Voisins lexicaux en contexte

Une mesure de la vision qu’à chaque occurrence des différents sens possibles

Sandrine Ollinger — 21 octobre 2025

Contexte scientifique

Rappel

- Projet ALUMCoCo
- enquête en ligne
- 420 contextes +/- 1 phrase par contexte
- 1 occurrence à annoter en sens par contexte
- 21 vocables (7 à 4 sens, 7 à 3 sens, 7 à 2 sens) décrits dans le RL-fr

Exemple

Question GS401213 - contexte c014

*Lisez ce court extrait (7/13) :

Connaître ou reconstruire très rapidement les résultats des tables d'**addition** (de 1 à 9) et les utiliser pour calculer une somme, une différence, un complément.

Quel est le sens d'**addition** dans ce texte ?

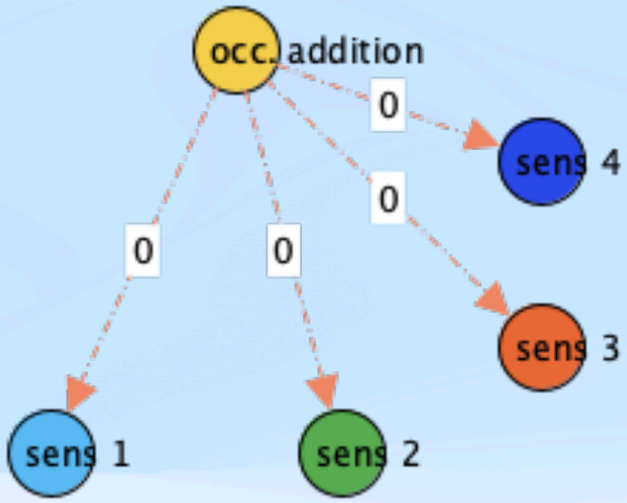
Si aucun sens ne vous convient, choisissez le plus proche.

1. Opération arithmétique symbolisée par le signe +. [*Il pose ses additions dans son cahier de brouillon.*]
2. Action d'ajouter. [*Le tissu est rendu plus souple par addition d'élasthanne.*]
3. Ce qu'on ajoute à quelque chose ; ajout. [*Ces additions calcaires sont des carbonates fins.*]
4. Note de la dépense qu'on a faite dans un restaurant. [*Elle demande l'addition au serveur.*]

- 1 occurrence de ADDITION
- 4 sens proposés
- Le répondant doit associer un de ces sens à l'occurrence

Graphe GS401213

$G_{GS401213}$



$[G_{GS401213}]$

	occ.	s1	s2	s3	s4
occ.	0	**1**	**1**	**1**	**1**
s1	0	0	0	0	0
s2	0	0	0	0	0
s3	0	0	0	0	0
s4	0	0	0	0	0

$[G_{GS401213\_SEM}]$

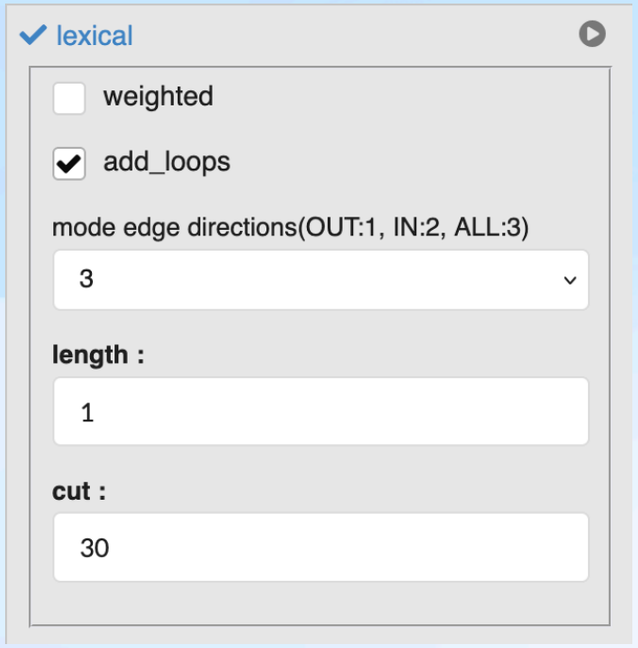
	occ.	s1	s2	s3	s4
occ.	0	**1**	**1**	**1**	**1**
s1	0	0	0	0	0
s2	0	0	0	0	0
s3	0	0	0	0	0
s4	0	0	0	0	0

