**Hackathon - sujet 4 - groupe 6**ligne horizontale

Aller dans : *Afficher* - *Afficher le plan du document* (ou Ctrl+Alt+A).

Organisation - logistique

Membres du **groupe n°6** :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Prénom | Nom | Ecole | Département | Mail |
| Sylvain | COMBETTES | Mines Nancy | Mathématiques Appliquées | sylvain.combettes8@etu.univ-lorraine.fr |
| Mohammed | EL YAAGOUBI | TELECOM NANCY | IAMD |  |
| Guillaume | GATTI | TELECOM NANCY | IAMD | guillaume.gatti@telecomnancy.eu |
| Zouhair | JANATI-IDRISSI | TELECOM NANCY | IAMD | zouhair.janati.idrissi@telecomnancy.eu |
| Ali | LABBAIZE | TELECOM NANCY | IAMD |  |

Comment on s’organise ?

* Adrien Coulet revient mardi et mercredi à 9h (pas jeudi).
* conversation sur [Slack](https://app.slack.com/client/TQEDN65RP/CQNBGPJKS)
* programmation sur [GitHub](https://github.com/AliLabbaize/Hackathon-NLP-2019)

Roadmap court terme :

1. Faire tourner BERT sur des données “faciles” puis sur PGxCorpus. En particulier : comprendre le format des données input de BERT. → Ali et Mohammed
2. Pipeline de la data : comprendre le format BRAT de PGxCorpus et comment le transformer au format pour qu’il soit compatible avec BERT. → Zouhair et Guillaume, voir ce [Drive](https://docs.google.com/document/d/1WJTlrLHOdXqHsOI7vUfd0Ct-K3hQacgiR5IZq8DnWuA/edit)
3. Recherche bibliographique : être capable d’expliquer NLP, BERT et PGxCorpus. → Sylvain, voir ce [Drive](https://docs.google.com/document/d/1u5HtzfkyXiXePYPYXgBPmsB1V2DjTSrwu8fx4xZedBM/edit#)

Descriptif du projet

Drive de présentation de tous les projets : [lien](https://docs.google.com/document/d/1a2GQ4MlJDy1AXKa6Z2_u__qfj10_52CV_WcKQAOFLFE/edit?ts=5dd2441a).

Entreprise/laboratoire : LORIA

*NER on PGxCorpus with BERT*

ou

*Amélioration des performances de la reconnaissances d’entités nommées biomédicales complexes avec PGxCorpus et BERT*

La pharmacogénomique est un domaine qui étudie l’impact des gènes sur la réponse aux médicaments. Dans ce domaine, l’essentiel des connaissances est contenu dans des textes de la littérature scientifique, et en particulier dans la base de données [PubMed](https://www.ncbi.nlm.nih.gov/pubmed/). Le volume important de la littérature biomédicale (plus de 3 milliards d’articles) fait que des approches de traitement automatique du langage naturel (TALN ou *NLP* pour l’anglais *Natural Language Processing*) capables de fouiller la littérature pour en extraire de façon synthétique les connaissances d’un domaine est très intéressants pour les chercheurs du domaine.

Dans les équipes Orpailleur, Capsid et Synalpe du Loria/Inria Nancy nous avons développé récemment un corpus appelé PGxCorpus qui est annoté manuellement pour le domaine de la pharmacogénomique. L’intérêt de ce corpus est de constituer un ensemble de données d’entraînement et d’évaluation pour des modèles d’apprentissage supervisés. Nous avons démontré l'intérêt de ce corpus pour des tâches de reconnaissances d’entités nommées et d’extraction de relation, mais avec des méthodes qui ne sont pas les meilleurs de l’état de l’art actuel.

