Utilisation d'outils de sémantique distributionnelle en linguistique

Illustration dans le domaine de la morphologie

Cécile Fabre

CLLE, Université de Toulouse & CNRS

Workshop Linguistique et Big Data
30/11/2017

Plan de la présentation

1. La sémantique distributionnelle : une histoire exemplaire du développement des *big data* en linguistique
2. Apports et limites des modèles de sémantique distributionnelle pour la linguistique
3. Illustration dans le domaine de la morphologie : l'étude des suffixes
4. Conclusion

L'approche distributionnelle

Principes

- Classer les mots selon leurs propriétés distributionnelles
- Hypothèse harrissienne : "difference of meaning correlates with difference of distribution" (Harris 1954)
- Ex : *parcours* et *itinéraire* sont proches parce qu'ils sont tous deux :
  - sujets des verbes *mener*, *traverser*
  - compléments du nom *étape*
  - modifiés par les adjectifs *spirituel*, *libre*
  - etc.
- Une sémantique fondée sur l'usage, sur la compilation des contextes d'apparition des mots dans des corpus

L'approche distributionnelle

- Méthode transposée en termes de calcul dans l'espace vectoriel :
  - Le calcul sémantique est fondé sur une représentation vectorielle des contextes des mots
  - Deux mots sont sémantiquement proches si leurs vecteurs sont proches
  - On peut alors calculer une mesure de similarité entre vecteurs (généralement, la valeur cosinus)

	baliser (obj)	traverser (suj)	changer (de)	étape (de)	incident (de)	spirituel (mod)	libre (mod)	...
parcours	0,6	0,3	0,4	0,6	0,7	0,8	0,8	...
itinéraire	0,7	0,4	0,5	0,6	0,1	0,2	0,2	...

L'approche distributionnelle

Quelques étapes

- Une catégorisation manuelle, réalisée à partir de collections de documents de petite taille, sur des textes très homogènes (Harris et al. 1991)
- Dès les années 1970 : automatisation partielle (regroupement des contextes)
- Années 1990 : automatisation intégrale de la procédure, sur corpus annotés (Grefensette 1994), (Habert et Nazarenko 1996)
- Années 2000 : multiplication des modèles, versions simplifiées de la procédure (corpus bruts). (Turney et Pantel 2010, Baroni et Lenci 2010)
- A partir de 2013 : modèles neuronaux plus performants sur le plan computationnel (Mikolov et al. 2013), (Levy et Goldberg 2014)

L'approche distributionnelle

Les *Word embeddings*

- Modèle "classique" : décompte des contextes, pondération, (éventuellement) réduction de dimension
- Modèle à base de réseaux de neurones : modèle prédit à partir du corpus : le système apprend à assigner des vecteurs similaires à des mots similaires
- Information compacte, dimensions réduites, pas directement interprétables
- Des outils de manipulation des vecteurs résultants : distance entre vecteurs, clustering, addition et soustraction...
- Un outil de référence : Word2Vec (Mikolov et al. 2013)

L'approche distributionnelle

- Une approche devenue dominante en TAL
- Dont la validité a été démontrée
  - Evaluation intrinsèque : adéquation de la mesure de voisinage pour rendre compte de relations sémantiques de type varié (synonymie, jugement de proximité sémantique, catégorisation)
  - Evaluation extrinsèque : contribution à différentes tâches de TAL
- Un nouvel outil pour la linguistique ?

Bénéfices

- En principe : "The more you can gather, the clearer and more accurate will be the picture that you get of the language" Sinclair (2004)
- Possibilité de tester à très large échelle et sur des données variées l'hypothèse distributionnelle
- Des méthodes de TAL arrivées à maturité et plus facilement intégrables : vers un outil de plus dans le bagage du linguiste ?
- De nombreuses questions linguistiques peuvent bénéficier de cet apport (Lenci 2008) :
  - La mise au jour des relations sémantiques (synonymie, antonymie...)
  - L'examen de la variation sémantique au fil du temps ou au gré des discours
  - Des questions linguistiques variées : degré de figement, typage sémantique des arguments, compositionnalité...

