Extraction d'associations symbiotiques entre organismes marins dans des corpus scientifiques

Samuel Girardeau, Harry Jandu

Encadré par Solen Quinou et Samuel Chaffron

Janvier 2021

Table des matières

Remerciements

Nous tenons à remercier Mme. Solen Quinou et M. Samuel Chaffron qui ont encadré ce projet, et nous ont accompagné tout la durée du projet.

1 Introduction

Notre problématique est :"Peut-on utiliser des approches à base de réseaux de neurones afin d'extraire des associations symbiotiques entre les organismes marins en utilisant les informations présentes dans les corpus scientifiques ?".

En effet, les organismes microbiens sont partout et jouent un rôle très important dans la régulation de l'écosystème. Et ils forment des communautés microbiennes complexes dont les interactions peuvent impacter les structures microbiennes et donc le fonctionnement de l'écosystème.

Le but de notre projet est d'utiliser les nouvelles méthodes de traitement automatique de la langue afin de détecter les interactions entre espèces planctoniques ainsi que pour déterminer le type de cette interaction (Comme par exemple le mutualisme, le commensalisme, le parasitisme ou encore la prédation). En effet, il existe plusieurs millions d'articles scientifiques sur le domaine biomédical et leur nombre continue à croître de plus en plus. Il semble donc intéressant de fouiller ces articles pour repérer les espèces qui sont mentionnées ainsi que les éventuelles associations qui peuvent avoir lieu entre celles-ci.

2 État de l'art

2.1 BERT

BERT [5] (Bidirectional Encoder Representations from Transformers) est un modèle de representation du langage qui a été conçu dans le but de pré-entrainer les représentations bidirectionnelles en prenant en compte le contexte de la gauche et de la droite en utilisant Attention[13]. Il a été pré-entrainé de façon non-supervisée sur les grands corpus comme Wikipédia. Il a eu de grands succès notamment dans les domaines de la classification[1] ou de la reconnaissance d'entités nommées[12].

2.2 BioBERT

BioBERT [8] est un variant de BERT qui a été pré-entraîné sur de grands corpus biomédicaux (4,5 milliards de mots provenant de résumés de PubMed et 13,5 Billions de mots provenant d'articles complets de PMC). Et ce pre-training a permis de grandement améliorer ses performances sur la reconnaissance d'entités nommées et l'extraction de relations dans le domaine biomédical. En effet, on remarque dans les résultats de l'article[8] (Tab 3) que BioBERT est capable de détecter automatiquement des espèces et de les annoter. Cependant, pour la tâche d'extraction de relations, il a été uniquement testé pour trouver des liens

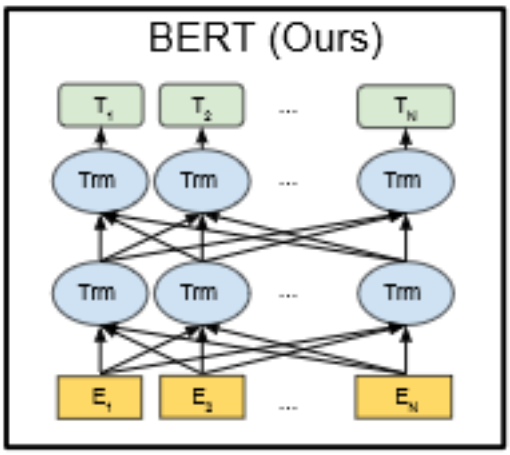

FIGURE 1 – Architecture du BERT. Source : https ://arxiv.org/pdf/1810.04805.pdf

entre les gènes et les maladies et entre les protéines et les produits chimiques. Il serait donc intéressant de le tester pour voir s'il permet aussi de détecter des relations entre espèces.

2.3 Recherche de relations entre gènes et maladies

La recherche de relations dans le domaine biomédical est un domaine où de nombreux articles ont déjà été publiés. Mais dans la plupart des cas, on s'intéressait davantage à trouver un lien entre les gènes et les maladies. C'est le cas de l'article[11] dont le but est d'extraire des triplets "gène- fonction de changement- maladie". Les relations définies par les fonctions de changement ont été classées en 4 types différents, le but étant de pouvoir détecter de manière automatique les triples intéressants et de déterminer le type de relation dont il est question.

