



**HAL**  
open science

# Représentations vectorielles et apprentissage automatique pour l'alignement d'entités textuelles et de concepts d'ontologie: application à la biologie

Arnaud Ferré

## ► To cite this version:

Arnaud Ferré. Représentations vectorielles et apprentissage automatique pour l'alignement d'entités textuelles et de concepts d'ontologie: application à la biologie. Intelligence artificielle [cs.AI]. Université Paris Saclay (COMUE), 2019. Français. NNT: 2019SACLS117. tel-02166253

**HAL Id: tel-02166253**

**<https://theses.hal.science/tel-02166253>**

Submitted on 26 Jun 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Représentations vectorielles et apprentissage automatique pour l'alignement d'entités textuelles et de concepts d'ontologie : application à la biologie

Thèse de doctorat de l'Université Paris-Saclay  
préparée à l'Université Paris-Sud

École doctorale n°580 Sciences et technologies de l'information et de la  
communication (STIC)  
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Orsay, le 24 mai 2019, par

**Arnaud Ferré**

## Composition du jury :

Alexandre Allauzen Professeur des Universités, Université Paris-Sud (LIMSI)	Président
Nathalie Aussenac Directrice de Recherche, CNRS (IRIT)	Rapporteuse
Emmanuel Morin Professeur des Universités, Université de Nantes (LS2N)	Rapporteur
Vincent Claveau Chargé de Recherche, CNRS (IRISA)	Examineur
Claire Nédellec Directrice de Recherche, INRA (MaIAGE)	Directrice de thèse
Pierre Zweigenbaum Directeur de Recherche, CNRS (LIMSI)	Co-Directeur de thèse



*« Les gens qui sont assez fous pour penser qu'ils peuvent changer le monde sont ceux qui le font. »*  
*Steve Job*

*« La connaissance des mots conduit à la connaissance des choses. »*  
*Platon*

*À mes parents.*



# Remerciements

Je voudrais remercier toutes les personnes qui ont contribué à l'aboutissement de ces trois années de travaux de thèse. Je vais essayer de faire une longue liste en n'oubliant personne, mais si vous m'avez soutenu ces dernières années d'une quelconque façon, comme partager avec moi une discussion, un problème, un conseil, un bon repas, une astuce ou histoire gastronomique, un thé vert japonais, une bière (ou plus), un whisky, un saké, une soirée « jeux de société », un bon film ou un mauvais, alors ces remerciements vous sont destinés.

Je remercie en premier mes directeurs de thèse, Claire Nédellec et Pierre Zweigenbaum, pour la confiance qu'ils m'ont accordée et pour m'avoir laissé libre d'explorer mes propres intuitions. Je pense que je n'aurais pas tant apprécié ces dernières années s'il en avait été autrement.

Je remercie Nathalie Aussenac et Emmanuel Morin d'avoir accepté d'être rapporteurs de mon travail de thèse, ainsi que Vincent Claveau et Alexandre Allauzen pour avoir accepté de l'examiner. La richesse de leurs remarques m'encourage à continuer mes travaux.

Je remercie l'Université Paris-Saclay qui a accepté de me financer sur ce sujet de thèse via leur Initiative Doctorale Interdisciplinaire. Merci tout particulièrement à tous les membres du Collège Doctoral, de l'école doctorale, de la DIRE, de La Diagonale, de la Vie Campus et de la SATT qui m'ont permis de participer à la construction et l'animation de notre Université. J'en connais à présent beaucoup sur l'organisation de la Recherche et le fonctionnement d'une Université.

Je remercie tous mes collègues de l'unité MaIAGE pour leur entrain, leur dynamisme et toutes les discussions scientifiques et culturelles du coin café telles que celles commençant par « est-ce que lécher la barre du métro est une bonne chose pour renforcer son système immunitaire ? ». Un merci tout particulier à ceux qui ont rythmé et fait vivre nos si nombreuses soirées jeux de société (Louise, Mahendra, Robert, Marc, Ba, Sandra, Cyprien, Estelle, Sam, Cedric, Julie, David, Charles, Valentin et Stephan).

