Réseau IN-OVIVE

Relier automatiquement des entités textuelles à des concepts d'une ontologie par apprentissage avec (presque) aucune donnée

Arnaud Ferré

BIBLIOME, MaIAGE INRA, Université Paris-Saclay	ILES, LIMSI CNRS, Université Paris-Saclay

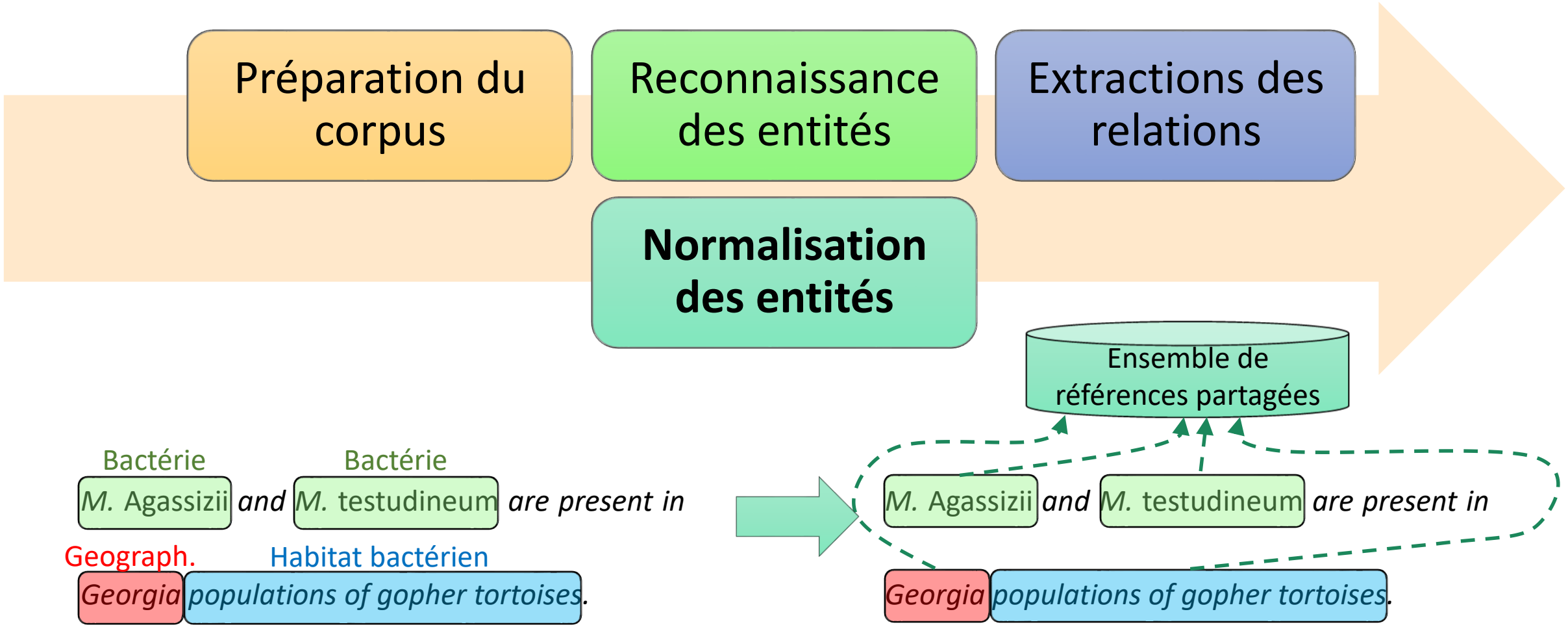
INRA SCIENCE & IMPACT

UNIVERSITÉ PARIS SUD — université PARIS-SACLAY

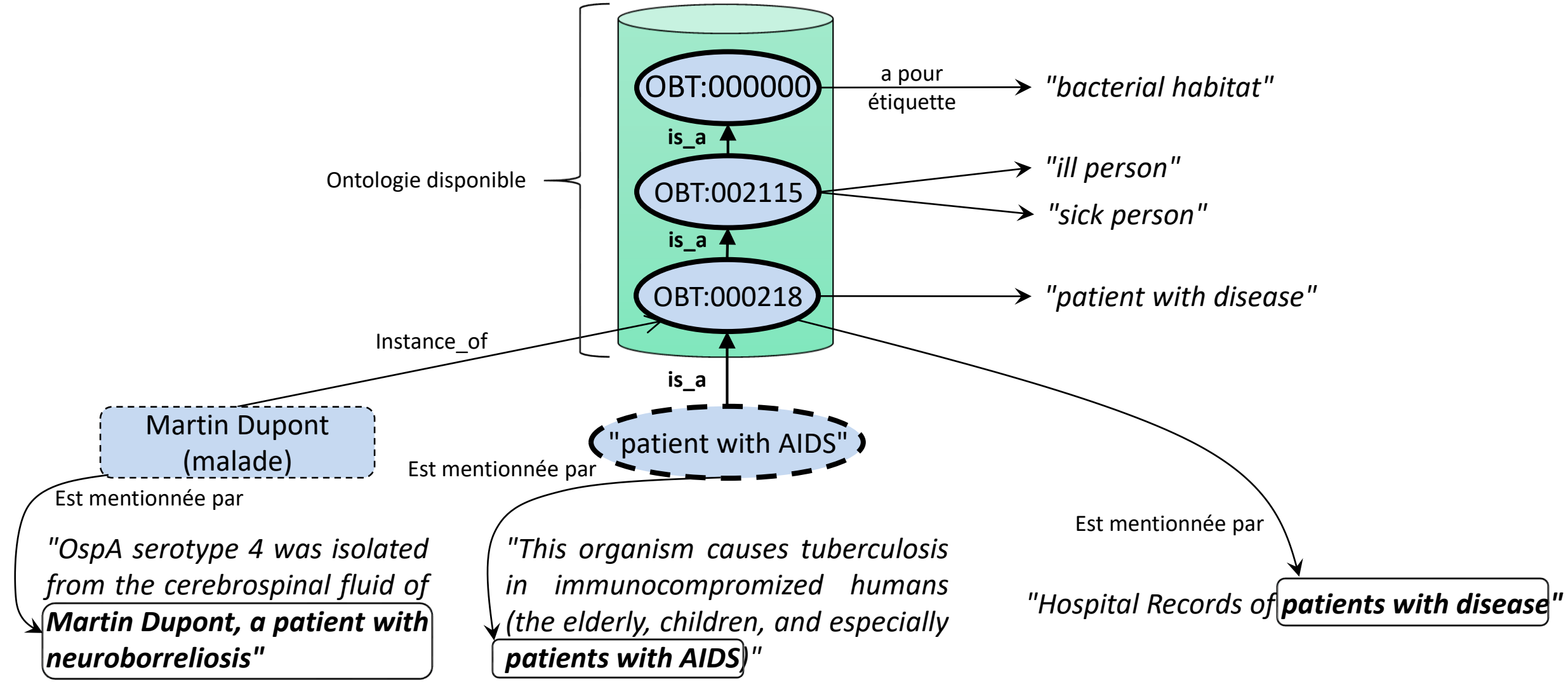
cnrs

INTRODUCTION

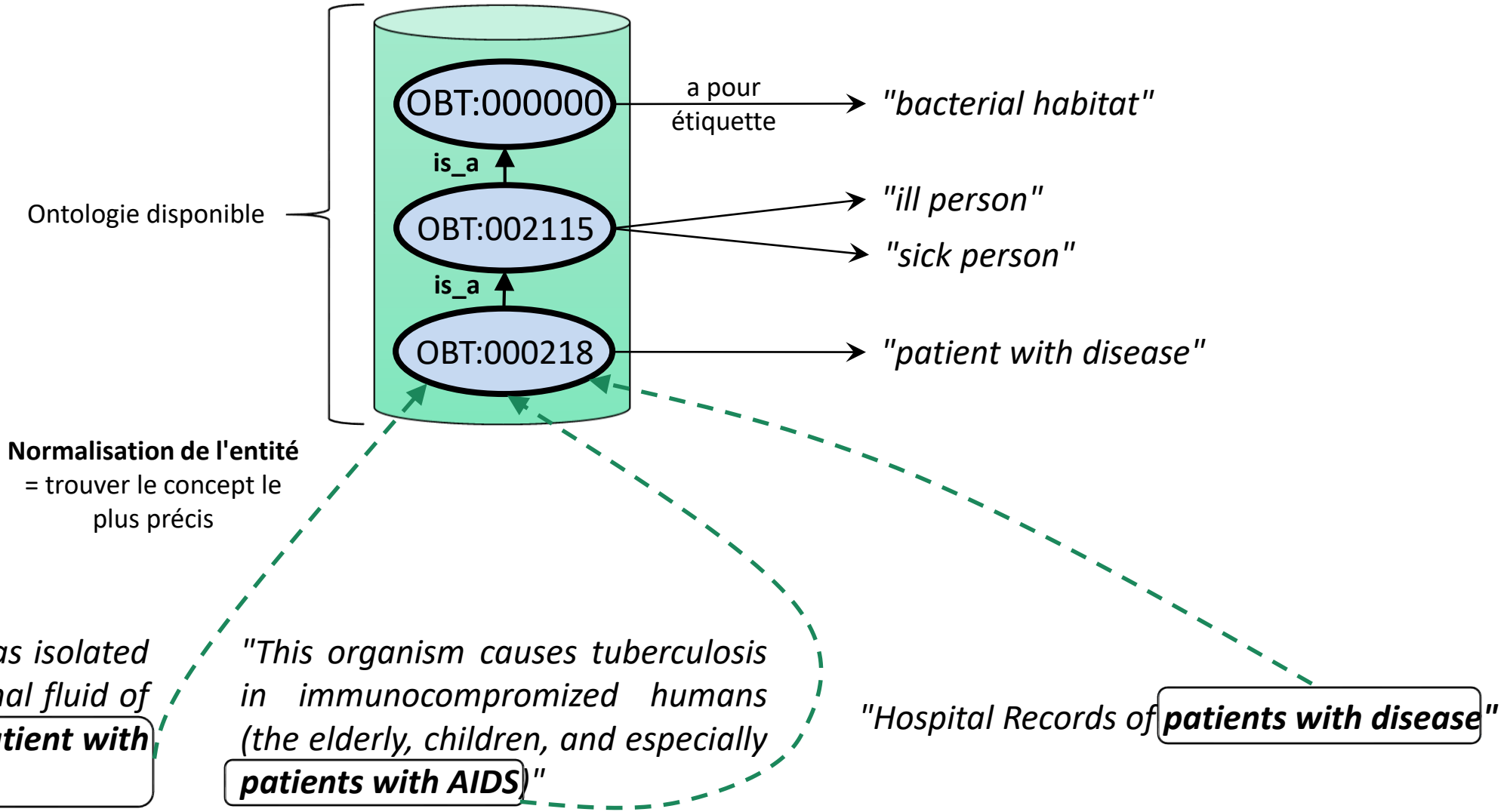
L'extraction d'information



Normalisation par les concepts d'une ontologie

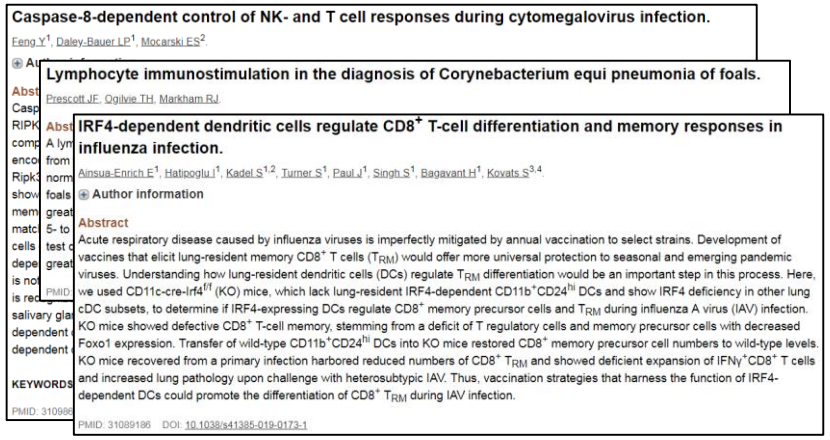


Normalisation par les concepts d'une ontologie



Intérêts de la normalisation d'entités

Extraction d'Information sans normalisation :



Caspase-8-dependent control of NK- and T cell responses during cytomegalovirus infection.

Feng Y[1], Daley-Bauer LP[1], Mocarski ES[2].

Lymphocyte immunostimulation in the diagnosis of Corynebacterium equi pneumonia of foals.

Prescott JF, Ogilvie TH, Markham RJ.

IRF4-dependent dendritic cells regulate CD8$^+$ T-cell differentiation and memory responses in influenza infection.

Ainsua-Enrich E[1], Hatipoglu I[1], Kadel S[1,2], Turner S[1], Paul J[1], Singh S[1], Bagavant H[1], Kovats S[3,4].

Author information

Abstract

Acute respiratory disease caused by influenza viruses is imperfectly mitigated by annual vaccination to select strains. Development of vaccines that elicit lung-resident memory CD8$^+$ T cells (T_{RM}) would offer more universal protection to seasonal and emerging pandemic viruses. Understanding how lung-resident dendritic cells (DCs) regulate T_{RM} differentiation would be an important step in this process. Here, we used CD11c-cre-Irf4$^{f/f}$ (KO) mice, which lack lung-resident IRF4-dependent CD11b$^+$CD24^{hi} DCs and show IRF4 deficiency in other lung cDC subsets, to determine if IRF4-expressing DCs regulate CD8$^+$ memory precursor cells and T_{RM} during influenza A virus (IAV) infection. KO mice showed defective CD8$^+$ T-cell memory, stemming from a deficit of T regulatory cells and memory precursor cells with decreased Foxo1 expression. Transfer of wild-type CD11b$^+$CD24^{hi} DCs into KO mice restored CD8$^+$ memory precursor cell numbers to wild-type levels. KO mice recovered from a primary infection harbored reduced numbers of CD8$^+$ T_{RM} and showed deficient expansion of IFNγ$^+$CD8$^+$ T cells and increased lung pathology upon challenge with heterosubtypic IAV. Thus, vaccination strategies that harness the function of IRF4-dependent DCs could promote the differentiation of CD8$^+$ T_{RM} during IAV infection.

