Approche « ontologico-vectorielle » pour l'extraction de termes d'intérêts dans la littérature scientifique

Arnaud Ferré
MaIAGE, INRA, Université Paris-Saclay
LIMSI, CNRS, Université Paris-Saclay

Limsi

cnrs
dépasser les frontières

université
PARIS-SACLAY

INRA
SCIENCE & IMPACT

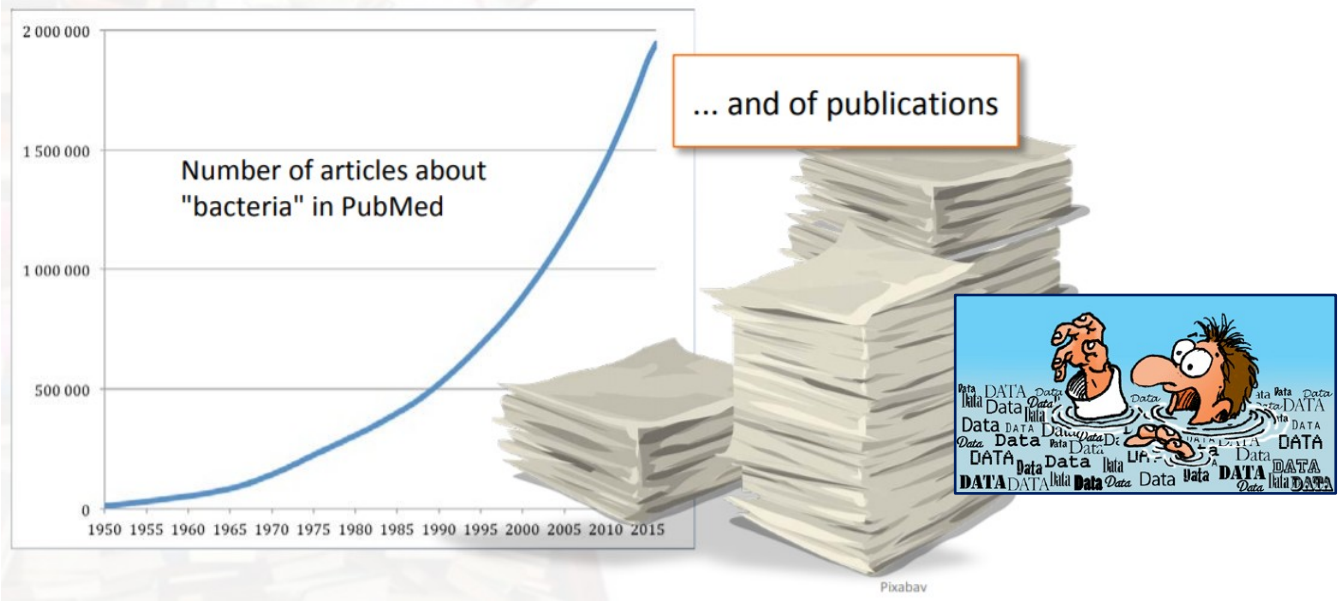
MaIAGE
Mathématiques et Informatique Appliquées du Génome à l'Environnement

INTRODUCTION

Problématique générale :

Beaucoup de connaissances sont sous forme non-structurée dans des textes (publications scientifiques, champs de texte dans une BDD, sites web, etc.)

Exemple :



INTRODUCTION

Définition :
L'**extraction d'information** vise à analyser du texte (*i.e.* **non-structuré**) pour interpréter et construire une représentation **formelle** d'une partie de l'information contenu dans celui-ci.

Les représentations formelles générées permettent de faire le lien avec d'autres sources de données formalisées.

INTRODUCTION

Reconnaissance d'entités nommées

Normalisation

Extraction de relations

[Bacteria] [Bacteria]

M. agassizii and M. Testudineum are present in

Georgia populations of gopher tortoises.

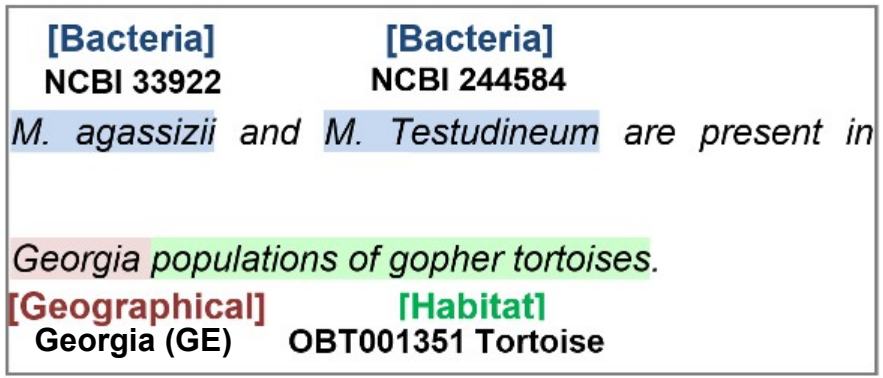
[Geographical] [Habitat]