Je remercie tous mes collègues du LIMSI, et en particulier ceux du groupe ILES, pour m'avoir tant appris et ainsi permis de prendre le train du TAL en marche. Un merci tout particulier à ceux qui m'ont accompagné durant les pauses café et autres afterworks (Sven, Julien, Rachel, Yuming, Zheng, Sanjay, Christopher, Vincent, Benjamin, Antoine, Marine et Arthur) et qui, par nos échanges, ont pu me remonter le moral à plusieurs occasions.

Je remercie les coordinatrices et managers du festival Pint of Science Ile-de-France, les membres de la Mission Arts, Culture, Sciences et Société, l'équipe derrière la Diagonale Paris-Saclay, les organisateurs de MT180 et tous les acteurs du projet Ma Thèse en BD. Grâce à vous, j'ai pu me former une solide expérience en vulgarisation scientifique, sortir le nez du guidon et prendre le recul nécessaire sur mes travaux.

Merci à Louise et à Gabriel pour m'avoir montré que l'on pouvait être chercheur permanent tout en conservant son humanité ! Merci aussi pour le soutien moral et pour toutes ces astuces de chercheur.

Merci à Robert pour m'avoir permis d'atteindre mes objectifs et d'être à la hauteur de mes ambitions. Sans lui, ce manuscrit n'aurait sans doute pas autant de résultats à présenter !

Merci à Swen pour toutes nos discussions, sérieuses ou non (parfois politiquement incorrectes !), et surtout, pour ne pas m'avoir laissé seul souffrir du syndrome de l'imposteur !

Merci à tous les « doctorants-boxeurs &co » du club de boxe française de Polytechnique, pour m'avoir gardé en forme sans trop me déformer ! Qu'aurait été l'esprit sans le corps ?!

Merci à tous mes amis d'enfance, de prépa, de l'ENSTA, de Master et d'ailleurs, dans la région comme au loin, pour votre soutien. C'est promis, j'arrête les études cette fois ! J'espère avoir plus de temps et l'esprit plus libre pour vous voir plus régulièrement à présent.

Merci à Renato pour le soutien et le suivi philosophique, psychologique et gastronomique ! S'il n'avait pas été là, j'aurais peut-être fini par me créer un ami imaginaire. Une page va se tourner, mais « *I'll be back* », comme dirait Arnold Schwarzenegger.

Merci à Clarisse d'avoir été, et de continuer à être, mon garde-fou. Sans elle, j'aurais sans doute perdu pieds et serais devenu un robot aujourd'hui, et je n'aurais donc pas pu arriver là où j'en suis.

Enfin, un grand merci à ma famille et surtout à mes parents pour l'éducation, la confiance, les encouragements, le soutien, et pour tout le reste, qu'ils ont pu me donner, de ma naissance à aujourd'hui. Les langues naturelles ne possèdent aucun mot assez fort pour vous exprimer ma gratitude.

