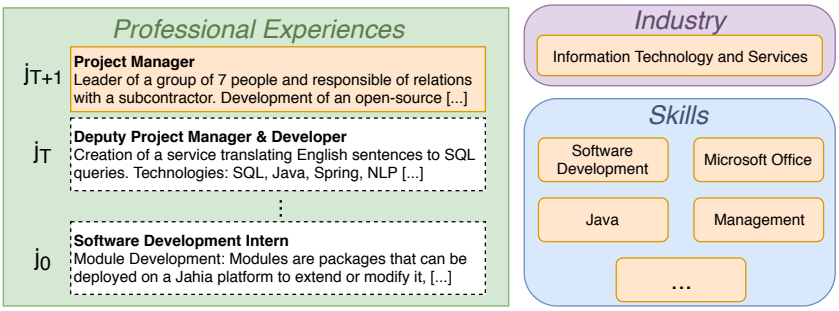
Analyse d'un corpus du CV, apprentissage de représentation sur des documents structurés et démêlage des facteurs explicatifs

Vincent Guigue et Clara Gainon de Forsan de Gabriac

December 2020

L'analyse des données textuelles a fortement évolué ces dernières années avec l'émergence des modèles de langues pré-entrainés [PNI+18, DCLT18]. Dans ce contexte, nous nous sommes intéressés à l'analyse d'un corpus de CV où chaque profil individuel est associé à une formation académique, des expériences professionelles, des compétences et un domaine industriel.



Les défis applicatifs sont nombreux : mieux structurer les jobs boards, aider les utilisateurs à remplir ou compléter les formulaires, suggérer des offres d'emploi à des personnes... Et, plus globalement, construire une représentation des entreprises du point de vue des employés qui la composent [DGG17, GdFdGGG20].

Sur le plan technique, la première question qui se pose est celle de l'adéquation des modèles de langue face à des données difficiles du fait des fautes d'orthographe mais surtout du style sténographique des descriptions. Le second challenge réside dans la structuration multi-échelles du document : les mots forment des expériences, les expériences forment un profil, les profils forment des entreprises. Les agrégations successives doivent permettre d'extraire des informations particulières pour représenter originalement les concepts de haut niveau sans tomber dans le sur-apprentissage.

L'idée de ce stage est de se focaliser sur un troisième défi : celui du démêlage des profils individuels et de l'apprentissage des descriptions de postes. On peut faire l'hypothèse qu'une description de poste correspond à différents éléments comme le secteur industriel (e.g. industrie automobile ou société de service en informatique), le service de rattachement (comptabilité, bureau d'étude), l'expérience (niveau hiérarchique) plus les spécificités liées à l'entreprise de rattachement et la manière d'écrire de l'utilisateur. L'enjeu du stage est de démontrer notre capacité à extraire ces différents aspects et apprendre des représentations pertinentes associées à un modèle génératif qui permette de trasnformer un intitulé en un autre à la manière de [LSS+19].

Contact :
vincent.guigue@lip6.fr

Références

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[DGG17] Charles-Emmanuel Dias, Vincent Guigue, and Patrick Gallinari. Passé, présent, futurs : induction de carrières professionnelles à partir de cv. In *CORIA*, pages 281–296, 2017.

[GdFdGGG20] Clara Gainon de Forsan de Gabriac, Vincent Guigue, and Patrick Gallinari. Resume : A robust framework for professional profile learning & evaluation. In *European Symposium on Artificial Neural Networks*, 2020.

[LSS+19] Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text rewriting. In *International Conference on Learning Representations*, 2019.

[PNI+18] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237, 2018.