INTRODUCTION

Reconnaissance d'entités nommées

Normalisation

Extraction de relations

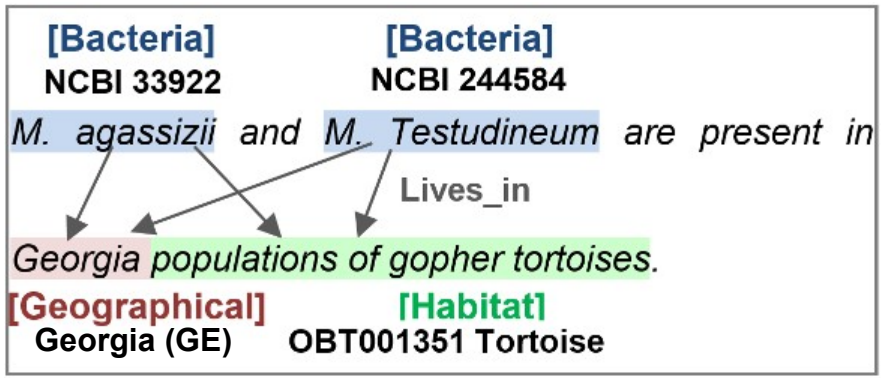


INTRODUCTION

Reconnaissance d'entités nommées

Normalisation

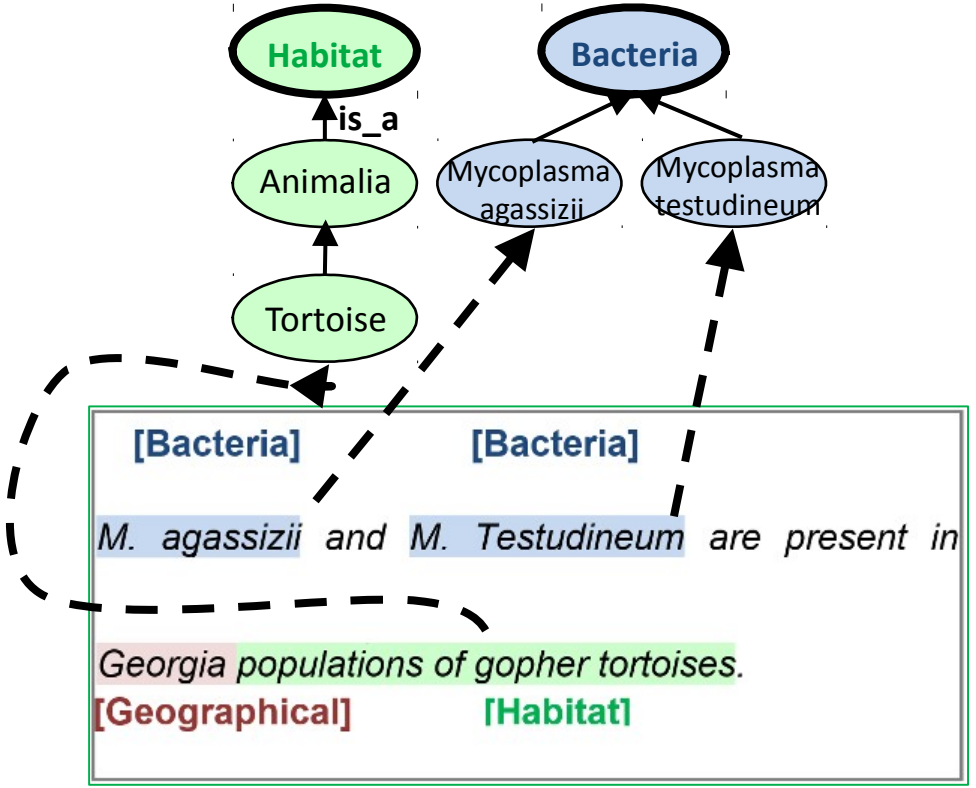
Extraction de relations



INTRODUCTION

Normalisation

Méthode :
- par des concepts d'une ontologie
- applicable à tout domaine

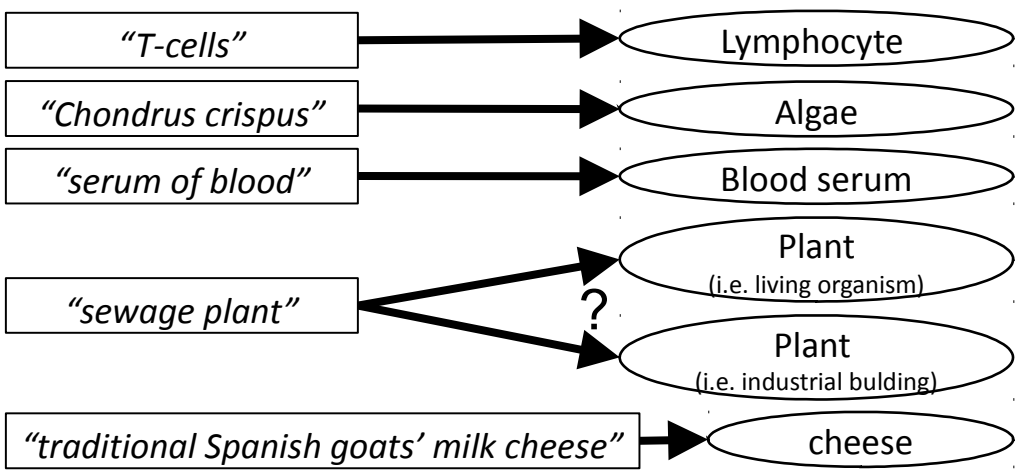


DIFFICULTÉS ET ETAT DE L'ART

- Forme et structure variables

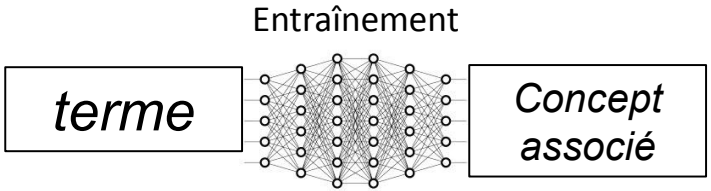
→ Approche par similarité de forme (*rule-based*, *dictionnary-based*, …) ✗

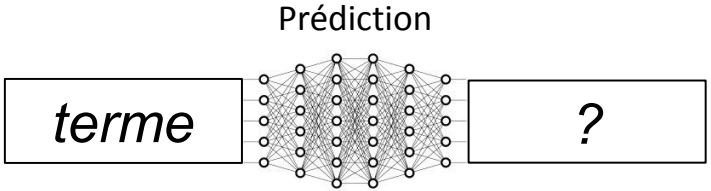
(précision+, rappel-, domaine-spécifique)



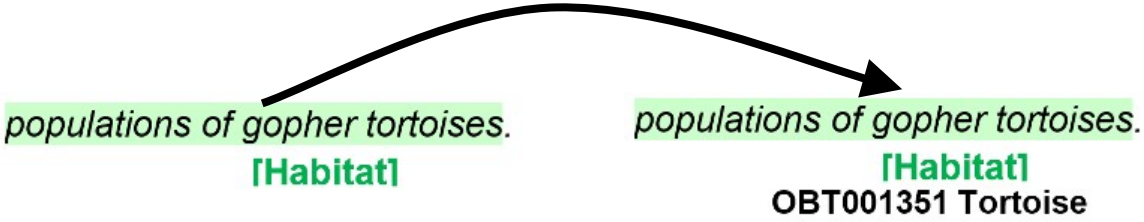
- Beaucoup de concepts-cibles, données annotées rares, couverture faible (i.e. peu de concepts utilisés), mots avec faible fréquence d'apparition dans les corpus
  - OntoBiotope : > 2000 concepts
  - UMLS : > 3 millions !

