# Mesurer la distance “linguistique” entre 2 corpus

## 1 Encadrants


- Mikaela Keller <mikaela.keller@univ-lille.fr>
- Remi Gilleron <remi.gilleron@inria.fr>


## 2 Mots clés

Machine Learning, Natural Language Processing, algèbre linéaire, programmation python

## 3 Contexte

Le plongement lexical (word embedding) [1] d'un mot m est un vecteur résumant la distribution des mots avec lesquels le mot m apparaît. Les plongements lexicaux sont utilisés avec succès dans beaucoup de problèmes de traitement automatique du langage. Une des raisons de ce succès est que ces vecteurs permettent d'obtenir une notion fine de distance entre mots. Étant donné que ces plongements sont le résultats de statistiques collectées sur un corpus particulier être capable de mesurer la distance entre plongements obtenus sur deux corpus différents nous permettrait d'ordonner des corpus. L'analyse diachronique du langage [2] (comment le langage évolue) ou la stylistique (qui a écrit le texte) sont des exemples de problèmes pour lesquels avoir une distance entre corpus pourraient être intéressant.

## 4 Problématique

Soit A une matrice de plongements lexicaux obtenue sur un corpus AA et B celle obtenue sur le corpus BB. Le problème de Procrustes orthogonal [3] consiste en une transformation de la matrice A par une matrice orthogonale R de façon a ce que A ressemble le plus possible à B.

$$R = argmin_{\Omega} \|\Omega A - B\|_F$$

Appliquer la transformation R permet une meilleure comparaison inter-corpus des vecteurs représentant les mots [2]. On voudrait savoir si l'écart irréductible $\|RA - B\|_F$ pourrait être utilisé pour caractériser la distance entre corpus.

# 5 Travail demandé

- Une étude empirique: simulation ou/et collecte de corpus plus ou moins différents et calcul de l'écart
- Peut-on prouver formellement que ce serait ou non une bonne mesure?
- Bibliographie d'autres méthodes d'alignement que le Procrustes par exemple: [4]

# 6 Bibliographie


- [1] Improving Distributional Similarity with Lessons Learned from Word Embeddings, ACL 2015
- [2] Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change, ACL 2016
- [3] A generalized solution of the orthogonal Procrustes problem, Psychometrika 1966
- [4] Manifold Alignment [`http://www-anw.cs.umass.edu/legacy/pubs/2011/wang_k_m_11.pdf`]