Pour cela, un corpus de 250 résumés provenant de PubMed a été annoté afin de repérer les gènes, les maladies ainsi que les fonctions de changement. Puis le corpus a été divisé en une partie de train (80%) et une partie de dev (20%). Pour repérer les entités nommées, ils ont utilisé Pubtator [1] qui au final s'est révélé moins efficace que ce à quoi ils s'attendaient. En effet, sur les 44 paires d'entités présentes, seulement 24 ont été correctement identifiées. L'erreur provenait principalement du fait que les entités étaient retrouvées mais pas avec une correspondance exacte. Il semble donc intéressant d'essayer de normaliser les entités nommées suivant une notation scientifique afin d'éviter les erreurs de "no perfect match".

1. *https : //www.ncbi.nlm.nih.gov/research/pubtator/*

Pour ce qui est de la détection du type de relation, ils ont utilisé une extension de SciBERT [2], entraînée afin d'accomplir 2 objectifs : prédire un mot du texte qui a été masqué (MASK) et deviner, en ayant 2 phrases du texte, si celles-ci sont consécutives ou non (NSP). Et les résultats obtenus ont été plutôt satisfaisants sur les données positives, c'est-à-dire les données qui contenaient une relation entre un gène et une maladie.

2.4 Extraction de relations entre traitements et maladies

Cet article[6] s'approche un peu de l'article précédent. Ici, l'objet de l'étude consiste à détecter des relations entre des traitements et des maladies mais aussi à déterminer la polarité de ces relations. En effet, ils ont classé les relations en 3 types :"*Cure, Prevent et Side Effect*". En effet, si l'on souhaite utiliser un traitement pour soigner une maladie, il est nécessaire de savoir si il est efficace et si il peut avoir des effets indésirables. Pour effectuer cette tâche, ils ont essayé 3 représentations différentes : la représentation à base de "Bag of words", celle basée sur le NLP et celle qui utilise "unified medical language system" (UMLS) qui contient un grand nombre de données lexicales sur le domaine biomédical. Et en cumulant ces 3 approches, ils ont obtenus des résultats de F-mesure extrêmement satisfaisants. (98,55% pour Cure et 100% pour Prevent).

2.5 Prédiction de traits métaboliques pour les microbes

L'étude suivante[7] s'intéresse comme nous aux espèces. Ici, elle s'intéresse plus précisément aux microbes. Et le but de l'étude était de prédire les traits métaboliques de ceux-ci. Pour cela, ils ont récupéré la description complète des espèces puis ils ont extrait le nom de chaque espèce depuis le texte complet. Puis ils ont définis deux traits particuliers à détecter : "Fermentative Metabolism" et "Acetate Production". Afin de détecter ces traits, ils ont utilisé l'approche des mots-clés. Ici, les mots-clés étaient "Ferment" et "Acetate". Lorsqu'ils détectaient un article contenant un de ces mots-clés ou un mot le contenant, l'article complet était ensuite lu afin de pouvoir définir si le trait était positif ou négatif. Le problème de cette méthode est qu'elle est efficace pour repérer si il y a un des traits qui est concerné dans le texte mais qu'elle ne peut pas déterminer la polarité du trait de manière automatique. Il est nécessaire de parcourir l'article par la suite. Pour autant, cette méthode peut être intéressante pour détecter les phrases qui sont intéressantes à étudier en recherchant des mots-clés qui représente une association symbiotique entre 2 organismes.