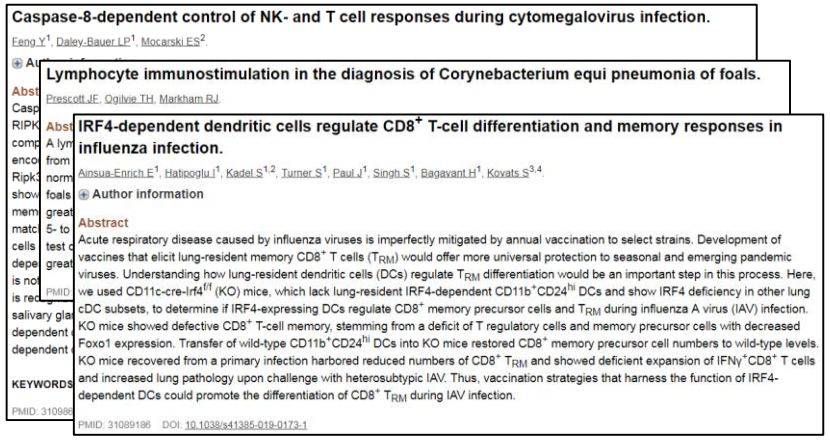
PMID: 31089186 DOI: 10.1038/s41385-019-0173-1

Lives_in	
Bactérie	**Habitat (mention)**
Listeria	*"CD20"*
Pseudomonas	*"CD20-positive cells"*
Corynebacterium	*"monoclonal B cells"*
Pseudomonas	*"T cells"*
Listeria	*"**Lymphoc**ytic"*

Requête : Bactéries infectieuses "*lymphocite*" ? ⟹ 0 ou 1 réponse

Intérêts de la normalisation d'entités

Dès qu'il est nécessaire de comparer des expressions textuelles entre elles ou avec des concepts :



Caspase-8-dependent control of NK- and T cell responses during cytomegalovirus infection.

Feng Y[1], Daley-Bauer LP[1], Mocarski ES[2].

Lymphocyte immunostimulation in the diagnosis of Corynebacterium equi pneumonia of foals.

Prescott JF, Ogilvie TH, Markham RJ.

IRF4-dependent dendritic cells regulate CD8$^+$ T-cell differentiation and memory responses in influenza infection.

Ainsua-Enrich E[1], Hatipoglu I[1], Kadel S[1,2], Turner S[1], Paul J[1], Singh S[1], Bagavant H[1], Kovats S[3,4].

Author information

Abstract

Acute respiratory disease caused by influenza viruses is imperfectly mitigated by annual vaccination to select strains. Development of vaccines that elicit lung-resident memory CD8$^+$ T cells (T_{RM}) would offer more universal protection to seasonal and emerging pandemic viruses. Understanding how lung-resident dendritic cells (DCs) regulate T_{RM} differentiation would be an important step in this process. Here, we used CD11c-cre-Irf4$^{f/f}$ (KO) mice, which lack lung-resident IRF4-dependent CD11b$^+$CD24^{hi} DCs and show IRF4 deficiency in other lung cDC subsets, to determine if IRF4-expressing DCs regulate CD8$^+$ memory precursor cells and T_{RM} during influenza A virus (IAV) infection. KO mice showed defective CD8$^+$ T-cell memory, stemming from a deficit of T regulatory cells and memory precursor cells with decreased Foxo1 expression. Transfer of wild-type CD11b$^+$CD24^{hi} DCs into KO mice restored CD8$^+$ memory precursor cell numbers to wild-type levels. KO mice recovered from a primary infection harbored reduced numbers of CD8$^+$ T_{RM} and showed deficient expansion of IFNγ$^+$CD8$^+$ T cells and increased lung pathology upon challenge with heterosubtypic IAV. Thus, vaccination strategies that harness the function of IRF4-dependent DCs could promote the differentiation of CD8$^+$ T_{RM} during IAV infection.

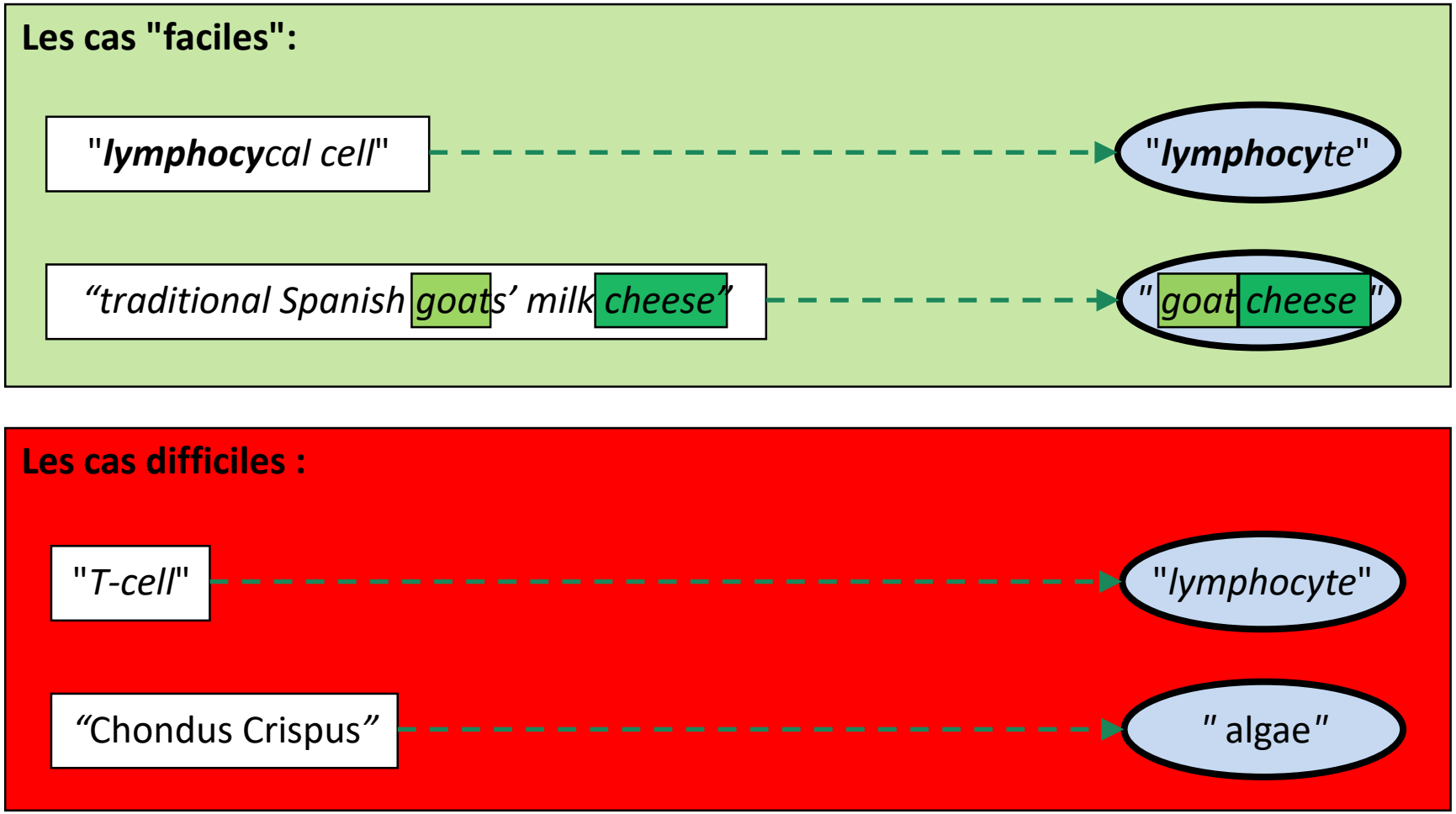
PMID: 31089186 DOI: 10.1038/s41385-019-0173-1