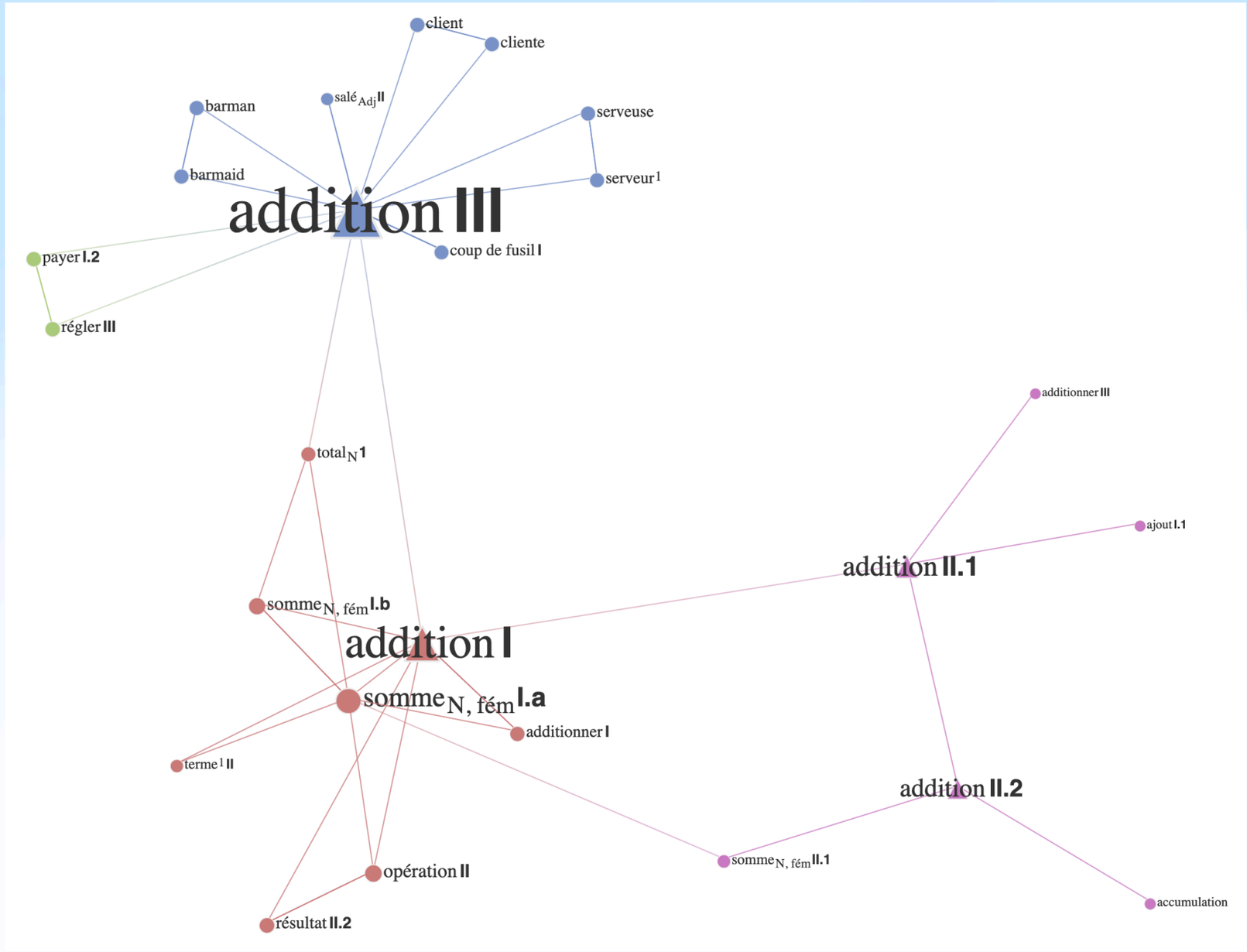
- Soit $G_{GS401213} = (V,E,w)$ un graphe multiple orienté de n sommets et m arcs où chaque arc $(i,j) \in E$ est pondéré par un poids sémantique $w(i,j) \in 0,1,2$
- Soit $[G_{GS401213}]$ la matrice d'adjacente de $G_{GS401213}$ telle que pour tout $i,j \in V$, $[G]_{i,j} = 1n$ où n correspond au nombre d'arcs $(i,j) \in E$.
- $G_{GS401213}$ compte 5 sommets, qui correspondent à l'occurrence *addition* dans le contexte c014 et à chacun des 4 sens d'ADDITION proposés pour l'annotation.
- $G_{GS401213}$ compte 4 arcs, de poids sémantique 0, qui correspondent l'association que nous provoquons entre l'occurrence et chacun des 4 sens proposés en énonçant la question.
- $[G_{GS401213\_SEM}]$ est appelé matrice d'adjacence sémantique de $G_{GS401213}$. Elle s'inspire de la matrice [SEM] de (Sinha et al. 2022). Chaque terme a_{ij} est égal au nombre d'arcs allant du sommet i au sommet j additionné du poids sémantique de chacun de ses arcs (1 arc de poids 0 entre occ. et le sens 1, …).