2.6 Comprendre la polarité des évènements dans les corpus biomédicaux

La recherche de la polarité est intéressante dans le cas de l'extraction de relations entre espèces. En effet, lorsqu'on a réussi à extraire la relation entre 2 espèces, on aimerait bien connaître la polarité de cette relation. Dans notre cas, on souhaite savoir si la relation est plutôt une relation bénéfique ou une relation

inhibitrice. L'article suivant[10] nous présente 2 méthodes permettant de détecter la polarité et leurs résultats. La première méthode présentée est la détection de polarité à l'aide d'une approche linguistique. Cette méthode consiste à extraire les mots qui sont liés aux deux acteurs participants à l'interaction puis à déduire à l'aide d'un lexique si ces mots sont polarisés. La deuxième méthode consiste à utiliser du Deep Learning, en utilisant l'approche du réseau de neurone récurrent, c'est-à-dire que la phrase est représentée comme une séquence de "word embeddings".

Quand on compare les 2 méthodes, on se rend compte que la méthode linguistique fonctionne mieux sur les cas simples. En effet, lorsque le mot indiquant la polarité est clair et facilement détectable, ce système fonctionne bien. Cependant, lorsqu'on se retrouve avec des cas plus complexes, comme par exemple le cas où on a plusieurs mots avec une polarité différente où lorsque le mot est légèrement différent de celui répertorié dans le lexique (Ex : "Inhibitable" au lieu d'"Inhibit"), la F-mesure chute drastiquement (F1 = 0.143 sur les cas complexes). Les réseaux de neurones, quant à eux, sont plus efficaces sur l'ensemble du texte. Cela peut s'expliquer par le fait que cette représentation arrive mieux à capter le sens global de la phrase plutôt que de s'intéresser uniquement aux mots polarisés. Et cela permet d'obtenir une F-mesure de 0.757 sur les cas complexes, ce qui fait un gain important par rapport à la méthode linguistique. Cependant, cette approche est beaucoup plus complexe à mettre en place. En effet, il est nécessaire de construire un "Bidirectional LSTM" et d'avoir une importante quantité de données afin de pouvoir effectuer un pre-training efficace.

2.7 Analyse de relations de coréférences dans la littérature biomédicale

L'article suivant[4] évoque les relations de coréférences dans la littérature biomédicale. Les coréférences sont intéressantes dans notre projet car il est fréquent dans un texte de se référer à une espèce citée précédemment en utilisant un pronom ou un groupe nominal. Par conséquent, le fait de résoudre les coréférences est une étape essentielle si l'on souhaite trouver des liens entre les entités.

Pour cette tâche, ils ont utilisés un dataset pour la résolution de coréférence pour les protéines. Et ils ont essayé 2 systèmes différents : le système de Stanford et le système TEES. Et il s'est avéré que le système TEES était beaucoup plus performant avec une F-mesure de 69% contre seulement 12% pour celui de Stanford. Cependant, même si la F-mesure est plutôt bonne, le rappel est très faible car le système a détecté seulement 37% de l'ensemble des coréférences. Cela peut s'expliquer par le fait que la détection de coréférences est une tâche qui reste encore aujourd'hui très complexe. En effet, une grande partie des coréférences sont des pronoms et ceux-ci sont difficiles à lier à leur antécédent car ils contiennent peu d'informations les liant à celui-ci.

Corpus	Type d'entité	Nombre de relations en train	Nombre de relations en test
GAD (Bravo et al. 2015) [3]	Gene-Disease	47970	5330
EUADR (Van Mulligen et al. 2012) [9]	Gene-Disease	3195	355

TABLE 1 – Les statistiques des relations dans les corpus

3 Constitution des corpus

3.1 Corpus existants

BioBERT utilise des corpus pour la Reconnaissance d'Entités Nommées [2] parce qu'ils sont fréquemment utilisés par plusieurs chercheurs. Nous allons utiliser les mêmes corpus [3] [14] pour pouvoir comparer les résultats de nos travaux avec les autres. Il y a deux corpus primaires utilisés pour l'extraction de relations, le EUADR[3] et le GAD (Gene-Disease Associations)[9]. Les corpus contiennent des phrases provenant des articles scientifique [3] et [9]. Chaque ligne est constituée forcément d'une phrase qui contien un gène et une maladie et un libellé qui indique s'il existe ou non une relation entre le gène et la maladie soit 1 pour oui et 0 pour non.

