Limiter l'impact des erreurs OCR sur les représentations distribuées de mots

Axel Jean-Caurant
axel.jean-caurant@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i) Université de La Rochelle, France

Cyrille Suire
cyrille.suire@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i) Université de La Rochelle, France

Vincent Courboulay
vincent.courboulay@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i) Université de La Rochelle, France

Jean-Christophe Burie
jean-christophe.burie@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i)
Université de La Rochelle, France

Les chercheurs en Humanités numériques intéressés par l'analyse de grands corpus textuels utilisent de nombreuses méthodes et outils issus de domaines informatiques comme le traitement du langage naturel (Piotrowski, 2012) ou l'analyse de réseaux (Lemercier, 2005). Des méthodes récentes fondées sur les réseaux de neurones présentent également un intérêt majeur. Word2Vec est une méthode qui a grandement facilité l'utilisation de tels modèles (Mikolov, 2013). Les différentes optimisations apportées permettent, très simplement, d'entraîner un modèle sur de grandes quantités de données en utilisant un simple ordinateur de bureau. Le code source a été largement diffusé et a rendu cette méthode très populaire, notamment parmi les chercheurs en Humanités numériques. Hamilton a par exemple montré l'intérêt de ces modèles pour analyser l'évolution de certains mots du langage au cours du temps (Hamilton, 2016). Ces méthodes peuvent également être utilisées à d'autres fins. En effet, de nombreux corpus utiles aux Humanités numériques sont issus de processus de reconnaissance de caractères (OCR). Malheureusement, ces processus génèrent très souvent des erreurs, en particulier quand les documents analysés sont de mauvaise qualité (documents anciens ou mal numérisés par exemple). Ces erreurs touchent notamment les entités nommées comme les noms de lieux ou de personnes, particulièrement intéressants pour les chercheurs (Gefen, 2015). Ces erreurs ont un impact majeur sur l'accès à l'information car elles peuvent empêcher d'accéder à toutes les occurrences d'un mot d'intérêt.

Dans ce poster, nous présentons la méthode que nous avons développée pour étendre l'usage de Word2Vec à l'identification des erreurs OCR dérivées d'entités nommées. Après avoir entraîné un modèle sur un corpus donné, chaque mot est associé à un vecteur représentatif. Il devient alors possible de comparer les vecteurs pour extraire des relations morphologiques ou sémantiques entre les mots. On peut par exemple calculer la distance cosinus qui sépare deux mots dans l'espace vectoriel du modèle. Si, au sein du corpus, ces mots apparaissent dans des contextes similaires, la distance qui les sépare sera faible. Or, une entité nommée, bien que mal reconnue par le processus OCR, apparaît souvent dans le même contexte que l'entité originale. En combinant cette distance, qui agit sur les vecteurs, avec une distance d'édition sur les mots, nous pouvons identifier des mots proches sémantiquement et qui possèdent beaucoup de caractères en commun. Cette analyse produit ainsi une liste de termes qui ont toutes les chances d'être des entités mal reconnues par le processus de reconnaissance de caractères.

Figure 1: Expérience menée par Bjerva et. al., qui présente les similarités entre différents personnages et quelques grands concepts. Plus une cellule est rouge, plus la similarité est importante.

Une fois les erreurs identifiées, il est possible de s'intéresser à une entité nommée particulière. Sur la base des résultats précédents, nous proposons la construction d'un nouveau vecteur associant le vecteur de l'entité originale et les vecteurs représentatifs des erreurs. Ce nouveau vecteur est le résultat de la combinaison linéaire des vecteurs du mot original et des erreurs OCR. Pour modérer l'importance des vecteurs dans la

combinaison, ces derniers sont pondérés selon le nombre d'occurrences du terme correspondant dans le corpus.

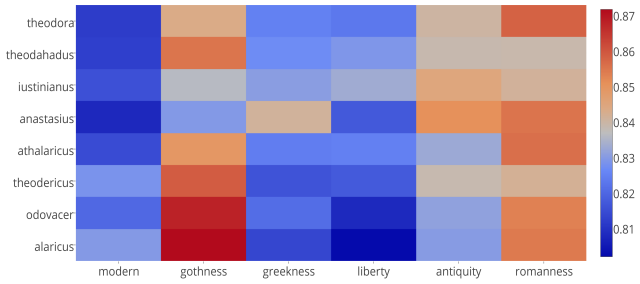

Figure 2: Reproduction de l'expérience menée par Bjerva et. al., avec notre modèle modifié.

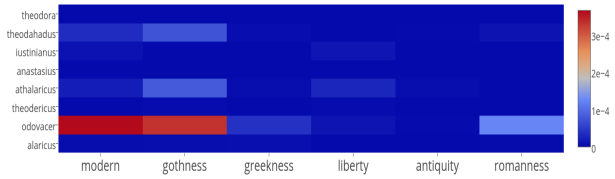

Figure 3: Comparaison des similarités Personne/Concept entre le modèle de Bjerva et. al. et notre modèle modifié. Chaque cellule représente la valeur absolue de la différence de similarité entre les deux modèles. Les cellules rouges présentent le plus de différences.

Nous avons expérimenté notre méthode en reproduisant l'expérience menée par Bjerva et. al. (Bjerva, 2015). Ces derniers se sont intéressés aux relations qu'entretiennent différentes personnalités du VI^{ème} siècle avec de grands concepts (Modernité, Liberté, Gothique, ...). Ils ont utilisé Word2Vec pour entraîner un modèle sur environ 11 000 textes latins, pour ensuite comparer les distances qui séparent les personnes des concepts dans l'espace vectoriel du modèle (voir figure 1). Nous avons utilisé notre méthode pour calculer, pour chaque personne d'intérêt, un nouveau vecteur représentatif prenant en compte les différentes erreurs OCR identifiées. Les distances entre personnes et concepts au sein de notre modèle modifié sont présentées dans la figure 2. Pour plus de clarté, les deux modèles sont comparés dans la figure 3. On peut par exemple observer qu'Odovacer, la personne pour qui les différences sont les plus grandes, est assez peu citée dans le corpus. Notre méthode a cependant identifié de nombreuses erreurs OCR qui ont révélé des informations inconnues au seul vecteur de l'entité originale.

La méthode présentée ici permet d'identifier de potentielles erreurs OCR sur les entités nommées au sein d'un corpus. La prise en compte de ces erreurs peut avoir un impact non négligeable sur le modèle et donc sur les analyses qui en découlent. Cela semble en particulier vrai pour les entités nommées peu présentes dans un corpus.

Bibliographie

Bjerva, J. and Praet, R. (2015). "Word Embeddings Pointing the Way for Late Antiquity." LaTeCH 2015: 53.

Gefen, A. (2015). Les enjeux épistémologiques des humanités numériques. Socio. La nouvelle revue des sciences sociales, 4: 61-74.

Hamilton, W. (2016). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1.

Lemercier, C. (2005). Analyse de réseaux et histoire. Revue D'histoire Moderne et Contemporaine(2): 88–112.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality." Advances in Neural Information Processing Systems. pp. 3111–3119

Piotrowski, M. (2012). Natural Language Processing for Historical Texts. (Synthesis Lectures on Human Language Technologies 17). San Rafael, CA: Morgan & Claypool.