Exemple

Etude en diachronie

- Le changement de sens est apprécié à partir du traitement d'un corpus diachronique en étudiant le degré de recouvrement des voisins distributionnels des mots calculés par période
- (Kim et al. 2014) : Utilisation de google books, 100 milliards de mots

mot	Voisins en :	
	1900	2009
cell	closet, dungeon, tent	phone, cordless, cellular
gay	cheerful, pleasant, brillant	lesbian, bisexual, lesbians

Difficultés

- Toute la sémantique n'est pas réductible à l'information distributionnelle
- Prise en compte d'une information linguistique généralement très pauvre (cooccurrences)
- Notion très floue de la proximité sémantique, agrégat de différents types de relation (ex : similarité / association, synonymie/antonymie)
- Difficulté à traiter la polysémie
- Priorité donnée au volume des données sur leur spécificité (Tanguy et al. 2015)
  - *trait* $\rightarrow$ *étiquette*, *feature*, *caractéristique* (articles en TAL)
  - *trait* $\rightarrow$ *expression*, *air* (romans)
- Des informations difficiles à utiliser, à interpréter (effet boîte noire)
  - "What does a point in vector space, where the dimensions are typically uninterpretable symbols, stand for ?" (Erk 2013)
- Difficulté à articuler approche quantitative et qualitative

Proximité sémantique des dérivés morphologiques

- En collaboration avec Marine Wauquier et Nabil Hathout (CLLE)
- Morphologie et sémantique distributionnelle (Kisselew et al. 2016, Lapesa et al. 2017...)
- Etude des nominalisations verbales : les verbes et leurs noms d'action et d'agent dérivés
- Ex : graver – graveur – gravure
- Questions examinées :
  - Degré de proximité sémantique entre les verbes et leurs noms d'action
  - Cartographie des familles dérivationnelles (cohésion vs éloignement)
  - Etude de l'instruction sémantique des suffixes

Un dispositif combinant :

- L'outil Word2Vec
- Des corpus : Wikipedia, Le Monde ($\sim$ 200 millions de mots)
- Une ressource lexicale recensant les unités morphologiques (Lexeur) : 5949 noms en -eur et leur famille dérivationnelle

Agent masc.	Agent fém.	Base	Cat	Action
balançeur	balanceuse	balancer	V	balancement, balançage, balance
balayeur	balayeuse	balayer	V	balayage, balaiement
baliseur	baliseuse	baliser	V	balisage, balisement

Extrait de Lexeur

Premières observations :

- Etude des triplets Nom d'agent - Verbe - Nom d'action
- 1945 triplets (sur 13 136) instanciés dans le modèle Word2Vec généré à partir du corpus Wikipedia
  - (paramètres par défaut, algorithme CBOW, 100 dimensions)
- Proximité moyenne des couples au sein des triplets

	P(AgV)	P(AgAc)	P(VbAc)
Wikipedia	0.25	0.29	0.39
Le Monde	0.25	0.28	0.34

Premières observations :

- Etude des triplets Nom d'agent - Verbe - Nom d'action
- 1945 triplets (sur 13 136) instanciés dans le modèle Word2Vec généré à partir du corpus Wikipedia
  - (paramètres par défaut, algorithme CBOW, 100 dimensions)
- Proximité moyenne des couples au sein des triplets

	P(AgV)	P(AgAc)	P(VbAc)
Wikipedia	0.25	0.29	0.39
Le Monde	0.25	0.28	0.34

Premières observations :

