Exploitations morphosémantiques des *embeddings* lexicographiques

Basilio Calderone, Nabil Hathout

CLLE/ERSS, CNRS & UT2J

Sémantique pour les ressources en morphologie dérivationnelle
8 juillet 2019, Toulouse

Table of Contents

WORK IN PROGRESS

Génèse : associer 2 thématiques CARTEL

- CARTEL a produit plusieurs dictionnaires extraits des Wiktionnaires :
  - **GLAWI** : français (Sajous and Hathout, 2015)
  - **ENGLAWI** : anglais
  - **GLAWIT** : italien (Calderone et al., 2017)
- Beaucoup dans CARTEL travaillent sur la description du sens au moyen d'embeddings
- Comment rapprocher ces deux thématiques ?

➜ En construisant des embeddings
à partir des descriptions lexicographiques

Embeddings

Représentations vectorielles du sens construites de sorte que les unités qui ont des propriétés proches soient représentées par des vecteurs proches.

Hypothèse distributionnelle (Harris, 1954; Firth, 1951)

Les mots dont les distributions sont différentes ont des sens différents.

À partir des propriétés distributionnelles des mots

- Les modèles count-based sont construits à partir des fréquences de cooccurrence en utilisant des mesures d'association et en procédant à des réductions de dimensions.
- Les modèles prédictifs sont construits en utilisant des réseaux de neurones.

Table of Contents

Dictionnaires

- Les **corpus** permettent de caractériser le sens des mots à partir des contextes de l'ensemble de leurs occurrences
- Les **dictionnaires** sont des artefacts qui décrivent différentes propriétés des mots
- Les **définitions** sont des descriptions synthétiques, cohérentes, et relativement complètes du sens des mots

Dictionnaires (2)

Définitions

- ✗ Les définitions présentent une **grande variabilité** dans la formulation du sens des mots
- ✔ Les définitions présentent de **fortes régularités** formelles
  - du fait de leur structure qui est idéalement de type
    genre prochain + *différence spécifique*
  - du fait de l'utilisation d'opérateurs lexicographiques comme
    celui celle qui V ; *action de V* ; *qualité de ce qui est A*
- Les mots dont les sens sont proches ont des définitions qui se ressemblent :
  - partage d'opérateurs
  - utilisation de mots «proches»

Dictionnaires (3)

Définitions «proches»

- **diagnosticable** : qui peut être diagnostiqué
- **refermable** : qu'on peut refermer
- **remboursable** : qui est susceptible d'être remboursé

Embeddings de définitions

Similarity vs relatedness (Hill et al., 2015)

- **relatedness** : relations entre notions associées
  - café : tasse, cuillère, machine, matin, pause, sucre, comptoir
- **similarity** : relations entre des notions proches **au sein de la même catégorie** (co-hyponymie)
  - café : thé, bière, coca, chocolat, eau, capuccino, vin

- Les embeddings de corpus tendent à capter des relations associatives (*relatedness*). Les mots associés à un mot donné apparaissent dans le contexte de ce dernier.
- Les embeddings de définitions devraient mieux capter les relations de similarité. Les mots de même catégorie tendent avoir des définitions similaires (Bosc and Vincent, 2018).

GLAWI

- 227027 lemmes d'entrées (sans distinction de catégorie).
- Les homographes de mots grammaticaux sont exclus.
- La description lexicographique d'un lemme est la concaténation de toutes ses définitions.

Définitions de *rédaction*

1. action de rédiger ou résultat de cette action.)
2. ensemble des rédacteurs d'un journal
3. exercice scolaire par lequel on enseigne aux enfants à rédiger

GLAWI (2)

Description lexicographique de *rédaction*

`action de rédiger ou résultat de cette action . ensemble des rédacteurs d' un journal . exercice scolaire par lequel on enseigne aux enfants à rédiger .`

Utiliser les définitions comme un corpus

- On voudrait que les contextes d'un mot soient ses définitions.

➜ Ajouter le lemme de l'entrée devant la description lexicographique et utiliser des fenêtres de taille suffisante

GLAWI (3)

Corpus de définitions

- rotondité caractère de ce qui est **rond** , **rondeur** . **grosseur** , **corpulence** .
- **rondeur** caractère d' une personne qui a de la franchise , qui est sans façon . caractère de ce qui est **rond** , de ce qui est sphérique , **circulaire** ou cylindrique . caractéristique d' un vin qui correspond à sa consistance , c' est-à-dire l' impression qu' il donne d' avoir dans la bouche un corps charnu . choses rondes , et particulièrement les parties du **corps** où se manifeste de l' **embonpoint** .

