Projet 2 DS Lab

Plongements lexicaux et un cas d'usage:

la traduction automatique

Eva BOUBA, Meriem DEHMAS, Matthieu SERFATY

Plongements lexicaux (Word Embeddings)

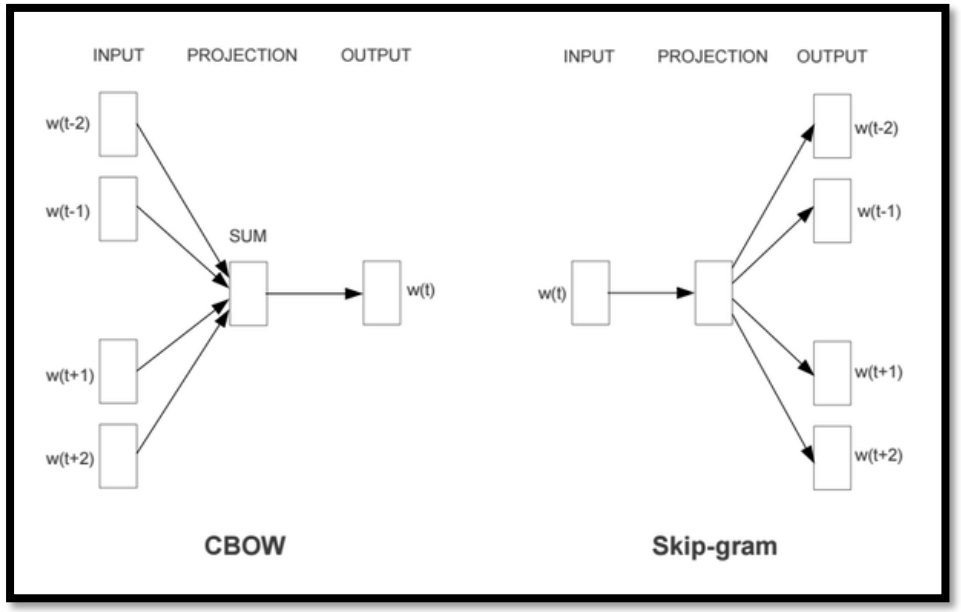
- **Objectif** : Obtenir une représentation vectorielle des mots qui capture leur sémantique.
- **Principe** : des mots de deux langues différentes apparaissent dans les mêmes contextes.
- Ces représentations sont les poids de couches de

réseaux de neurones, tels que:

CBOW ⟶ Permet de prédire un mot en fonction d'un contexte.

Skip-gram ⟶ Permet de prédire un contexte en fonction d'un mot.

BERT/ camemBERT ⟶ Entrainé de manière non supervisée sur diverses tâches d'NLP.



Traduction automatique

- **Traduction** $\longrightarrow$ Associer le vecteur d'un mot source au vecteur de sa traduction.

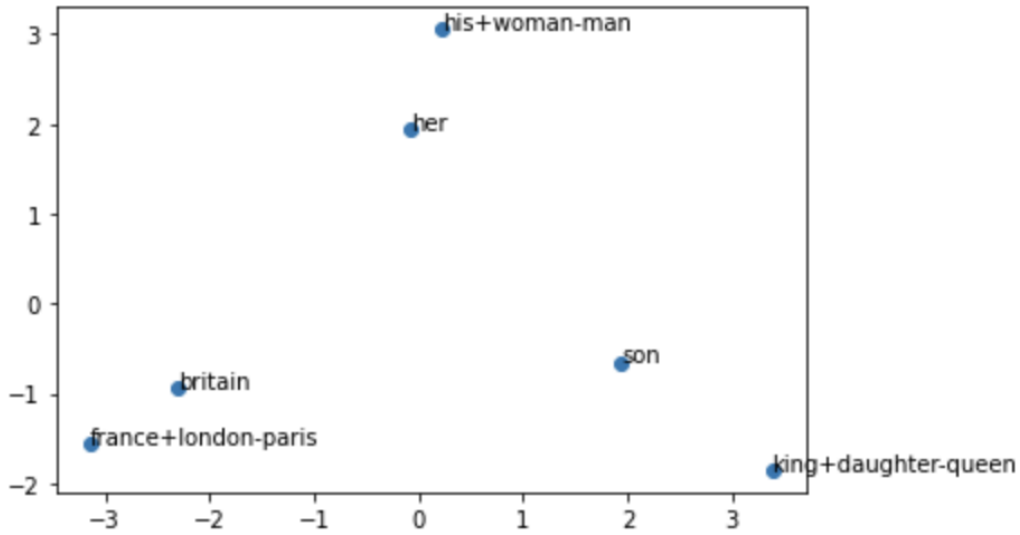


Figure 2, Propriété de composabilité des embeddings,

Astuce : Les mots de différentes langues présentent des structures similaires.

