Soft Skills NER

Projet de Statistiques Appliquées

Note de Synthèse

Martin Bordes, Efflam Fouques Duparc, Victor Michel, Adrien Servière

Mai 2022

1 Présentation du projet

En collaboration avec Pôle Emploi, ce projet de Statistiques Appliquées est encadré par Morgane Hoffmann et Bruno Crépon. Il a pour sujet les *soft skills* dans les offres de travail proposées par l'agence Pôle Emploi. Selon la définition donnée par Pôle Emploi, les *soft skills*, également appelés compétences non-techniques ou compétences humaines, sont l'ensemble des manières d'agir et d'interagir dans un contexte professsionnel.

Pôle Emploi possède une très importante base de données comportant toutes les offres d'emploi et tous les CV postés sur leur site depuis près de 10 ans. Cela concerne des secteurs d'activités très variés, des postes très différents et des zones géographiques recouvrant l'ensemble du territoire français. Afin de pouvoir travailler sur les *soft skills*, Pôle Emploi nous a fourni une base de données contenant plus de 200 000 offres. Chaque ligne correspond à une offre et contient 150 features. La feature qui nous intéressera le plus dans notre étude est le texte libre écrit par l'entreprise afin de décrire le poste, les missions et le profil recherché.

L'objectif de notre projet est de trouver une façon d'extraire et de regrouper les compétences humaines des offres d'emploi pour comparer ces catégories à celles qui existent déjà à Pôle Emploi. Nous allons aussi estimer le "prix" de ces compétences en calculant la hausse de salaire qu'elles impliquent. Nous nous basons sur les textes descriptifs des offres pour notre étude.

2 Modèle

2.1 Description et préparation du modèle utilisé

Le modèle le plus efficace pour le type de tâches qui nous intéresse est un BERT (*Bidirectional Encoder Representations for Transformers*), un outil de NLP (*Natural Language Processing*) développé par Google AI en 2018. Comme nos données textuelles sont en français, nous travaillerons avec la version française de BERT : CamemBERT.

Le BERT reçoit en entrée des embeddings : vecteur unique associé à chaque mot de la phrase. On ajoute ensuite à ce vecteur la place du mot dans la phrase et le segment de phrase dans lequel le mot se situe. On place alors les vecteurs dans des encodeurs (entre 6 et 12 ici). Chaque vecteur passe dans un mécanisme d'attention et de position donnant un poids aux mots et aux relations entre les mots. Le BERT renvoie alors les parties qui ont une attention maximale.

La première étape de préparation du modèle est la labellisation d'offres d'emploi. Durant cette

étape, on sélectionne manuellement les *soft skills* dans un échantillon d'offres afin que notre modèle puisse ensuite s'entraîner sur cet échantillon et étendre ce qu'il aura appris de ces offres à l'ensemble de la base de données. Nous avons ainsi labellisé 1200 offres choisies en stratifiant par secteur d'activité. Ce travail a été réalisé sur Doccano, une application open-source.

2.2 Entraînement du modèle

L'entraînement de notre modèle comporte deux parties : le pré-entraînement (*pre-training*) et l'ajustement (*fine-tuning*). Ainsi, notre modèle pourra classifier binairement chaque token en *soft skill* ou non à partir des pharses tokenisées.

Afin de prendre en main le BERT, nous avons commencé par entraîner un modèle de MLM (*Masked-Language Modelling*) sur les 200 000 offres. Pour un BERT et comme souvent en NLP, la quantité de données prime sur le nombre d'époques, ce qui nous permet de n'utiliser que 3 époques. Cette étape permet de faire assimiler au BERT le type de langage très spécifique employé dans les descriptions d'offre. Le MLM consiste à masquer une fraction des tokens de chaque séquence pour ensuite essayer de prédire les tokens masqués grâce aux mots qui l'entourent.

Pour le *fine-tuning*, nous n'avons finalement pas utilisé le travail effectué dans la partie de *pre-training* avec le MLM à cause d'un problème lié au transfert des pondérations du pré-entraînement vers le *fine-tuning*. L'ajustement s'est alors fait sur un BERT de base déjà pré-entraîné, qui s'est avéré performant. Dans cette partie, on adapte le BERT au problème de la classification binaire des *soft skills* grâce aux offres précédemment labellisées.

3 Résultats

Nous avons utilisé pour estimer la pertinence des résultats de notre modèle des statistiques qui montrent l'absence de sur-apprentissage ou *overfitting* et de sous-apprentissage ou *underfitting*. Notre modèle se montre pertinent pour trouver presque tous les *soft skills* dans chaque annonce tout en évitant de comptabiliser comme compétence non-technique des mots inadéquats. Ainsi, le *recall* et la *precision* de notre modèle sont proches de 1, valeur maximale atteignable.

Dans cet exemple, nous avons pris au hasard une offre contenant 1024 tokens, c'est-à-dire un découpage de l'offre en 1024 entités. La matrice de confusion suivante montre la qualité de notre modèle sur cet exemple :

classe	0	1
0	1004	5
1	1	14

Enfin, pour les fonctions de *loss* et d'*accuracy*, on remarque que la "perte" de notre modèle diminue avec l'entraînement alors que sa précision augmente. Ces statistiques sont bonnes car elles indiquent que le modèle s'approche de la meilleure performance réalisable.