Lives_in

Bactérie	Habitat (mention)	Habitat (concept)
Listeria	*"CD20"*	OBT:001342: lymphocyte
Pseudomonas	*"CD20-positive cells"*	OBT:001342: lymphocyte
Corynebacterium	*"monoclonal B cells"*	OBT:001342: lymphocyte
Pseudomonas	*"T cells"*	OBT:001342: lymphocyte
Listeria	*"Lymphocytic"*	OBT:001342: lymphocyte

Requête : Bactéries infectieuses OBT:001342: lymphocyte ? ⟹ 5 réponses

Difficulté : variation de forme



MetaMap (Aronson, 2001)
ToMap (Golik et al., 2011)
(Claveau, 2013)
BOUN (Tiftikci et al., 2016)

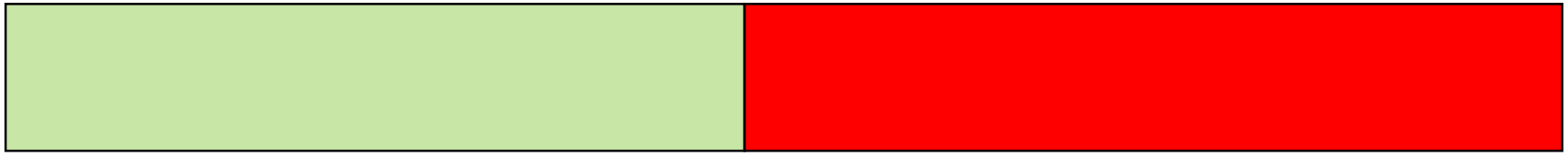
(+ homonymies
ex : "*avocat*" désigne-t-il un métier ou un fruit ?)

Difficulté : variation de forme

Les tâches de normalisation "faciles":

(ex : normalisation de mentions de bactéries)

Les tâches de normalisation difficiles :

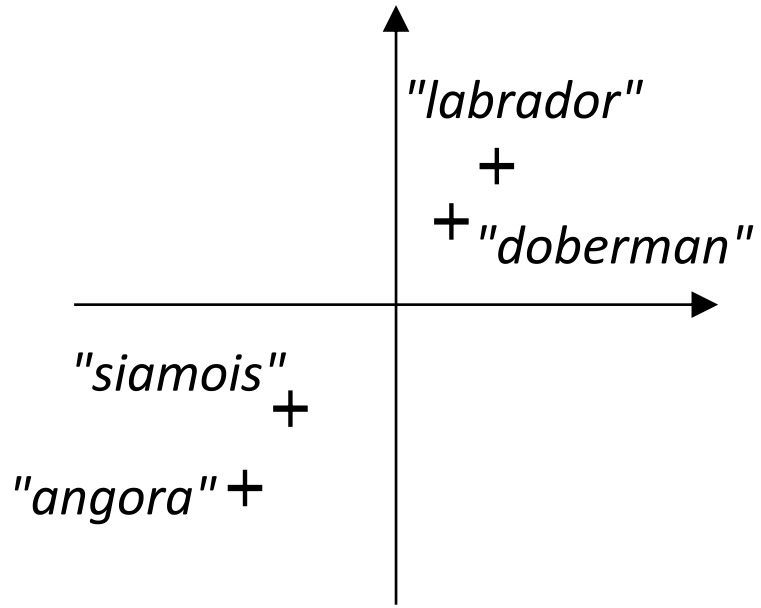


(ex : normalisation de mentions d'habitats bactériens)

ÉTAT DE L'ART

Approche fondée sur des représentations vectorielles

Une réponse au problème de variation de forme : Les espaces distributionnels.



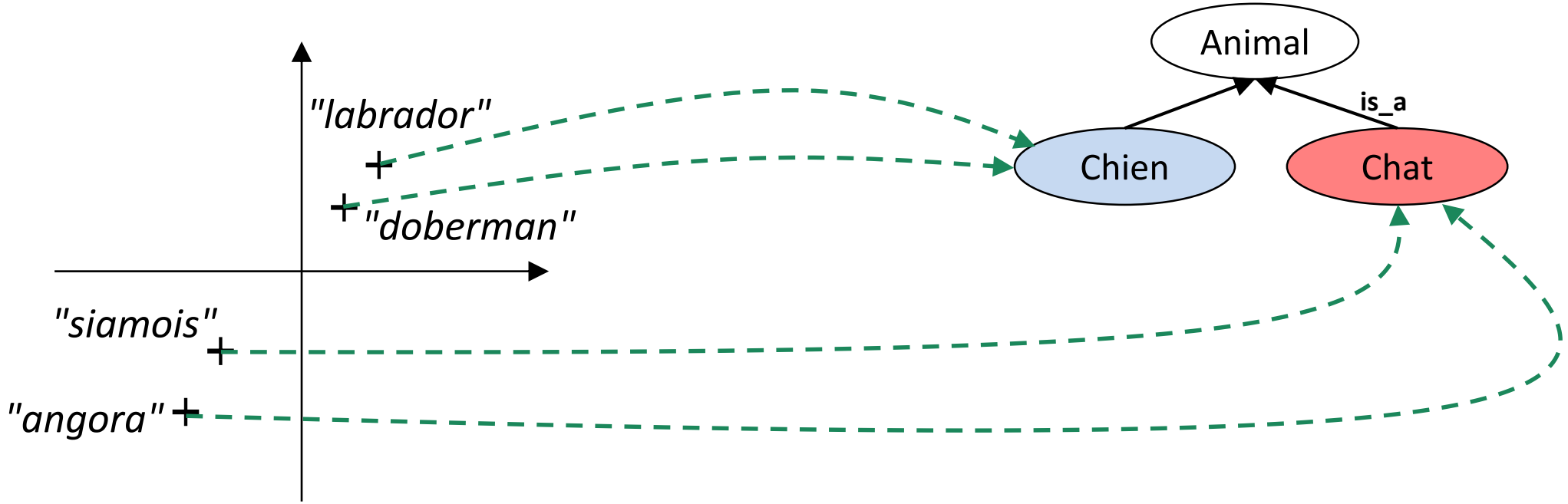
Représentations distributionnelles :
- (Hinton et al., 1986)

Puis plongements lexicaux :
- **Word2Vec** (Mikolov et al. 2013)
- Glove (2013), FastText (2017), ELMo (2018)

Puis "modèles intégrés et connectables" :
- BERT (Devlin et al. fin 2018)
- XLNet (2019)

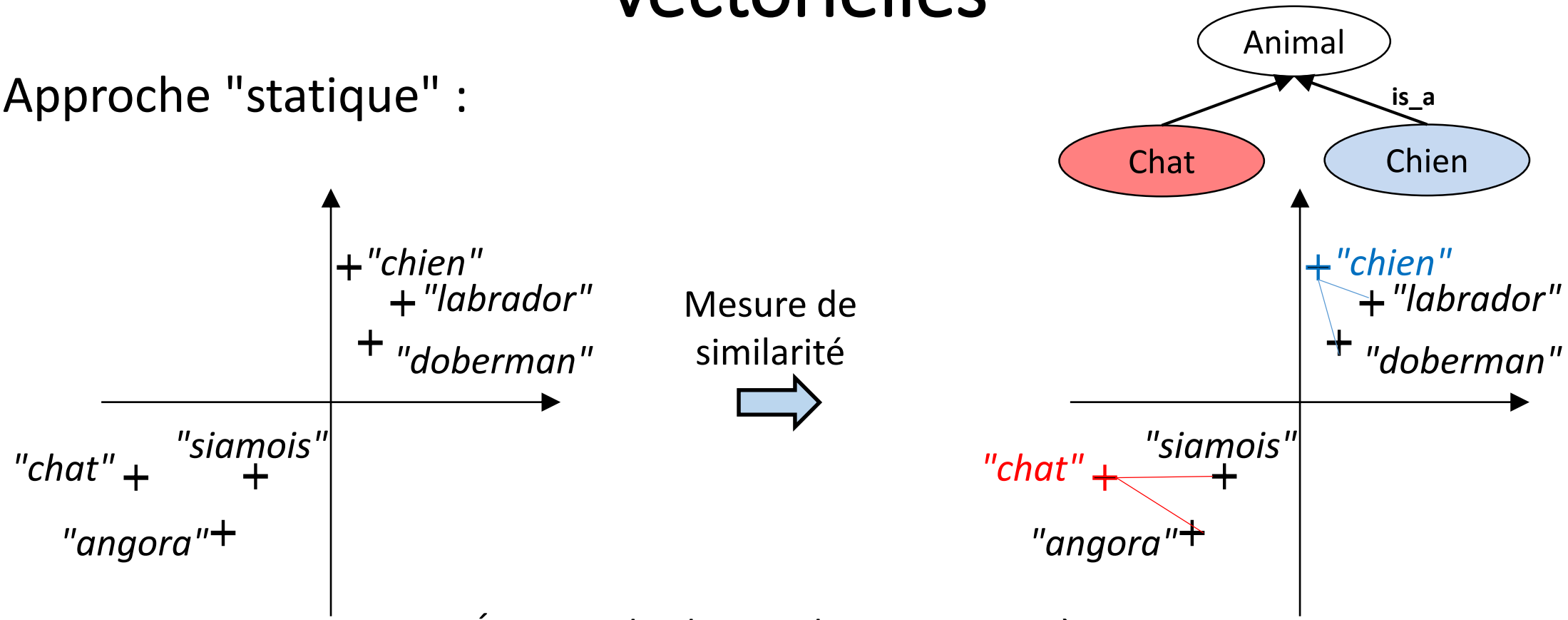
Approche fondée sur des représentations vectorielles