- Cartographie des familles dérivationnelles

Agent	Verbe	Action	P(AgVb)	P(AgAc)	P(VbAc)
danseur	danser	danse	0,64	0,67	0,67
graveur	graver	gravure	0,51	0,64	0,63
ouvreur	ouvrir	ouvrage	0,04	0,04	0,01
accepteur	accepter	acception	<0,01	0,06	0,01
batteuse	battre	battement	0,06	0,01	0,01
échangeur	échanger	échange	<0,01	0,21	0,46
fonceur	foncer	fonçage	0,21	0,04	0,01

Premières observations :

- Familles homogènes et cohésives

Agent	Verbe	Action	P(AgVb)	P(AgAc)	P(VbAc)
danseur	danser	danse	0,64	0,67	0,67
graveur	graver	gravure	0,51	0,64	0,63
ouvreur	ouvrir	ouvrage	0,04	0,04	0,01
accepteur	accepter	acception	<0,01	0,06	0,01
batteuse	battre	battement	0,06	0,01	0,01
échangeur	échanger	échange	<0,01	0,21	0,46
fonceur	foncer	fonçage	0,21	0,04	0,01

Premières observations

- Familles "désunies"

Agent	Verbe	Action	P(AgVb)	P(AgAc)	P(VbAc)
danseur	danser	danse	0,64	0,67	0,67
graveur	graver	gravure	0,51	0,64	0,63
ouvreur	ouvrir	ouvrage	0,04	0,04	0,01
accepteur	accepter	acception	<0,01	0,06	0,01
batteuse	battre	battement	0,06	0,01	0,01
échangeur	échanger	échange	<0,01	0,21	0,46
fonceur	foncer	fonçage	0,21	0,04	0,01

Premières observations

- Familles hétérogènes, "intrus"

Agent	Verbe	Action	P(AgVb)	P(AgAc)	P(VbAc)
danseur	danser	danse	0,64	0,67	0,67
graveur	graver	gravure	0,51	0,64	0,63
ouvreur	ouvrir	ouvrage	0,04	0,04	0,01
accepteur	accepter	acception	<0,01	0,06	0,01
batteuse	battre	battement	0,06	0,01	0,01
échangeur	échanger	échange	<0,01	0,21	0,46
fonceur	foncer	fonçage	0,21	0,04	0,01

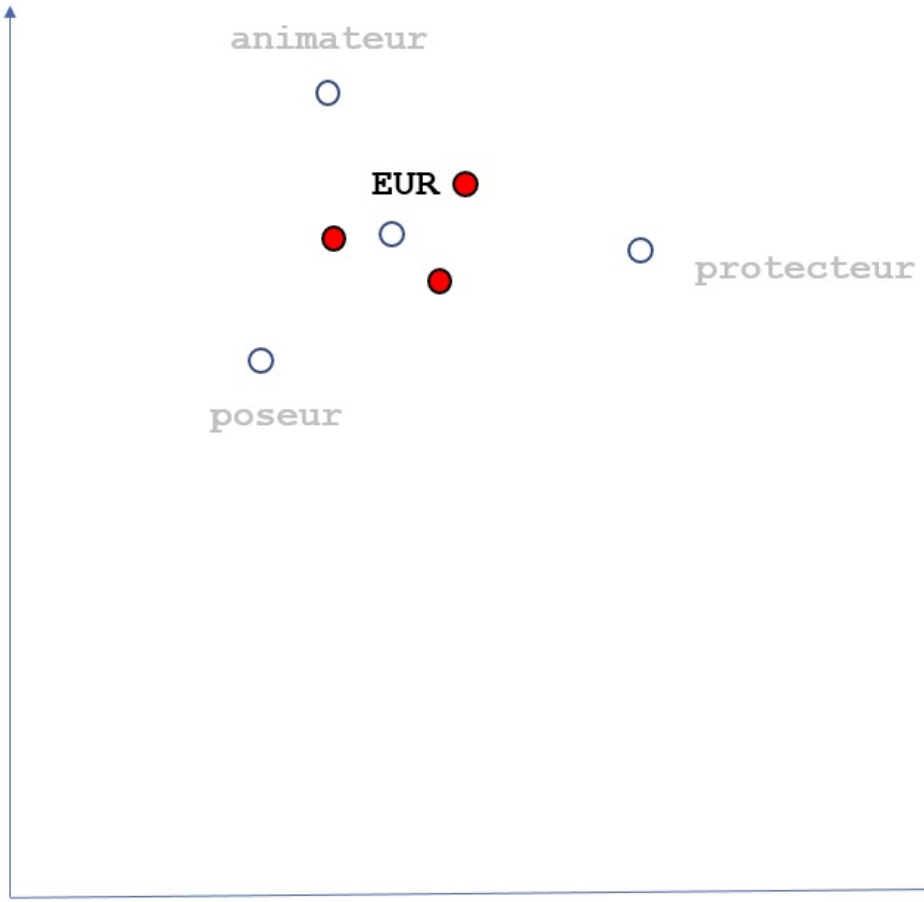
Représenter l'instruction sémantique d'un suffixe ?