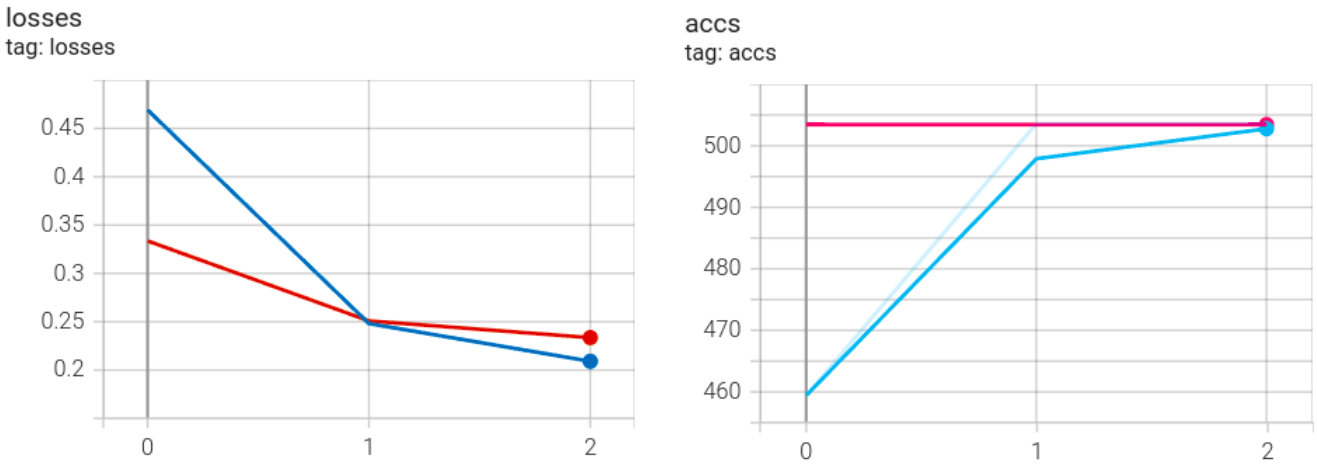

Figure 1: Fonctions de loss et d'accuracy du modèle

Pour permettre de visualiser de manière originale les résultats obtenus, nous avons construit une application Streamlit permettant d'accéder aux sorties de notre algorithme depuis n'importe quel espace de travail.

4 Analyse

4.1 Identification des grands groupes de compétences

Pôle Emploi a groupé en 14 catégories principales les *soft skills*. Notre objectif a été de vérifier la véracité de ces catégories. Nous avons donc utilisé plusieurs méthodes de *clustering* pour catégoriser les compétences interpersonnelles renvoyées par notre modèle.

Certaines méthodes se sont avérées inefficaces ou peu cohérentes dans leurs sorties. Par exemple, lorsque nous avons utilisé la méthode de clusterisation avec DBSCAN, nous obtenions comme sortie que 4 clusters qui ne nous ont pas paru pertinents.

Finalement, nous utilisons une méthode exploitant le calcul de similarité du *Word2Vec*. Nous

associons à chaque *soft skill* détecté la catégorie de compétences dont il est le plus proche d'après la similarité de *Word2Vec*. Ce clustering apparaît empiriquement bien plus cohérent. On compare ensuite la base de données initiale comprenant les compétences humaines renseignées directement dans l'offre, hors du texte, à la base à laquelle on a ajouté les compétences humaines détectés par notre modèle dans le texte et classé dans une des 14 catégories. Le BERT détecte ainsi des offres contenant un *soft skill* d'une certaine catégorie dans le texte mais non renseigné dans les features prévues pour les *soft skills*. Cela va de 3836 à 43688 compétences supplémentaires selon les catégories, sur le total de 200 000 offres.

4.2 Existe-t-il un "Wage premia" pour les *soft skills*

Nous avons choisi de régresser le salaire sur l'ensemble des compétences humaines pour estimer la différence de salaire causée par ces compétences. Cela montre que toutes les compétences humaines n'ont pas la même valeur mais que toutes sont significatives dans cette régression. Les différences de salaire induites par la plupart des *soft skills* sont minimes, de l'ordre de 0.1 à 0.5%. Cependant, deux compétences ressortent : la persévérance et le leadership. Leur présence dans une offre semble induire une augmentation de salaire de près de 2% et 3% respectivement, par rapport à des offres ne spécifaint pas la recherche de ces compétences. Ces valeurs ne paraissent pas absurdes car les métiers faisant appel au leadership et à la persévérance sont des métiers de direction ou d'expertise, les deux catégories de métier les plus rémunératrices.

5 Conclusion

La mission qui nous avait été donné par nos encadrants a pu être remplie grâce au BERT. En effet, notre modèle s'est avéré précis et adapté à notre étude. Cela nous a notamment permis d'analyser l'impact des *soft skills* sur le salaire après les avoir catégoriser. Une potentielle prochaine étape pourrait être d'approfondir notre analyse, par exemple en cherchant un moyen d'améliorer la catégorisation des *soft skills*.