Prof. Amal Zouaq
Chargés de laboratoire : Dominique Piché ; Charles-Olivier Favreau ; Fabrice Charbonneau

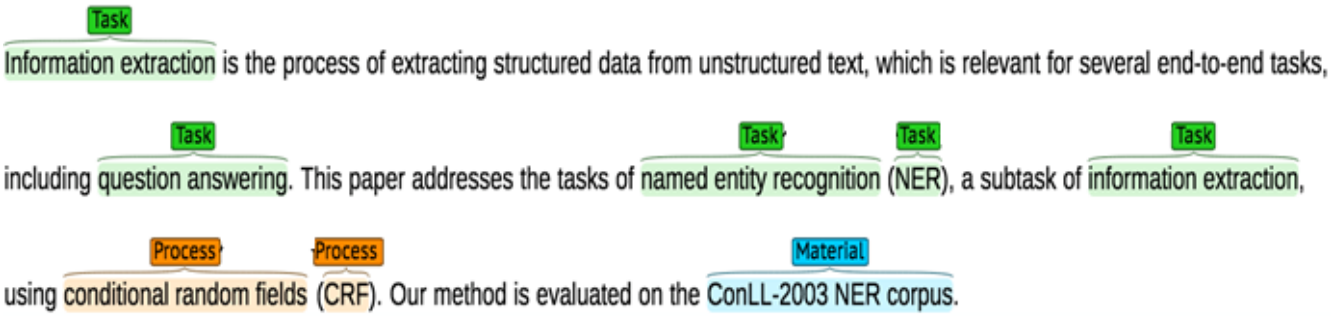
École Polytechnique de Montréal
Département Génie Informatique et Génie Logiciel
INF8460 – Traitement automatique de la langue naturelle

TP3 INF8460
Automne 2021

1. DESCRIPTION

Dans ce TP, vous aurez la tâche d'extraire des mots-clés à partir de textes (sous-tâche A) ainsi que leur type (sous-tâche B). Le problème est un problème de prédiction de séquence à séquence. Pour les sous-tâches A et B, on veut prédire un label en sortie par jeton (token) en entrée.

Exemple de sortie pour A et B :



Sous-tâche (A): Identification de mots-clé
Exemple de format de soumission attendu :

DocID	TokenID	Token	Tag
S0370269304007567	S0370269304007567-0	the	O
S0370269304007567	S0370269304007567-1	oxidation	U
S0370269304007567	S0370269304007567-2	of	O

*Pour la soumission sur Kaggle, on soumet **uniquement** les colonnes TokenID et Tag

Sous-tâche (B) : Classification des mots-clés en types prédéfinis. Dans cette tâche, vous aurez non seulement à identifier les mots-clés mais aussi à leur assigner un type.

Exemple de format de soumission attendu :

DocID	TokenID	Token	Tag
S0370269304007567	S0370269304007567-0	the	O
S0370269304007567	S0370269304007567-1	oxidation	U-P
S0370269304007567	S0370269304007567-2	of	O

*Pour la soumission sur Kaggle, on soumet **uniquement** les colonnes TokenID et Tag

Dans ce projet, vous avez le choix de vous attaquer à une seule sous-tâche (A) ou à effectuer les tâches A et B. Dans le cas où vous ne résoudrez que la sous-tâche A, votre projet obtiendra une note maximale de 80%.

Les équipes qui s'attaqueront aux tâches A et B pourront obtenir une note maximale de 100%.

2. LIBRAIRIES PERMISES

- HuggingFace
- Keras
- ScikitLearn
- NLTK
- SPACY
- Pour toute autre librairie, demandez à votre chargé de laboratoire via le forum du cours sur Moodle

3. INFRASTRUCTURE

- Vous avez accès aux GPU du local L-4818. Dans ce cas, vous devez utiliser le dossier temp (voir le tutoriel VirtualEnv.pdf)
- Vous pouvez aussi utiliser l'environnement Google Colab : https://colab.research.google.com/

4. ECHÉANCE

Fin de la session. La date précise sera indiquée sur Moodle.