Tableau 1 décrit le nombre total des relations dans les fichiers train et test respectivement. D'ailleurs, selon la Figure 2, nous constatons que les nombres de relations dans le corpus GAD sont équilibrés mais ce n'est pas le cas pour le corpus EUADR.

Note : Nous n'avons pas utilisé les données de validation fournies. Nous avons créé nos propres données de validation lors des entraînements, en utilisant les données d'entraînement.

4 Méthodologie

4.1 Pré-traitements

Les données fournis par les auteurs de BioBERT avait un pré-traitement important. Les entités nommés cible ont été remplacées par les balises pré-définies comme @GENE$ et @DISEASE$. Comme chaque phrase est soit positive (relation existe) soit negative (relation n'existe pas), la relations d'extraction dans ce cas peut être modélise comme un problème de classification binaire.

Les autres pré-traitements effectués sont décrits ci-dessous :

1. Ajouter le jeton spécial [CLS] pour indiquer à BERT que c'est un problème de classification

2. $https://fr.wikipedia.org/wiki/Reconnaissance_d\%27entit\%C3\%A9s_nomm\%C3\%A9es$
3. $https://drive.google.com/open?id=1-jDKGcXREb2X9xTFnuiJ36PvsqoyHWcw$

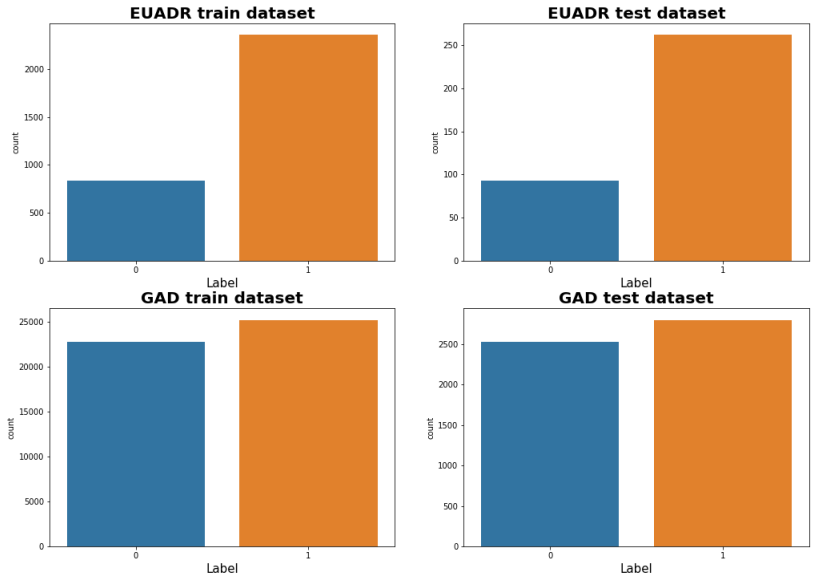

FIGURE 2 – Nombre de relations dans les corpus. Bleu : Négatives, Orange : Positives

2. Tronquer les phrases à une taille fixe (BERT necéssite une taille maximale de 512)
3. Ajouter le jeton spécial [PAD] à la fin des phrases qui sont inférieures à la taille maximale
4. Transformer toutes les mots en jetons et les jetons en vecteur numériques
5. Ajouter les masques d'attention (ceci est soit 0 soit 1, elle indique au BERT quelles sont les phrases et quelles sont les [PAD])

Remarque : La fonction encode_plus [4] fournie par HuggingFace fait les pré-traitements mentionnés ci-dessus pour nous.

4.2 Modèle

Pour notre travail, nous avons décidé d'utiliser BERT et son variant BioBERT afin de comparer les résultats entre les deux modèles. Le modèle BioBERT à la même structure que BERT et ils ne diffèrent que dans le pré-entraînement. La Figure 1 montre l'architecture de BERT. Nous allons décrire ci-dessous les différentes couches du modèle :

4. *https* : //*huggingface.co/transformers/internal/tokenization*$_u$*tils.html#transformers.tokenization*$_u$*tils*$_b$*ase.PreTrain*

1. Couche d'entrée qui prend un vecteur de taille inférieure ou égale à 512
2. Couche d'embedding qui transforme chaque entrée à un vecteur de taille 768
3. 12 couches d'encodeurs dont chacune contient en lui-même une couche d'attention et un réseau de neurones à propagation avant
4. La dernière couche ajoutée est une couche de classification (binaire pour notre problématique)

4.3 Entraînement

Nous avons choisi d'utiliser Google Colab [5] pour entraîner nos modèles parce que c'est gratuit, facile à utiliser et contient un GPU. Les hyper-paramètres sont décrit dans la Table 2.