- Vecteur de référence des mots d'une classe
- Une représentation abstraite : un point dans un espace vectoriel
  - Vecteur moyen pour représenter le sens prototypique des mots comportant un suffixe

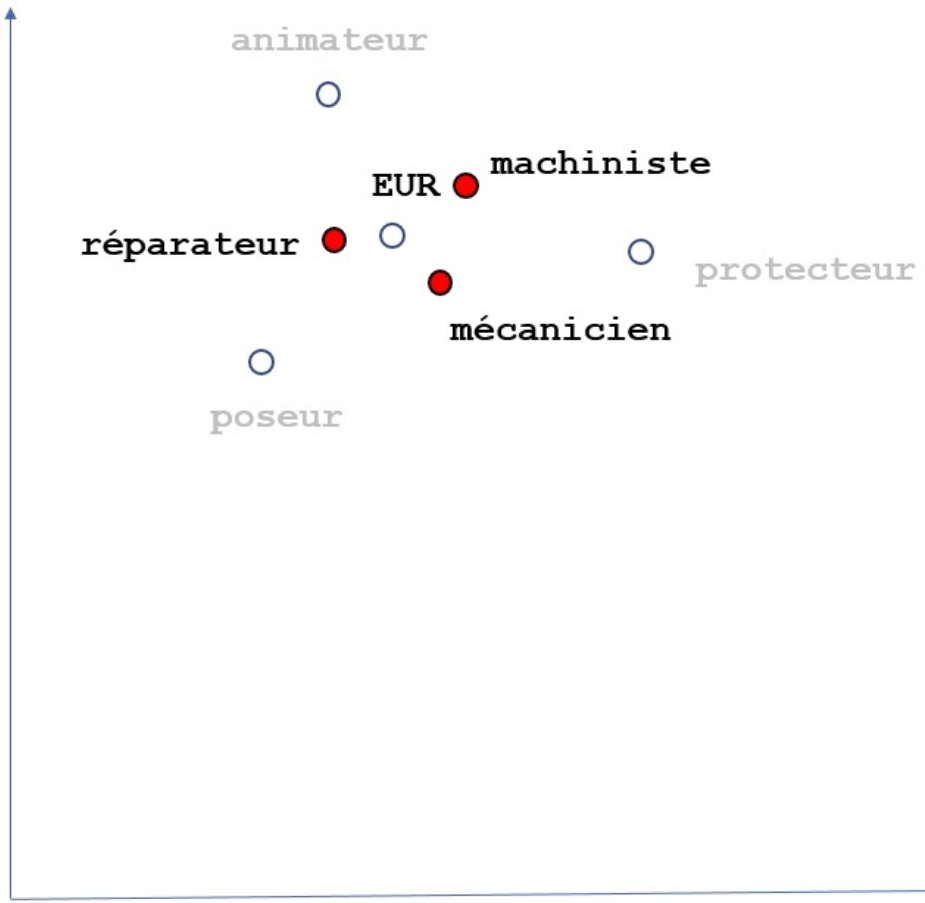
$$\overrightarrow{SUFF} = \frac{\overrightarrow{\text{Nsuff}_1} + \overrightarrow{\text{Nsuff}_2} + ... + \overrightarrow{\text{Nsuff}_n}}{n}$$