- Corpus de 4.7 millions mots seulement

Word2Vec

Modèle Word2Vec

- Taille maximale de la fenêtre = 15 mots ;
- Fréquence minimale des mots = 1 ;
- Sous-échantillonnage = 0.1 ;
- SkipGram ;
- 100 dimensions

Table of Contents

Le modèle adopté

Notre modèle et l'idée sous-jacente

- Obtenir des représentations vectorielles (*embeddings*) des mots définis dans un dictionnaire à partir de leur définitions
  - Les *embeddings* doivent tenir compte des séquences de mots pour capter les patrons récurrents et cohérents dans les définitions
  - *We expect the embeddings of a word to represent its* ***meaning compactly*** *(Bosc and Vincent, 2018)*
- Il s'agit d'un modèle **self-contained** :
  il est entraîné seulement sur les données d'un dictionnaire

Le modèle adopté

Notre modèle et l'idée sous-jacente

- Obtenir des représentations vectorielles (*embeddings*) des mots définis dans un dictionnaire à partir de leur définitions
  - Les *embeddings* doivent tenir compte des séquences de mots pour capter les patrons récurrents et cohérents dans les définitions
  - *We expect the embeddings of a word to represent its* ***meaning compactly*** *(Bosc and Vincent, 2018)*
- Il s'agit d'un modèle **self-contained** :
  il est entraîné seulement sur les données d'un dictionnaire

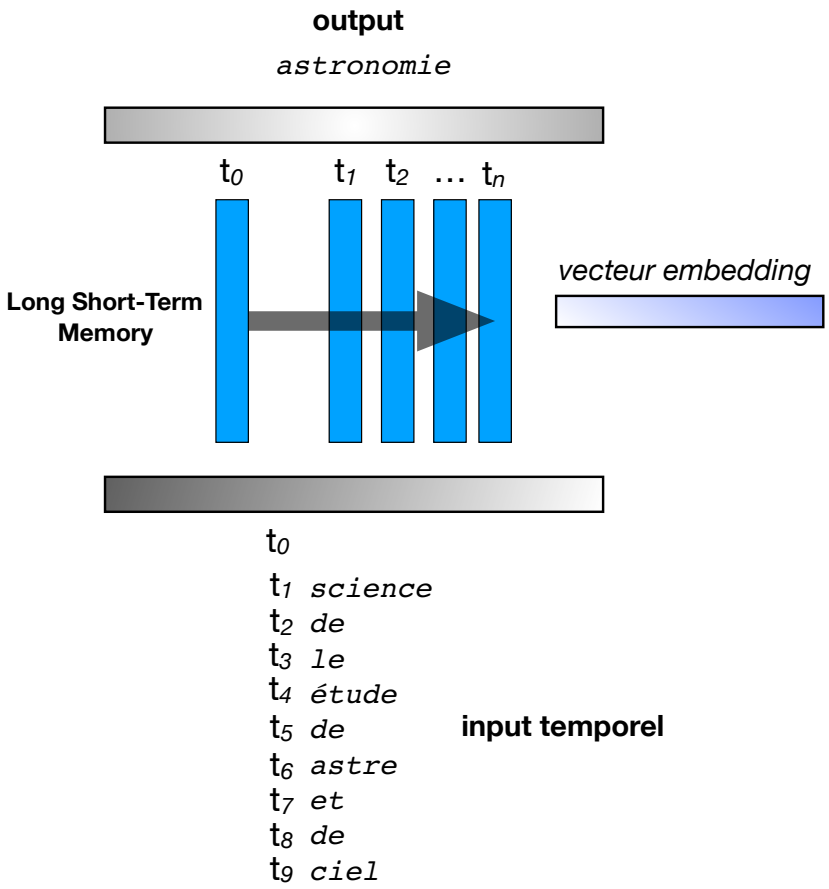
Traiter la dimension temporelle : le Long Short-Term Memory

- L'idée des LSTM est de permettre au réseau « d'oublier » ou de ne pas prendre en compte certaines observations passées afin de pouvoir donner du poids aux informations importantes
- Le système apprendre à faire des prédictions, à ignorer, à oublier et à sélectionner

Le modèle adopté (II)

Mapping "word sequences to word"

- Définition $\mapsto$ LSTM $\mapsto$ Mot défini
- {science de l'étude des astres et du ciel} $\mapsto$ LSTM $\mapsto$ {astronomie}



Récupèrer la mémoire LSTM

- Une fois le modèle entraîné, récupèrer l'état d'activation de la mémoire LSTM $c_{t_{def}}$ (**meaning compactly**)
- L'état d'activation de $\boxed{\text{LSTM}}$ est l'*embedding* ($c_{t_{astronomie}}$)
- Tester la similarité de l'*embedding* $c_{t_{astronomie}}$ par rapport aux autres *embeddings* (voisinage, distribution du voisinage, etc.)

Table of Contents

Expériences et résultats

Les paramètres du modèle : la mémoire LSTM

- Nous avons utilisé les *embeddings* W2V endogènes présentés précédemment (modèle *Skip-Gram* avec *Negative Sampling* (SGNS) ; 100 dimensions ; fenêtre de 15 mots ; fréquence minimale = 1 ; sous-échantillonnage = 0.1)
- L'état d'activation de la mémoire LSTM est l'*embedding* $c_{t_{astronomie}}$
  - Taille de la mémoire LSTM = 100 dimentions (la mémoire LSTM a 100 neurones)
  - Taille du batch = 32 definitions (nous avons entraînés d'autres modèles avec tailles différentes : 1, 16, 64, 128)
  - Nombre d'époques = 50 (d'autres modèles ont été entraînés avec des nombres d'époques différents : 1, 10, 100, 1000)

Analyse du voisinage

Les voisins de *lavable*

qui pouvoir être laver.

