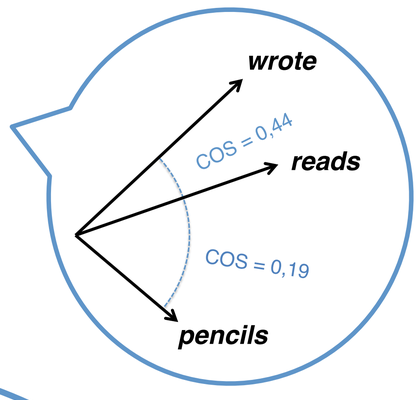
La séparation des composantes lexicale et flexionnelle des vecteurs de mots

François Lareau, Gabriel Bernier-Colborne et Patrick Drouin
Observatoire de linguistique Sens-Texte

Université de Montréal

problématique

En sémantique distributionnelle, le sens des mots se modélise par des vecteurs qui représentent leur distribution en corpus. Le cosinus de l'angle entre deux vecteurs indique la similarité sémantique des mots qu'ils modélisent.

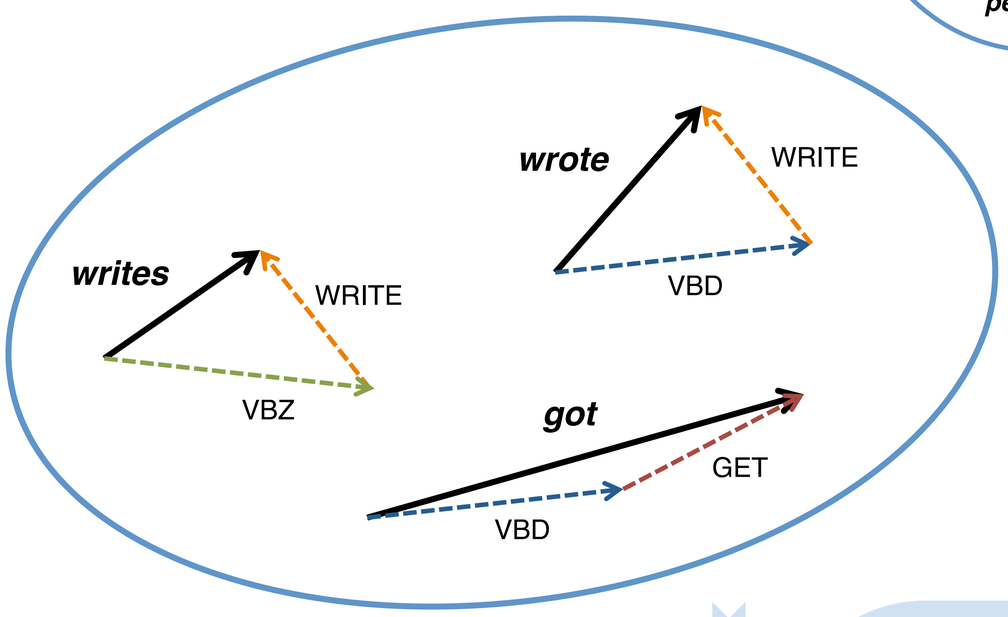


On observe des aberrations:

Mot 1	Mot 2	Similarité
seemed	*seems*	0,72
appeared	*appears*	0,66
seemed	*appeared*	0,72
seems	*appears*	0,81

Comment éliminer le bruit morphosyntaxique pour comparer les sens lexicaux?

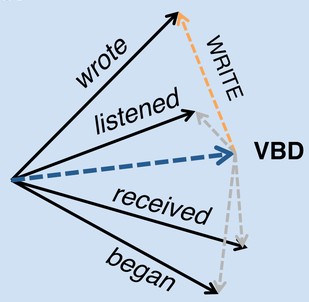
En décomposant les vecteurs!