- Approximation du sens représenté grâce aux voisins les plus proches

Calcul du vecteur moyen



Calcul du vecteur moyen



Interprétation du vecteur moyen par l'examen des voisins les plus proches

-EUR

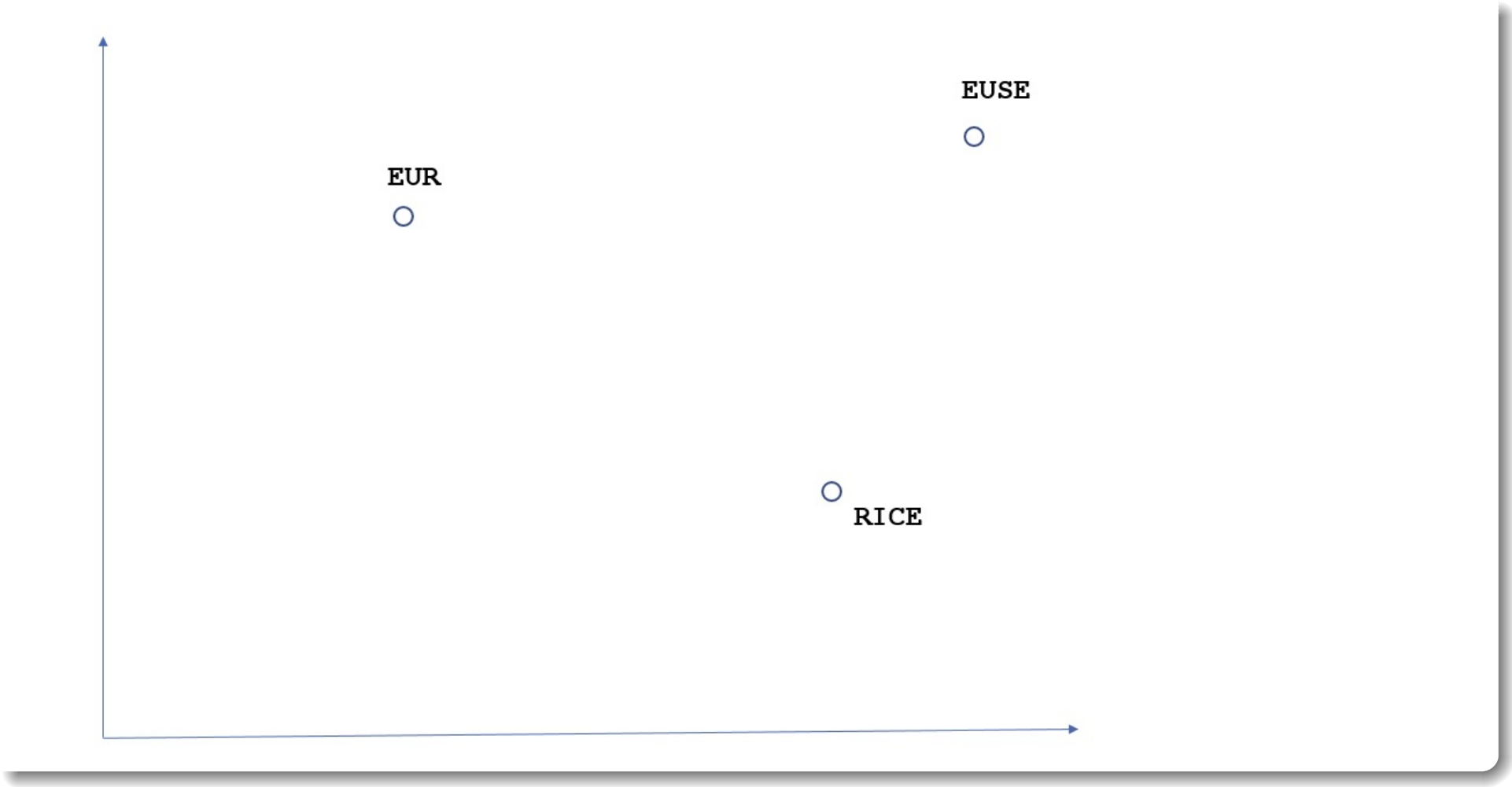
réparateur, sèche-cheveux, soudeur, armurier, minuteur, wattman, conducteur, laborantin, machiniste, mécanicien, plombier, tournevis, stéthoscope, client, mécano, coursier, déménageur, manomètre, aspirateur, soigneur, extincteur, vendeur, installateur, toiletteur, mélangeur, cric, ampèremètre, goniomètre, débogueur, technicien, ramasse-miettes, contacteur, descendeur, dépresseur, tune-o-matic, leurre, télérupteur, coupe-ongles, égoutier, microphone, juge-arbitre, opticien, nettoyeur, adaptateur, grappin, détecteur, ordinateur

