BERT et autres Transformers pour décrypter un CV

Thème 6 : Intelligence artificielle

BEL-LETOILE Justine, jbel-letoile@hellowork.com, HelloWork, Pessac

Résumé : Les avancées récentes en Traitement automatique des langues (ou NLP) s'appuient souvent sur des réseaux de neurones de type Transformer, faisant bon usage de l'attention ou de l'apprentissage par transfert. Est-ce que ces méthodes sont applicables à un problème de reconnaissance d'entités nommées (ou NER) dans un contexte précis, celui de l'emploi ? Le Datalab du groupe HelloWork s'est intéressé à l'utilisation de tels réseaux pré-entraînés pour du parsing de CVs.

Mots clés : Intelligence Artificielle, NLP, NER, Bert, Transformer

1. Introduction

HelloWork (ex Jobijoba) utilise l'intelligence artificielle pour faciliter la mise en relation entre les candidats et les recruteurs. Avec CV Catcher, notre module de parsing de CVs, il s'agit par exemple d'extraire l'explicite et l'implicite contenu dans le texte des profils reçus. Une manière d'extraire l'explicite peut consister en une solution de recherche d'entités nommées, pour reconnaître les éléments clés d'un CV : nom, prénom, métiers, lieux, diplômes, etc.
L'état de l'art en NLP évolue très rapidement, et dernièrement la littérature sur le NER implique souvent une variation autour des modèles de type Transformer, pré-entraînés sur de larges corpus de texte. Comment appliquer ces méthodes à de l'analyse de CVs ?

2. Méthodologie

Une fois que l'on définit l'analyse du CV comme une tâche de NER et que l'on dispose d'un dataset orienté emploi, il reste plusieurs points à développer pour adapter les mécanismes généraux à notre cas d'usage. Qu'apportent les systèmes s'appuyant sur des représentations de type word embeddings ? Quelles catégories de modèles sont possibles pour ce problème ? Comment les intégrer à notre pipeline d'apprentissage et de traitement des CVs ?
Nous avons réalisé une comparaison des performances de plusieurs modèles pré-entraînés, à commencer par BERT, l'un des premiers à être démocratisé. Il s'agit de réfléchir au compromis entre qualité d'extraction des entités et temps passé sur l'entraînement et/ou l'inférence.

3. Originalité / perspective

A notre connaissance, il existe de nombreux exemples d'utilisation de modèles type BERT ou assimilés pour des problèmes de classification de séquences (classification de documents, sentiment analysis), de questions-réponses, ou encore de traduction, mais moins pour des travaux au niveau du token (à l'exception du part-of-speech tagging).
De plus, nous proposons une vue d'ensemble des modèles applicables au français.

- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.. CoRR, abs/1910.01108.
- Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D. & Sagot, B. (2020). CamemBERT: a Tasty French Language Model. ACL.