Objectif : classer les vecteurs de mentions dans un concept de l'ontologie



Approche fondée sur des représentations vectorielles

Approche "statique" :

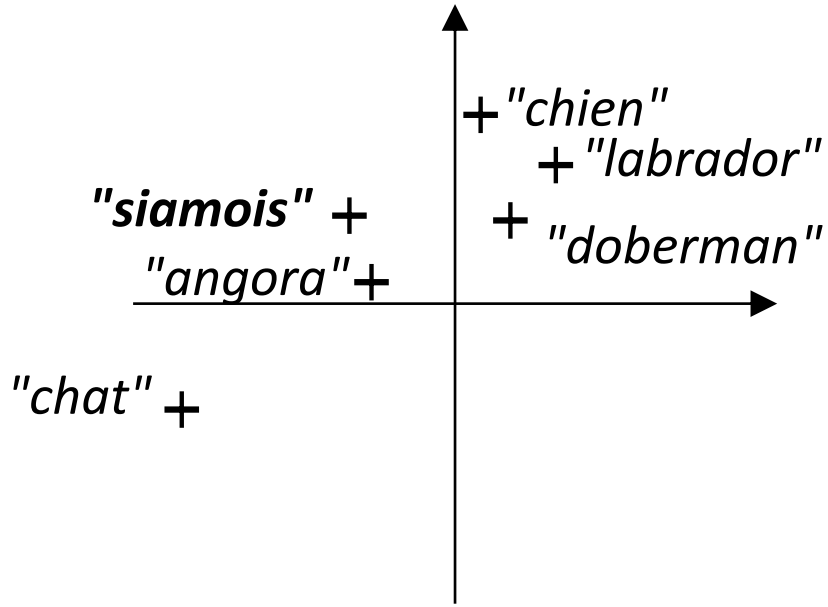


Étiquette la plus proche par rapport à une mesure particulière

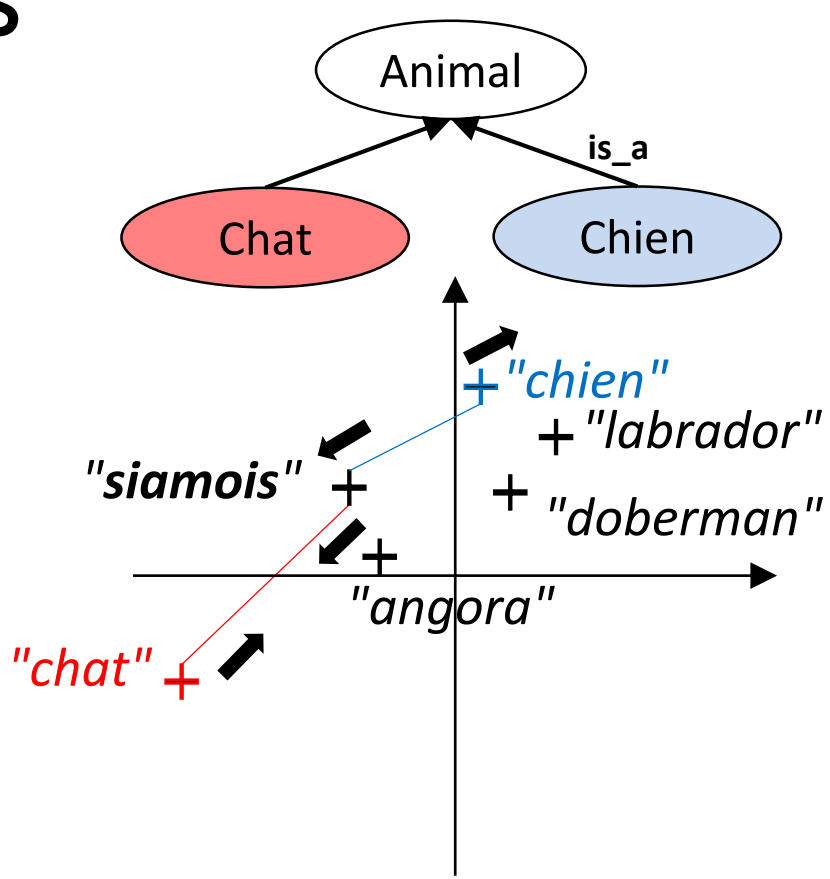
BOUNEL (Karadeniz et al., 2019)

Approche fondée sur des représentations vectorielles

Approche par apprentissage (régression pour apprendre une mesure de similarité adaptée) :



Mesure de similarité

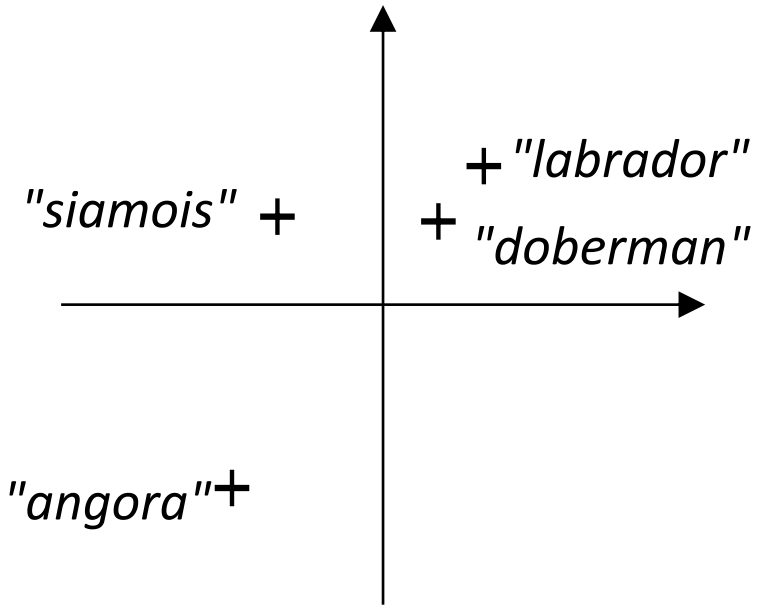


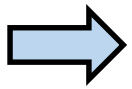
Fonction *sim* telle que :
sim("siamois", "chat") < sim("siamois", "chien")

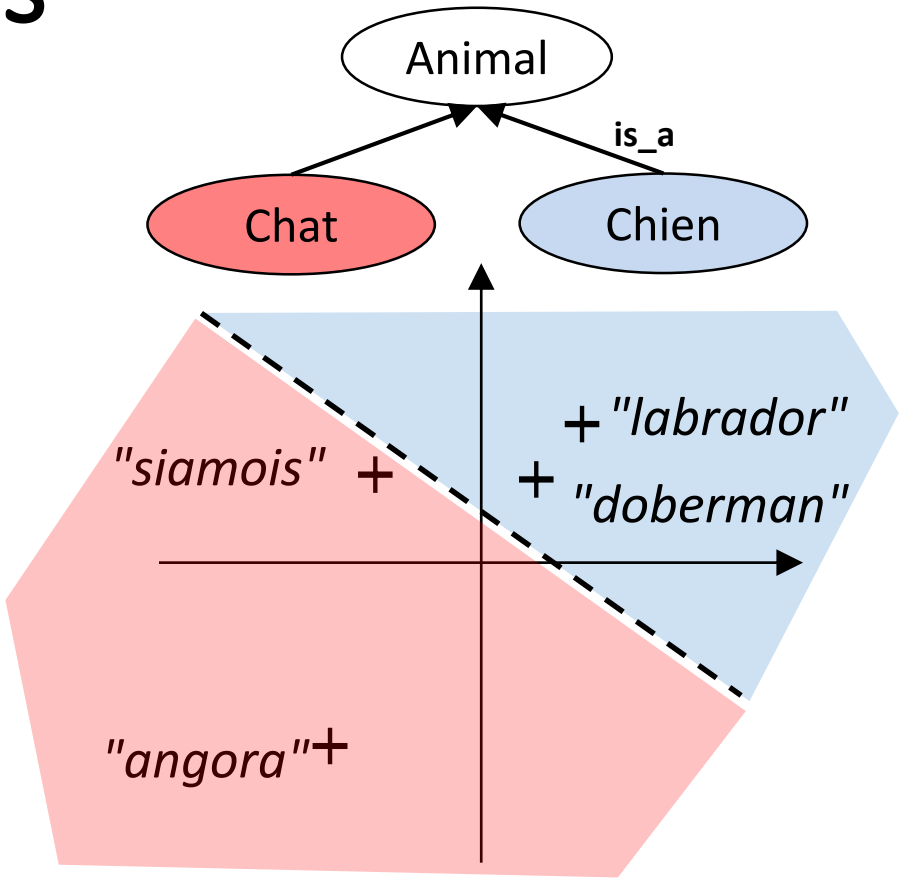
Avec des représentations non-distributionnelles :
Dnorm (Leaman et al., 2013)
TaggerOne (Leaman and Lu, 2016)

Approche fondée sur des représentations vectorielles

Approche par apprentissage (classification) :