$\downarrow$

Transformation linéaire

$$\min_{W \in \mathbb{R}^{d_2 \times d_1}} \sum_i \|W \cdot x_i - y_i\|^2$$

Traduction supervisée

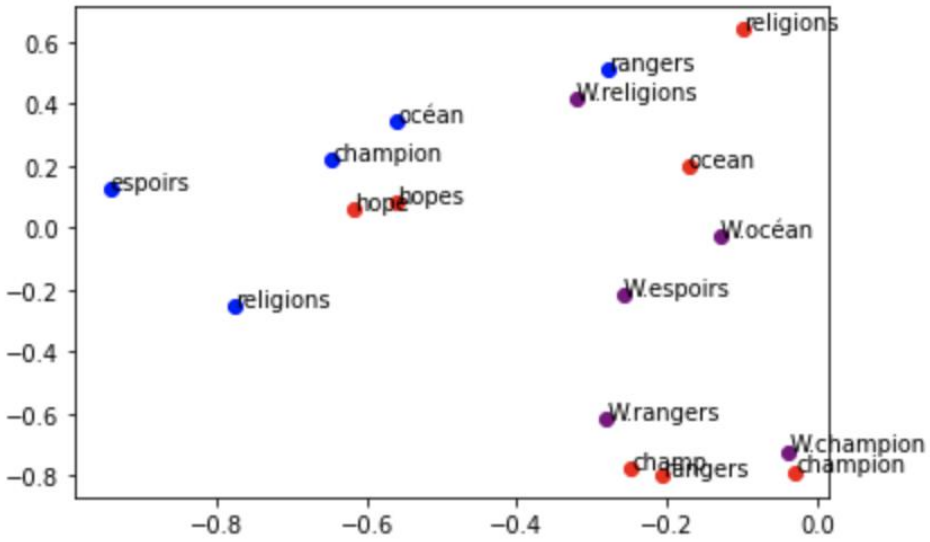
- Deux cas possibles :
  - Connaissance d'un dictionnaire.
  - Il existe des homogrammes dans les deux langues.
- Apprentissage grâce au dictionnaire

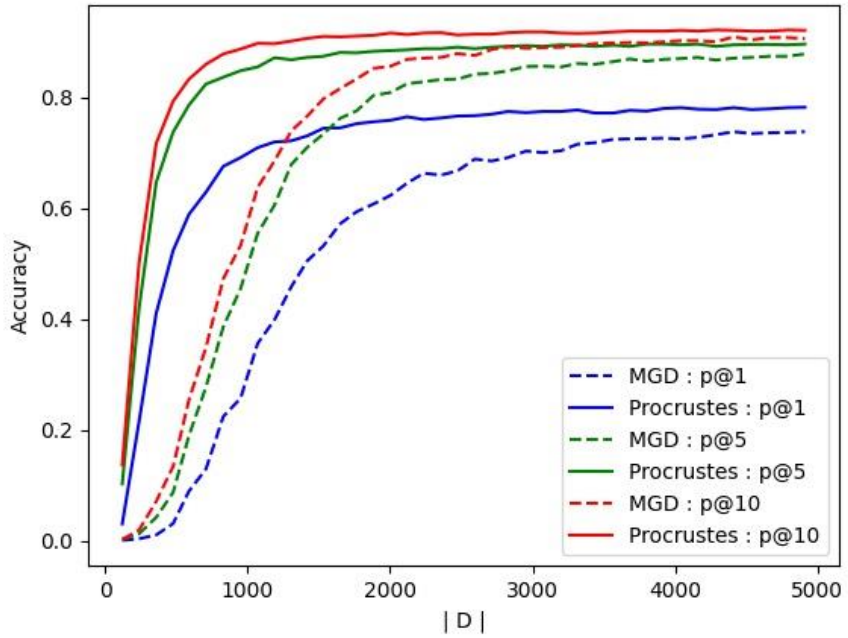
$$W^* = \underset{W \in \mathbb{R}^{d_2 \times d_1}}{\operatorname{argmin}} \sum_{i \in \mathcal{D}} \|W \cdot x_i - y_i\|^2$$