# Table des matières

REMERCIEMENTS.....	5
TABLE DES MATIÈRES .....	7
<b>CHAPITRE 1 - INTRODUCTION .....</b>	<b>11</b>
1.1. CADRE DE LA THÈSE .....	13
1.2. MOTIVATIONS.....	14
1.3. PROBLÉMATIQUE .....	14
1.4. CONTRIBUTIONS.....	19
1.5. PLAN.....	20
<b>CHAPITRE 2 - CONTEXTE.....</b>	<b>23</b>
2.1. INTRODUCTION.....	25
2.2. REPRÉSENTATIONS SÉMANTIQUES POUR L'EXTRACTION D'INFORMATION .....	25
2.2.1. <i>Introduction.....</i>	<i>25</i>
2.2.2. <i>De la donnée, à l'information, à la connaissance .....</i>	<i>26</i>
2.2.3. <i>Ontologie : un modèle formel et structuré de connaissances.....</i>	<i>27</i>
2.2.4. <i>Espaces sémantiques .....</i>	<i>29</i>
2.3. L'EXTRACTION D'INFORMATION .....	32
2.3.1. <i>Introduction.....</i>	<i>32</i>
2.3.2. <i>La préparation de corpus .....</i>	<i>32</i>
2.3.3. <i>La reconnaissance d'entités.....</i>	<i>33</i>
2.3.4. <i>L'extraction de relations .....</i>	<i>35</i>
2.3.5. <i>La normalisation d'entités nommées.....</i>	<i>36</i>
2.4. DIFFICULTÉS DE LA NORMALISATION D'ENTITÉS .....	37
2.4.1. <i>Introduction.....</i>	<i>37</i>
2.4.2. <i>Variabilité des formes .....</i>	<i>38</i>
2.4.3. <i>Normalisation multiple .....</i>	<i>39</i>
2.4.4. <i>Expression homonymique.....</i>	<i>40</i>
2.4.5. <i>Insuffisance de données annotées.....</i>	<i>40</i>
2.5. TÂCHE DE NORMALISATION BACTERIA BIOTOPE.....	41
2.5.1. <i>Introduction.....</i>	<i>41</i>
2.5.2. <i>Présentation de Bacteria Biotope 2016.....</i>	<i>42</i>
2.5.3. <i>Intérêt scientifique de la tâche.....</i>	<i>43</i>
2.5.4. <i>Tâche de normalisation d'habitat bactérien.....</i>	<i>43</i>
2.6. BILAN .....	45
<b>CHAPITRE 3 - NORMALISATION D'ENTITÉS FONDÉE SUR UNE ONTOLOGIE : ÉTAT DE L'ART.....</b>	<b>47</b>
3.1. INTRODUCTION.....	49
3.2. NORMALISATION FONDÉE SUR LA SIMILARITÉ DE FORME ENTRE EXPRESSIONS D'UN DICTIONNAIRE ET EXPRESSIONS TEXTUELLES .....	50
3.2.1. <i>Introduction.....</i>	<i>50</i>
3.2.2. <i>Limitations des appariements exacts.....</i>	<i>50</i>
3.2.3. <i>Présentation générale.....</i>	<i>51</i>
3.2.4. <i>État de l'art.....</i>	<i>53</i>
3.2.5. <i>Limitations.....</i>	<i>56</i>
3.2.6. <i>Évaluation de méthodes.....</i>	<i>57</i>
3.3. CONSTRUCTION D'ESPACES SÉMANTIQUES.....	57
3.3.1. <i>Introduction.....</i>	<i>57</i>