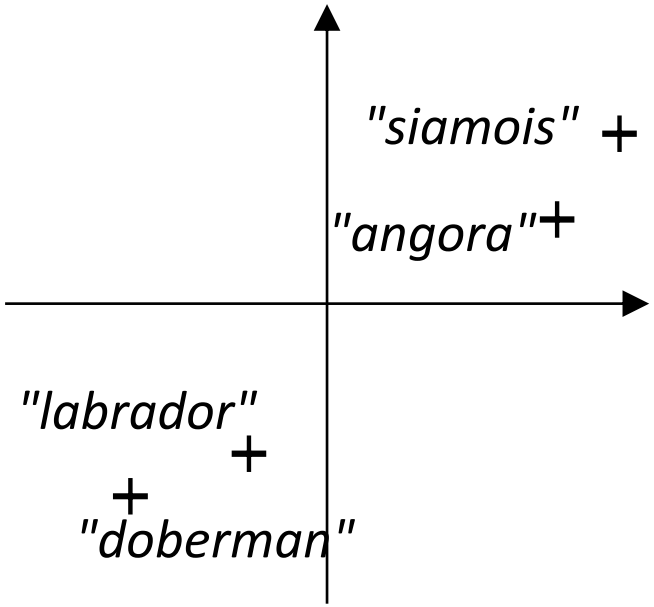


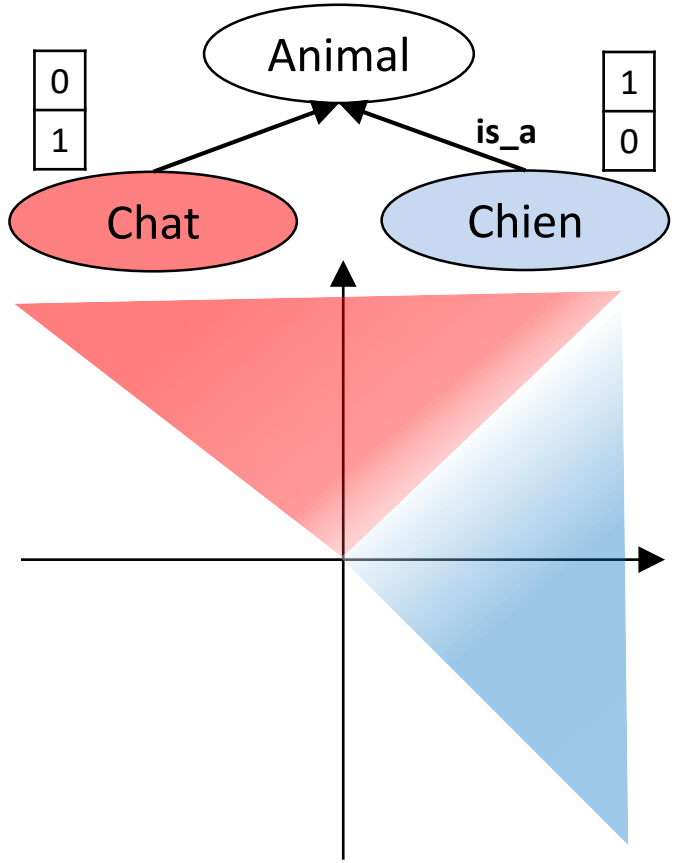




Approche fondée sur des représentations vectorielles

Approche par apprentissage (classification) :

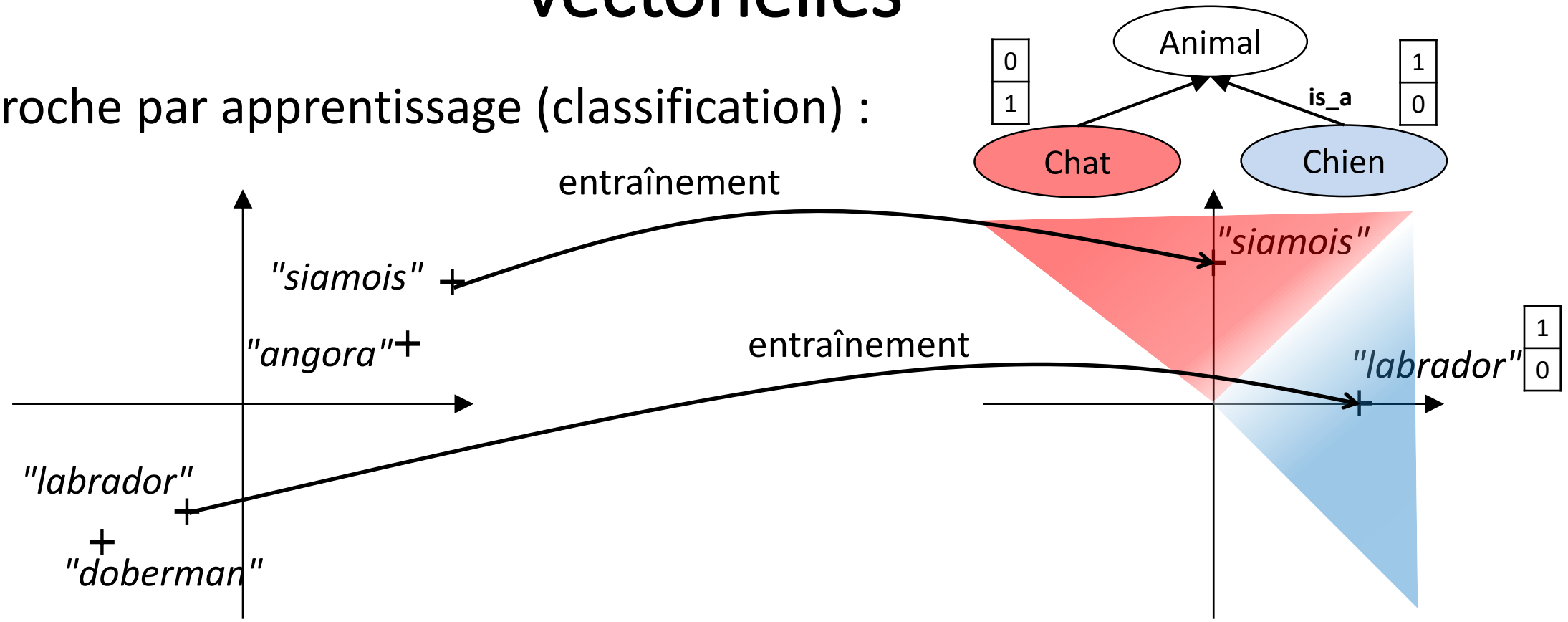




(Limsopatham and Collier, 2016)

Approche fondée sur des représentations vectorielles

Approche par apprentissage (classification) :



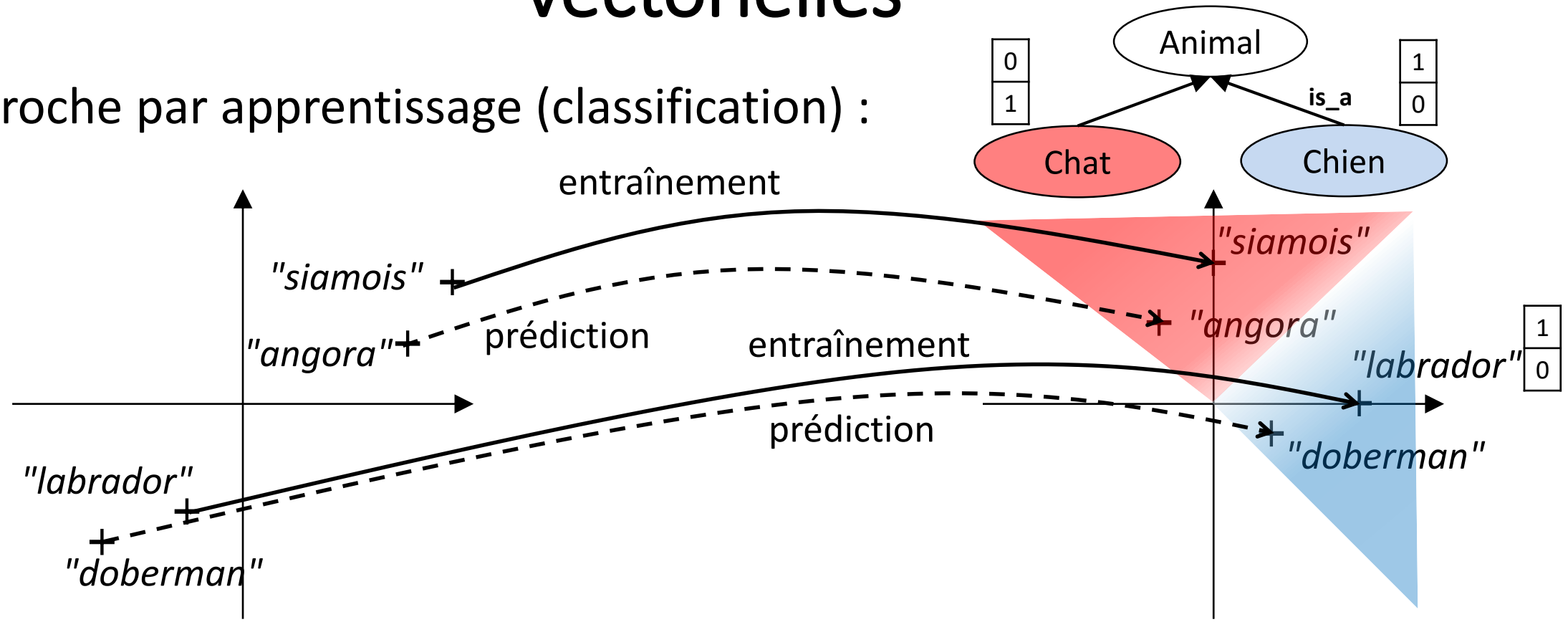
Projection p telle que :

$$\overrightarrow{\text{"angora"}} \approx \overrightarrow{\text{"siamois"}} \Longrightarrow p(\text{"angora"}) \approx p(\text{"siamois"})$$

(Limsopatham and Collier, 2016)

Approche fondée sur des représentations vectorielles

Approche par apprentissage (classification) :