→ Approche par classification "directe" supervisée ✗





MATÉRIEL



Description :

- Challenge Bacteria Biotope de BioNLP Shared Task 2016
- Objectif : Normalisation de mentions d'habitat bactérien
- Données : Sur des titres + résumés d'articles scientifiques en biologie

3 Corpus :

- Corpus d'entraînement : termes d'habitat bactérien + concepts associés (< 1500 associations annotées manuellement)
- Corpus élargi non-annoté du domaine biomédical (PubMed)
- Corpus de test : juste les termes d'habitat bactérien

MÉTHODE : Représentation vectorielle One-Hot

Vocabulaire de n termes :

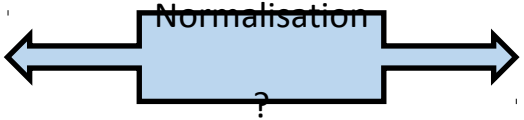
“cat”

“persian”

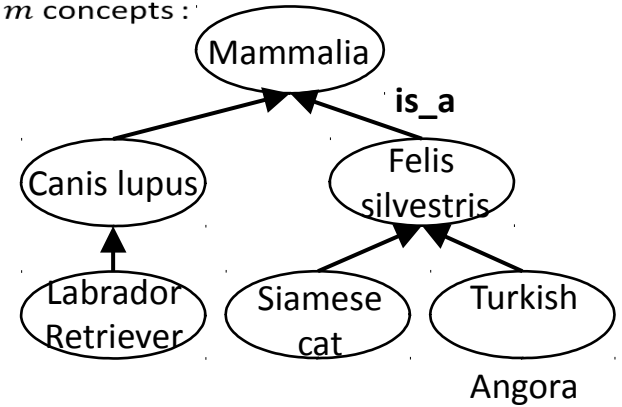
. . .

“dog”

“labrador”



Ontologie de m concepts :



MÉTHODE : Représentation vectorielle One-Hot

Vocabulaire de n termes :

"cat"

"persian"

. . .

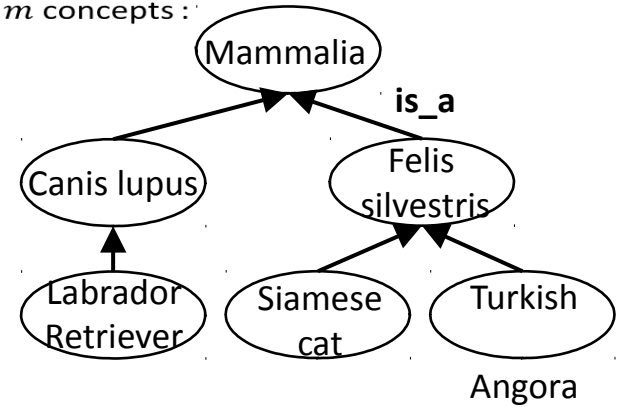
"dog"

"labrador"

Normalisation

?

Ontologie de m concepts :



One-hot pour mots :

n dimensions

1	0	0	0	
0	1	0	0	
.	.	.	.	
.	.	.	.	
.	.	.	.	
0	0	1	0	
0	0	0	1	
$\overrightarrow{cat}$	$\overrightarrow{persian}$	$\overrightarrow{dog}$	$\overrightarrow{labrador}$	

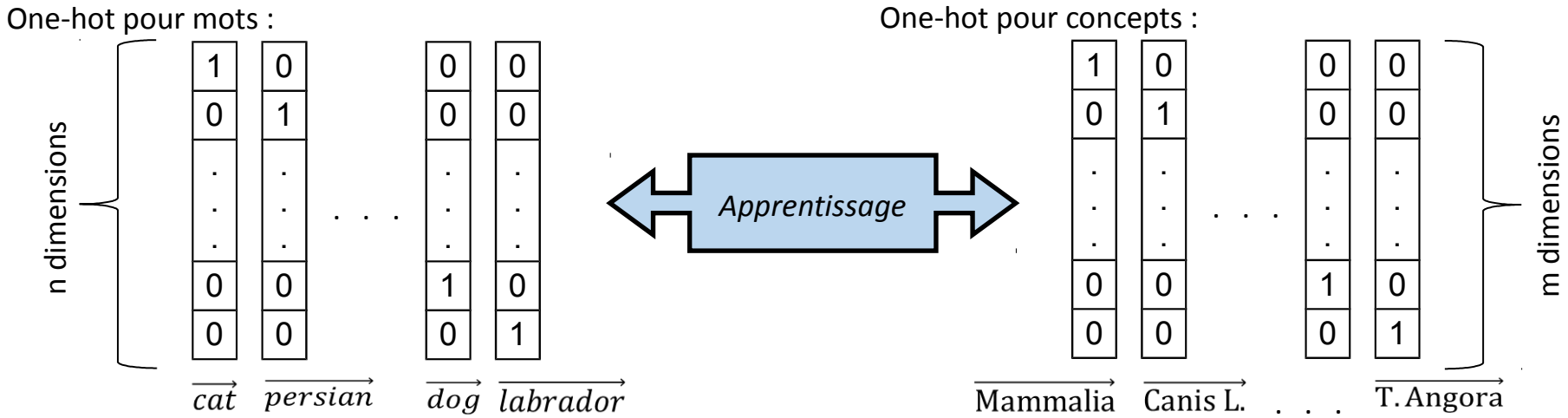
Apprentissage

One-hot pour concepts :

1	0	0	0	
0	1	0	0	
.	.	.	.	
.	.	.	.	
.	.	.	.	
0	0	1	0	
0	0	0	1	
$\overrightarrow{\text{Mammalia}}$	$\overrightarrow{\text{Canis L.}}$	. . .	$\overrightarrow{\text{T. Angora}}$	

