**Data Pipeline**

*par Zouhair et Guillaume*ligne horizontale

Ressources

* Language-Independent Named Entity Recognition | [lien](https://www.clips.uantwerpen.be/conll2003/ner/?fbclid=IwAR2CltYFHbIEMbDjbyQhXste_JfiLKKlCFLFBMR1QuB6v9rhVZ94wI1-xEk)

Le format BRAT du dataset PGxCorpus

Ressources

* *annotation examples in brat* | <https://brat.nlplab.org/examples.html>
* *brat standoff format* | <https://brat.nlplab.org/standoff.html>

Le dataset PGxCorpus est au format BRAT. Pour télécharger le corpus, on va sur le [figshare](https://figshare.com/s/9d315cec6bb629d04210) et on extrait le dossier PGxCorpus qui contient **1892** **éléments** : des fichiers **ann** et **txt**. Chaque fichier ann correspond à un fichier txt (du même nom). Un fichier txt contient une phrase et ann contient les annotations de cette phrase. L’outil le plus simple pour ouvrir ces fichiers est un éditeur de texte.

Exemple :

|  |
| --- |
| *Fichier 1225820\_10.txt*  This present study provides an additional piece of support to the hypothesis that the ChF1 and ChS1 are alleles determining the synthesis of usual and atypical cholinesterase together with the likelihood of ChU1 ChD1 heterozygotes having occasional suxamethonium apnoea . |

|  |
| --- |
| *Fichier 1225820\_10.ann*  T1 Gene\_or\_protein 86 90 ChF1  T2 Gene\_or\_protein 95 99 ChS1  T3 Phenotype 128 140;160 174 synthesis of cholinesterase  T4 Gene\_or\_protein 160 174 cholinesterase  T5 Disease 263 269 apnoea  T6 Chemical 249 262 suxamethonium  T7 Limited\_variation 207 230 ChU1 ChD1 heterozygotes  T8 Gene\_or\_protein 212 216 ChD1  T9 Pharmacodynamic\_phenotype 249 269 suxamethonium apnoea  T11 Gene\_or\_protein 207 211 ChU1  R1 influences Arg1:T1 Arg2:T3  R2 influences Arg1:T2 Arg2:T3  R3 isAssociatedWith Arg1:T7 Arg2:T9 |

Chaque ligne du fichier *1225820\_10.ann* commençant avec un T est un tag (une annotation) de NER (reconnaissance d'entité nommée). La première ligne surlignée en couleur cyan signifie que le mot *ChF1*, situé entre les offset 86 et 90, appartient à la classe *Gene\_or\_protein* et il s’agit du tag référencé par T1. Un offset correspond à une position. Les lignes commençant par R signifient une relation : on ne s'intéresse pas aux relations ici mais uniquement aux tags.

Pour visualiser le format BRAT, on peut aller sur le site <https://pgxcorpus.loria.fr/> du projet PGxCorpus. Pour visualiser la phrase correspondante à l’exemple plus haut, on va sur <https://pgxcorpus.loria.fr/#/PGxCorpus/1225820_10>. On vérifie que *ChF1* a bien le tag *Gene* et on ignore les relations (dans le cadre de ce hackathon).

Le format IOB de l’algorithme BERT NER

Ressources

* [Wikipedia](https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging))
* Voir le 2.6 du chapitre 7 de la documentation NLTK | [lien](https://www.nltk.org/book/ch07.html)
* *Introduction to the CoNLL-2003 Shared Task:Language-Independent Named Entity Recognition* | [lien](https://www.aclweb.org/anthology/W03-0419.pdf?fbclid=IwAR2me80hVOQjFFNqmCVKd72S1aKaH8f_hpm9hf4DSUVjv_3cZN43ZZRQ8k8) → article scientifique qui explique comment IOB fonctionne
* Alphabetical list of part-of-speech tags used in the Penn Treebank Project | [lien](https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html) → signification de la première colonne des données input du bret, POS

Les données input de [BERT NER](https://github.com/kamalkraj/BERT-NER) sont au **format IOB** pour **I**nside-**O**utside-**B**eginning (tagging), et font figurer **4 colonnes**. On peut visualiser les données input de BERT NER qui sont au format IOB sur le GitHub dans le dossier [*data / test.txt*](https://github.com/kamalkraj/BERT-NER/blob/dev/data/test.txt) par exemple. Voici les 17 premières lignes :

|  |
| --- |
| *Fichier test.txt*  -DOCSTART- -X- -X- O  SOCCER NN B-NP O  - : O O  JAPAN NNP B-NP B-LOC  GET VB B-VP O  LUCKY NNP B-NP O  WIN NNP I-NP O  , , O O  CHINA NNP B-NP B-PER  IN IN B-PP O  SURPRISE DT B-NP O  DEFEAT NN I-NP O  . . O O  Nadim NNP B-NP B-PER  Ladki NNP I-NP I-PER |

Pour la ligne surlignée en cyan, on a en première colonne le mot auquel on s'intéresse (soit *JAPAN*) et en dernière colonne le tag correspondant (soit *LOC* pour localisation). On peut lire le vrai texte de la manière suivante : *SOCCER: JAPAN GET LUCKY WIN, CHINA IN SURPRISE DEFEAT*.

On trouve généralement la dernière colonne qui est associé à la classe sous la forme B-class ou I-class et cela signifie dans le cas de B-class (beginning) qu’on est au début d’une expression composé de plusieurs termes ou qu’on devant un seul terme. D’un autre côté, la I-class (inside) signifie que le mot associé appartient à une expression mais ce n’est pas le premier mot. Un exemple pour cela est Nadim Ladki c’est une expression qui appartient à la classe PER, on a donc le premier terme de cette expression qui a la classe B-PER et le deuxième qui a la classe I-PER.

Le format IOB de l’algorithme BioBERT

[BioBERT](https://github.com/MeRajat/SolvingAlmostAnythingWithBert/tree/master/biobert_ner) utilise également le format IOB mais ne prend que le mot et la classe à la fin, donc seulement **2 colonnes**. BioBERT est déjà entraîné sur des dataset pour la médecine donc semble plus adapté que BERT dans notre cas.

<https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html>

<https://github.com/kamalkraj/BERT-NER/blob/dev/data/test.txt>

<https://www.biorxiv.org/content/biorxiv/early/2019/09/16/534388.full.pdf>