Sous-graphe ADDITION RL-fr

vue dans Spiderlex

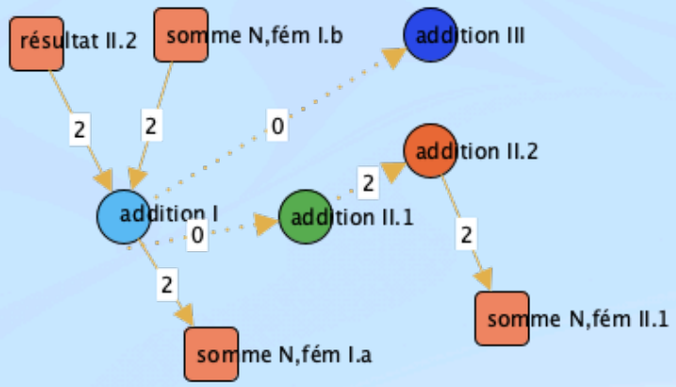
- Espace lexical





Graphe ADDITION$_{RL\text{-}FR}$

$G_{ADDITIONRL\text{-}FR}$



$[G_{ADDITIONRL\text{-}FR}]$

	I	II.1	II.2	III	ré	so I.a	so I.b	so II.1
I	0	**1**	0	**1**	0	**1**	0	0
II.1	0	0	**1**	0	0	0	0	0
II.2	0	0	0	0	0	0	0	**1**
III	0	0	0	0	0	0	0	0
ré	**1**	0	0	0	0	0	0	0
so I.a	0	0	0	0	0	0	0	0
so I.b	**1**	0	0	0	0	0	0	0
so II.1	0	0	0	0	0	0	0	0

