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

Encadre par:
Solen Quinou
Samuel Chaffron

Présenté par:
Samuel Girardeau
Harry Jandu

Plan

1. Introduction
2. Travails précédents
3. Méthodologie
4. Résultats
5. Limitations
6. Conclusion

Introduction

Extraction des relations

- Tâche de classification des entités nommés dans les corpus biomédicaux
- Nous allons utiliser les approches neuronales pour réaliser cette tâche

Travail précédents

LS2N - Université de Nantes

- Réalisé en Juin 2018 par David Kerbrat
- Approche par étude de cooccurrences
  - Observation des mots qui apparaissent dans le même contexte

État de l’art - BERT

- Bidirectional Encoder Representations from Transformers (Devlin et al. 2019)
- Modèle pré-entraîné pour pouvoir sur plusieurs corpus comme Wikipédia
- Prend en considération le contexte du mots avec Attention (Vaswani et al. 2017)
- Deux modèles principales
  - BERT-base (L=12, H=768, A=12)
  - BERT-large (L=24, H=1024, A=16)

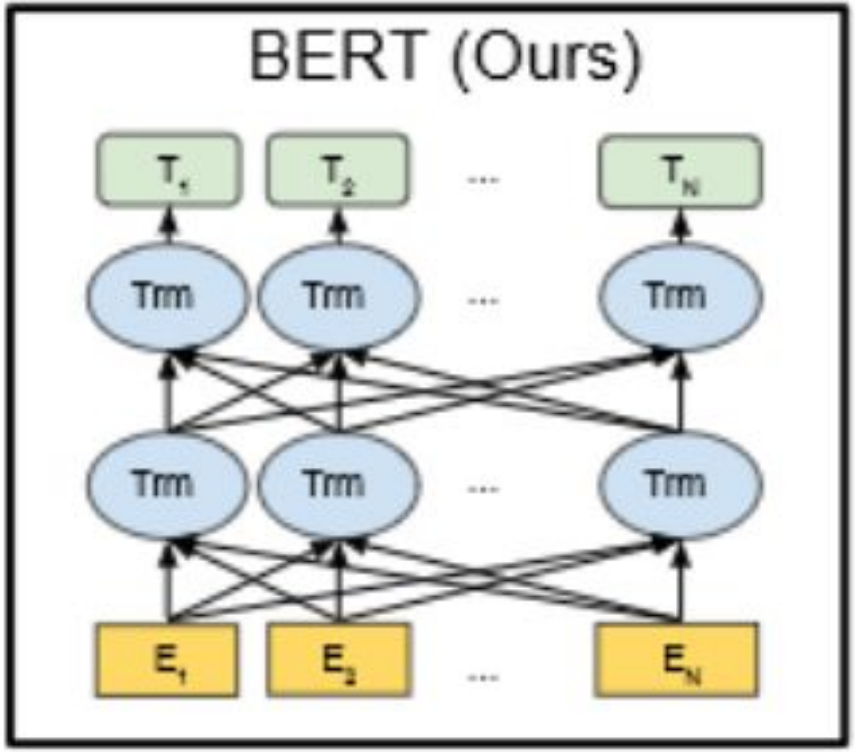


Figure 1 : Architecture du BERT

État de l’art - BioBert

- BioBERT: a pre-trained biomedical language representation model for biomedical text mining (Jinhyuk et al. 2019)
- Basé sur le modèle BERT
- Pré-entraîné sur les grands corpus biomédicaux
  - Wikipédia (Général)
  - BooksCorpus (Général)
  - PubMed Abstracts (Biomédical)
  - PMC Full-text articles (Biomédical)
- Utilisation de BERT-base pour pré-entraînement