m dimensions

MÉTHODE : Représentation vectorielle One-Hot



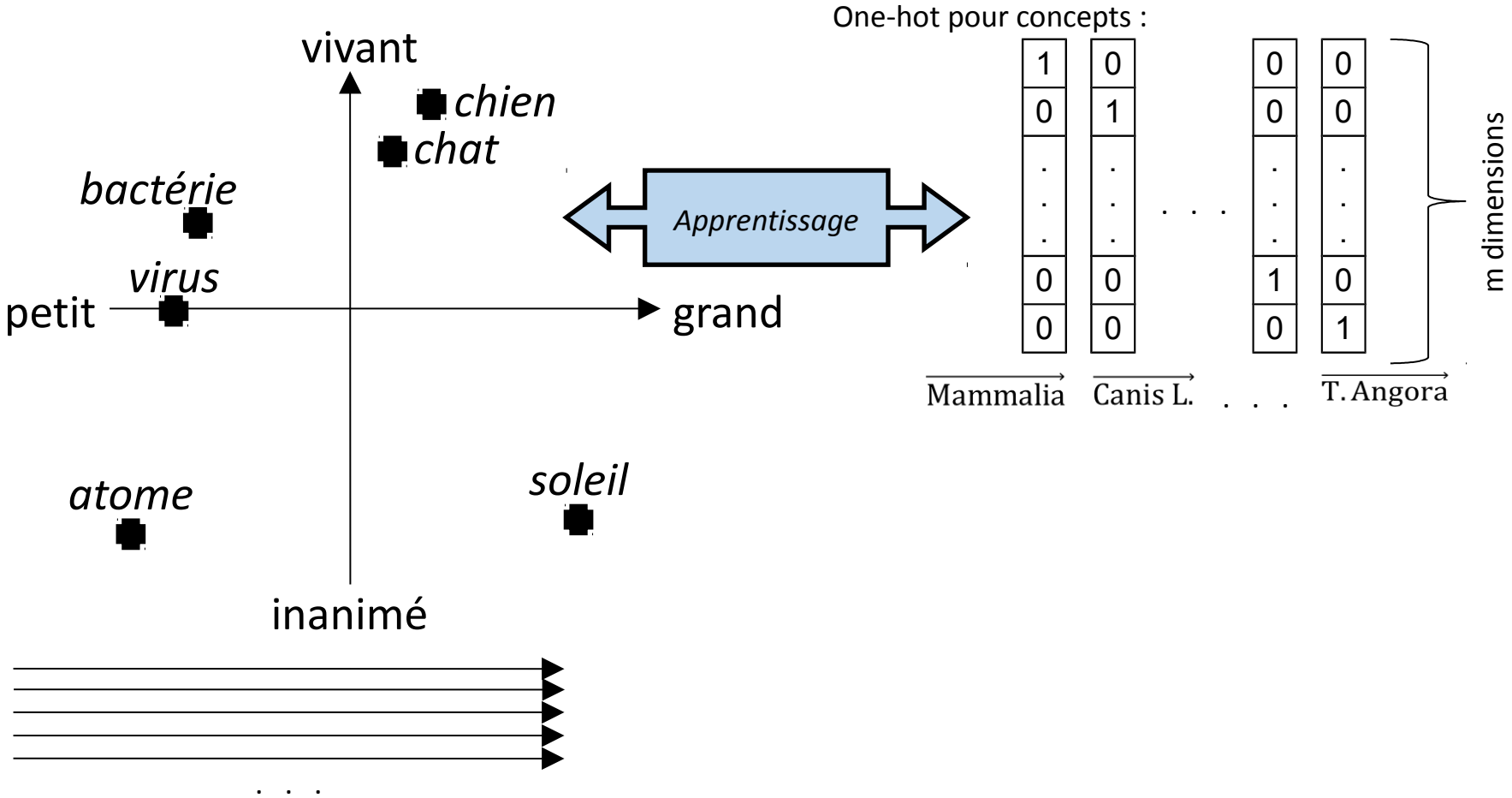
Avantage :

- Représentation vectorielle = nécessaire pour la plupart des algorithmes d'apprentissages

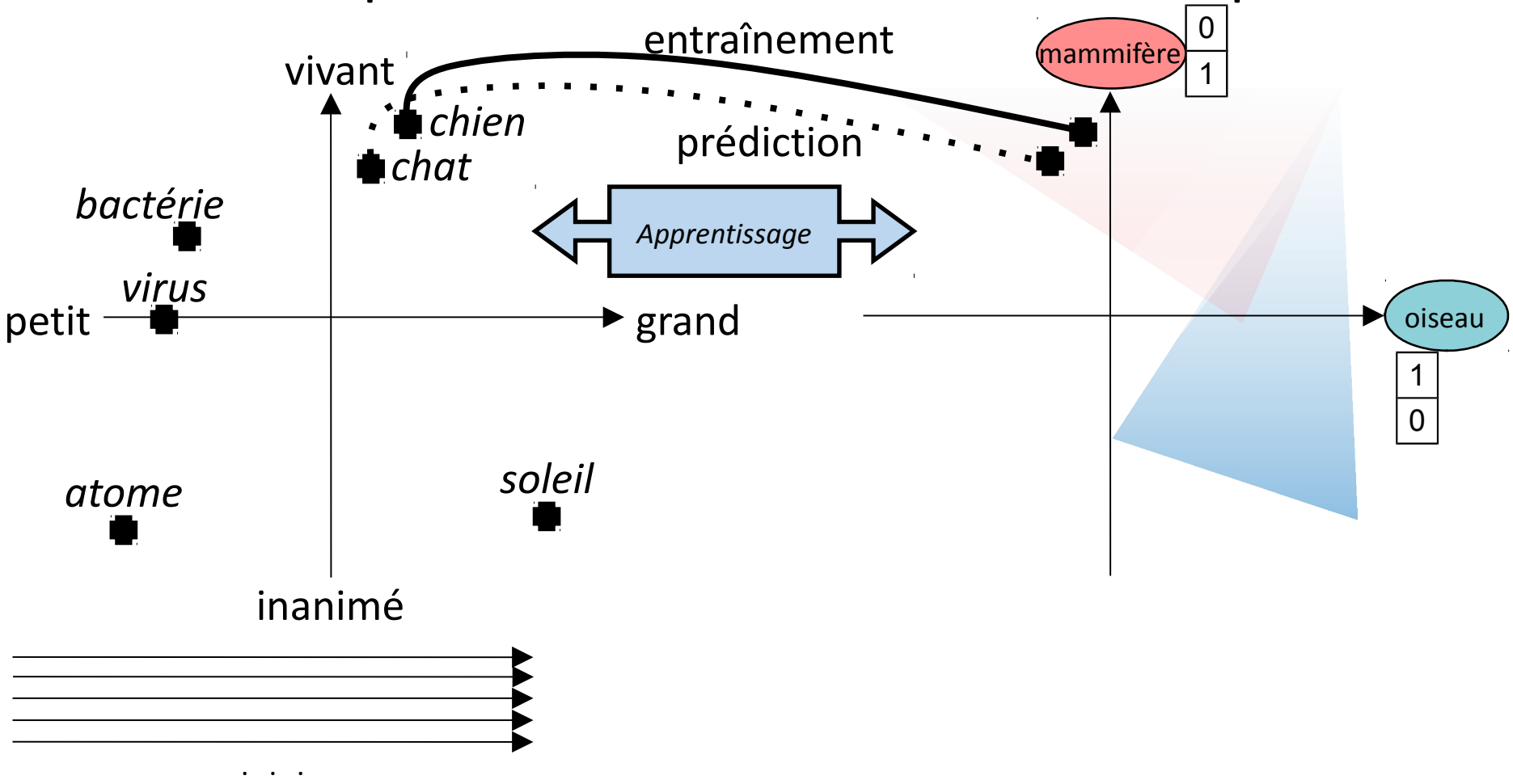
Inconvénients :