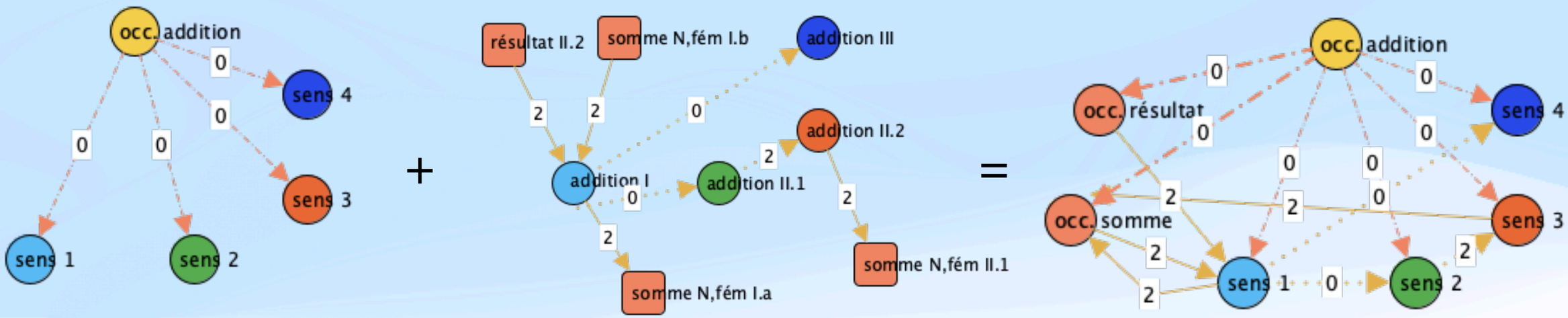
$[G_{ADDITIONRL\text{-}FR\_SEM}]$

	I	II.1	II.2	III	ré	so I.a	so I.b	so II.1
I	0	**1**	0	**1**	0	**3**	0	0
II.1	0	0	**3**	0	0	0	0	0
II.2	0	0	0	0	0	0	0	**3**
III	0	0	0	0	0	0	0	0
ré	**3**	0	0	0	0	0	0	0
so I.a	0	0	0	0	0	0	0	0
so I.b	**3**	0	0	0	0	0	0	0
so II.1	0	0	0	0	0	0	0	0

- Vue simplifiée du sous-graphe de ADDITION dans le RL-fr
- Sélection des 4 lexies qui forment le vocable ADDITION et des voisins de ces nœuds qui partagent leur lemme avec une des formes du contexte c014.
- Ici certaines relations ont un poids sémantique supérieur à 0, les matrices $[G_{ADDITIONRL\text{-}FR}]$ et $[G_{ADDITIONRL\text{-}FR\_SEM}]$ sont donc différentes.

Projection du RL-fr sur c014

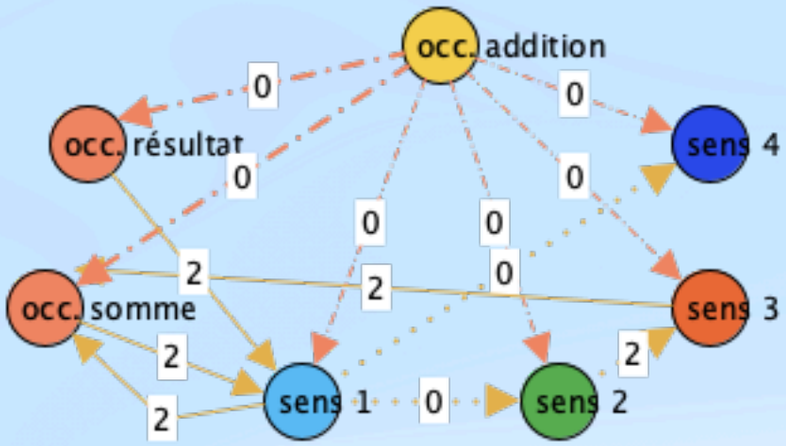
$G_{GS401213}$ + $G_{ADDITIONRL\text{-}FR}$ = $G_{GS40123RL\text{-}FR\_TEMP}$



- On revient à la numérotation adoptée pour l'enquête (ADDITION I → sens 1, ADDITION II.1 → sens 2, ADDITION II.2 → sens 3, ADDITION III → sens 4)
- On perd la distinction entre les différents sens de $SOMME_{N,fém}$, puisque le contexte n'est pas désambiguïsé
- On ajoute des liens de cooccurrence de poids sémantique 0 entre les occurrences.

Matrices associées à $G_{GS40123RL\text{-}FR\_TEMP}$

$G_{GS40123RL\text{-}FR\_TEMP}$



$[G_{GS40123RL\text{-}FR\_TEMP}]$

	occ. addition	s1	s2	s3	s4	occ. résultat	occ.somme
occ. addition	0	**1**	**1**	**1**	**1**	**1**	**1**
sens 1	0	0	**1**	0	**1**	0	**1**
sens 2	0	0	0	**1**	0	0	0
sens 3	0	0	0	0	0	0	**1**
sens 4	0	0	0	0	0	0	0
occ. résultat	0	**1**	0	0	0	0	0
occ. somme	0	**1**	0	0	0	0	0