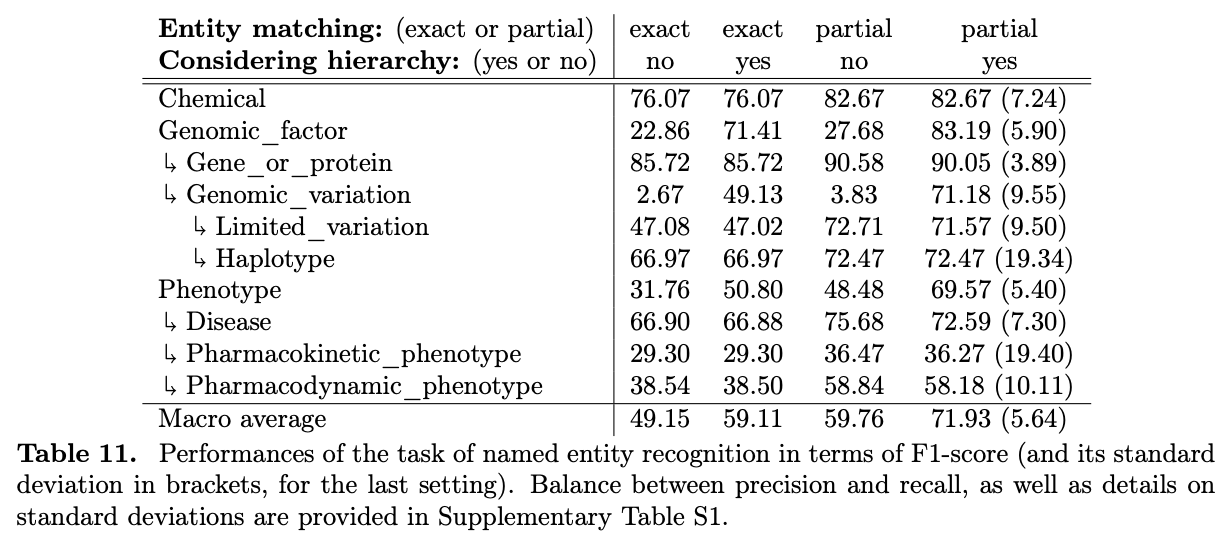
Nous proposons aux participants du Hackathon IAMD 2019 qui choisiront ce projet de développer un modèle (ou des modèles) de reconnaissances d’entités nommées (ou *NER* pour *Named Entity Recognition*), entraîné avec PGxCorpus avec les dernières méthodes de l’état de l’art pour obtenir de meilleurs performances que les nôtres.

Nous pensons notamment que le modèle *BERT* (*Bidirectional Encoder Representations from Transformers*) pourrait aider à atteindre de meilleurs performances.

Liens :

* PGxCorpus peut être exploré de façon visuelle [ici](https://pgxcorpus.loria.fr/) et est téléchargeable [ici](https://figshare.com/s/9d315cec6bb629d04210).
* L’article scientifique qui le décrit est [ici](https://www.biorxiv.org/content/10.1101/534388v3).
* Une implémentation de BERT pour la tâche de reconnaissances d’entités nommées est sur [GitHub](https://github.com/kamalkraj/BERT-NER).
* Notre expérience en Lua sur [GitHub](https://github.com/practikpharma/PGxCorpus/tree/master/baseline_experiment).

Les performances à battre :



Une roadmap possible est :

* Reproduire les expériences faites avec BERT [ici](https://github.com/kamalkraj/BERT-NER).
* Adapter le format BRAT de PGxCorpus pour le rendre compatible avec l’expérience précédente.
* Évaluer la reconnaissances des entités nommées de PGxCorpus avec BRAT.
* Voir si le fine-tuning est important ?
* Explorer d’autres voies d’amélioration comme l’enrichissement du corpus pour les types d’entités les moins fréquents.