5. KAGGLE

Le TP3-projet se fera sous forme d'une compétition Kaggle. Vous devrez utiliser l'environnement Kaggle pour l'évaluation de vos approches.

Pour tester votre système au fur et à mesure, vous aurez le droit à 4 soumissions par jour sur Kaggle. Vous verrez deux types de résultats sur votre « private leaderboard » et votre « public leaderboard » :

- Le « public leaderboard » est calculé sur approximativement 30% des données de test, choisies aléatoirement par Kaggle. Ce score est public et est calculé sur la même tranche de données pour tous les participants.
- Le « private leaderboard » est calculé sur approximativement 70% des données de test et n'est visible qu'à la fin de la compétition. Le résultat final sera basé sur ce leaderboard. Si aucune soumission n'est choisie, la soumission avec le meilleur score sur le « public leaderboard » sera utilisée pour calculer le score sur le « private leaderboard ».

Pour l'évaluation sur Kaggle, vous devrez soumettre un fichier de données submission_test_A.csv et/ou submission_test B.csv du même format que le fichier *sample_submission.csv* (disponible sur le site de la compétition et Moodle). Submission_test_(A ou B).csv devra contenir pour chaque ligne de votre ensemble

de test, la réponse retournée par votre approche, selon le format indiqué dans la compétition. Vous devrez aussi générer le fichier Submission_val_(A ou B).csv sur l'ensemble de validation.

Dans le projet INF8460, notre classement comprendra deux leaderboards distincts :

- Un leaderboard pour les modèles de la sous-tâche A
- Un leaderboard pour les modèles de la sous-tâche B

Chaque équipe doit soumettre ses résultats à un des leaderboards (A) ou aux deux (A et B), selon son choix.

6. DESCRIPTION DES DONNEES ET METRIQUES D'EVALUATION

Le corpus est un corpus de 500 paragraphes extraits de publications scientifiques dans les domaines de l'informatique, de la physique et de la science des matériaux. 3 types de documents sont fournis dans des fichiers zip pour les ensembles d'entrainement et de validation :

- Les paragraphes dans les fichiers .txt
- Les documents entiers dont les paragraphes ont été extraits dans des fichiers .xml
- Les annotations qui indiquent les mots-clés dans des fichiers .ann. Veuillez ne tenir compte que des lignes marquées Ti (T1, T2, etc.)

Notez qu'il vous est possible de travailler uniquement avec les fichiers .txt et .ann (à vous de voir si les xml seront utiles).

Pour l'ensemble de test, seuls les fichiers .txt et .xml sont fournis.

L'ensemble d'entrainement consiste en 350 documents, l'ensemble de validation et de test contiennent chacun 100 documents.
Nous vous fournissons également 3 csv :

- train.csv
- val.csv
- test.csv

Ces fichiers contiennent tous les jetons (tokens) de chaque document.txt. Vous aurez à compléter la colonne tag au format demandé ci-dessous.

La sortie de votre modèle sera comparée à notre ensemble de référence de test sur Kaggle. Le leaderboard Kaggle utilisera la métrique f1-score.

7. ETAPES DU TP

A partir du notebook inf8460_A21_TP3 qui est distribué, vous devez réaliser les étapes suivantes. (Notez que les cellules dans le squelette sont là à titre informatif, il est fort probable que vous rajoutiez des sections au fur et à mesure de votre TP).

7.1. Etat de l'art (10%)

Décrivez en deux paragraphes, dans une cellule du notebook, l'état de l'art pour la reconnaissance de mots clé et leur annotation. Utilisez le service Google Scholar. Voici quelques mots-clé (non exhaustifs) : Named Entity recognition, NER, entity typing.

Quelles sont les meilleures techniques de l'état de l'art ?

