Méthodes d'évaluation pour la substitution de vecteurs de mots

Stanislas Morbieu [1,2], François Role [1], Mohamed Nadif [1]

[1] LIPADE, Université Paris Descartes
[2] Kernix

XXVI èmes Rencontres de la Société Francophone de Classification

Plan

Représentation vectorielle de mots

Mots hors vocabulaire

Similarité entre les vecteurs originaux et les substituts

Comportement sur des tâches de similarité et d'analogie

Idée principale

"You shall know a word by the company it keeps." John R. Firth (1957)

Idée principale

"You shall know a word by the company it keeps." John R. Firth (1957)

Je visite la place_Stanislas à Nancy .

mot cible
contexte

Je visite la place_Stanislas à Nancy .

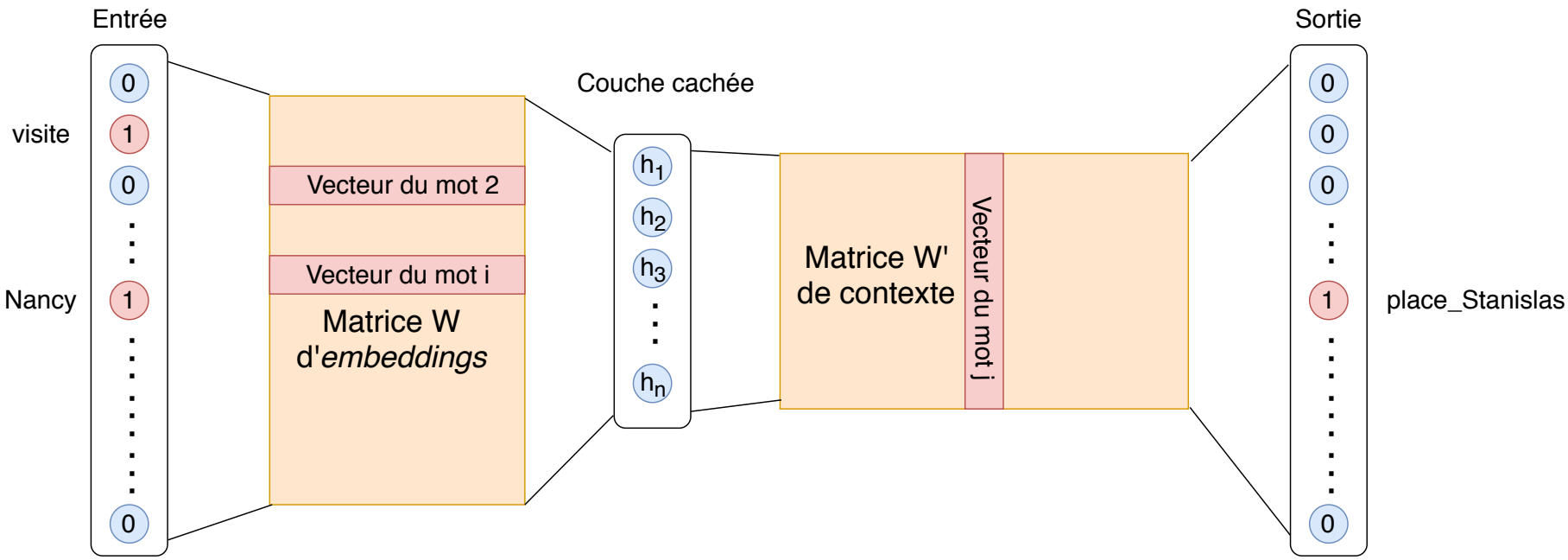


Figure 1: Word2vec CBOW [Mikolov 2013]