- **Très** grande dimension des vecteurs (performance…)
- Apprentissage « par-cœur »

MÉTHODE : Représentation vectorielle sémantique



MÉTHODE : Représentation vectorielle sémantique



MÉTHODE : Sémantique distributionnelle

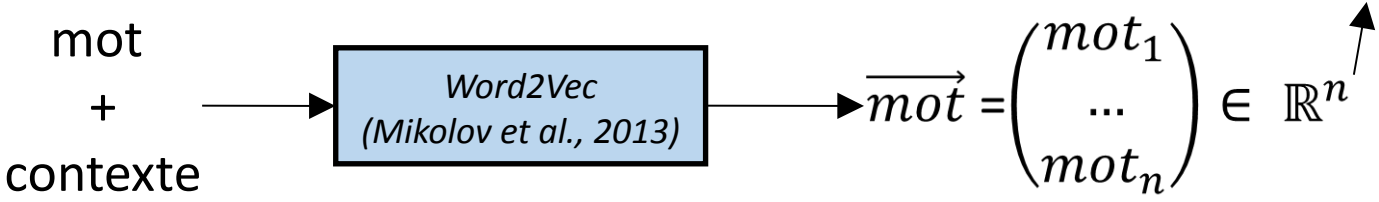
Idée 1 : si on connaît ses contextes d'apparition, on comprend le sens d'un mot

Exemple : « Un chien ronge un _____. »

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

Méthodes permettant de vectoriser le sens d'un mot :

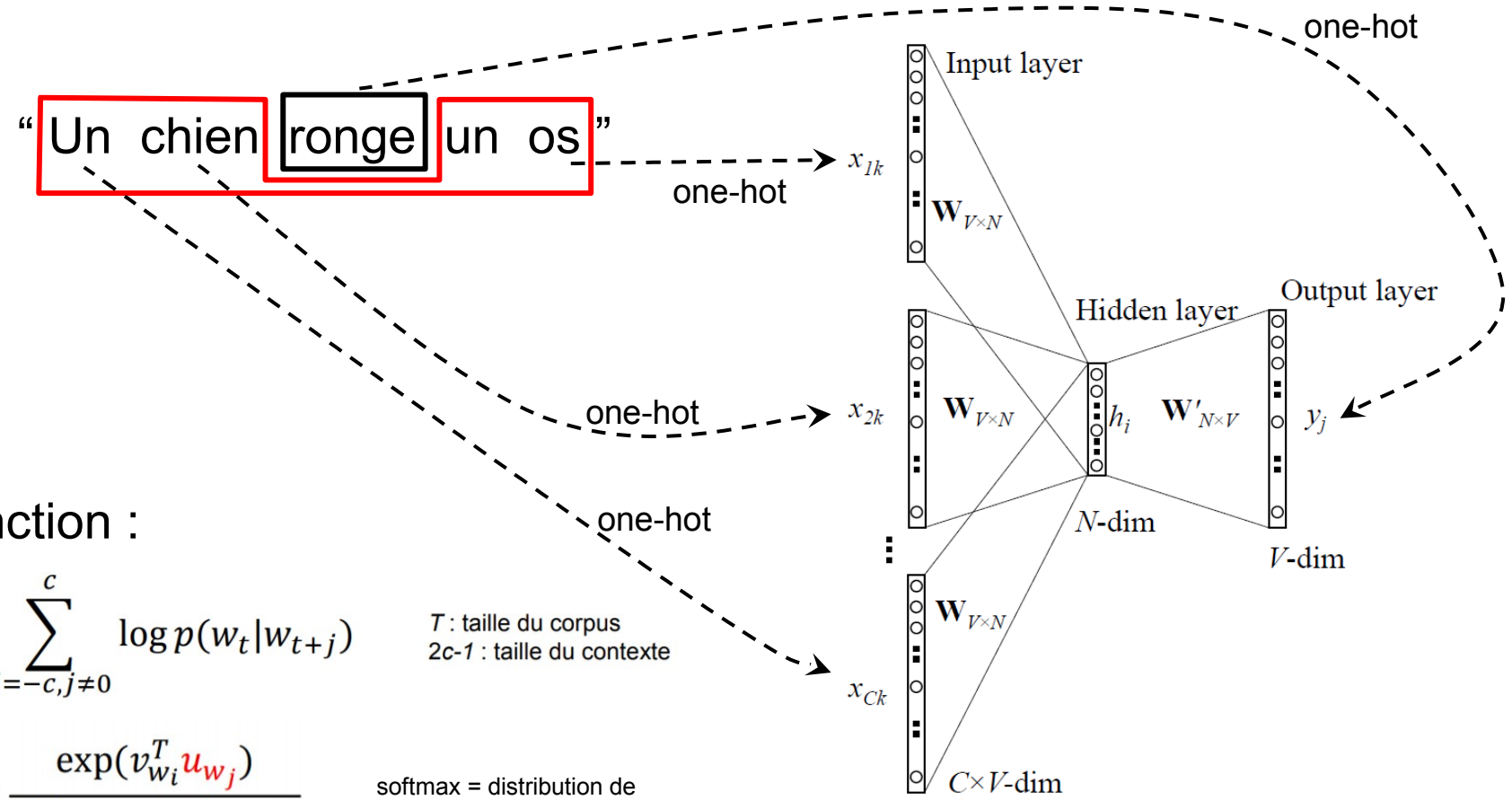
n est un paramètre à choisir (souvent performant entre 100 et 300)



- Proximité spatiale ≈ similarité sémantique
- Compositionalité :

$$\overrightarrow{roi} - \overrightarrow{homme} + \overrightarrow{femme} = \overrightarrow{reine} \dots$$