Analyse du voisinage

Les voisins de *lavable*

qui pouvoir être laver.

R.	W2V	LSTM	R.	W2V	LSTM
1	déblanchi	charriable	11	crameux	satinable
2	cardée	recousable	12	paraison	usinable
3	assiettée	allégeable	13	démerger	arrachable
4	dégorgeage	triturable	14	effilure	plâtrable
5	atomisation	vendangeable	15	billig	châtrable
6	bouillotter	déployable	16	pivotable	ourdissable
7	brosseur	calcinable	17	lécythe	redressable
8	crêpière	chaussable	18	buffeter	scissile
9	seringuer	arrosable	19	têtée	défrichable
10	goulleh	entrecoupable	20	bigophone	cintrable

Analyse du voisinage

Les voisins de *brocantable*

qui pouvoir être brocanter.

Analyse du voisinage

Les voisins de *brocantable*

qui pouvoir être brocanter.

R.	W2V	LSTM	R.	W2V	LSTM
1	increvable	recousable	11	arbitrable	optable
2	suicidable	capitalisable	12	ahuri	désolable
3	incristallisable	décachetable	13	enjailleur	arbitrable
4	insapidité	piégeable	14	intransitiver	escomptable
5	inobéissant	dégringolable	15	inoculable	apeurable
6	accusable	accoutrable	16	dépunaiser	transperçable
7	discompté	rejouable	17	asynchronie	trébuchable
8	alcalifiable	hypothécable	18	aperceptif	enracinable
9	concupiscible	ensevelissable	19	urnapouiller	ressuscitable
10	adultérateur	manufacturable	20	absalonien	incarcérable

Analyse du voisinage

Les voisins de *adressage*

action de adresser.

Analyse du voisinage

Les voisins de *adressage*

action de adresser.

R.	W2V	LSTM	R.	W2V	LSTM
1	shell	confiage	11	branchement	déchargeage
2	mesure	invitement	12	wikipedia	incitation
3	interagir	copiage	13	indexation	essayage
4	indexer	annoncement	14	qualitatif	recréation
5	réimpression	itérer	15	suivi	mettage
6	html	encadrage	16	discret	guidage
7	chiffrement	collectage	17	mot-clé	télé
8	associatif	répertoriage	18	module	comparution
9	hypertexte	cachement	19	suivi	dévoilement
10	en_plus_de	récompensement	20	report	souvenance

Analyse du voisinage

Les voisins de *affinage*

action de affiner.

Analyse du voisinage

Les voisins de *affinage*

action de affiner.

R.	W2V	LSTM	R.	W2V	LSTM
1	broyage	tassage	11	pressurage	emmagasinage
2	lavage	graissage	12	pulvériser	lessivage
3	chauffant	perforage	13	malt	collure
4	vernisser	malaxation	14	surnager	amassage
5	panade	rinçage	15	égouttage	concassage
6	mouture	puisement	16	cornue	recuisson
7	trempe	puisage	17	humecter	pétrissage
8	pocher	délayement	18	diluer	vaporisation
9	séchage	décapsulage	19	refroidir	désinfection
10	pressage	réchauffage	20	fine	épluchage

Analyse du voisinage

Les voisins de *affinement*

action de affiner ou état de ce qui être affiner.

Analyse du voisinage

Les voisins de *affinement*

action de affiner ou état de ce qui être affiner.

R.	W2V	LSTM	R.	W2V	LSTM
1	labile	étamage	11	circulateur	encollage
2	agroupé	étayage	12	illisibilité	barriérage
3	salinisation	écrasement	13	voltaïsation	scellement
4	applicabilité	calfatage	14	cisaillage	lamellisation
5	ennoiement	plâtrage	15	poiret	incision
6	aridification	amincissement	16	corneau	enroulement
7	impréméditation	aplanissement	17	ennuyance	bassinement
8	effritement	encombrement	18	éthérisme	défilage
9	diamagnétique	obturation	19	malik	aération
10	affrérissement	évasement	20	individuer	emballement

Analyse du voisinage

Les voisins de *boulangerie*

fabrication et vente de pain. il se dire également de fonds de un boulanger. lieu où être fabriquer et vendre le pain.

Analyse du voisinage

Les voisins de *boulangerie*

fabrication et vente de pain. il se dire également de fonds de un boulanger. lieu où être fabriquer et vendre le pain.

R.	W2V	LSTM	R.	W2V	LSTM
1	savonnerie	gargote	11	azyme	burtoire
2	brasserie	soupe	12	savonnier	bertoire
3	burette	salmigondis	13	malt	pain
4	fondue	parfumerie	14	rôti	chaussure
5	provende	café	15	entremets	repas
6	industriellement	peausserie	16	huche	potage
7	gruyère	cuisiner	17	pailler	brasserie
8	viennoiserie	beurrerie	18	moudre	entremets
9	frite	dessert	19	potage	tabletterie
10	charcutier	cuisine	20	levain	chandellerie

Analyse du voisinage

Les voisins de *tapisserie*

grand ouvrage faire à métier avec de le laine, de le soie, etc., et servir à revêtir et à orner le muraille de un salle, de un chambre, etc. ouvrage faire à le aiguille sur du canevas , avec de le laine, de le soie, etc. papier colorer ou orner de motif que on utiliser pour recouvrir le mur de un pièce. étoffe, tissu servir à couvrir et à orner le muraille de un salle, de un chambre, etc. on dire plutôt aujourd'hui " tenture ".