méthodologie

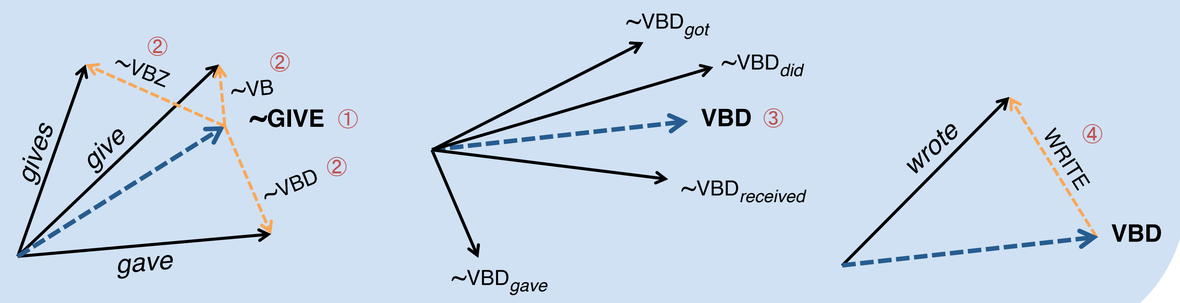
Méthode A

On fait la moyenne des vecteurs des formes associées à la même flexion pour obtenir la composante flexionnelle, puis on la soustrait de chaque vecteur pour en connaître la composante lexicale.



Méthode B

1. On fait la moyenne des vecteurs des trois formes associées à un même vocable pour obtenir une approximation de leur composante lexicale.
2. On soustrait du vecteur de chaque forme sa composante lexicale approximative pour obtenir une approximation de sa composante flexionnelle.
3. On fait la moyenne des composantes flexionnelle approximatives.
4. On obtient par soustraction les vecteurs lexicaux.



Deux jeux de données
20 verbes choisis
762 verbes non choisis

Les plus proches voisins des vecteurs flexionnels sont-ils des verbes qui portent cette flexion?

VB	VBZ	VBD
OPTION_ONE	*sees*	*moved*
tofind	*creates*	*arrived*
Bedbugs_Bite_Act	*finds*	*turned*
through_emergency_recapitalizion	*introduces*	*returned*
outguess_God	*initiates*	*escorted*
contacting_Melissa_Medalie	*gets*	*shaken_awake*
Yourself_Loan_Modification	*uses*	*pushed*
Eat_Spaghetti_Dinner	*pushes*	*chased*
Tiger_Wear_Necktie	*sustains*	*hauled*
unleash_cyberattacks	*interprets*	*untouchable_Gio_Gonzalez*

(vecteurs flexionnels calculés sur 762 verbes avec la méthode B)

évaluation

Les vecteurs lexicaux obtenus à partir des trois formes fléchies d'un même vocable sont-ils plus proches les uns des autres que les vecteurs originaux?

Expérience	Gain	CTRL
B-20-**20**	**0,140**	-0,035
B-762-**20**	0,086	-0,040
A-20-**20**	0,077	-0,017
A-762-**20**	0,050	-0,044
B-762-762	0,047	-0,048
A-20-762	0,013	-0,048
A-762-762	0,009	-0,049
B-20-762	0,008	-0,047

On applique le même traitement à des vecteurs choisis au hasard dans le modèle

La similarité moyenne passe de 0,634 à 0,774

Mot	Gain	CTRL
be / is / was	0,192	-0,028
get / gets / got	0,190	-0,033
give / gives / gave	0,181	-0,026
bring / brings / brought	0,171	-0,010
have / has / had	0,167	-0,024
receive / receives / received	0,159	-0,024
ask / asks / asked	0,158	-0,063
begin / begins / began	0,156	-0,034
speak / speaks / spoke	0,155	-0,050
send / sends / sent	0,148	-0,017
hear / hears / heard	0,144	-0,057
meet / meets / met	0,136	-0,032
tell / tells / told	0,132	-0,065
become / becomes / became	0,131	-0,015
continue / continues / continues	0,131	-0,034
follow / follows / followed	0,127	-0,017
understand / understands / understood	0,101	-0,035
seem / seems / seemed	0,099	-0,010
write / writes / wrote	0,085	-0,072
***live** / **lives** / lived*	0,036	-0,046
Moyenne	**0,140**	**-0,035**

VB = forme verbale nue
VBD = forme verbale en *–ed*
VBZ = forme verbale en *–s*

Article Affiche

VOIR L'ARTICLE POUR LES RÉFÉRENCES