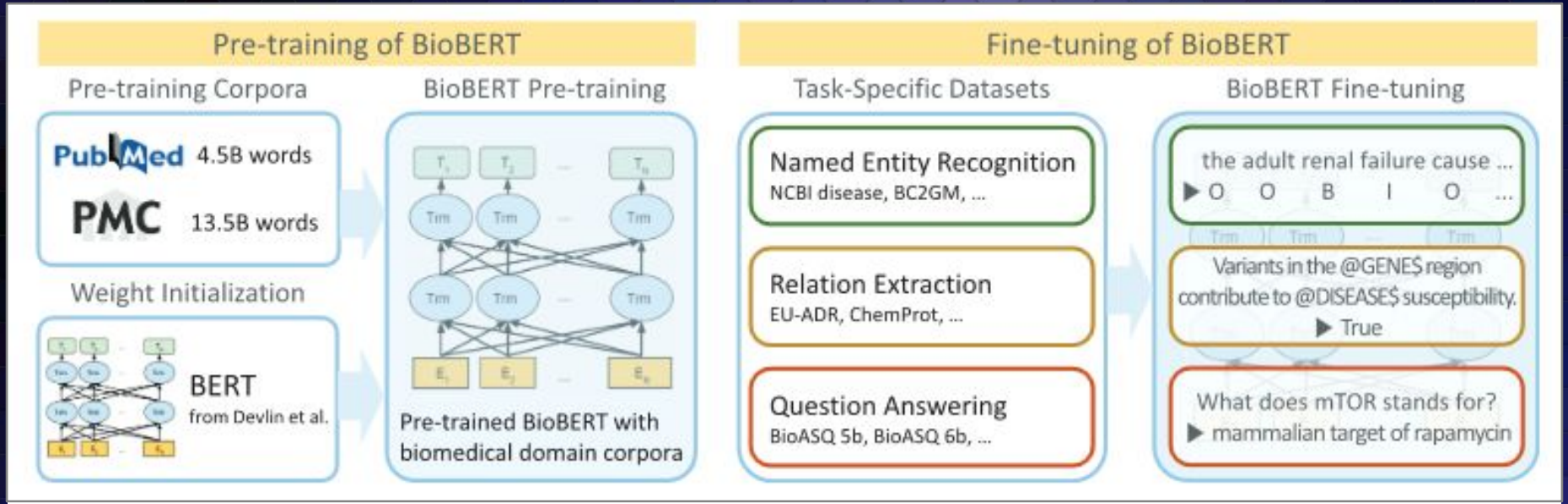


Figure 2 : Pré-entraînement et Fine Tuning de BioBert

Méthodologie

Notre approche

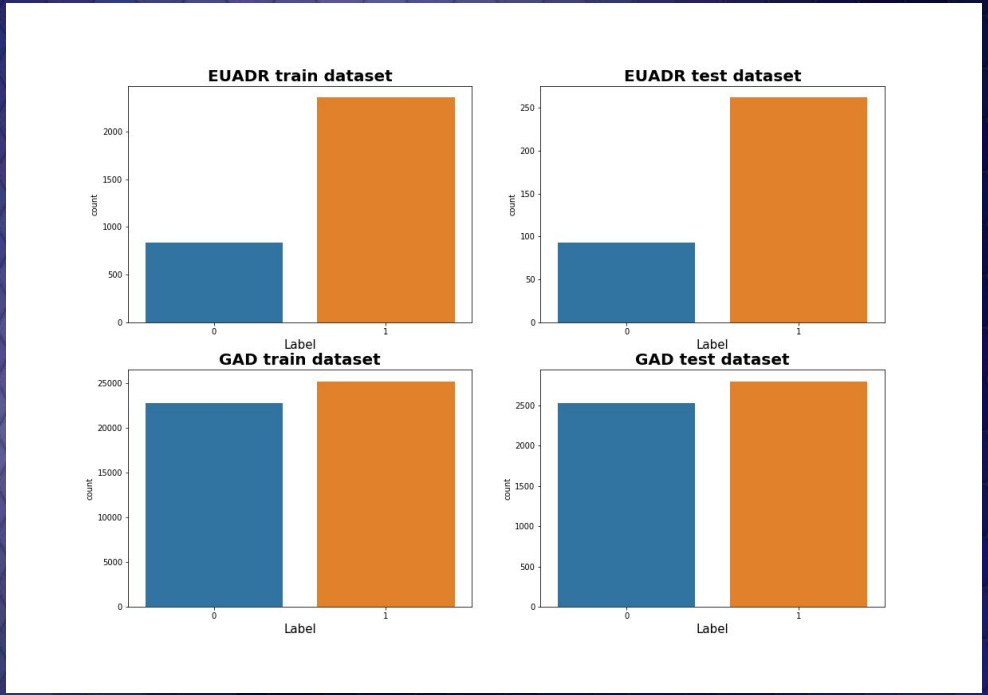
- Utilisation du modèle BERT-base pour la classification des séquences
- Modèle pré-entrainé fourni par HuggingFace pour PyTorch
  - https://huggingface.co/transformers/model_doc/bert.html
- Implémentation PyTorch en utilisant Google Colab
  - Colab facile à utiliser
  - GPU gratuit fourni

Corpus

- Deux corpus principales utilisé par BioBERT
  - European Union database of Suspected Adverse Drug Reaction Reports - EUADR
    - Erik M. Van Mulligen et al. “The EU-ADR corpus : Annotated drugs,diseases, targets, and their relationships”. In :Journal of biomedical in-formatics45 (avr. 2012), p. 879-84.doi:10.1016/j.jbi.2012.04.004
  - Gene Disease Associations - GAD
    - 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
- Corpus pour l’extraction des relations
- Chaque ligne contient
  - Un gène
  - Une maladie
  - Libellé : 1 pour positive et 0 pour négative

Taille des données

Corpus	Train	Test
GAD	47,970	5,330
EUADR	3,195	355



Pré-traitements

- Pré-traitements déjà effectué
  - Anonymisation : remplacement des entités nommées ciblés
  - Exemple : serine position 986 of @GENE$ may be an independent genetic predictor of angiographic @DISEASE$
- Conversion des mots en vecteurs de taille maximale 128
- Ajout de jeton [CLS] au début de chaque phrase et [PAD] pour les phrases inférieures à 128
- Ajout des masques d'attention
  - Valeur binaire

Modèle

Couche de sortie
(768 -> 2)

12 couches encodeurs
empilés
(768 -> 768)

Couche d’embedding
(128 -> 768)

Couche d’entrée
(taille 128)

Hyperparamètres d’entraînement

Hyper-paramètres	GAD	EUADR
Taille max des séquences	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

Résultats

Relation	Corpus	Metric	BERT-base-cased (ours)	BERT (état de l'art)	BioBERT V1.1 (+ PubMed)
Gene-Disease	GAD	Precision	**99.71%**	79.21%	77.32%
		Recall	**98.93%**	89.25%	82.62%
		F1-Score	**99.32%**	83.93%	79.83%
Gene-Disease	EUADR	Precision	**98.05%**	76.43%	77.86%
		Recall	95.80%	**98.01%**	83.55%
		F1-Score	**96.91%**	85.35%	79.74%

Limitations

Niveau hardware

- GPU fourni par Google Colab ne permet pas de faire des très gros calculs
- Nous avons pas réussi à exécuter le code fourni par les auteurs de BioBert
  - https://github.com/dmis-lab/biobert

Conclusion

Nouveau état de l’art

- Nos résultats ont dépassé l’état de l’art existent
- Travail prévu
  - Utilisation d’autres corpus
    - BB 2019 - https://sites.google.com/view/bb-2019/dataset/corpus-statistics

Merci pour votre attention