Analyse du voisinage

Les voisins de *tapisserie*

grand ouvrage faire à métier avec de le laine, de le soie, etc., et servir à revêtir et à orner le muraille de un salle, de un chambre, etc. ouvrage faire à le aiguille sur du canevas , avec de le laine, de le soie, etc. papier colorer ou orner de motif que on utiliser pour recouvrir le mur de un pièce. étoffe, tissu servir à couvrir et à orner le muraille de un salle, de un chambre, etc. on dire plutôt aujourd'hui " tenture ".

R.	W2V	LSTM	R.	W2V	LSTM
1	tenture	ornement	11	couturier	entoiler
2	orfèvrerie	dessin	12	bijou	étiquette
3	broderie	tenture	13	menuiserie	ardoise
4	dais	plafonner	14	reliure	cachet
5	canevas	chemise	15	graveur	écusson
6	parure	serpillière	16	heaume	estampe
7	soierie	toilette	17	vannerie	apposer
8	rideau	couverte	18	bure	reliure
9	brocher	véla	19	vignette	échantillon
10	tiroir	ganse	20	tailleur	paillasse

Analyse du voisinage

Les voisins de *patiemment*

avec patience.

Analyse du voisinage

Les voisins de *patiemment*

avec patience.

R.	W2V	LSTM	R.	W2V	LSTM
1	antiasthmatique	persévéramment	11	antihémorroïdal	imprudemment
2	assimilabilité	arrogamment	12	antipoison	spiritoso
3	chalara	miséricordieusement	13	odorer	fougueusement
4	dujardin	incommodément	14	revitalisation	naïvement
5	hygiocérame	infâmement	15	magnétostriction	animeusement
6	bouleversé	ambitieusement	16	révulser	majestueusement
7	renauder	inadvertamment	17	anaplérotique	maestoso
8	nocébo	dextrement	18	hypniatre	obstinément
9	pseudotuberculosis	fervemment	19	dicoccon	effrontément
10	refleurir	prudemment	20	antirouille	voluptueusement

Analyse du voisinage

Les voisins de *arabisant*

celui, celle qui se adonner à le étude de le arabe.

Analyse du voisinage

Les voisins de *arabisant*

celui, celle qui se adonner à le étude de le arabe.

R.	W2V	LSTM	R.	W2V	LSTM
1	louanger	cabaliste	11	hasardeur	coléoptériste
2	armoriste	mythologue	12	gabeur	architectonographe
3	vétilleur	islamisant	13	typomanie	cosmogoniste
4	agréeur	talmudiste	14	liseur	prophétisme
5	étalier	prédestinateur	15	grelotteux	commentateur
6	affronteur	sunna	16	archiviste	moniste
7	endetteur	médiéviste	17	bouquinier	cabalistique
8	antidreyfusard	olmécologue	18	conciliateur	rubricaire
9	bouquiniste	voltairienne	19	charadiste	antimusulmane
10	louanger	runologue	20	mandoliniste	interniste

Analyse du voisinage

Les voisins de *arabisant*

celui, celle qui se adonner à le étude de le arabe.

R.	W2V	LSTM	R.	W2V	LSTM
1	louanger	cabaliste	11	hasardeur	coléoptériste
2	armoriste	mythologue	12	gabeur	architectonographe
3	vétilleur	islamisant	13	typomanie	cosmogoniste
4	agréeur	talmudiste	14	liseur	prophétisme
5	étalier	prédestinateur	15	grelotteux	commentateur
6	affronteur	sunna	16	archiviste	moniste
7	endetteur	médiéviste	17	bouquinier	cabalistique
8	antidreyfusard	olmécologue	18	conciliateur	rubricaire
9	bouquiniste	voltairienne	19	charadiste	antimusulmane
10	louanger	runologue	20	mandoliniste	interniste

- cabaliste : celui qui être savant dans le cabale de juif.
- islamisant : celui , celle qui étudier le islam.

Analyse du voisinage

Les voisins de *sémanticien*

linguiste spécialiste de le étude de le sémantique.

Analyse du voisinage

Les voisins de *sémanticien*

linguiste spécialiste de le étude de le sémantique.