Interprétation du vecteur moyen par l'examen des voisins les plus proches

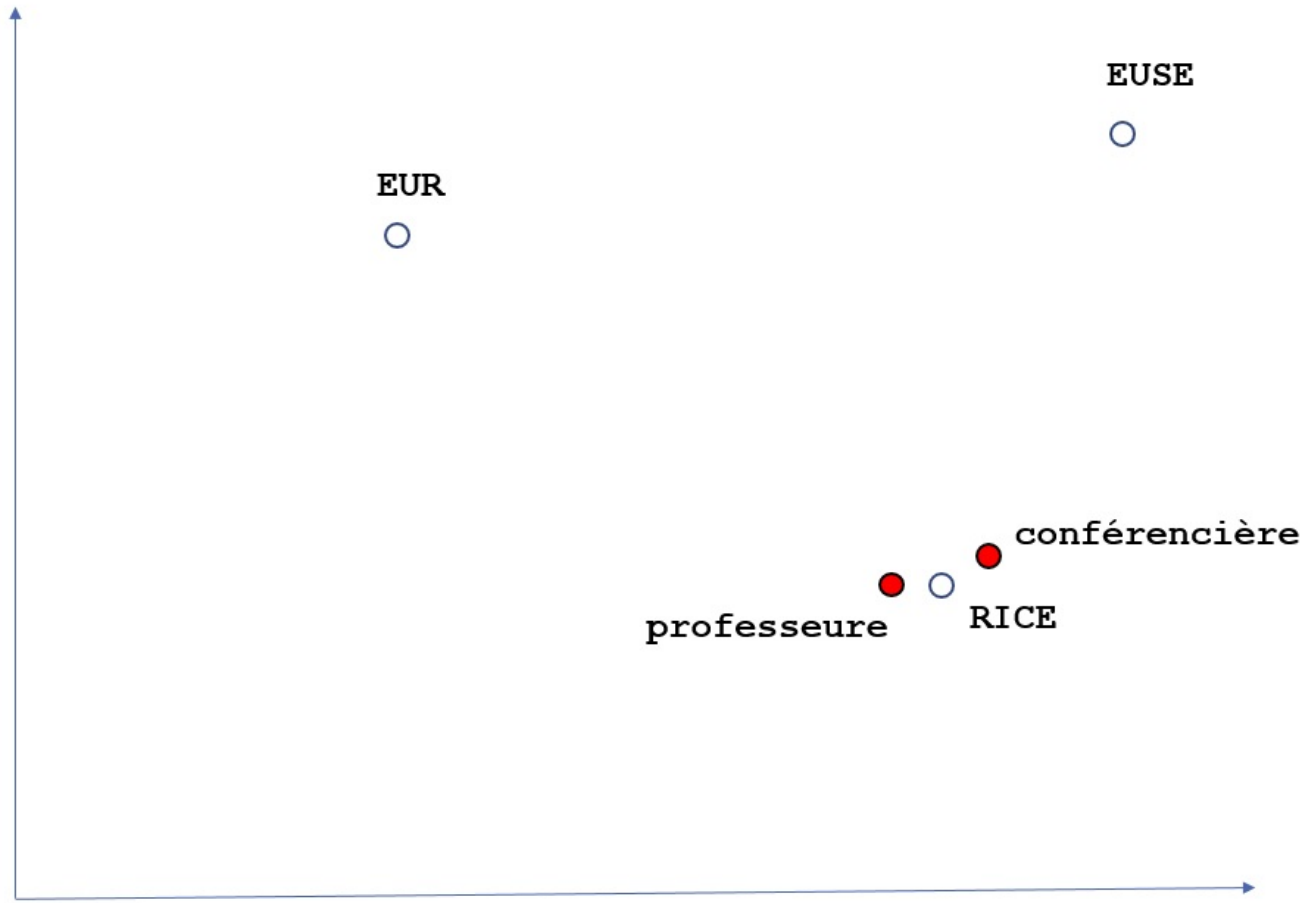
-EUR : des noms d'agent et d'instrument

réparateur, sèche-cheveux, soudeur, armurier, minuteur, wattman, conducteur, laborantin, machiniste, mécanicien, plombier, tournevis, stéthoscope, client, mécano, coursier, déménageur, manomètre, aspirateur, soigneur, extincteur, vendeur, installateur, toiletteur, mélangeur, cric, ampèremètre, goniomètre, débogueur, technicien, ramasse-miettes, contacteur, descendeur, dépresseur, tune-o-matic, leurre, télérupteur, coupe-ongles, égoutier, microphone, juge-arbitre, opticien, nettoyeur, adaptateur, grappin, détecteur, ordinateur