3.3.2.	<i>Représentations creuses et non-sémantiques : les représentations 1-parmi-N.....</i>	58
3.3.3.	<i>Des représentations creuses et sémantiques : les sacs-de-mots TF-IDF.....</i>	58
3.3.4.	<i>Des représentations creuses et non-fondées sur la forme : les sacs-de-mots distributionnels.....</i>	60
3.3.5.	<i>Des représentations continues et de faible dimension : les plongements lexicaux.....</i>	62
3.3.6.	<i>Compositionnalité et représentation vectorielle d'expression.....</i>	63
3.3.7.	<i>Adaptation de plongements lexicaux à des connaissances externes.....</i>	64
3.4.	NORMALISATION FONDÉE SUR DES REPRÉSENTATIONS SÉMANTIQUES .....	67
3.4.1.	<i>Introduction.....</i>	67
3.4.2.	<i>Approche par calcul de scores entre représentations sémantiques .....</i>	67
3.4.3.	<i>Approche par apprentissage supervisé.....</i>	68
3.4.4.	<i>État de l'art .....</i>	70
3.5.	BILAN .....	75
<b>CHAPITRE 4 - CONSTRUCTION D'ESPACES SÉMANTIQUES .....</b>		<b>77</b>
4.1.	INTRODUCTION.....	79
4.2.	MÉTHODE DE REPRÉSENTATION VECTORIELLE DE CONCEPTS D'UNE ONTOLOGIE HIÉRARCHIQUE : ANCESTRY.....	81
4.2.1.	<i>Introduction.....</i>	81
4.2.2.	<i>Étude du modèle .....</i>	82
4.2.3.	<i>Limites du modèle Ancestry.....</i>	85
4.2.4.	<i>Rapprochement entre représentations de concepts fils et parents : Ancestry+ .....</i>	86
4.2.5.	<i>Tentative de densification et de diminution du nombre de dimensions .....</i>	88
4.3.	CONSTRUCTION INITIALE DE L'ESPACE DISTRIBUTIONNEL DES EXPRESSIONS.....	90
4.3.1.	<i>Introduction.....</i>	90
4.3.2.	<i>Choix de méthode et de corpus d'entraînement .....</i>	91
4.3.3.	<i>Choix des hyper-paramètres de Word2Vec et de prétraitements.....</i>	92
4.3.4.	<i>Construction de plongements lexicaux pour des expressions multi-mots.....</i>	95
4.4.	BILAN .....	96
<b>CHAPITRE 5 - MÉTHODES DE NORMALISATION D'ENTITÉS.....</b>		<b>97</b>
5.1.	INTRODUCTION.....	99
5.2.	APPRENTISSAGE D'UNE PROJECTION CONSERVANT LA STRUCTURE DE L'ESPACE SOURCE ET NORMALISATION.....	101
5.2.1.	<i>Introduction.....</i>	101
5.2.2.	<i>Détermination d'une projection linéaire entre espaces sémantiques.....</i>	101
5.2.3.	<i>Effets de la régression linéaire sur la projection .....</i>	103
5.2.4.	<i>Prédiction de normalisation par calcul de distance.....</i>	105
5.3.	EXPÉRIMENTATIONS AVEC DIFFÉRENTES SOURCES D'EXEMPLES .....	106
5.3.1.	<i>Introduction.....</i>	106
5.3.2.	<i>Paramètres expérimentaux.....</i>	108
5.3.3.	<i>Approche mixte par intégration d'une méthode fondée sur l'étude de similarité de forme : HONOR.....</i>	108
5.3.4.	<i>Approche par apprentissage supervisé faible .....</i>	109
5.3.5.	<i>Approche mixte et par apprentissage faiblement supervisé : WSEP-HONOR.....</i>	113
5.4.	DISCUSSION ET EXPLORATIONS .....	114
5.5.	DIFFUSION DES MÉTHODES .....	116
5.6.	BILAN .....	118
<b>CHAPITRE 6 - CONCLUSION ET PERSPECTIVES .....</b>		<b>121</b>
6.1.	ADAPTABILITÉ À D'AUTRES JEUX DE DONNÉES .....	124
6.2.	ADAPTATION DE L'ESPACE DISTRIBUTIONNEL ET AMÉLIORATION DE LA PROJECTION .....	124

6.3. APPLICATION À D'AUTRES TÂCHES EN EXTRACTION D'INFORMATION .....	125
<b>PUBLICATIONS DURANT LA THÈSE .....</b>	<b>127</b>
PUBLICATIONS LIÉES À LA THÈSE .....	127
PUBLICATIONS NON DIRECTEMENT LIÉES À LA THÈSE .....	128
<i>Représentation de connaissances en sciences du vivant</i> .....	128
<i>Bioinformatique et médiation scientifique</i> .....	128
<b>BIBLIOGRAPHIE .....</b>	<b>129</b>



**Chapitre 1**  
-  
**Introduction**



## 1.1. Cadre de la thèse

Cette thèse s'inscrit dans le domaine de l'intelligence artificielle (IA) et plus spécifiquement dans celui du Traitement Automatique des Langues (TAL). L'IA peut être définie comme l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence. Or, comprendre une langue naturelle, au sens de langue humaine, est en particulier une forme d'intelligence, que l'on nomme l'intelligence linguistique dans la théorie de psychologie des intelligences multiples (Gardner, 2011). Le TAL est le domaine de l'IA qui vise à simuler cette intelligence chez des machines. Ce domaine d'étude traite pour cela divers aspects de la langue et du discours, à tous les niveaux (lexical, morphologique, syntaxique, sémantique, ...).

