés de PubMed.

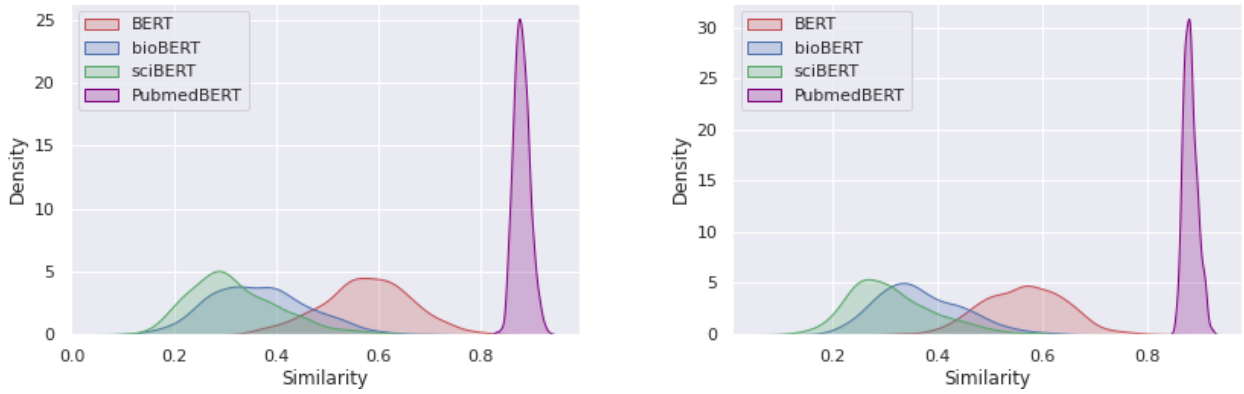
3 Méthodologie

Afin d'identifier si une relation est présente dans une phrase et la nature de cette relation, le cas échéant, la similarité cosinus est calculée entre la représentation sémantique de la phrase considérée (qui contient au moins une entité nommée) et la représentation sémantique de chacune des 129 relations issues de l'ontologie. La relation correspondant au score de similarité le plus élevé est ensuite attribuée à la phrase. Si la phrase ne contient aucune entité nommée du domaine biomédical, on considère qu'elle ne contient aucune relation de l'ontologie.

4 Expérimentations et résultats

Évaluation des scores de similarité sémantique Dans un premier temps, nous comparons la distribution des scores de similarité, sur chacun des 4 sous-corpus, selon la représentation sémantique des différents modèles considérés. La figure 1 montre la densité des scores de similarité sémantique, pour chaque modèle considéré, sur le sous-corpus contenant les phrases avec 1 entité nommée et sur celui contenant les phrases avec 4 entités nommées. En effet, les distribution des scores sont très

similaires, pour chaque modèle, sur chacun des 4 sous-corpus. Le nombre d'entités présentes dans une phrase affecte peu le score de similarité, ce qui induit que seule la présence des entités, sans prendre en considération leur nombre, peut être considérée comme un indicateur de la présence d'une relation dans une phrase. En effet, il est probable d'avoir des entités nommées dans une phrase sans qu'elles ne participent à la relation exprimée, comme pour la citation d'exemples dans une phrase. Le modèle PubMedBERT donne les scores les plus hauts, se situant entre 0,83 et 0,93. Ceci peut être expliqué par la sensibilité du modèle au vocabulaire de spécialité, puisqu'il est entièrement entraîné sur des textes du domaine biomédical. Les distributions de SciBERT et BioBERT restent assez proches, malgré les différences entre les deux modèles en termes de techniques de pré-entraînement et de corpus d'entraînement, et sont plus faibles que les scores obtenus avec le modèle BERT.



(a) Phrases avec 1 entité nommée

(b) Phrases avec 4 entités nommées

FIGURE 1 – Densité des scores de similarité cosinus par modèle et sur deux sous-corpus

Évaluation pour l'identification des relations Dans un second temps, nous donnons des premiers résultats obtenus sur un corpus annoté manuellement, afin de discuter de l'approche proposée pour identifier les relations en lien avec le microbiome humain. La table 1 présente les scores d'*accuracy*, pour chacun des 4 modèles, sur le petit sous-corpus annoté manuellement, en comparant la relation associée à la similarité cosinus la plus élevée avec la relation annotée manuellement.

Modèle	BERT	BioBERT	SciBERT	PubMedBERT
Accuracy	0,37	0,62	0,37	0,50

TABLE 1 – Accuracy des modèles sur les 16 phrases annotées manuellement

Sur ce petit corpus de 16 phrases, BioBERT obtient les meilleurs résultats. Alors que BioBERT et SciBERT obtenait des distributions de scores de similarité proches, SciBERT obtient des résultats nettement moins bons (similaires à ceux obtenus pour BERT, qui n'a pas été spécifiquement entraîné sur des données biomédicales). Le modèle PubMedBERT obtient des résultats entre ceux de BioBERT et ceux de SciBERT, dû a son entraînement exclusif sur des données biomédicales issues d'articles scientifiques. Le modèle BioBERT semble être celui qui allie le mieux l'encodage de l'information sémantique avec la connaissance du domaine de spécialité. Il convient de noter que les résultats des modèles sont souvent proches les uns des autres. Cela signifie que même si le modèle se trompe sur la relation attribuée à la phrase, la relation prédite reste cohérente avec ce qui est exprimé dans la phrase

et ne contredit pas la réalité. Ceci dit, dans certains cas, l'erreur peut être causée par la présence d'un terme de spécialité dans la phrase et dans la relation, ce qui pousse le modèle à les associer même si ce n'est pas la bonne relation. Nous en déduisons que l'amélioration de ces résultats consisterait à réduire le nombre élevé de relations extraites d'OHMI afin d'alléger le processus d'annotation et à opter pour des formulations plus simples pour améliorer la précision des résultats.