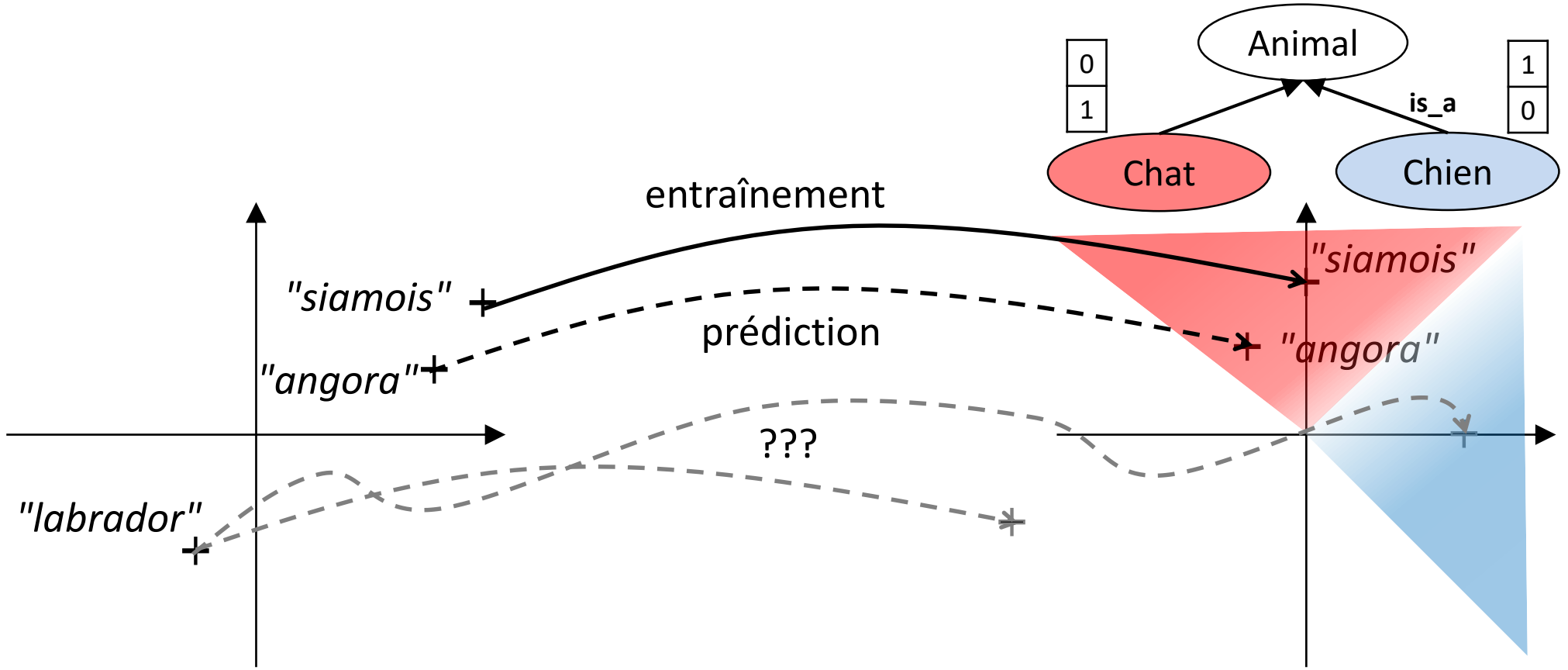
Projection p telle que :

$$\overrightarrow{\text{"angora"}} \approx \overrightarrow{\text{"siamois"}} \Longrightarrow p(\text{"angora"}) \approx p(\text{"siamois"})$$

(Limsopatham and Collier, 2016)

CONTRIBUTIONS

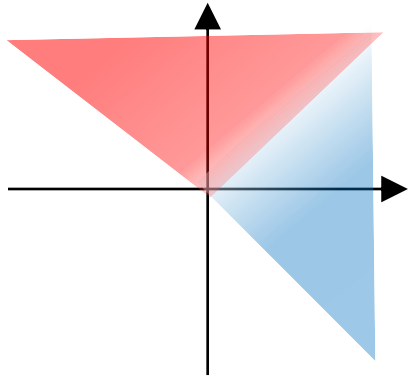
Difficulté abordée : "*zero/few/single-shot learning*"



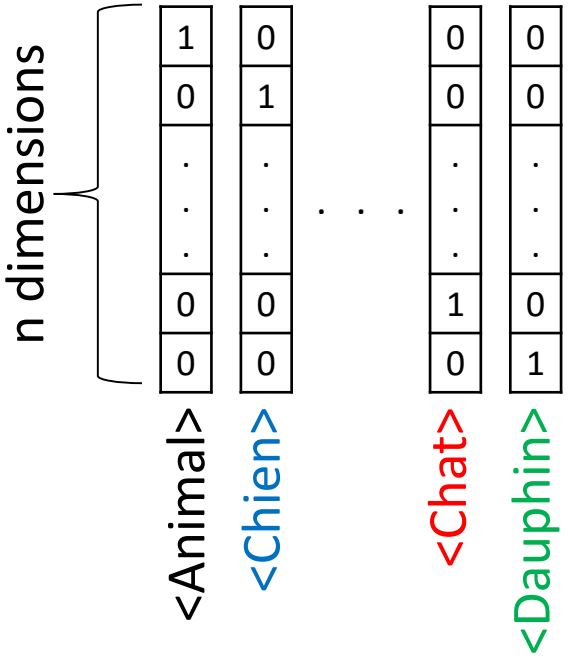
Problème de classification avec des classes n'apparaissant pas dans les exemples d'entraînement

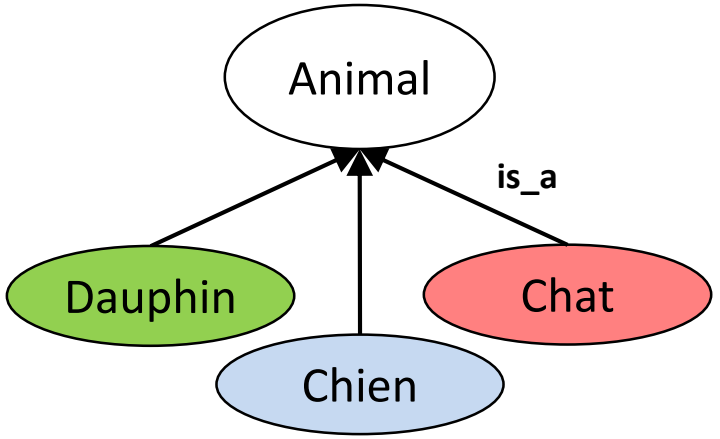
(Hugo Larochelle et al., 2008, "Zero-data learning of new tasks.")

Et si...

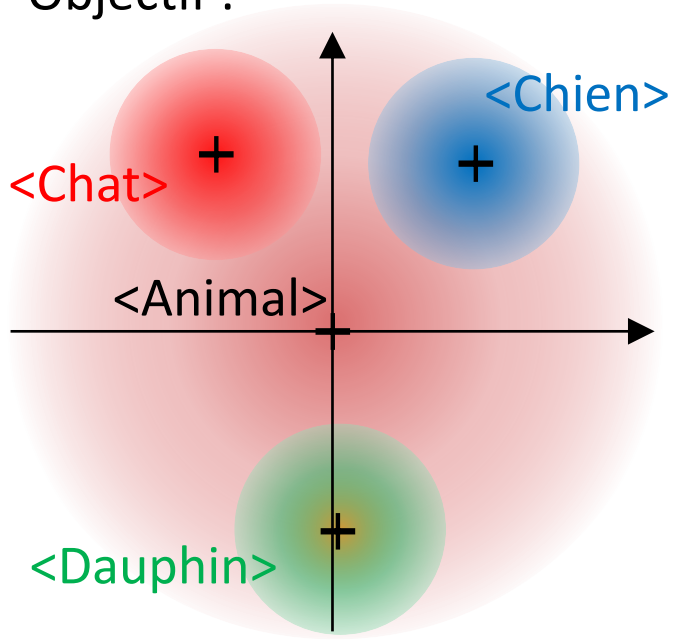


1-Parmi-N pour **n** concepts :



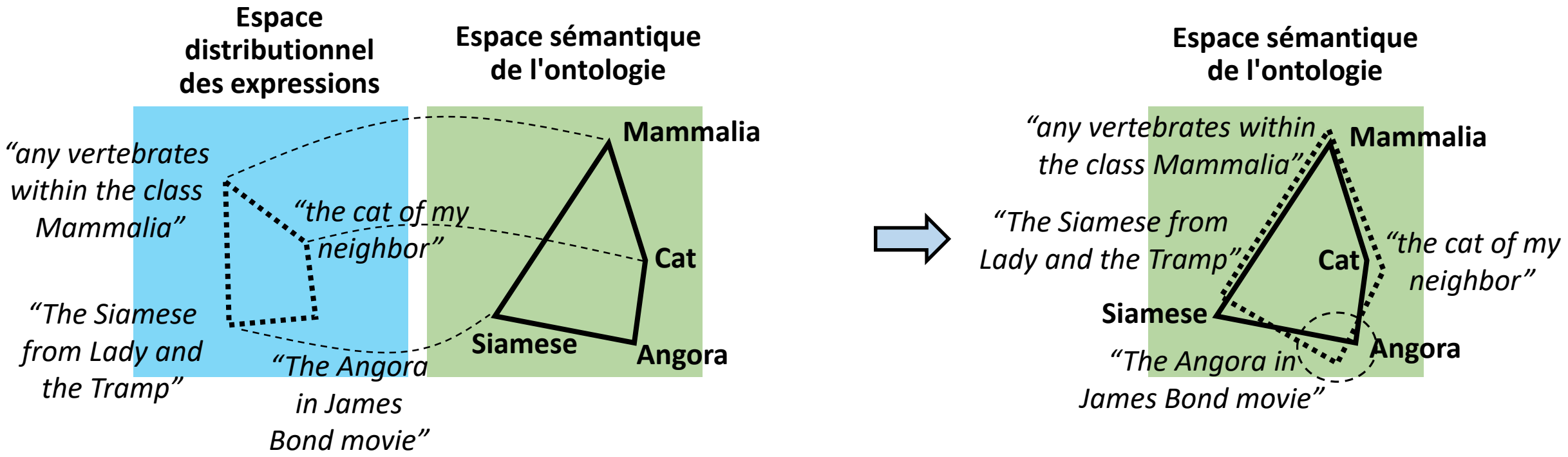


Objectif :

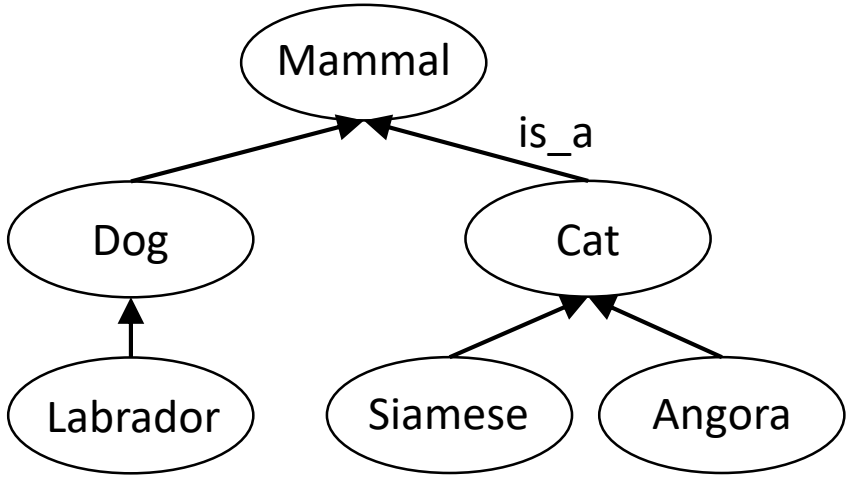


CONTES (“*CONcept-TErm System*”)

Hypothèse : Si les 2 espaces ont une structure similaire, alors il est possible de déterminer une projection qui permettra d'obtenir des prédictions pertinentes.