Contacts :

* Adrien Coulet, Loria, [adrien.coulet@loria.fr](mailto:adrien.coulet@loria.fr)
* Joël Legrand (à confirmer), Loria, [joel.legrand@loria.fr](mailto:joel.legrand@loria.fr)

Présentation du projet par Adrien Coulet (lundi 18/11)

Il s’agit d’un projet fouille de texte par deep learning (NLP) : extraire information intéressante. Le but du hackathon est de s'intéresser au modèle BERT de Google. Le modèle BERT pour NER est implémenté en PyTorch sur un [GitHub](https://github.com/kamalkraj/BERT-NER). NER signifie reconnaissance d’entité nommée et RE signifie *relation extraction*. Le but de NER est de structurer nos données.

NER (*Named Entity Recognition*) ou reconnaissance d’entité nommée

* approche par dictionnaire : très complet et fonctionne très bien pour des articles scientifiques pour la médecine car le nom de médicament n’a pas d'ambiguïté, sémantique lexicale avec hiérarchie (mammifère - félin - chat)
* sémantique distributionnelle, il faut étiqueter (humain ou semi-supervisé) les termes qui décrivent un chat → <https://pgxcorpus.loria.fr> chaque fichier est une phrase annotée (humain) au format BRAT, dictionnaire avec tous les mots qui se trouvent dans le voisinage (matrices creuses) avec des embeddings pour la description d’une entité

personnalisation de la médecine → pharmaco-génomique ou pharmacogenomics (PGx)

* comment les gènes impactent la réponse au médicament
* 0,2% de variation de version des gènes (séquence de nucléotides) entre les individus qui influent que la réponse
* regarder si traitement fonctionne et est supportable (attention aux effets secondaires)
* note culturelle : regarder aussi l’alimentation, par exemple les brocolis interagissent aussi avec les médicaments
* médicament et version des gènes (facteur génétique) influent sur la réponse au traitement, phénotype (quelque chose qu’on observe comme les rougeurs sur la peau) → fouiller du texte pour trouver des relations ternaires, on a annoté des relations binaires, sous-types dans les facteurs génétiques
* ces articles sont des connaissances expérimentales mais pas validées

Notre objectif est d’améliorer l’article en entraînant BERT sur PGxCorpus (NER). Regarder le format input, BERT pour le NER. BERT est juste le modèle. BERT utilise les CNN pour NLP. Regarder aussi BioBERT.

Comment regarder la performance de la fouille de texte ? On mesure la qualité de ce qu’on a annoté. On va regarder avec les matrices de confusion :

* précision = TP/(TP+FP)
* rappel (recall) = TP/(TP+FN)
* F-measure = formule harmonique entre la précision et le rappel, si 1 les deux termes sont hauts
* AUC-ROC

Il faut se donner de la flexibilité : une annotation n’est pas entièrement fausse par exemple.

Réunion avec Adrien Coulet et Joël Legrand (mardi 19/11)

Avec Adrien Coulet (matin)

* Il est possible de faire tourner BERT NER en CPU sur nos ordinateurs en changeant les paramètres, pas forcément besoin des GPU de Google Colab.
* Vérifier 10 phrases par exemple pour PGxCorpus sur BERT NER.
* On peut commencer directement avec BioBERT plutôt que BERT.
* On peut utiliser la [libraire NLP](https://nlp.stanford.edu/software/) de Stanford pour obtenir le tagging à partir de phrases. → Pour l’instant on ne l’a pas utilisée.
* On se concentre uniquement sur le NER, ou au moins on commence par le NER et on le fait bien.
* Difficultés de NER à ne pas traiter pour l’instant :
  + beaucoup de types différents
  + entités discontinues
  + entités imbriquées
* Note : proposition de stage de fin d'étude sur le transfert learning avec Adrien Coulet et Joël Legrand. Enrichir un corpus avec d’autres corpus, en particulier avec PGxCorpus.

Avec Joël Legrand (après-midi)

* voir la partie de [Sylvain](https://docs.google.com/document/d/1u5HtzfkyXiXePYPYXgBPmsB1V2DjTSrwu8fx4xZedBM/edit#)