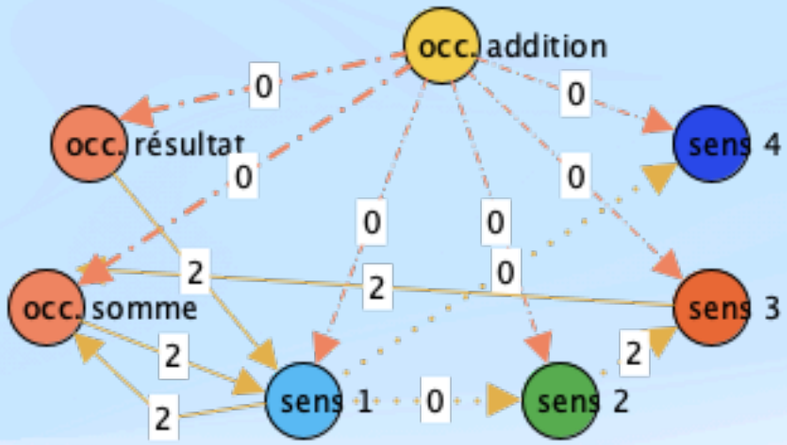
$[G_{GS40123RL\text{-}FR\_TEMP\_SEM}]$

	occ. addition	s1	s2	s3	s4	occ. résultat	occ.somme
occ. addition	0	**1**	**1**	**1**	**1**	**1**	**1**
sens 1	0	0	**1**	0	**1**	0	**3**
sens 2	0	0	0	**3**	0	0	0
sens 3	0	0	0	0	0	0	**3**
sens 4	0	0	0	0	0	0	0
occ. résultat	0	**3**	0	0	0	0	0
occ. somme	0	**3**	0	0	0	0	0

Réduction de $G_{GS40123RL\text{-}FR\_TEMP}$

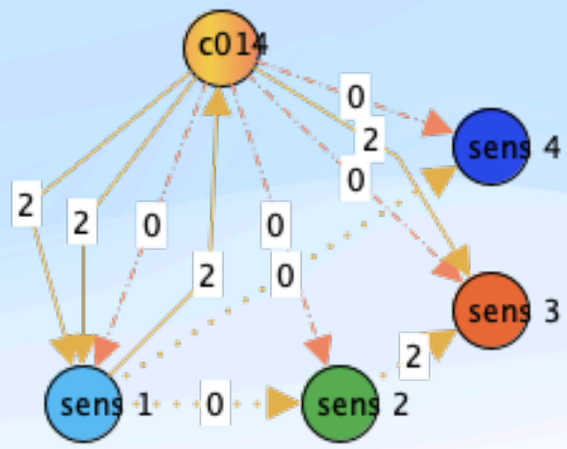
Fusion des nœuds occurrences

$G_{GS40123RL\text{-}FR\_TEMP}$



→

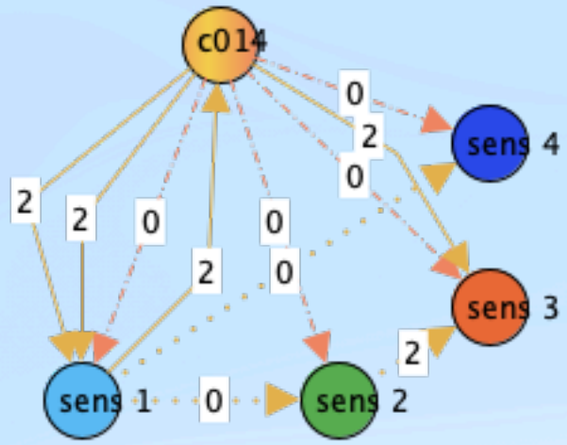
$G_{GS40123RL\text{-}FR}$



- Fusion des nœuds des différentes occurrences en un nœud unique c014 qui représente le contexte
- Les relations de cooccurrences disparaissent
- Les autres relations persistent

Matrices associées à $G_{GS40123RL\text{-}FR}$

$G_{GS40123RL\text{-}FR}$



$[G_{GS40123RL\text{-}FR}]$

	c014	s1	s2	s3	s4
contexte c014	0	**3**	**1**	**2**	**1**
sens 1	**1**	0	**1**	0	**1**
sens 2	0	0	0	**1**	0
sens 3	0	0	0	0	0
sens 4	0	0	0	0	0

$[G_{GS40123RL\text{-}FR\_SEM}]$

	c014	s1	s2	s3	s4
contexte c014	0	**7**	**1**	**4**	**1**
sens 1	**3**	0	**1**	0	**1**
sens 2	0	0	0	**3**	0
sens 3	0	0	0	0	0
sens 4	0	0	0	0	0

Dans ce cas précis, on voit que le contexte entretient un lien privilégié avec le sens 1.

Comment passer d’une matrice à une mesure exploitable dans modèle de stat prédictive ?