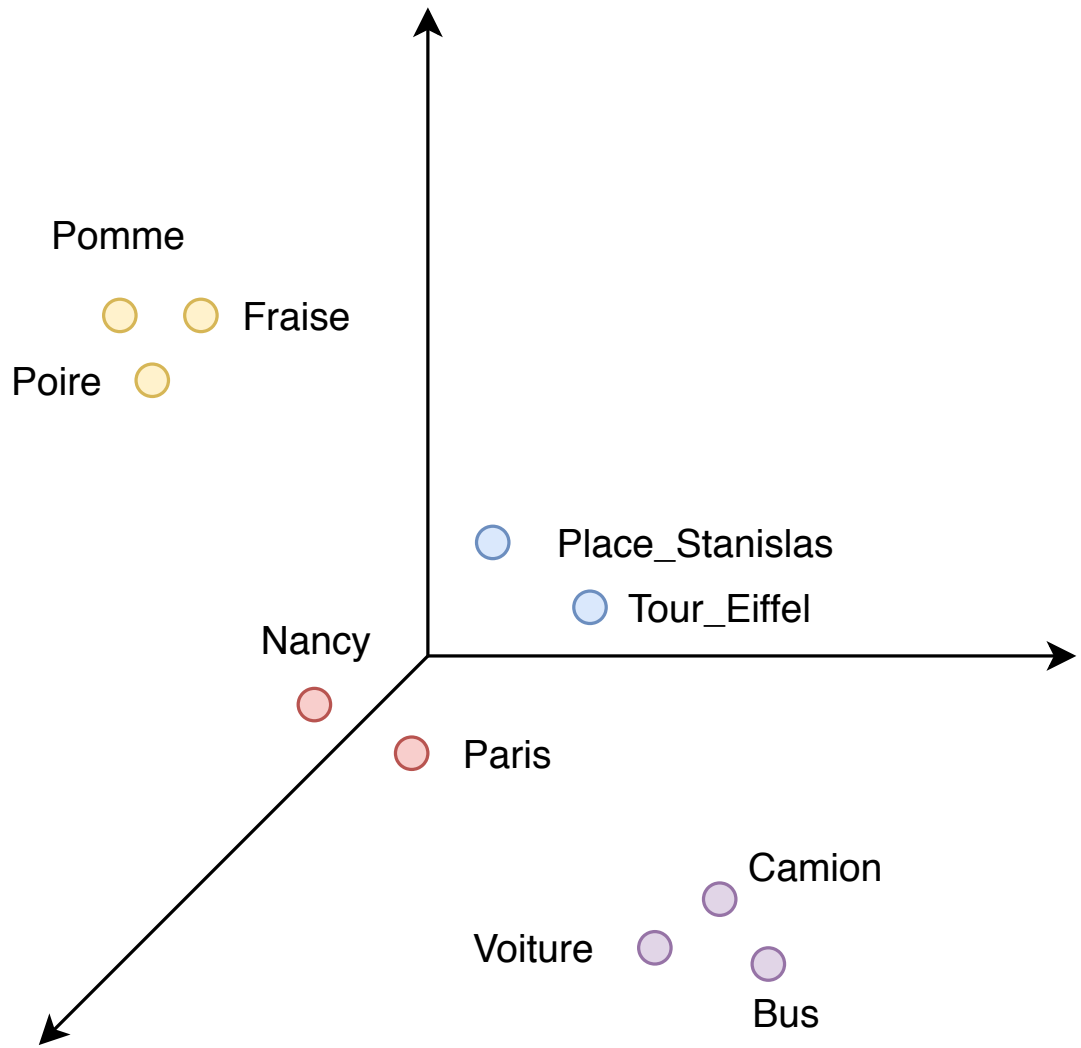


Figure 2: Représentation des mots dans l'espace vectoriel.