R.	W2V	LSTM	R.	W2V	LSTM
1	anthropozoologue	syntacticien	11	piagétien	épileptologue
2	moliériste	morphologue	12	glaciologue	frontologie
3	victimologie	psychiatre	13	anatomopathologiste	médianité
4	arthurianisme	helminthologiste	14	catalographie	testologie
5	andragogue	atlantologie	15	écoblanchiment	marchéage
6	delphinologue	argotiste	16	atlantiste	bouddhologie
7	islamologie	myrmécologie	17	préphilatélie	trippeux
8	antisémitique	ancestrologie	18	généticien	cinéphile
9	assyriologue	palynologie	19	azérisant	anthropologue
10	anthroponymie	étymologisme	20	baïen	islamologie

Table of Contents

ACP des embeddings

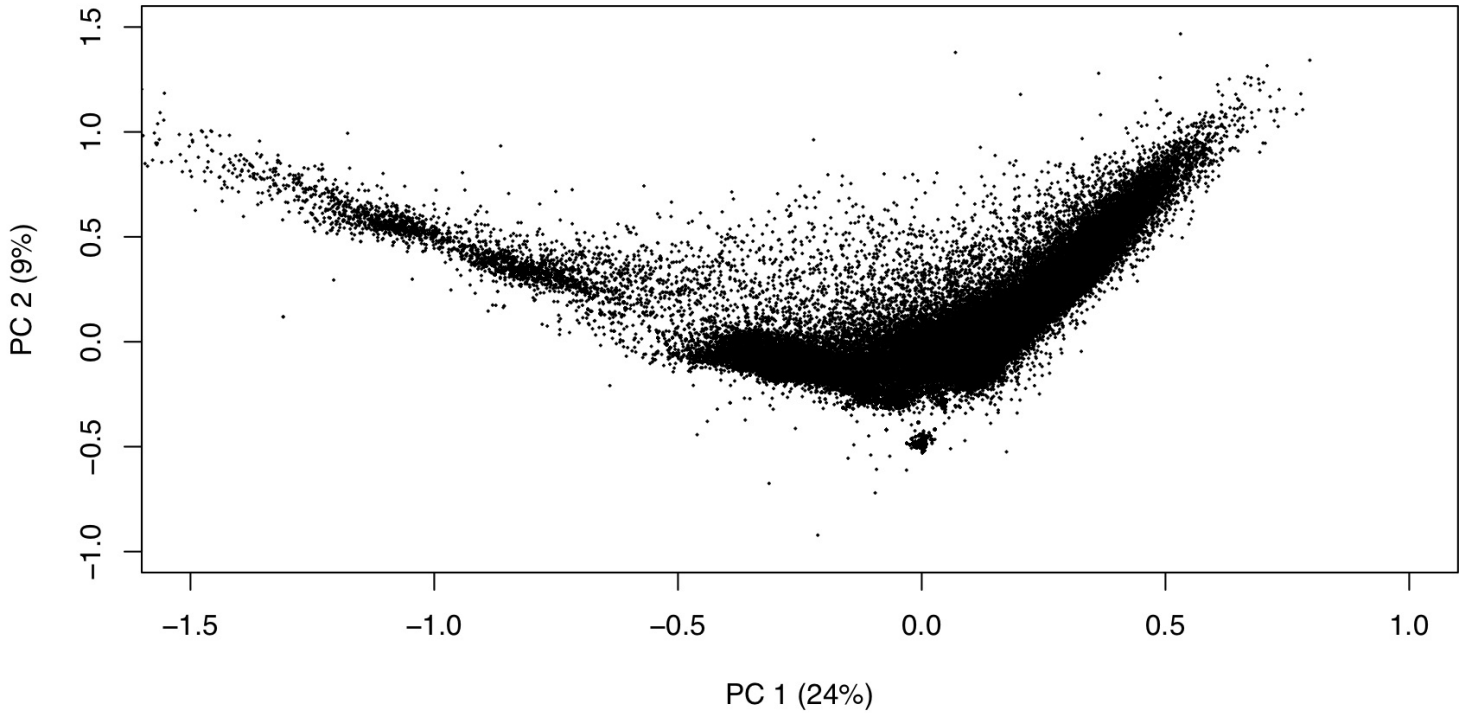
ACP des embeddings

- Reduction de dimensionalité : de 100 à 2 composantes
- Visualisation des données après projection sur les composantes principales