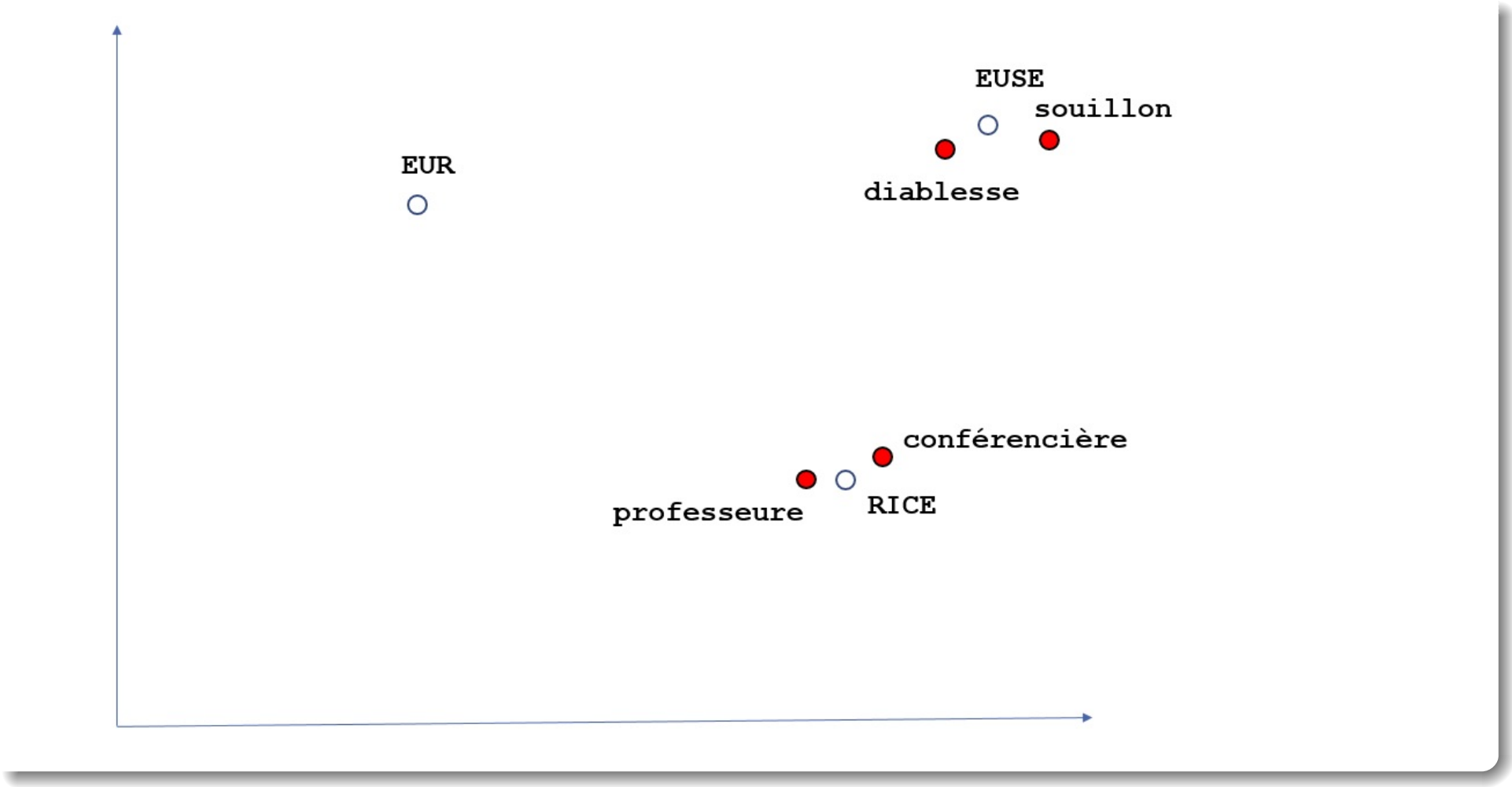
Calcul du vecteur moyen



Calcul du vecteur moyen



Calcul du vecteur moyen



Interprétation du vecteur moyen par l'examen des voisins les plus proches

-RICE : des métiers et des noms de femme

professeure, co-fondatrice, cofondatrice, herzigova, directrice, pharmacienne, traductrice, chercheure, saint-lucienne, venhard, éducatrice, co-directrice, fondatrice, laury, conférencière, comédienne, blogueuse, gogean, desmarais-rondeau, assistante, mageina, musicienne, vyghen, ingénieure, gérante, mammamia, spaziani, anska, bouhenni, slávka, joano, séménoff, herzigová, shrier, dartonne, warmus, présentatrice, bourgeois-leclerc, tonietti, otternaud, directrice-adjointe, guirous, saller, sculptrice, tshiteya, naymark, écrivaine, rajskub, pomfresh, fadeïeva

-EUSE : des qualificatifs peu valorisés voire dépréciatifs

gitane, trulle, chauffeuse, manucure, soubrette, trapéziste, coiffeuse, chocolatière, cuisinière, minouche, salopette, allumeuse, barancey, herzigova, souillon, diablesse , cochonne, vericel, serveuse, sorokina, stroyberg, naymark, rivale, corré, venhard, sarbel, kajmak, râblure, fédora, montalant, poulaine, stripteaseuse, catzéflis, minijupe, rosine, mariée, ptereleotris, tallier, irma, suffel, cover-girl, épicière, marie-olivier, javotte, kerny, basquaise, emilienne, estragnat, tigresse

Un outil prometteur :

- Manipuler l'information distributionnelle de façon variée
- Traiter aisément des masses importantes de données
- Concevoir des procédures nouvelles d'exploration des données

A manipuler avec discernement :

- Opacité, manque de contrôle sur les résultats
- Une question à résoudre : comment articuler procédures quantitatives et analyses qualitatives ?