Ancestry

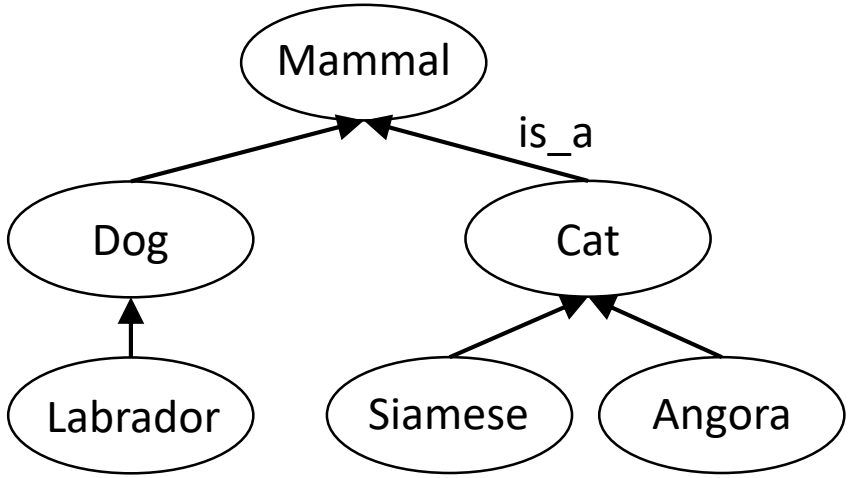


$$\forall k \in [\![1,n]\!], v_{c_k} = \left(w_{c_k}^0, \ldots, w_{c_k}^i, \ldots, w_{c_k}^n \right)$$

$$w_{c_k}^i = \begin{cases} 1 & si\ i = k \\ 1 & si\ c_i\ ancêtre\ de\ c_k \\ 0 & sinon \end{cases}$$

(Arnaud Ferré et al., 2017, "Representation of complex terms in a vector space structured by an ontology for a normalization task. ")

Ancestry

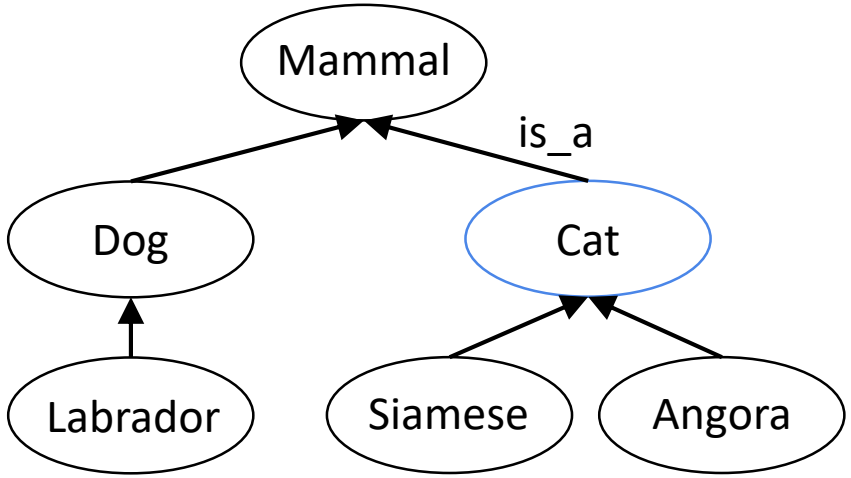


$$\forall k \in [\![1, n]\!], v_{c_k} = \left(w_{c_k}^0, \ldots, w_{c_k}^i, \ldots, w_{c_k}^n \right)$$

$$w_{c_k}^i = \begin{cases} 1 & si\ i = k \\ 1 & si\ c_i\ ancêtre\ de\ c_k \\ 0 & sinon \end{cases}$$

One-Hot	Mammal	Dog	Cat	Labrador	Siamois	Angora
$\overrightarrow{Mammal}$	1	0	0	0	0	0
$\overrightarrow{Dog}$	0	1	0	0	0	0
$\overrightarrow{Cat}$	0	0	1	0	0	0
$\overrightarrow{Labrador}$	0	0	0	1	0	0
$\overrightarrow{Siamese}$	0	0	0	0	1	0
$\overrightarrow{Angora}$	0	0	0	0	0	1

Ancestry



$$\forall k \in [\![1,n]\!], v_{c_k} = \left(w_{c_k}^0, \ldots, w_{c_k}^i, \ldots, w_{c_k}^n \right)$$

$$w_{c_k}^i = \begin{cases} 1 & si\ i = k \\ 1 & si\ c_i\ ancêtre\ de\ c_k \\ 0 & sinon \end{cases}$$

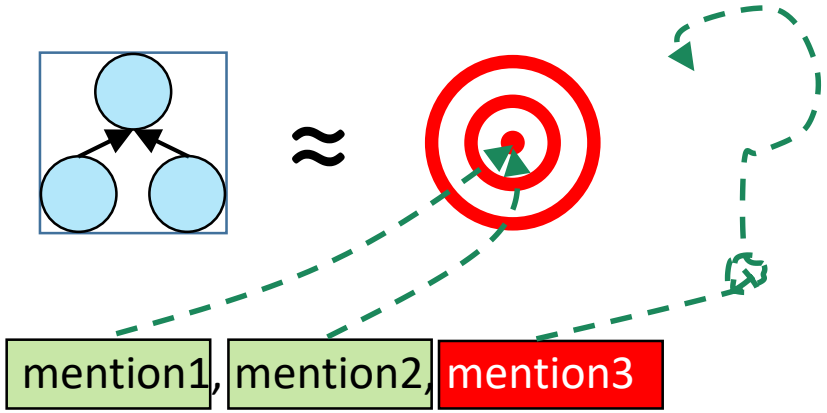
Ancestry	Mammal	Dog	Cat	Labrador	Siamois	Angora
$\overrightarrow{Mammal}$	1	0	0	0	0	0
$\overrightarrow{Dog}$	**1**	1	0	0	0	0
$\overrightarrow{Cat}$	**1**	0	1	0	0	0
$\overrightarrow{Labrador}$	**1**	**1**	0	1	0	0
$\overrightarrow{Siamese}$	**1**	0	**1**	0	1	0
$\overrightarrow{Angora}$	**1**	0	**1**	0	0	1

Similarité à Cat	**Similarité cosinus**
Cat	1,00
Siamese	0.82
Angora	0.82
Mammal	0.71
Dog	0.50
Labrador	0.41

(Ferré, 2017, RECITAL)

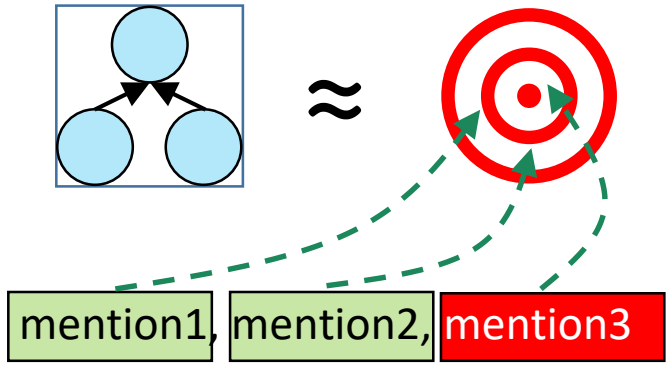
Méthodes mixtes : "Hierarchical Ontological NORmalization"

- Méthode mixte par intégration d'une méthode fondée sur l'étude de similarité de forme.



ToMap
(similarité de forme)

≠



CONTES
(approche distributionnelle + apprentissage)

(Arnaud Ferré, Louise Deléger, et al., 2018, LREC)

Résultats

	Nécessite des exemples annotés	Fondée sur la similarité de forme	Fondée sur un espace vectoriel	Fondée sur un espace distributionnel	Score global de similarité
Méthode de base (lemmatisation + appariement exact)	N	O	N	N	0,54
ToMap	N	O	N	N	0,61
ToMap (spécialisée au domaine)	N	O	N	N	0,66
BOUN (Tiftikci et al., 2016)	N	O	O	N	0,62
Turku (Mehryary et al., 2017)	N	O	O	N	0,63
BOUNEL (Karadeniz et al., 2019)	N	N	O	O	0,66
CONTES (Ferré et al. 2017)	O	N	O	O	0,61
HONOR (Ferré et al. 2018)	O	O	O	O	**0,74**
WSEP-CONTES (Ferré et al. 2018)	N	N	O	O	0,59
WSOT-CONTES (en cours de publication)	N	N	O	O	0,63
WSOT-HONOR (en cours de publication)	N	O	O	O	**0,73**
Full-CONTES (en cours de publication)	O	N	O	O	**0,69**
Full-HONOR (en cours de publication)	O	O	O	O	**0,76**

CONCLUSION ET PERSPECTIVES

CONCLUSION

Intégration en cours dans l'application logicielle d'extraction d'information pour Florilège : base de connaissance sur les habitats et phénotypes microbien.