MÉTHODE : Sémantique distributionnelle



Cost function :

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=-c, j\neq 0}^{c} \log p(w_t|w_{t+j})$$

T : taille du corpus
$2c$-1 : taille du contexte

$$p(w_i|w_j) = \frac{\exp(v_{w_i}^T u_{w_j})}{\sum_{k=1}^{N}\exp(v_{w_k}^T u_{w_j})}$$

softmax = distribution de probabilité sur le vocabulaire

MÉTHODE : Sémantique distributionnelle

cell	Similarité
HCE cell	0.9999
13C-labeled cell	0.9998
parietal cell	0.9989
Schwann cell	0.9965
CD8+ T cell	0.9770
PMN cell	0.9669
macrophage cell	0.9473

seawater	Similarité
sediments	0.7696
sediment sample from a disease free fish farm	0.7499
fish farm sediments	0.7342
subterranean brine	0.7320
lagoon on the outskirts of the city of Cagliari	0.7128
petroleum reservoir	0.7095
marine environments	0.7077
marine bivalves	0.6896
sediment samples from diseased farms	0.6870
urine sediments	0.6819
petroleum	0.6576
subterranean environment	0.6497
fresh water	0.6494
fresh water supply	0.6395
Seafood	0.6390
marine	0.6366

MÉTHODE : Sémantique distributionnelle

Avantages :

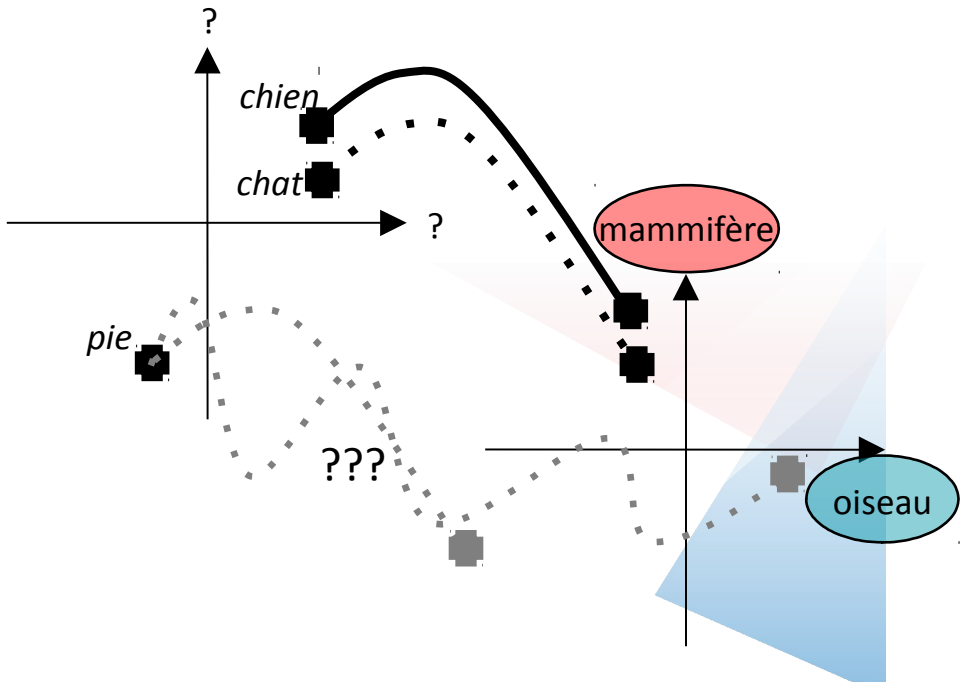
- Représentation vectorielle = nécessaire pour la plupart des algorithmes d'apprentissages
- Permet de faire des prédictions pertinentes de normalisation pour des termes non-rencontrés pendant l'entraînement
- Dimension des vecteurs plus dense (< 1000)

Inconvénient :

- Axes non-interprétables

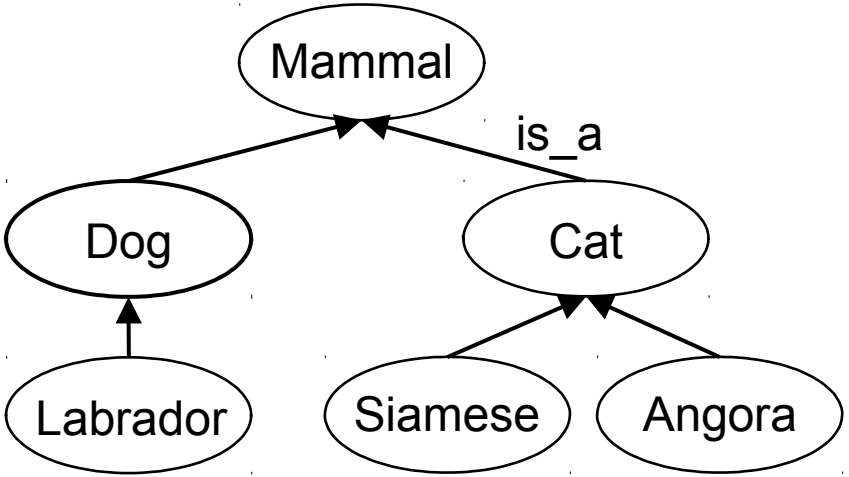
Question :

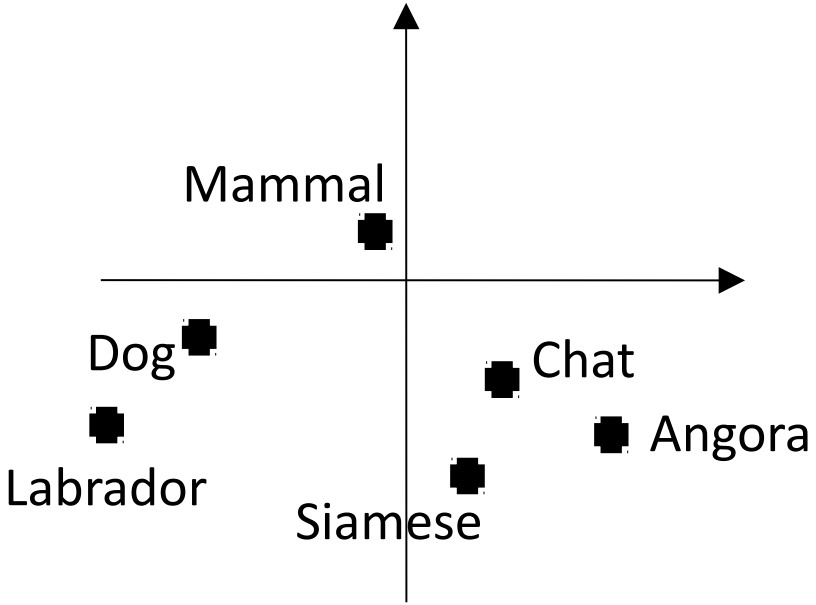
Comment obtenir des prédictions pertinentes dans le cas où les concepts à prédire n'ont pas été rencontrés lors de l'entraînement ?