Hyper-paramètres	**GAD**	**EUADR**
Taille maximale des sequences	128	128
Epochs	2	3
Optimizer	AdamW	AdamW
Loss Function	Binary Cross Entropy	Binary Cross Entropy
Batch size	32	32
Learning rate	3e-5	3e-5
Weight Decay	1e-8	1e-8

TABLE 2 – Hyper-paramètres pour l'entraînement du modèle

Ces hyper-paramètres sont pareil que celle utilisé par les auteurs de BioBert avec l'exception de nombre d'epochs. Ils ont choisi d'affiner leur modèle avec 20 epochs et nous croyons que cela a sur-entraîne le réseau. Les auteurs de BERT ont recommandé d'utiliser entre 2-4 epochs pour le fine-tuning.

L'entraînement a pris moins de 20 minutes pour le corpus EUADR et moins de 40 minutes pour le corpus GAD avec un GPU Tesla T4 fourni par Google Colab.

Remarque : Le choix d'epochs a été fait selon la taille des données. Comme GAD contient plus de données que EUADR, nous avons utilisé moins d'epochs afin de prévenir le sur-entraînement. Une raison pour cela est parce que les modèles convergent moins rapidement quand les données sont nombreux. [6]

5 Résultats

Les résultats de l'extraction des relations de chacun des modèles sont montrés dans la Table 5. Notre modèle BERT a eu les meilleurs résultats dans les deux

5. *https* : *//colab.research.google.com/*
6. *https* : *//stackoverflow.com/questions/*35050753/*how* − *big* − *should* − *batch* − *size* − *and* − *number* − *of* − *epochs* − *be* − *when* − *fitting* − *a* − *model* − *in* − *keras*

corpus comparé à l'état de l'art et BioBERT (les autres résultats ont été pris de l'article [8]).

Relation	Corpus	Metric	BERT-base-cased (ours)	BERT (SOTA)	BioBert V1.1 (+ PubMed)
Gene-Disease	GAD	Precision	**99.71%**	79.21%	77.32%
		Recall	**98.93%**	89.25%	82.68%
		F1	**99.32%**	83.93%	79.83%
Gene-Disease	EUADR	Precision	**98.05%**	76.43%	77.86%
		Recall	95.80%	**98.01%**	83.55%
		F1	**96.91%**	85.35%	79.74%

TABLE 3 – Comparaison des résultats sur les données et les différents modèles

Le modèle BERT (SOTA) a été pré-entraîne sur les données anglaises Wikipédia et BooksCorpus pendant. Ceci est le même modèle que nous avons utilisé pour affiner notre modèle. D'autre coté, BioBert V1.1 (+ PubMed) a été pré-entraîne sur les données PubMed en utilisant les même hyper-paramètres que BERT. BioBert a utilisé les vecteurs Bert-base pre-entraîné due à la complexité de Bert-large.

5.1 Limitations

Nous avons rencontré quelques limitations qui nous ont empêchées de comparer au mieux les résultats de notre mise en œuvre du modèle BERT. Les limitations sont décrites ci-dessous :

1. Le GPU fourni par Google Colab ne permet pas de faire de très gros calculs donc nous avons dû utilisé les séquences de taille 128 afin d'assurer un entraînement rapide et faisable.
2. Nous n'avons pas pu faire tourner le code fourni par BioBert [7] et donc nous avons utilisé les résultats déjà fournis dans leur article.

6 Conclusion

Parmi les différentes approches présentées, on remarque que la détection de relations dans des corpus biomédicaux a déjà été au sujet de nombreux articles scientifiques. Ces approches se basent principalement sur l'utilisation de réseaux neuronaux et de modèles d'apprentissage automatique comme BERT afin de détecter les relations ainsi que la polarité de celles-ci.