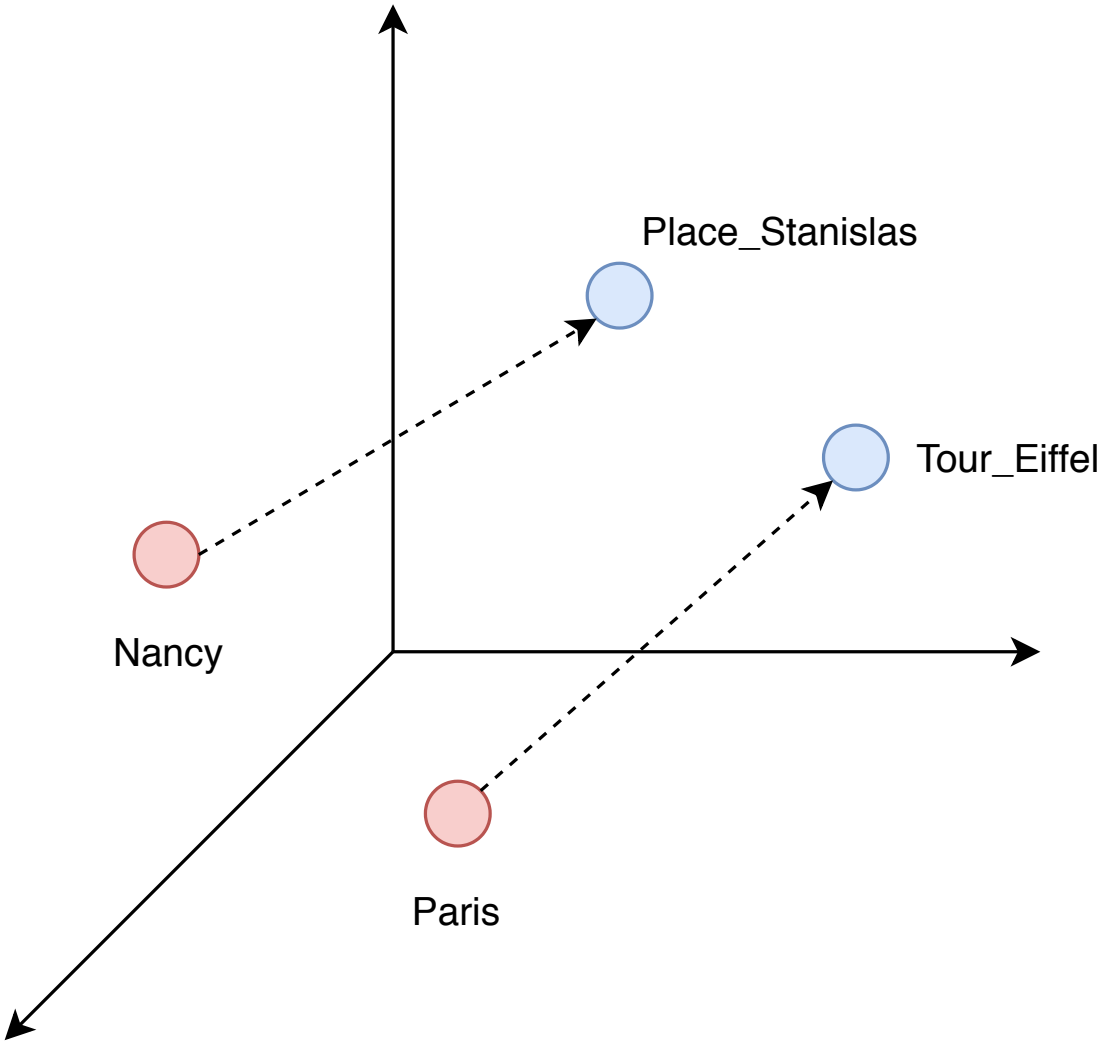
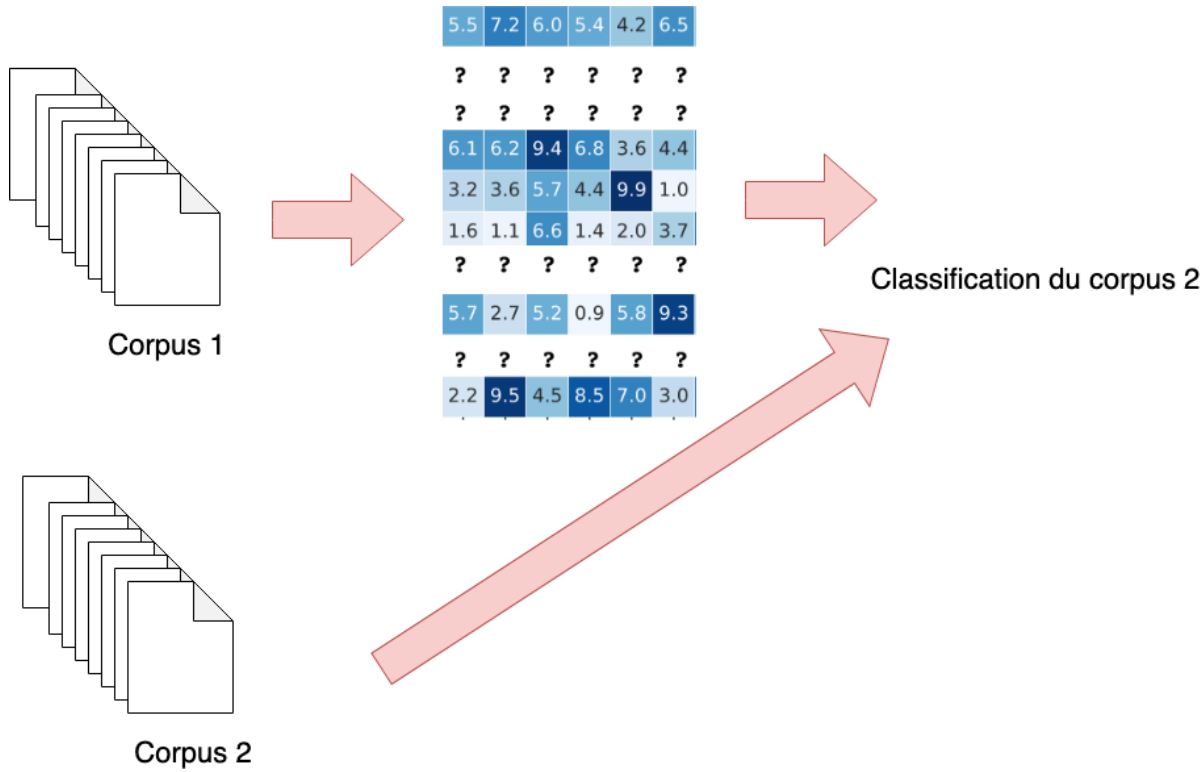


Figure 3: Représentation des analogies dans l'espace vectoriel.



Ordre de grandeur

Près de 61% des mots du jeu de données NG20 [a] sont absents dans le modèle Word2vec entraîné sur le corpus Google News.

[a] http://qwone.com/~jason/20Newsgroups/

Gestion des mots hors vocabulaire

Différentes méthodes :

– Initialiser les vecteurs inconnus avec des **zéros ou** des valeurs **aléatoires**.

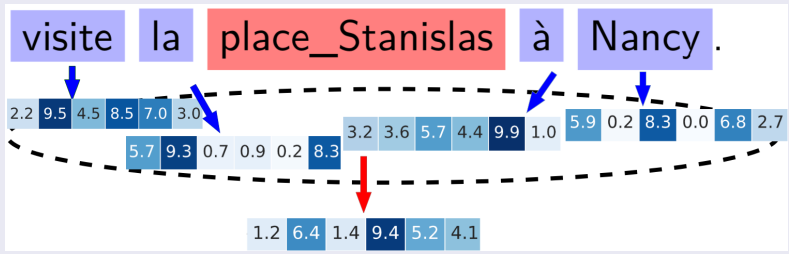
$$[?, ?, ?, ..., ?] \rightarrow [0, 0, 0, ..., 0]$$

Gestion des mots hors vocabulaire

Différentes méthodes :

- Initialiser les vecteurs inconnus avec des **zéros ou** des valeurs **aléatoires**.
- Prendre le **centroïde des vecteurs des mots** d'une fenêtre **de contexte**.