MÉTHODE : Représentation sémantique de concept

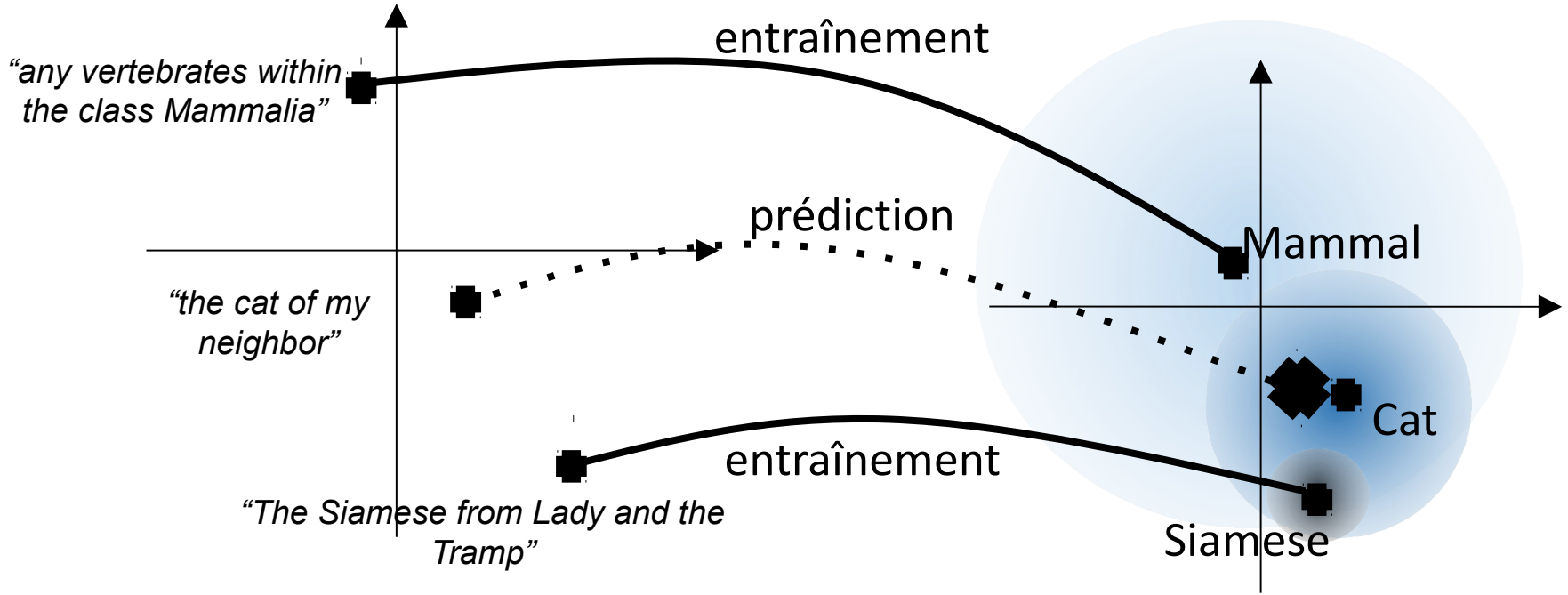
Idée : Adopter une représentation vectorielle sémantique pour les concepts également





MÉTHODE : Représentation sémantique de concept

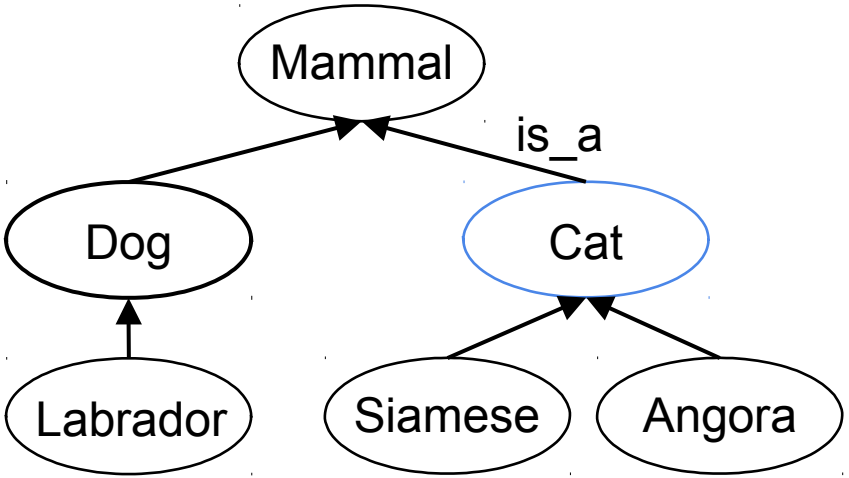
Idée : Adopter une représentation vectorielle sémantique pour les concepts également



Comment peut-on construire de tels vecteurs ?

Quelle information est la plus exploitable ?