Nous avons utilisé la même approche avec BERT et nous croyons avoir atteint un nouveau état de l'art dans le problématique d'extraction des relations. Cependant, ces sont des résultats sur deux corpus qui ont été pre-traité d'avance.

7. *https* : *//github.com/dmis − lab/biobert − pytorch*

Pour pouvoir assurer que notre modèle marche bien sur ce genre de tache, il est imperative de tester sur d'autres corpus comme le BB2019 [8].

Références

[1] Ashutosh Adhikari et al. *DocBERT : BERT for Document Classification*. 2019. arXiv : 1904.08398 [cs.CL].

[2] Iz Beltagy, Kyle Lo et Arman Cohan. *SciBERT : A Pretrained Language Model for Scientific Text*. 2019. arXiv : 1903.10676 [cs.CL].

[3] Paulina Bravo et al. "Conceptualising patient empowerment : A mixed methods study". In : *BMC health services research* 15 (juil. 2015), p. 252. doi : 10.1186/s12913-015-0907-z.

[4] Miji Choi, Karin Verspoor et Justin Zobel. "Analysis of Coreference Relations in the Biomedical Literature". In : *Proceedings of the Australasian Language Technology Association Workshop 2014*. Melbourne, Australia, nov. 2014, p. 134-138. url : https://www.aclweb.org/anthology/U14-1019.

[5] J. Devlin et al. "BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding". In : *NAACL-HLT*. 2019.

[6] Oana Frunza et Diana Inkpen. "Extraction of Disease-Treatment Semantic Relations from Biomedical Sentences". In : *BioNLP@ACL*. 2010.

[7] Timothy J. Hackmann. "Using neural networks to mine text and predict metabolic traits for thousands of microbes". In : *bioRxiv* (2020). doi : 10.1101/2020.09.29.319335. eprint : https://www.biorxiv.org/content/early/2020/09/30/2020.09.29.319335.full.pdf. url : https://www.biorxiv.org/content/early/2020/09/30/2020.09.29.319335.

[8] Jinhyuk Lee et al. "BioBERT : a pre-trained biomedical language representation model for biomedical text mining". In : *Bioinformatics* 36.4 (sept. 2019), p. 1234-1240. issn : 1367-4803. doi : 10.1093/bioinformatics/btz682. eprint : https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf. url : https://doi.org/10.1093/bioinformatics/btz682.

[9] Erik M. van Mulligen et al. "The EU-ADR corpus : Annotated drugs, diseases, targets, and their relationships". In : *Journal of biomedical informatics* 45 (avr. 2012), p. 879-84. doi : 10.1016/j.jbi.2012.04.004.

8. *https* : //*sites.google.com*/*view*/*bb* − 2019/*dataset*

[10] Enrique Noriega-Atala et al. “Understanding the Polarity of Events in the Biomedical Literature : Deep Learning vs. Linguistically-informed Methods”. In : *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*. Minneapolis, Minnesota : Association for Computational Linguistics, juin 2019, p. 21-30. doi : 10.18653/v1/W19-2603. url : https://www.aclweb.org/anthology/W19-2603.

[11] Ashok Thillaisundaram et Theodosia Togia. “Biomedical relation extraction with pre-trained language representations and minimal task-specific architecture”. In : *BioNLP-OST@EMNLP-IJNCLP*. 2019.

[12] Tejas Vaidhya et Ayush Kaushal. “IITKGP at W-NUT 2020 Shared Task-1 : Domain specific BERT representation for Named Entity Recognition of lab protocol”. In : *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)* (2020). doi : 10.18653/v1/2020.wnut-1.34. url : http://dx.doi.org/10.18653/v1/2020.wnut-1.34.

[13] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv : 1706.03762 [cs.CL].

[14] Xuan Wang et al. *Cross-type Biomedical Named Entity Recognition with Deep Multi-Task Learning*. 2018. arXiv : 1801.09851 [cs.IR].