- Codes partagés sur GitHub (3 contributeurs)

  https://github.com/ArnaudFerre/CONTES

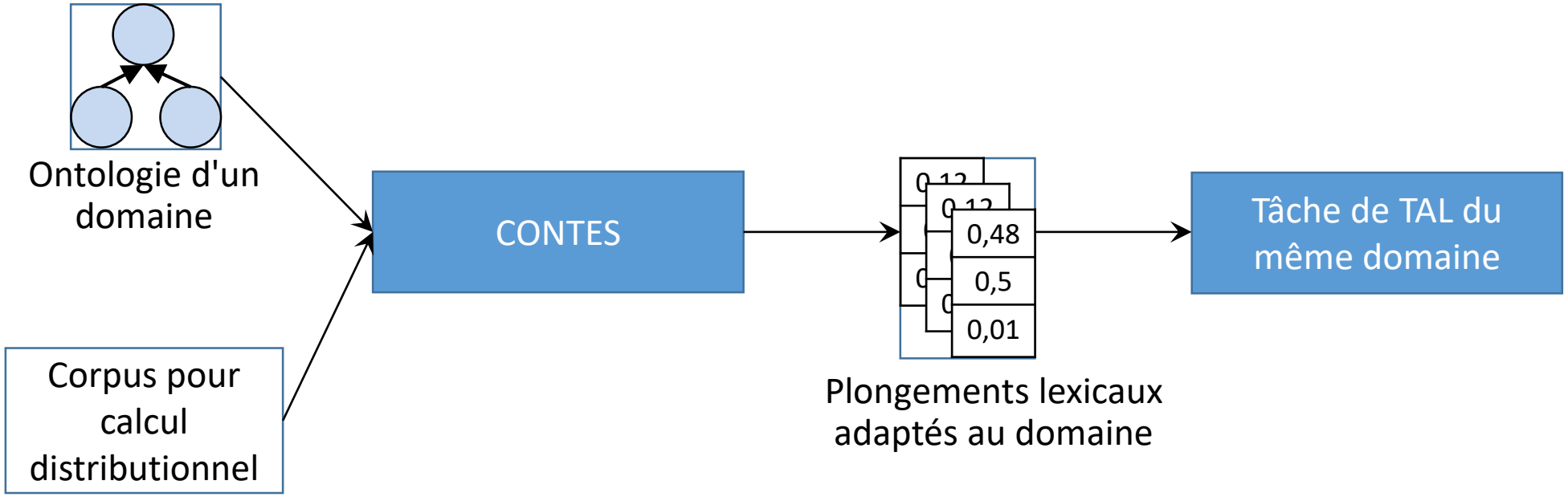
- Intégration dans la suite logicielle libre AlvisNLP/ML

  https://bibliome.github.io/alvisnlp/

(Falentin et al., 2017, Florilege: a database gathering microbial phenotypes of food interest)

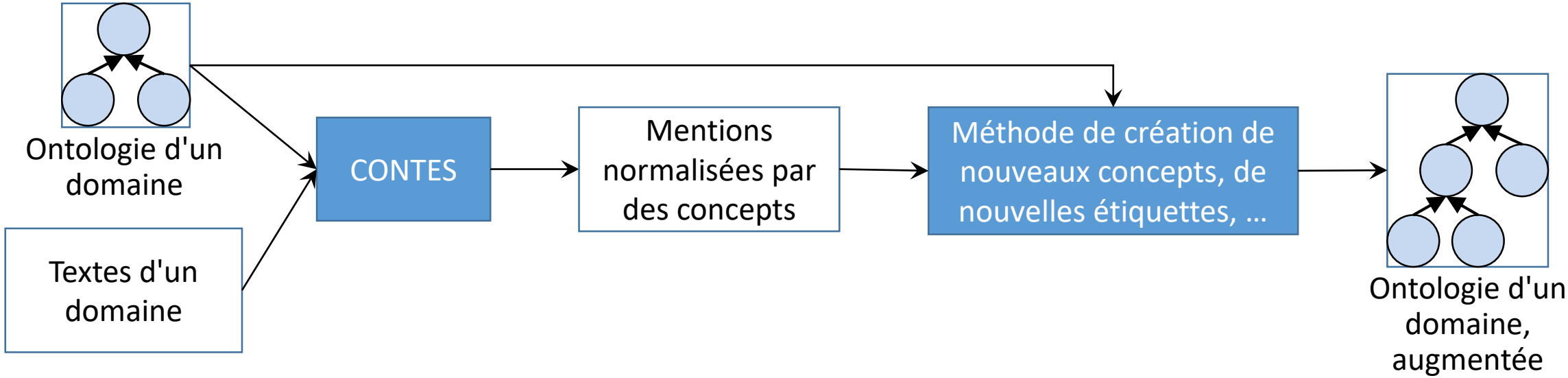
PERSPECTIVES

- Évaluer la pertinence des plongements lexicaux produits par la méthode pour la résolution d'autres problèmes du même domaine (reconnaissance d'entité, extraction de relation, etc.)



PERSPECTIVES

- Évaluer la pertinence des plongements lexicaux produits par la méthode pour la résolution d'autres problèmes du même domaine (reconnaissance d'entité, extraction de relation, etc.)
- Compléter automatique une ontologie à partir de texte :



Merci pour votre attention

Approche fondée sur des représentations vectorielles

Idée 1 : si ses contextes d'apparition sont connus, on peut comprendre le sens d'un mot.

Exemple : "*Un chien ronge un _____.*"

(Harris, 1954)
(Firth, 1957)

Approche fondée sur des représentations vectorielles

Idée 1 : si ses contextes d'apparition sont connus, on peut comprendre le sens d'un mot.

Exemple : "*Un chien ronge un _____.*"

Idée 2 : si deux mots partagent les mêmes contextes, ils auront un sens proche.

Exemple : "*Un chien ronge un bâton.*" $\Longrightarrow$ "*bâton"* et *"os"* désignent des entités qui peuvent être rongées par un chien, donc similaires.

(Harris, 1954)
(Firth, 1957)

Approche fondée sur des représentations vectorielles

Méthodes par sacs-de-mots distributionnels

"Un chien ronge un bâton."

"Un chien ronge un os."

Vocabulaire (sans mots-outils) :
{"chien", "ronge", "bâton", "os"}

	"chien"	*"ronge"*	*"bâton"*	*"os"*
"chien"	0	2	1	1
"ronge"	2	0	1	1
"bâton"	1	1	0	0
"os"	1	1	0	0

Approche fondée sur des représentations vectorielles

Méthodes par sacs-de-mots distributionnels

"Un chien ronge un bâton."

"Un chien ronge un os."

Vocabulaire (sans mots-outils) :
{"chien", "ronge", "bâton", "os"}

	"chien"	*"ronge"*	*"bâton"*	*"os"*
"chien"	0	2	1	1
"ronge"	2	0	1	1
"bâton"	1	1	0	0
"os"	1	1	0	0

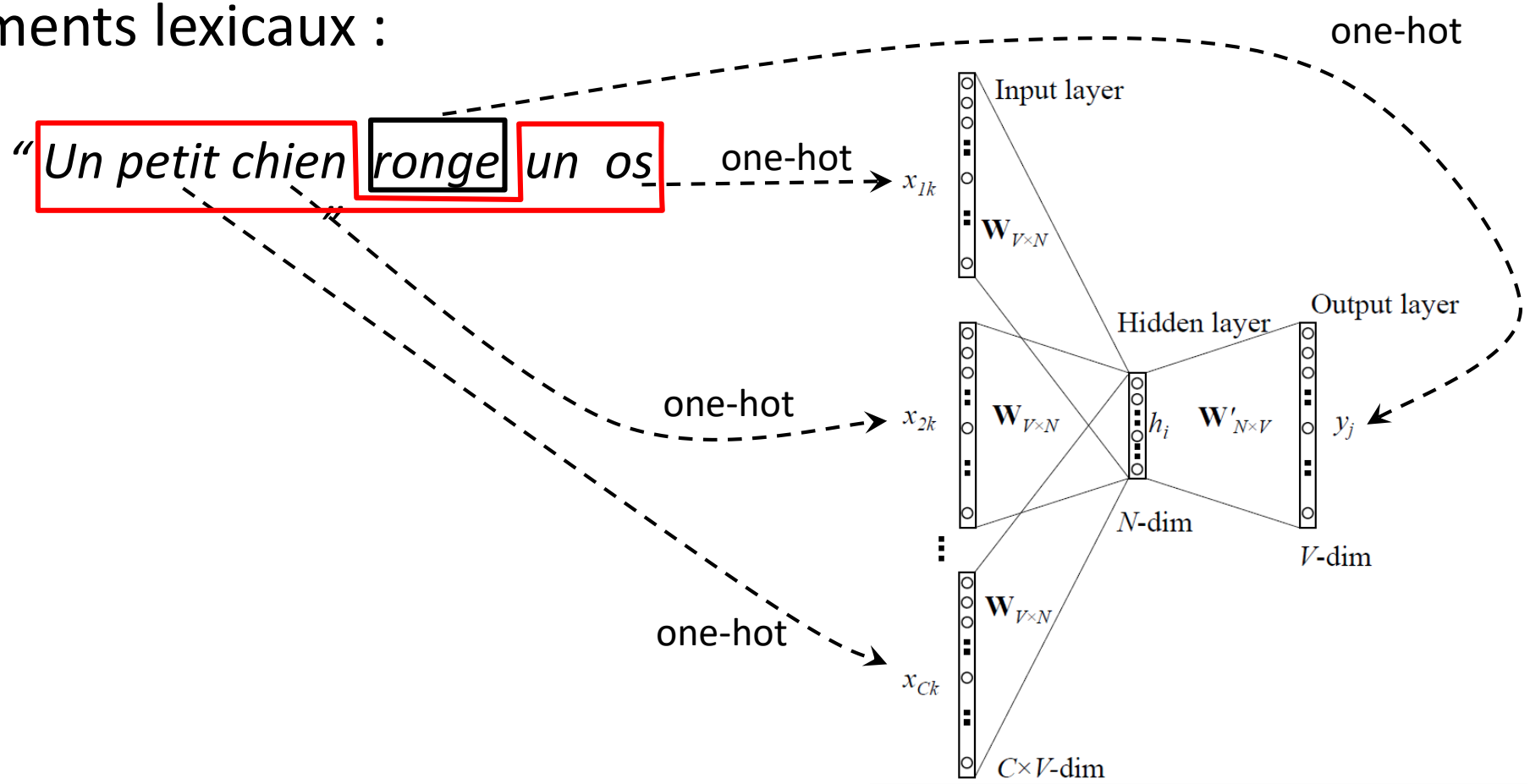
Distance euclidienne :

	"chien"	*"ronge"*	*"bâton"*	*"os"*
"chien"	0	2,83	2,45	2,45
"ronge"	2,83	0	2,45	2,45
"bâton"	2,45	2,45	0	**0**
"os"	2,45	2,45	**0**	0

(Hinton et al., 1986)
(Pollack, 1990)
(Deerwester et al., 1990)
(Elman, 1991)

Approche fondée sur des représentations vectorielles

Plongements lexicaux :



Word2Vec (CBOW) (Mikolov et al. 2013)

Approche fondée sur des représentations vectorielles

Représentation des expressions multi-mots principalement par composition :

Si : expression = (mot_1, mot_2, … , mot_n)
Alors : $\overrightarrow{expression} = f(\overrightarrow{mot_1}, \overrightarrow{mot_2}, \dots, \overrightarrow{mot_n})$

Majoritairement :

$$\overrightarrow{expression} = \frac{\sum_{k=1}^{n} \overrightarrow{mot_k}}{n}$$

(Mitchell and Lapata, 2010)

Approche fondée sur des représentations vectorielles

Limite :

Compromis entre taille du corpus d'entraînement et spécialisation au domaine
+ Connaissances du corpus ≠ connaissances de l'ontologie (experts)
⟹ Espaces sémantiques inadaptés à une tâche spécifique de normalisation.