7.2. Extraction d'information

7.2.1. Sous-tâche A : Identification des mots-clés (65%)

Dans la sous-tâche A, la tâche est de prédire les types BILOU pour les jetons en dehors (Outside), au début (Beginning) ou dans un mot-clé (Inside). Voir plus de détails ci-dessous.

a) Format (5%) : En partant des fichiers csv de l'ensemble d'entrainement (train) et de validation (val), générez un fichier train_A_bilou.csv et val_A_bilou.csv de format de type BILOU en complétant la colonne tag des fichiers train.csv et val.csv:

B :'Beginning' : Premier jeton d'une entité composée de multiples jetons
I : 'Inside' : Jeton interne d'une entité composée de multiples jetons
L : 'Last' : Dernier jeton d'une entité composée de multiples jetons
O : 'Outside': Un jeton qui ne représente pas une entité
U : 'Unit' : Une entité composée d'un seul jeton

Ici, entité réfère aux mots-clés que vous devez identifier.

b) Modèle (50%) :

- Entrainez un modèle à reconnaitre les mots-clés à partir de votre fichier train_A_bilou.csv .
- Votre modèle doit retourner les tags pour l'ensemble de validation et de test en suivant le format indiqué dans *sample_submission.csv*
- Vous utiliserez le fichier val_A_bilou.csv pour déterminer vos hyper-paramètres optimaux, et effectuer une évaluation de votre modèle.

c) Évaluation (10%) :

Vous devez calculer les métriques de précision, rappel et F1-score de votre modèle sur l'ensemble de validation. Votre fichier val_A_bilou.csv avec la colonne Tag complétée constitue alors votre référence.

- Affichez une table comparant les performances de votre modèle pour chaque tag (B, I, L, O, U) ainsi que la performance globale sous forme de moyenne micro.
- Générez les réponses de votre modèle sur l'ensemble de validation et stockez-les dans le fichier submission_val_A.csv, sur le modèle du fichier sample_submission.csv
- Générez les réponses de votre modèle sur l'ensemble de test et stockez-les dans le fichier submission_test_A.csv sur le modèle du fichier sample_submission.csv
- **Kaggle :** Vous devez soumettre votre fichier de soumission submission_test_A.csv sur Kaggle, où il sera évalué. Ce fichier doit être généré avec votre meilleur modèle au moment de sa

soumission pour évaluation sur Kaggle. Notez que vous pourrez faire des soumissions jusqu'à la date de remise et vous comparer aux performances des autres équipes.

7.2.2. Sous-tâche B : Identification des mots-clés et de leurs types (85%)

a) Format (5%) : En partant des fichiers csv de l'ensemble d'entrainement (train) et de validation (val), générez un fichier train_B_bilou.csv et val_B_bilou.csv de format de type BILOU en complétant la colonne tag des fichiers train.csv et val.csv:

B :'Beginning' : Premier jeton d'une entité composée de multiples jetons
I : 'Inside' : Jeton interne d'une entité composée de multiples jetons
L : 'Last' : Dernier jeton d'une entité composée de multiples jetons
O : 'Outside': Un jeton qui ne représente pas une entité
U : 'Unit' : Une entité composée d'un seul jeton

Ici, entité réfère aux mots-clés que vous devez identifier.
Dans ce cas, il vous faudra créer une annotation B_P, B_M, B_T pour indiquer le début (Beginning) d'un type *Process, Material* et *Task*, et ainsi de suite pour les autres tags I, L et U. Les jetons sans types et qui ne représentent pas de mot-clé auront le tag O.

b) Modèle de typage (70%)

Dans la sous-tâche B, vous devez implémenter un modèle pour prédire les types O, M, P, T pour les jetons qui sont en dehors d'un mot-clé (Outside), ou qui font partie d'un type *Material*, *Process* ou *Task* (les types à annoter).

- Votre modèle doit retourner les tags pour l'ensemble de validation et de test en suivant le format indiqué dans *sample_submission.csv*
- Vous utiliserez le fichier val_B_bilou.csv pour déterminer vos hyper-paramètres optimaux, et effectuer une évaluation de votre modèle.

c) Évaluation (10%) :

Vous devez calculer les métriques de précision, rappel et F1-score de votre modèle sur l'ensemble de validation. Votre fichier val_B_bilou.csv avec la colonne Tag complétée constitue alors votre référence.