ACP des embeddings

ACP des embeddings

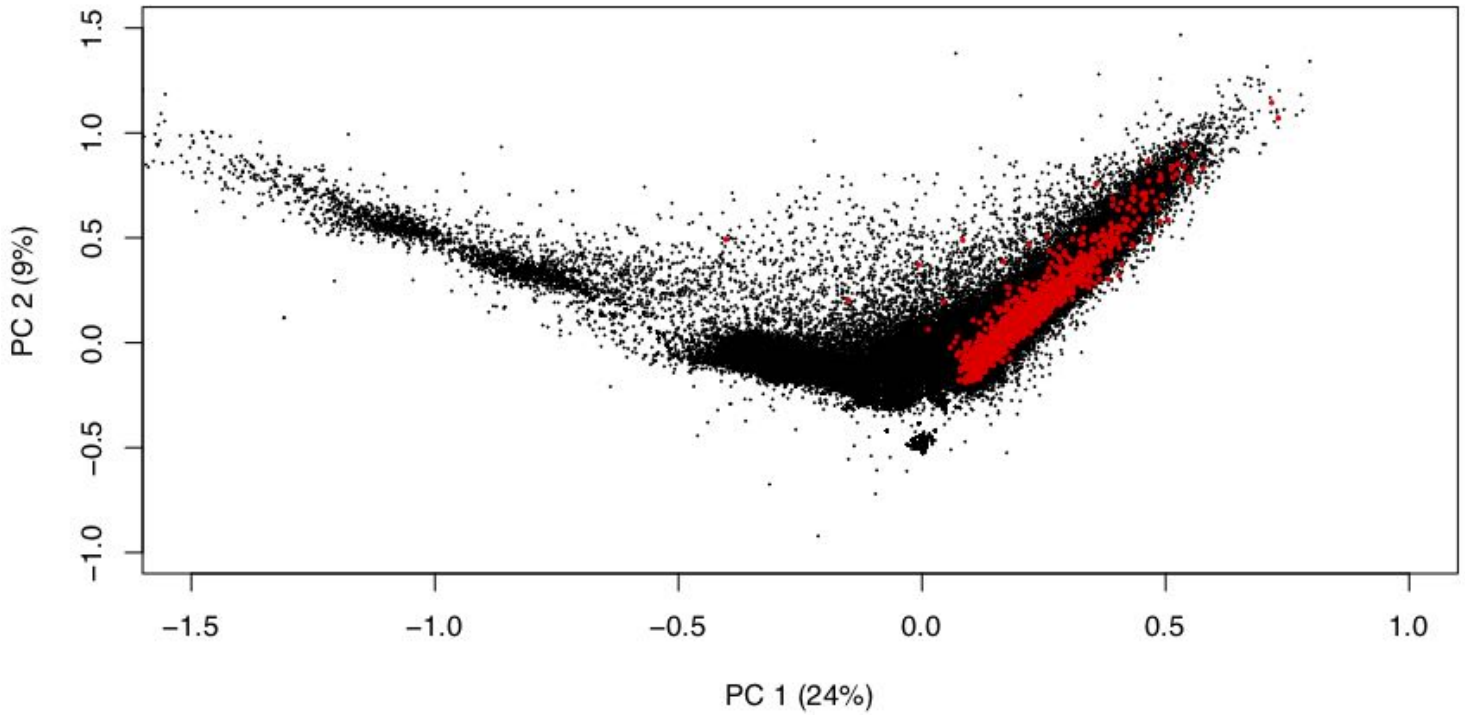
- Reduction de dimensionalité : de 100 à 2 composantes
- Visualisation des données après projection sur les composantes principales



Focus sur les noms d'agent

ACP des embeddings

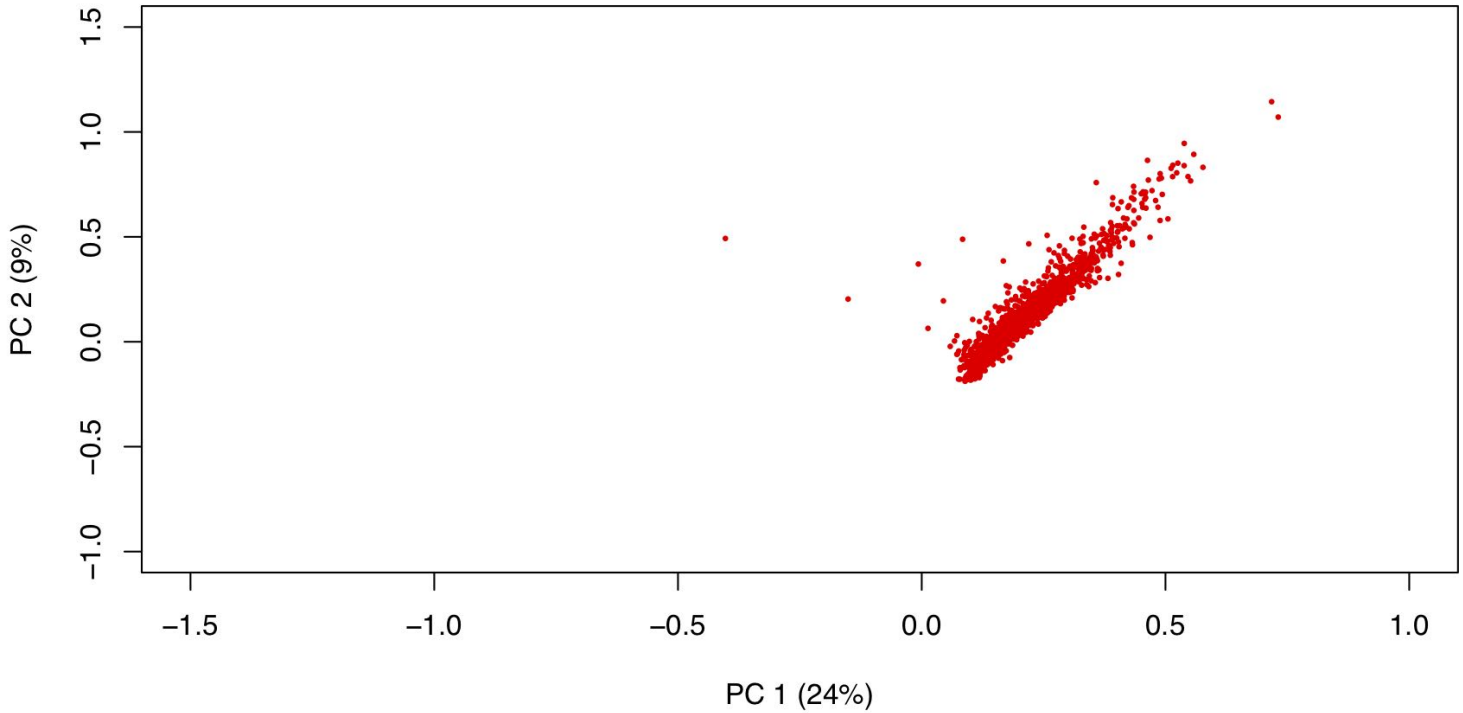
- Sélection des noms d'agent (suffixés en *-eur*) sur la base de Démonette.
- 1950 noms d'agent attestés