$$[?, ?, ?, ..., ?] \rightarrow [0, 0, 0, ..., 0]$$

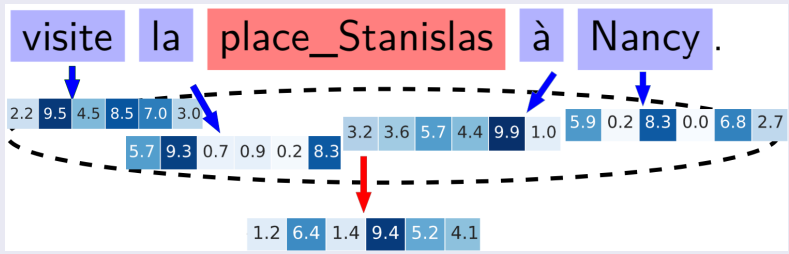


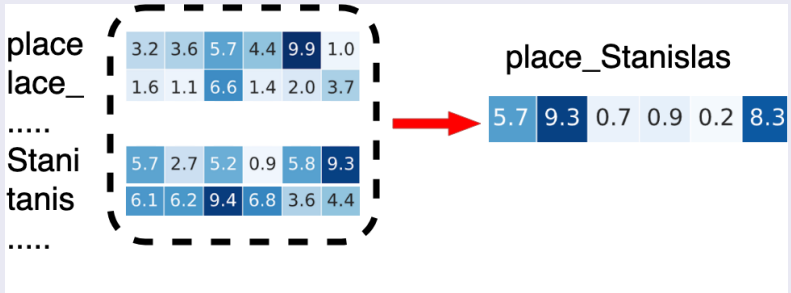
Gestion des mots hors vocabulaire

Différentes méthodes :

- Initialiser les vecteurs inconnus avec des **zéros ou** des valeurs **aléatoires**.
- Prendre le **centroïde des vecteurs des mots** d'une fenêtre **de contexte**.
- Calculer des **vecteurs de sous-mots** (vecteurs de caractères ou n-grams) et les assembler pour former un vecteur de substitution. [Bojanowski 2017]

$$[?, ?, ?, ..., ?] \rightarrow [0, 0, 0, ..., 0]$$



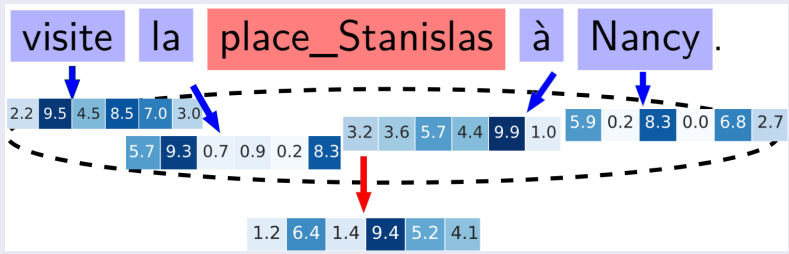


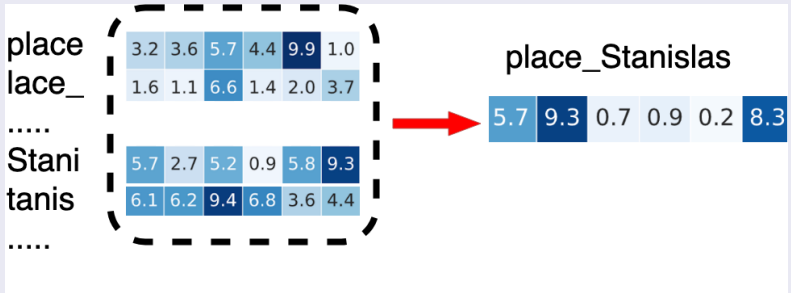
Gestion des mots hors vocabulaire

Différentes méthodes :

- Initialiser les vecteurs inconnus avec des **zéros ou** des valeurs **aléatoires**.
- Prendre le **centroïde des vecteurs des mots** d'une fenêtre **de contexte**.
- Calculer des **vecteurs de sous-mots** (vecteurs de caractères ou n-grams) et les assembler pour former un vecteur de substitution. [Bojanowski 2017]
- Exploiter des **données auxiliaires** (terminologies, thésaurus, bases de données lexicales telles que WordNet, etc.). [Pilehvar 2016]

$$[?, ?, ?, ..., ?] \rightarrow [0, 0, 0, ..., 0]$$





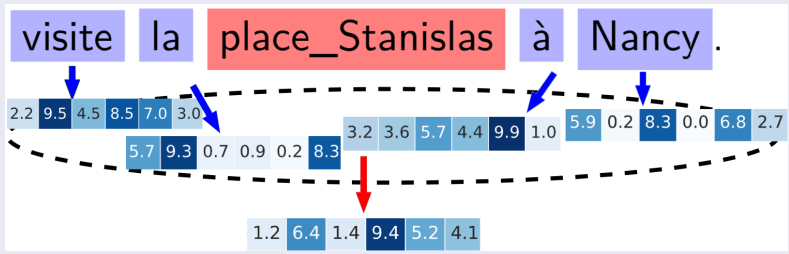
Gestion des mots hors vocabulaire

Différentes méthodes :

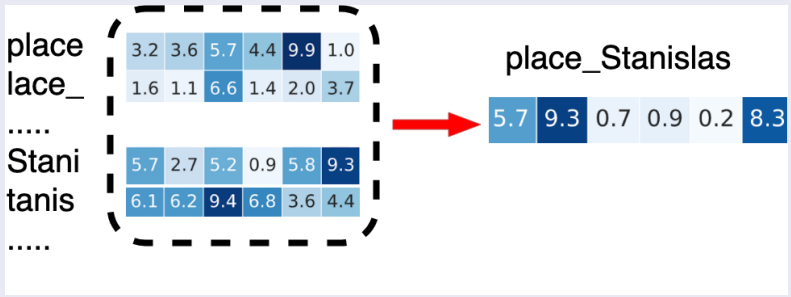
- Initialiser les vecteurs inconnus avec des **zéros ou** des valeurs **aléatoires**.

$$[?, ?, ?, ..., ?] \rightarrow [0, 0, 0, ..., 0]$$

- Prendre le **centroïde des vecteurs des mots** d'une fenêtre **de contexte**.