Tentative

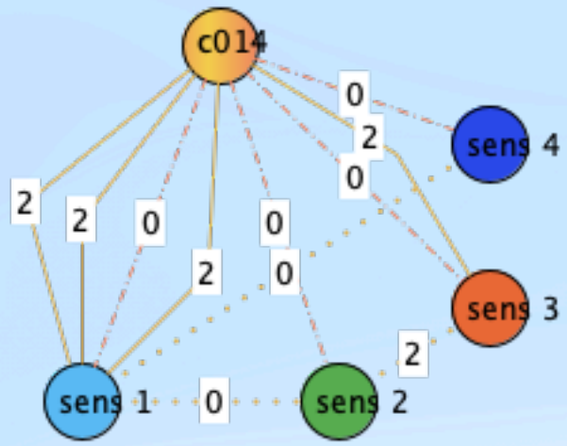
Entropie de Shannon

1. Supprimer orientation des arcs
2. Sélectionner vecteur du contexte (notion de « vision », inspiration Desalle et al. 2014)
3. Calculer entropie de Shannon (1948) des sens

→ On obtient entropy_neighbours

Suppression de l'orientation des arcs

$G_{GS40123RL\text{-}FR}$'



[$G_{GS40123RL\text{-}FR}$']

	c014	s1	s2	s3	s4
contexte c014	0	**4**	**1**	**2**	**1**
sens 1	**4**	0	**1**	0	**1**
sens 2	**1**	**1**	0	**1**	0
sens 3	**2**	0	**1**	0	0
sens 4	**1**	**1**	0	0	0

[$G_{GS40123RL\text{-}FR}$'_SEM]

	c014	s1	s2	s3	s4
contexte c014	0	**10**	**1**	**4**	**1**
sens 1	**10**	0	**1**	0	**1**
sens 2	**1**	**1**	0	**3**	0
sens 3	**4**	0	**3**	0	0
sens 4	**1**	**1**	0	0	0

Sélection du vecteur du contexte

- Soit le $G_{GS40123RL\text{-}FR}$' = (*V,E,w*) un graphe multiple non orienté pondéré par un poids sémantique *w*(*i*,*j*) ∈ 0,1,2, [$G_{GS40123RL\text{-}FR}$'] sa matrice d'adjacence de taille 5x5 et [$G_{GS40123RL\text{-}FR}$'_SEM] sa matrice d'adjacence sémantique.

- On note [$G_{GS40123RL\text{-}FR}$']$_{i,j}$ la valeur située à la i^{ème} ligne et à la j^{ième} colonne de la matrice [$G_{GS40123RL\text{-}FR}$']

- On note [$G_{GS40123RL\text{-}FR}$']$_i$ le vecteur ligne ([$G_{GS40123RL\text{-}FR}$']$_{i,1}$, [$G_{GS40123RL\text{-}FR}$']$_{i,2}$... [$G_{GS40123RL\text{-}FR}$']$_{i,5}$) d'un sommet i ∈ *V et* [$G_{GS40123RL\text{-}FR}$'_SEM]$_i$ son vecteur ligne sémantique.

- [$G_{GS40123RL\text{-}FR}$']$_{c014}$ = [0,4,1,2,1]

- [$G_{GS40123RL\text{-}FR}$'_SEM]$_{c014}$ = [0,10,1,4,1]

Calcul entropie (1)

à partir de $\mathcal{V}(G_{\text{GS40123RL-FR'\_SEM}}, \text{c014})$

- Objectif : mesurer le degré de « choix » impliqué dans la sélection de l'évènement « attribuer un sens à l'occurrence d'ADDITION proposée dans le cadre de la question GS40123»; l'incertitude provoquée par l'évocation des différents sens dans le contexte.