- Méthode Procruste itératif : rajout d'une contrainte d'orthogonalité

$$W^* = \underset{W \in \mathbb{R}^{d \times d}, WW^T = I_d}{\operatorname{argmin}} \|WX - Y\|^2 = UV^T; USV^T = SVD(X^TY)$$

Traduction supervisée

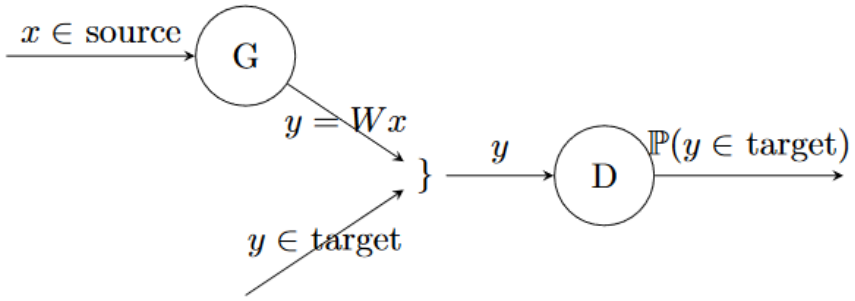




P@x : On considère qu'une prédiction est juste si la traduction fait partie des x plus proches voisins.

Traduction non-supervisée

- Utilisation d'un réseau adversarial génératif .
  - Mettre en compétition deux réseaux : un GÉNÉRATEUR et un DISCRIMINANT.



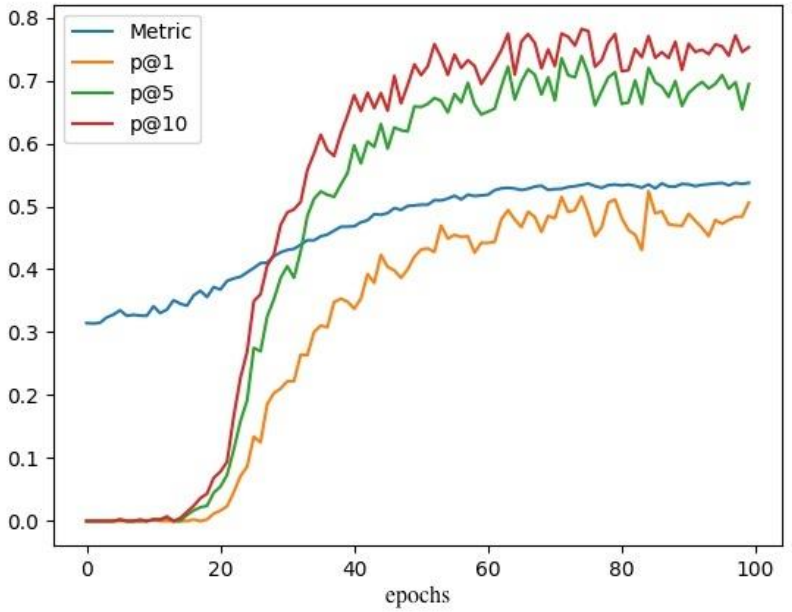


Figure 5. Une « métrique non supérvisée » et corrélée avec la précision de la traduction est définie pour évaluer la performance du réseau.

Résultats des différents traducteurs

	fr→en			en→fr			fr→es			fr→de		
	p@1	p@5	p@10	p@1	p@5	p@10	p@1	p@5	p@10	p@1	p@5	p@10
MGD	0.737	0.878	0.909	0.745	0.887	0.911	0.796	0.916	0.940	0.700	0.869	0.906
Procuste	0.780	0.896	0.921	0.788	0.905	0.931	0.819	0.937	0.955	0.753	0.895	0.925
GAN	0.652	0.813	0.847	0.686	0.846	0.849	0.677	0.835	0.868			