Compte tenu du grand nombre de relations à prendre en considération lors de l'annotation, à ce stade, un unique annotateur a effectué la tâche d'annotation, dans le but d'obtenir plus de visibilité sur la complexité de la tâche en l'état actuel. Quelques exemples sont fournis dans la table 2.

Phrases	**Annotations**
Moreover, airway microbiota composition and greater bacterial diversity were significantly correlated with bronchial hyperresponsiveness, including the relative abundance of specific microbiota belonging to bacterial families within the Proteobacteria.	microbe susceptibly expanded in respiratory airway of human with disease
Paradoxically, some degree of innate immune recognition of commensal bacteria is essential for normal development and function of the mucosal and peripheral immune system.	microbial population phenotype in host
Bronchoscopic studies indicate that the lungs of healthy people who smoke are inhabited by diverse types of bacteria in relatively small numbers and that this microbiome changes with disease.	microbe susceptibly depleted in respiratory airway of human with disease

TABLE 2 – Exemples de phrases avec leur relation manuellement annotée

5 Conclusion

Nous avons évalué l'utilisation de l'ontologie OHMI pour détecter les relations présentes dans les phrases des articles scientifiques, en calculant la similarité cosinus entre les relations définies dans l'ontologie et les phrases des articles. Le modèle BERT et trois variantes biomédicales ont été utilisés pour obtenir les représentations des relations et des phrases. Ces modèles ont été comparés sur un corpus construit à partir d'articles scientifiques complets issus de la plateforme ISTEX, dont une petite sous-partie a été annotée manuellement.

Cette étude a permis d'observer que le modèle BioBERT obtenait les meilleurs résultats et semblait être le plus adapté aux articles scientifiques considérés, en alliant connaissances biomédicales et informations sémantiques, pour identifier la relation présente dans une phrase. Pour pouvoir identifier plus précisément la relation de l'ontologie OHMI, présente dans une phrase donnée, il sera nécessaire de réduire le nombre de relations considérées. En effet, les 129 relations sélectionnées sont trop nombreuses, ce qui complexifie la tâche d'identification de la relation présente ainsi que la tâche d'annotation (l'annotation d'une phrase pouvait prendre jusqu'à 30 minutes).

Références

BADA M., ECKERT M., EVANS D., GARCIA K., SHIPLEY K., SITNIKOV D., BAUMGARTNER W. A., COHEN K. B., VERSPOOR K., BLAKE J. A. *et al.* (2012). Concept annotation in the craft corpus. *BMC bioinformatics*, **13**(1), 1–20.

BELTAGY I., LO K. & COHAN A. (2019). Scibert : A pretrained language model for scientific text. *arXiv preprint arXiv :1903.10676*.

CONSORTIUM G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, **32**(suppl_1), D258–D261.

CUXAC P. & THOUVENIN N. (2017). Archives numériques et fouille de textes : le projet istex. *Atelier TextMine, EGC 2017 (Extraction et Gestion des Connaissances), Grenoble, France, January 24*, **27**, 2017.

DEGTYARENKO K., DE MATOS P., ENNIS M., HASTINGS J., ZBINDEN M., MCNAUGHT A., ALCÁNTARA R., DARSOW M., GUEDJ M. & ASHBURNER M. (2007). Chebi : a database and ontology for chemical entities of biological interest. *Nucleic acids research*, **36**(suppl_1), D344–D350.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

EILBECK K., LEWIS S. E., MUNGALL C. J., YANDELL M., STEIN L., DURBIN R. & ASHBURNER M. (2005). The sequence ontology : a tool for the unification of genome annotations. *Genome biology*, **6**(5), 1–12.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, **3**(1), 1–23.

HARRIS Z. S. (1954). Distributional Structure. *WORD*, **10**(2-3), 146–162. DOI : 10.1080/00437956.1954.11659520.

HE Y., WANG H., ZHENG J., BEITING D. P., MASCI A. M., YU H., LIU K., WU J., CURTIS J. L., SMITH B. *et al.* (2019). Ohmi : the ontology of host-microbiome interactions. *Journal of biomedical semantics*, **10**, 1–14.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240. DOI : 10.1093/bioinformatics/btz682.

LI J., SUN Y., JOHNSON R. J., SCIAKY D., WEI C.-H., LEAMAN R., DAVIS A. P., MATTINGLY C. J., WIEGERS T. C. & LU Z. (2016). Biocreative v cdr task corpus : a resource for chemical disease relation extraction. *Database*, **2016**.

NEUMANN M., KING D., BELTAGY I. & AMMAR W. (2019). Scispacy : Fast and robust models for biomedical natural language processing. *CoRR*, **abs/1902.07669**.

PANKAJ S. & GAUTAM A. (2022). Augmented bio-sbert : Improving performance for pairwise sentence tasks in bio-medical domain. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, p. 43–47.

REIMERS N. & GUREVYCH I. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv :1908.10084*.

ZHOU D., ZHONG D. & HE Y. (2014). Biomedical relation extraction : from binary to complex. *Computational and mathematical methods in medicine*, **2014**.