MÉTHODE : Représentation sémantique de concept



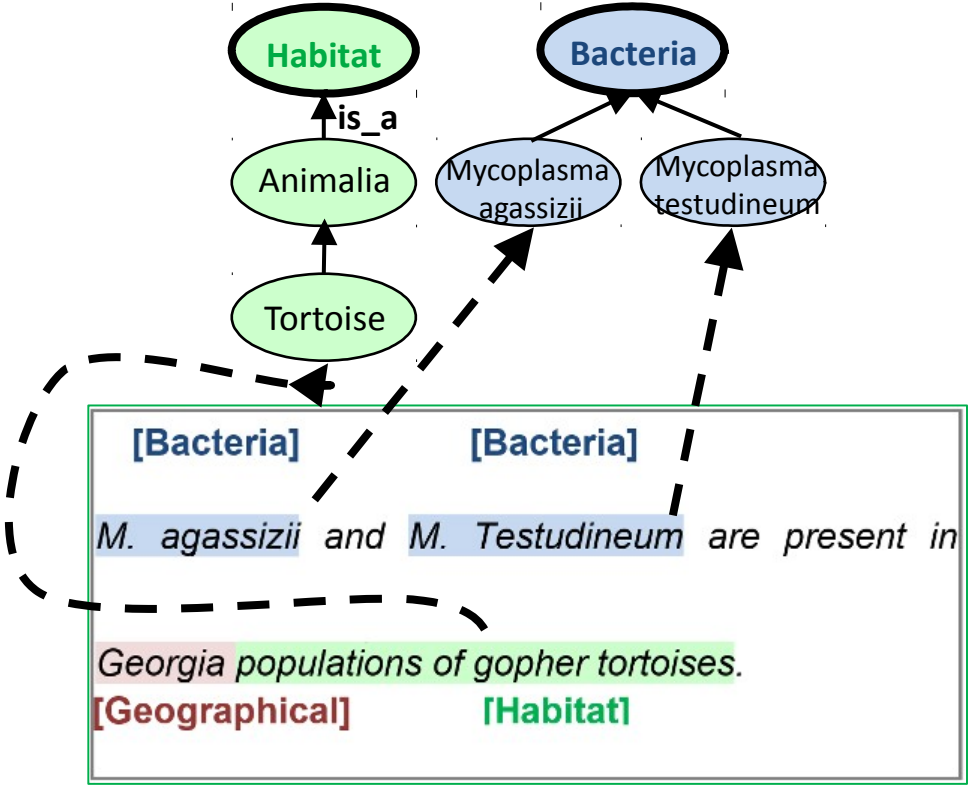
$$\forall k \in [\![1,n]\!], v_{c_k} = \left(w^0_{c_k}, \ldots, w^i_{c_k}, \ldots, w^n_{c_k} \right)$$

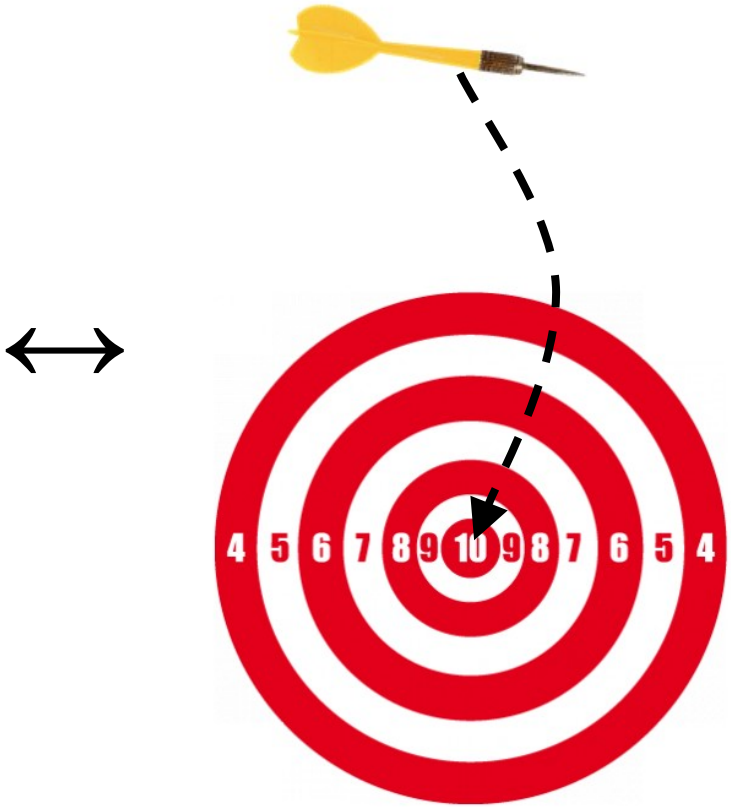
$$w^i_{c_k} = \begin{cases} 1 & si\ i = k \\ 1 & si\ c_i\ ancêtre\ de\ c_k \\ 0 & sinon \end{cases}$$

Ex : $v_{Cat} = [1,0,1,0,0,0]$

Cat	**similarité**
Cat	1,00
Siamese	0.82
Angora	0.82
Mammal	0.71
Dog	0.50
Labrador	0.41

RÉSULTATS :





Mesure de similarité de Wang et al. (2007)

RÉSULTATS :

VS

ToMap

(Golik Wiktoria, Warnier Pierre, Nédellec Claire, 2011)

Normalise plutôt très bien les termes qui possèdent une tête sémantique de forme similaire à celle d'un label (≈ 50%), mais pas du tout ceux qui n'en possèdent pas.

CONTES

(Ferré Arnaud, Zweigenbaum Pierre, Nédellec Claire, 2017)

Normalise l'ensemble des termes plutôt bien et en particulier ceux ne possédant pas de similarité de forme avec les labels, mais est moins précis pour les termes qui en possèdent.

RÉSULTATS

intégration

Equipe	Score
HONOR	**0.73**
ToMap (avec règles)	**0.66**
Turku (2017)	0,63
BOUN (2016)	0,62
ToMap (sans règles)	**0.61**
CONTES (2017)	**0,61**
LIMSI (2016)	0,44
Baseline	0,32

Combining rule-based and embedding-based approaches to normalize textual entities with an ontology
A Ferré, L Deléger, P Zweigenbaum, C Nédellec - LREC 2018

Representation of complex terms in a vector space structured by an ontology for a normalization task
A Ferré, P Zweigenbaum, C Nédellec - BioNLP 2017

CONCLUSION

Normalisation

Méthode :
- par des concepts d'une ontologie
- applicable à tout domaine

Méthode encourageante pour normaliser avec :

- Peu de données d'entraînement
- Un potentiel à répondre au problème de variabilité de forme
- Seulement 2 types d'informations utilisées (sémantique distributionnelle + information hiérarchique)
- Possibilité de rendre plus interprétable les espaces générés

-> Méthode pouvant servir à peupler automatiquement une ontologie à partir de textes

PERSPECTIVES

- Densifier les vecteurs de concepts (UMLS > 3 millions de concepts)
- Intégrer plus d'information linguistique (notamment syntaxique)
- Quid d'une fonction d'apprentissage non-linéaire ?
- Évoluer vers une méthode non-supervisée qui n'a pas besoin de reconnaissance d'entités nommées

-> S'appuyer sur cette approche pour enrichir (instances mais aussi concepts et relations) automatiquement une ontologie à partir de texte

- Merci de votre attention.
- Des questions?