- Calculer des **vecteurs de sous-mots** (vecteurs de caractères ou n-grams) et les assembler pour former un vecteur de substitution. [Bojanowski 2017]



- Exploiter des **données auxiliaires** (terminologies, thésaurus, bases de données lexicales telles que WordNet, etc.). [Pilehvar 2016]

Évaluation

Évalution des vecteurs de mots :

- **sur des tâches en aval** : classification de documents (dont analyse de sentiment), reconnaissance d'entités nommées, étiquetage morpho-syntaxique, traduction automatique, etc.
- **intrinsèque** : tâches de similarité et d'analogie.

$\rightarrow$ Évaluation intrinsèque des méthodes de substitution.

Évaluation

- **Corpus** : text8 [1] pour l'entrainement et le calcul des substituts.
- **Paramètres** : par défaut de FastText et l'implémentation de Gensim pour Word2vec. Fenêtre de 2 pour calculer des centroïdes.

[1] http://mattmahoney.net/dc/textdata

Similarités original - substitut

Objectif

Mesurer à quel point un vecteur construit (le substitut) est proche du vecteur réel du mot (original).

Méthode

- Sélection aléatoire d'un ensemble de mots
- Pour chaque mot :
  - calcul du vecteur de substitution
  - similarité cosinus entre le vecteur de substitution et l'original
- Statistiques descriptives de la distribution des similarités cosinus

Similarités original - substitut

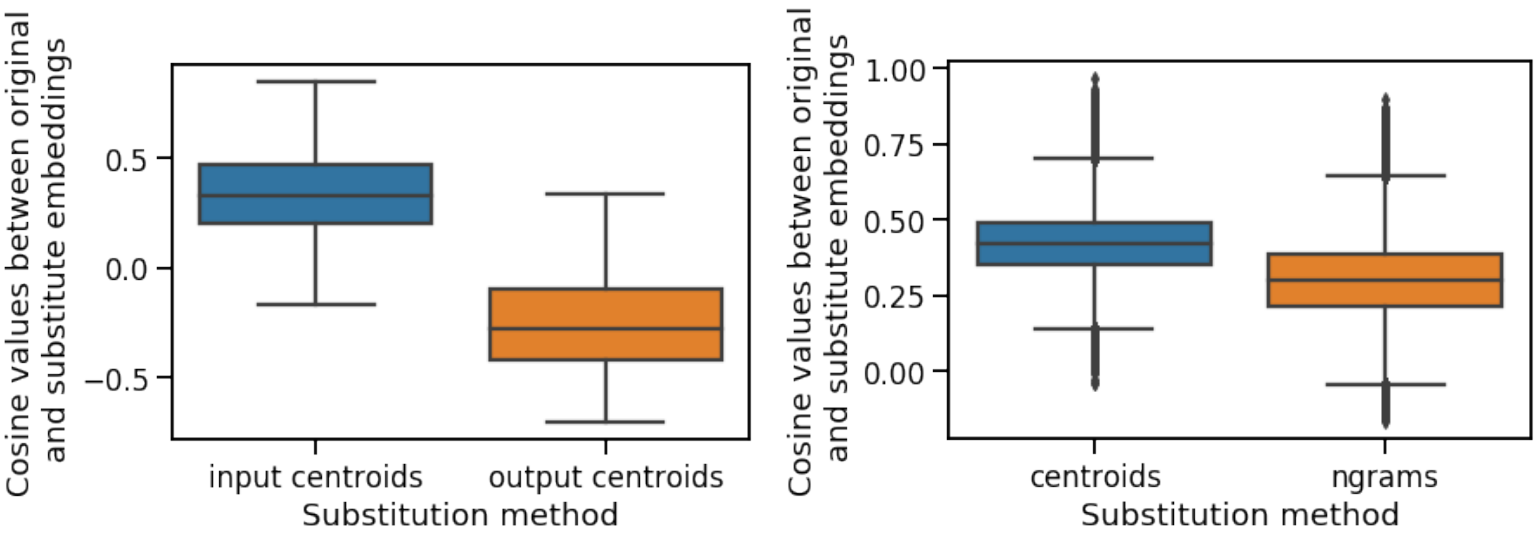


Figure 4: Similarité cosinus entre les vecteurs substituts et originaux pour Word2vec (gauche) et FastText (droite).

Similarités paire d'originaux - paire de substituts

Objectif

Mesurer à quel point deux mots dont les représentations vectorielles sont proches sont aussi proches dans l'espace des substituts.