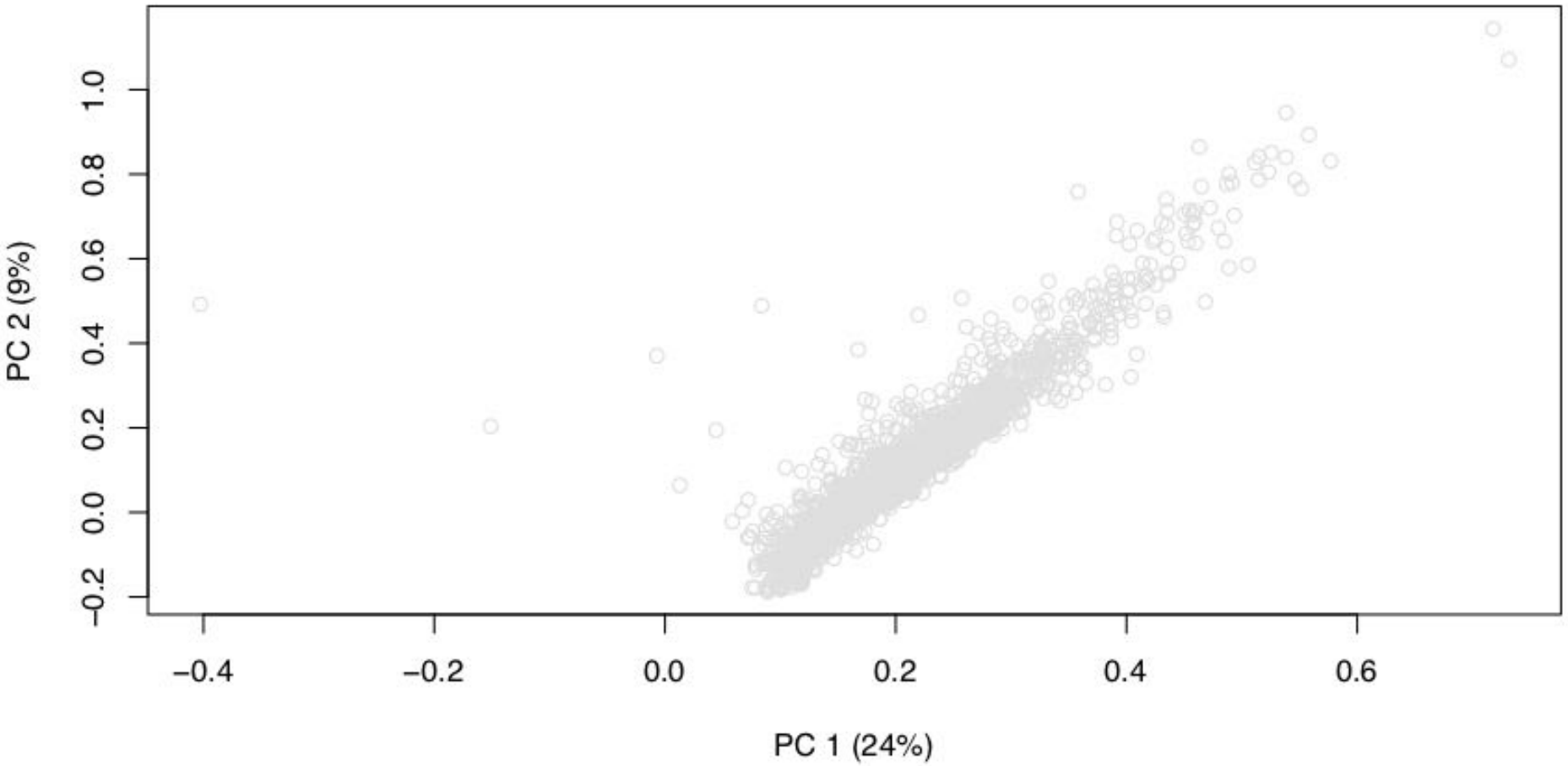
Focus sur les noms d'agent

ACP des embeddings

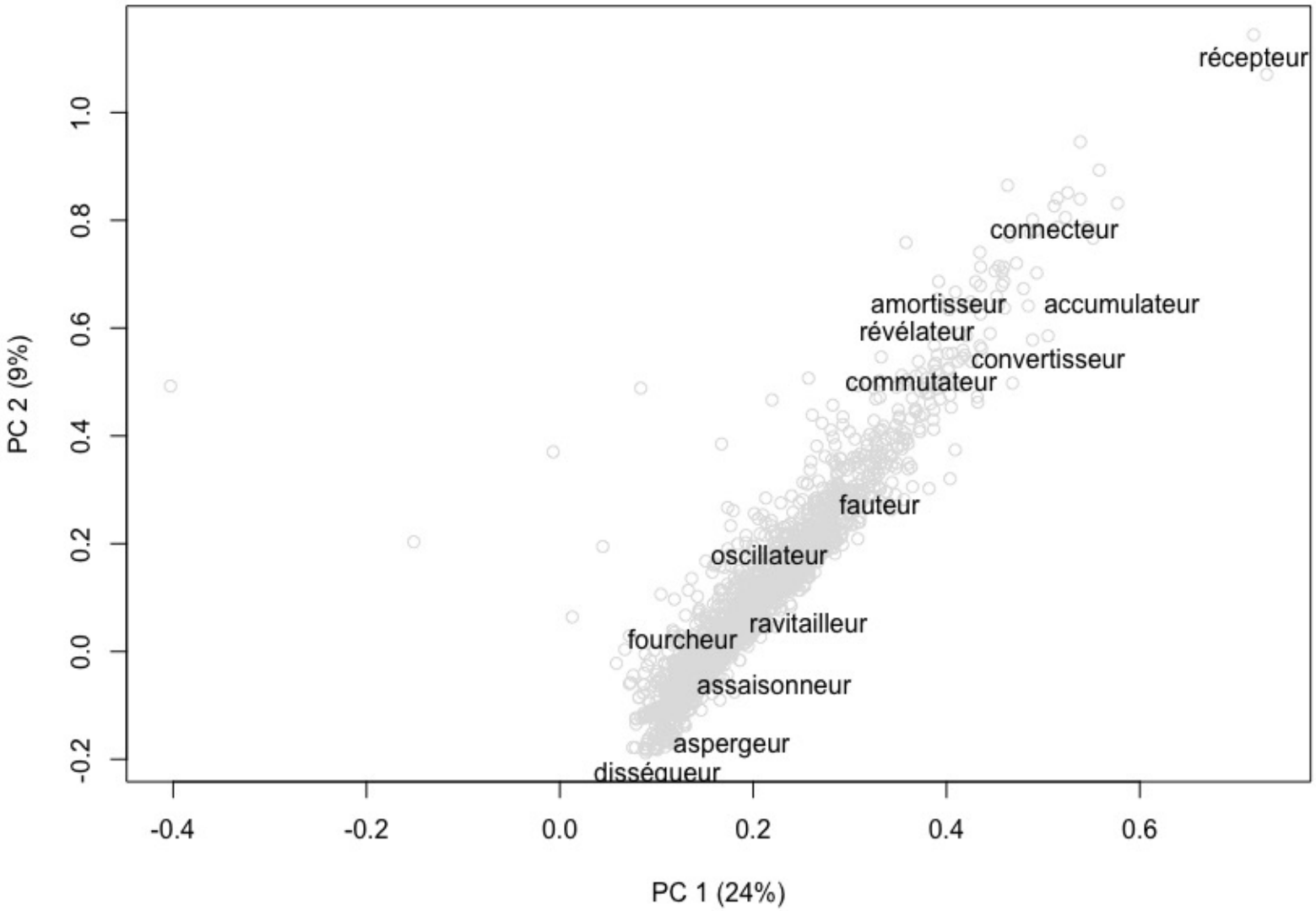
- Sélection des noms d'agent (suffixés en *-eur*) sur la base de Démonette.
- 1950 noms d'agent attestés



Focus sur les noms d'agent



Focus sur les noms d'agent



Conclusion

À cours terme

- Créer un décodeur pour que la sortie du LSTM soit la définition elle-même et non le mot défini pour permettre de générer de nouvelles définitions
  - {alcoholic, fermented, juice, of, grapes, used, as, beverage} $\Longrightarrow$ {alcoholic, fermented, juice, of, grapes, used, as, beverage} et NON $\Longrightarrow$ {wine}
- Construire un modèle qui prédit la catégorie *Unique Beginners* qui correspond à une définition

À long terme

- Comment construire des paradigmes sémantiques à partir des définitions : comment connecter et superposer les définitions ?

Références

Bosc, T. and P. Vincent (2018). Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1522–1532.

Calderone, B., M. Pascoli, F. Sajous, and N. Hathout (2017). Hybrid Method for Stress Prediction Applied to GLAFF-IT, a Large-scale Italian Lexicon. In J. Gracia, F. Bond, J. P. McCrae, P. Buitelaar, C. Chiarcos, and S. Hellmann (Eds.), *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017*, Cham, pp. 26–41. Springer International Publishing.

Firth, J. R. (1951). Modes of meaning. In *Papers in linguistics 1934-1951 (1957)*. Oxford University Press.

Harris, Z. (1954). Distributional structure. *Word 10*(2-3), 146–162.

Hill, F., R. Reichart, and A. Korhonen (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics 41*(4), 665–695.

Références (2)

Sajous, F. and N. Hathout (2015, august). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*, Herstmonceux, England, pp. 405–426.