- Affichez une table comparant les performances de votre modèle pour chaque tag (B, I, L, O, U) ainsi que la performance globale sous forme de moyenne micro.
- Générez les réponses de votre modèle sur l'ensemble de validation et stockez-les dans le fichier submission_val_B.csv, sur le modèle du fichier sample_submission.csv
- Générez les réponses de votre modèle sur l'ensemble de test et stockez-les dans le fichier submission_test_B.csv sur le modèle du fichier sample_submission.csv
- **Kaggle :** Vous devez soumettre votre fichier de soumission submission_test_B.csv sur Kaggle, où il sera évalué. Ce fichier doit être généré avec votre meilleur modèle au moment de sa soumission pour évaluation sur Kaggle. Notez que vous pourrez faire des soumissions jusqu'à la date de remise et vous comparer aux performances des autres équipes.

Quelques pistes pour débuter les sous-tâches A et B : Bi-LSTM avec plongements lexicaux, CRF, BERT, apprentissage profond.

7.3. Conclusion (5%)

Indiquez, dans une cellule, vos conclusions sur la tâche : qu'est-ce qui fonctionne ? qu'est-ce qui ne fonctionne pas ? quel type de pré-traitement vous a donné les meilleurs résultats ? quelles architectures ?

8. LIVRABLES

Vous devez remettre sur Moodle un zip contenant :

1- *Le code* : Un Jupyter notebook en Python qui contient le code tel que soumis dans l'environnement Kaggle implanté avec les librairies disponibles pour ce cours (Python, Keras, NLTK, scikitLearn, etc.) ainsi que votre fichier de soumission de données de test. Le notebook doit contenir le résultat de l'exécution de toutes les cellules. Le code doit être exécutable sans erreur et accompagné des commentaires appropriés dans le notebook de manière à expliquer les différentes fonctions et étapes dans votre projet. Nous nous réservons le droit de demander une démonstration ou la preuve que vous avez effectué vous-mêmes les expériences décrites. *Attention, en aucun cas votre code ne doit avoir été copié de projets potentiellement existants*.

2- *Le html du notebook une fois qu'il est exécuté*

3- Un fichier *requirements.txt* doit indiquer toutes les librairies / données nécessaires. Les critères de qualité tels que la lisibilité du code et des commentaires sont importants.

4- Un lien *GoogleDrive (ou autre)* vers les modèles nécessaires pour exécuter votre notebook si approprié

5- Les fichiers *train_A_bilou.csv* et *val_A_bilou.csv et/ou train_B_bilou.csv* et *val_B_bilou.csv*

6- Le fichier *submission_val_A.csv et/ou submission_val_B.csv pour l'ensemble de validation*

7- Le fichier *submission_test_A.csv et/ou submission_test_B.csv pour l'ensemble de test*

8- Un document *contributions.txt :* Décrivez brièvement la contribution de chaque membre de l'équipe. Tous les membres sont censés contribuer au développement. Bien que chaque membre puisse effectuer différentes tâches, vous devez vous efforcer d'obtenir une répartition égale du travail. En particulier, tous les membres du projet devraient participer à la conception du projet et participer activement à la réflexion et à l'implémentation du code.

EVALUATION

Votre TP sera évalué sur les points suivants :

Critères :

1. Performance de votre modèle
2. Implantation correcte et efficace
3. Exécution du code sans exceptions
4. Qualité du code
5. Commentaires clairs et informatifs
6. Aspect novateur ; recherche à partir de l'état de l'art

CODE D'HONNEUR

Règle 1: Le plagiat de code est bien évidemment interdit.

Règle 2: Vous êtes libres de discuter des idées et des détails de mise en œuvre avec d'autres équipes. Cependant, vous ne pouvez en aucun cas consulter le code d'une autre équipe INF8460, ou incorporer leur code dans votre TP.

Règle 3: Vous ne pouvez pas partager votre code publiquement (par exemple, dans un dépôt GitHub public) tant que le cours n'est pas fini.