Méthode

- Sélection aléatoire d'un ensemble de paires de mots
- Pour chaque paire de mots :
  - calcul des vecteurs de substitution
  - calculer S_1, le cosinus entre les deux vecteurs originaux
  - calculer S_2, le cosinus entre les deux vecteurs de substitution
  - calculer $S_1 - S_2$
- Statistiques descriptives de la distribution de $S_1 - S_2$

Similarités paire d'originaux - paire de substituts

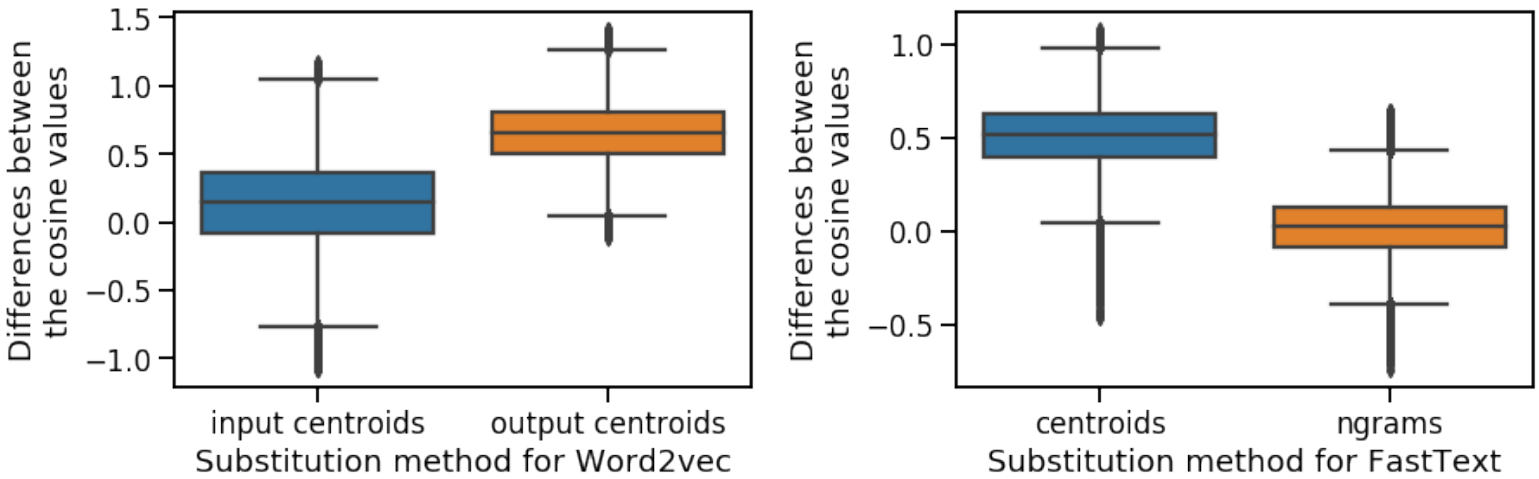


Figure 5: Différence entre la similarité cosinus obtenue avec les vecteurs originaux et celle obtenue avec les substituts, pour Word2vec (gauche) et FastText (droite).

Tâches de similarité

À chaque paire de mots est associé un score de similarité :

computer	keyboard	7.62
Jerusalem	Israel	8.46
planet	galaxy	8.11

Évaluation

Score de **corrélation de rang** entre le classement donné par l'évaluation humaine et celui donné par les vecteurs.

Tâches de similarité

Table 1: Scores de similarité quand les vecteurs de chaque paire sont issus du même espace : celui des vecteurs originaux ou celui des substituts.

classement 1	classement 2	wordsim	mturk	mc	rg
w2v_wc	humain	0.3501	0.1941	**0.3658**	**0.2654**
w2v_cc	humain	**0.3936**	**0.2236**	0.3562	0.2562
w2v_wc	w2v	**0.6551**	**0.4879**	**0.7018**	**0.6479**
w2v_cc	w2v	0.6212	0.4705	0.6155	0.5615
ft_ngrams	humain	0.3280	**0.4036**	**0.5000**	**0.4628**
ft_wc	humain	**0.3393**	0.2554	0.4760	0.4243
ft_ngrams	ft	0.4701	**0.4962**	0.5452	**0.5779**
ft_wc	ft	**0.5332**	0.4506	**0.7810**	0.5430

Tâches de similarité

Table 2: Scores de similarité quand, pour chaque paire, les deux vecteurs sont l'original et le substitut.

classement 1	classement 2	wordsim	mturk	mc	rg
w2v_wc / w2v	humain	**0.3678**	**0.2904**	**0.2488**	**0.3740**
w2v_cc / w2v	humain	0.2652	0.2623	0.1640	0.0992
w2v_wc / w2v	w2v	**0.6009**	**0.5443**	**0.6654**	**0.5996**
w2v_cc / w2v	w2v	0.4855	0.4033	0.5167	0.1142
ft_wc / ft	humain	0.3568	0.3535	0.4613	0.4717
ft_ngrams / ft	humain	**0.4346**	**0.4673**	**0.5100**	**0.5089**
ft_wc / ft	ft	0.5919	0.6209	0.6275	0.7108
ft_ngrams / ft	ft	**0.6628**	**0.6977**	**0.8251**	**0.7613**