- Soit $t_{i,j}$ un terme de la matrice d'adjacence sémantique $[G_{\text{GS40123RL-FR'\_SEM}}]_{\text{c014}}$

- On définit la probabilité d'attribuer 1 sens *k* donné de la manière suivante :

  - $p_{(\text{sens } 1)} + p_{(\text{sens } 2)} + p_{(\text{sens } 3)} + p_{(\text{sens } 4)} = 1$

  - $p_{(\text{sens } k)} = t_{\text{c014,sens } k} / (t_{\text{c014,sens } 1} + t_{\text{c014,sens } 2} + t_{\text{c014,sens } 3} + t_{\text{c014,sens } 4})$

Calcul entropie (2)

à partir de $\mathcal{V}(G_{GS40123RL\text{-}FR'\_SEM}, c014)$

- Mesure entropie Shannon(1948) :
  - $H = - \sum p_i \log p_i$
- Choix log base 2
- Cas précis de c014
  - $[G_{GS40123RL\text{-}FR'\_SEM}]_{c014} = [0,10,1,4,1]$
  - $H = 1,4238$

$p_{(sens\ 1)}$	$p_{(sens\ 2)}$	$p_{(sens\ 3)}$	$p_{(sens\ 4)}$	$p_{(sens\ 1)} \log_2 (p_{(sens\ 1)})$	$p_{(sens\ 2)} \log_2 (p_{(sens\ 2)})$	$p_{(sens\ 3)} \log_2 (p_{(sens\ 3)})$	$p_{(sens\ 4)} \log_2 (p_{(sens\ 4)})$	$\sum$	***H***
0,625	0,0625	0,25	0,0625	-0,4238	-0,25	-0,5	-0,25	-1,4238	**1,4238**

Question existentielle

- Est-ce que cette proposition permet bien de mesurer l'influence des voisins lexicaux dans la sélection d'un sens à attribuer à l'occurrence présente ?

Limites rencontrées

- Selon la définition de Shannon :
  - Si tous les p_i sont égales, $p_i = 1/n$, alors H doit être une fonction monotone croissante de n
  - → Plus de choix = plus d'incertitude
  - → $H_{max}(p_{\text{sens 1}}, p_{\text{sens 2}}, p_{\text{sens 3}}, p_{\text{sens 4}}) > H_{max}(p_{\text{sens 1}}, p_{\text{sens 2}}, p_{\text{sens 3}}) > H_{max}(p_{\text{sens 1}}, p_{\text{sens 2}})$
  - Comment comparer l'entropie des contextes de vocables à 4 sens avec l'entropie des contextes de vocables à 3 sens et avec l'entropie des contextes de vocables à 2 sens ?
  - Normalisation possible ?
- Conséquence dans modèle stat prédictives : forte corrélation avec autre mesure d'entropie (basée sur proba de rencontrer chacun des sens dans 100 contextes tirés au hasard dans grand corpus)

Normalisation Shannon

équitabilité de Pielou

- Pielou (1966) : H/H_{max} → *On obtient evenness_neighbours*
- Ne dépend plus du nombre d'espèces
- Varie entre 0 et 1 :
  - 0 → un sens est prédominant → l'annotation est simplifiée
  - 1 → tous les sens sont également présents (ou absents) → l'annotation est plus difficile

$p_{(sens\ 1)}$	$p_{(sens\ 2)}$	$p_{(sens\ 3)}$	$p_{(sens\ 4)}$	$p_{(sens\ 1)} \log_2(p_{(sens\ 1)})$	$p_{(sens\ 2)} \log_2(p_{(sens\ 2)})$	$p_{(sens\ 3)} \log_2(p_{(sens\ 3)})$	$p_{(sens\ 4)} \log_2(p_{(sens\ 4)})$	$\sum$	H	H_{max} $log_{2(4)}$	$\boldsymbol{H/H_{max}}$
0,625	0,0625	0,25	0,0625	-0,4238	-0,25	-0,5	-0,25	-1,4238	1,4238	2	**0,7119**

Biblio

- Desalle Y, Navarro E, Chudy Y, et al. (2014) BACANAL : Balades Aléatoires Courtes pour ANAlyses Lexicales. In: *TALN-RECITAL 2014 Workshop SemDis 2014 : Enjeux actuels de la sémantique distributionnelle (SemDis 2014: Current Challenges in Distributional Semantics)*, Marseille, France, July 2014, pp. 206–217.
- Pielou EC (1966) Species-diversity and pattern-diversity in the study of ecological succession. *Journal of Theoretical Biology* 10(2): 370–383.
- Shannon CE (1948) A Mathematical Theory of Communication. *Bell System Technical Journal* 27(3): 379–423.
- Sinha A, Ollinger S and Constant M (2022) Word Sense Disambiguation of French Lexicographical Examples Using Lexical Networks. In: *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, Gyeongju, Republic of Korea, October 2022, pp. 70–76.