Tâches d'analogie

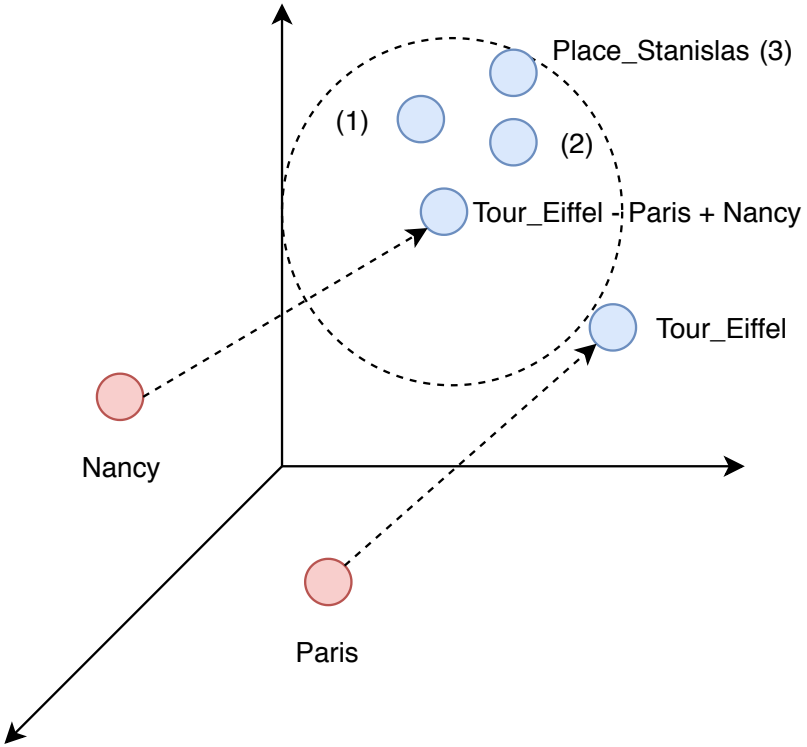


Figure 6: Analogies dans l'espace vectoriel.

Une tâche (sur 14) : ensemble de questions $w_1 - w_2 + w_3 = ?$

Score par question : nombre de vecteurs plus proches de $w_1 - w_2 + w_3$ que du vecteur cible.

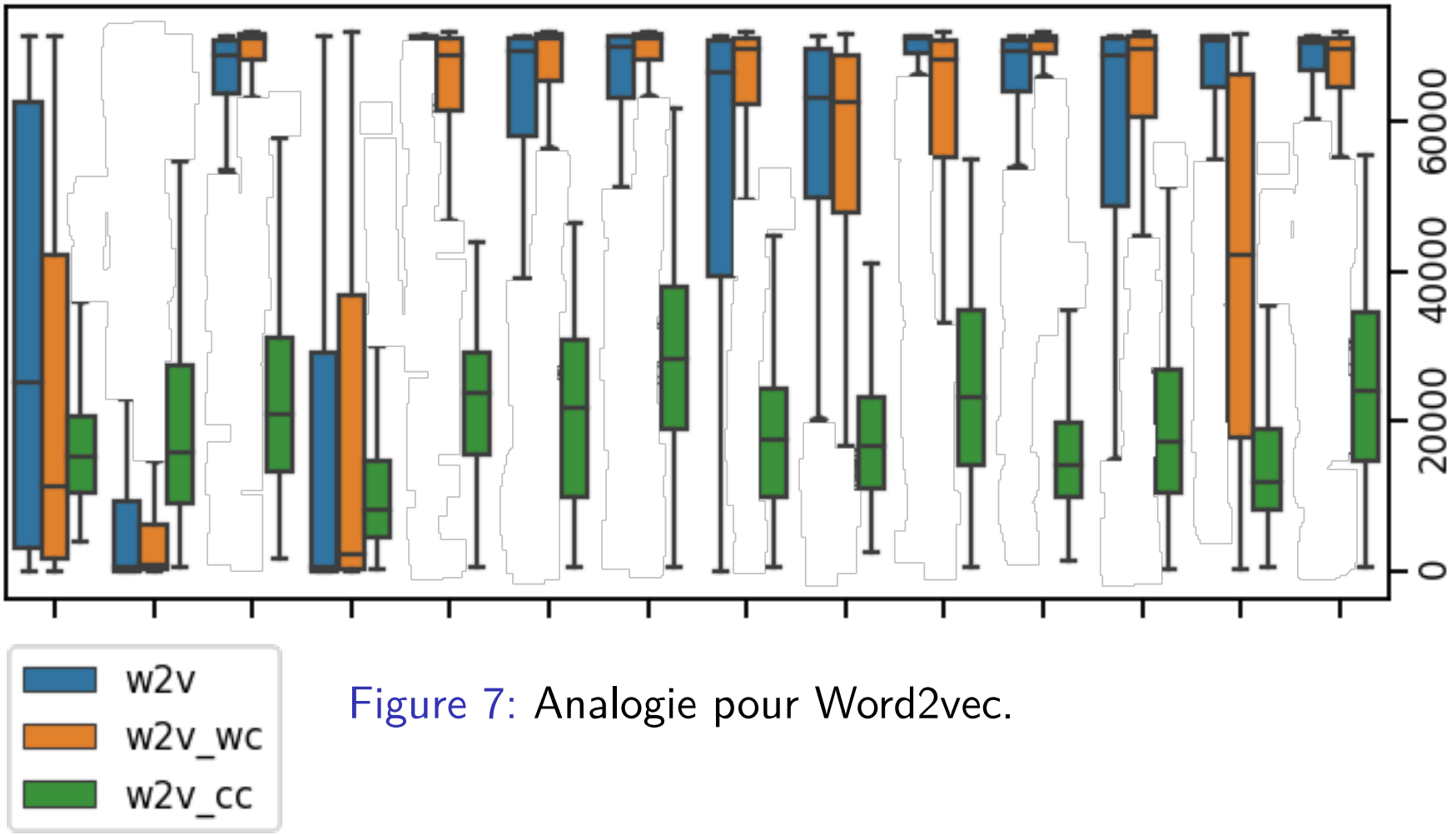


Figure 7: Analogie pour Word2vec.

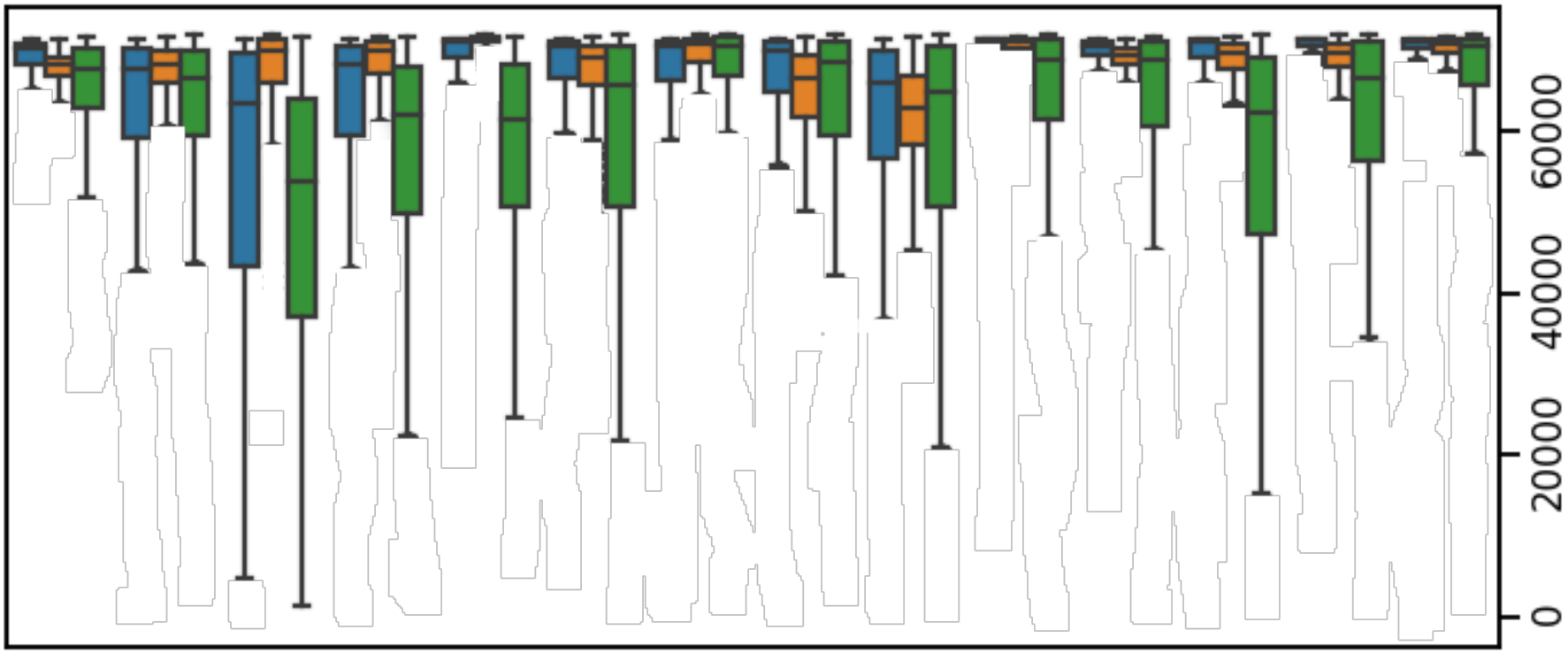


Figure 8: Analogie pour FastText.

Conclusion

- Pas de méthode universellement meilleure.
- Résultats variables sur les différentes tâches :
  - Similarité : pour Word2vec, la matrice d'entrée donne de meilleurs résultats que la matrice de sortie.
  - Analogie :
    - pour Word2vec, la matrice de sortie donne des résultats avec moins de variabilité que la matrice d'entrée.
    - pour FastText, la méthode intégrée donne de meilleurs résultats au prix d'une plus grande variabilité.

Perspectives

Utiliser des méthodes d'ensemble.

Bibliographie

- Bojanowski P. et al. Enriching word vectors with sub-word information. *TACL* 2017.
- Mikolov T. et al. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 2013.
- Pilehvar M. et al. Inducing embeddings for rare and unseen words by leveraging lexical resources. *ACL* 2017.
- Chu, C. et S. Kurohashi. Paraphrasing out-of-vocabulary words with word embeddings and semantic lexicons for low resource statistical machine translation. *LREC* 2016.

Précisions pour FastText

Deux méthodes de substitution :

- Centroide des mots du contexte (comme pour Word2vec)
- Méthode intégrée : à partir des représentations de ngrams

Entraînement

Les mots étudiés sont remplacés par le hash du mot :

- Une mise à jour des poids lorsque le mot est rencontré (représenté par le hash) ne met pas à jour les poids de ses sous-mots réels.
- Le hash du mot est suffisamment long pour que FastText puisse opérer (assez de sous-mots).