Parmi les applications du TAL, le grand public connaît essentiellement la correction orthographique, la reconnaissance de la parole ou encore les moteurs de recherche documentaire. En anglais, le terme le plus répandu est NLP (*Natural Language Processing*). Même si la liste suivante n'est pas exhaustive, voici certains des domaines d'étude habituels en TAL :

- L'extraction d'information (EI) : L'EI vise à extraire et à structurer automatiquement un ensemble d'informations précises apparaissant dans des textes écrits en langue naturelle. L'objectif principal est alors de produire ou de compléter des bases de données ou des bases de connaissances (Cowie and Wilks, 2000).
- La recherche d'information (recherche documentaire) : l'objectif de la recherche d'information est de permettre de trouver un ensemble de documents pertinents, pour une question ou un problème donné, dans une large collection (Singhal and others, 2001). La recherche d'information peut également s'appuyer sur une étape d'EI en amont, ou inversement.
- La génération automatique de texte : domaine visant à faire générer automatiquement du texte à une machine à partir d'informations structurées ou d'un texte incomplet (Shannon, 1948; Yngve, 1961). À l'inverse de l'EI, ce domaine peut donc viser à améliorer la communication de la machine vers l'être humain.
- La traduction automatique : l'objectif de ce domaine est d'automatiser complètement le processus de traduction d'une langue source vers une langue cible (Weaver, 1955; Hutchins, 2005).

Les travaux décrits dans ce document pourraient trouver des applications dans plusieurs domaines du TAL, mais ils se concentrent principalement sur des contributions au domaine de l'extraction d'information.

## 1.2. Motivations

Les processus cognitifs sont ceux qui permettent à un être humain d'acquérir des nouvelles connaissances. Ces processus s'appuient sur la comparaison et l'analyse de multiples informations appartenant à un même domaine de connaissance (Rowley, 2007). Par exemple, si un être un humain fait les observations suivantes :

- Plusieurs corbeaux observés sont noirs,
- Aucun corbeau observé n'est d'une autre couleur.

Alors, selon le raisonnement inductif théorisé par Aristote, il pourra concevoir par induction, c'est-à-dire à partir d'exemples, la connaissance que tous les corbeaux sont noirs. Ce raisonnement est d'ailleurs au fondement de l'apprentissage automatique. Pour que l'humain ou la machine puissent acquérir de nouvelles connaissances, ils doivent donc commencer par recueillir des informations.

Or, l'humanité recueille en permanence une quantité gigantesque d'informations qu'elle entrepose sous diverses formes, telles que des livres ou des fichiers numériques, pour faciliter leur accès à un plus grand nombre. Pour rendre possible leurs analyses à grande échelle, une partie est entreposée sous la forme de données structurées, c'est-à-dire sous la forme de données interprétables par des machines. Néanmoins, la majorité de ces données sont non structurées en moyenne, 31% à 85% (Russom, 2007), donc non directement interprétables par des machines. La quantité et la complexité de ces données rendent d'ores-et-déjà impossible une structuration manuelle des informations qui y sont décrites. Parmi elles, une partie importante est composée de textes rédigés en langue naturelle. En conséquence, des besoins importants en systèmes capables d'extraire automatiquement les informations contenues dans ces données textuelles ont émergés.

Le domaine de l'extraction d'information (EI) vise à répondre à ces besoins en produisant des méthodes capables d'analyser des données textuelles numériques, d'en extraire des informations ciblées et de les représenter sous une forme structurée prédéfinie (Dupont et al., 2002). Ces informations deviennent alors directement exploitables par d'autres programmes ou par des humains. Depuis les débuts de l'EI dans les années 80 (Grishman and Sundheim, 1996), le domaine a vu l'apparition de plusieurs approches. Dans cette thèse nous nous intéresserons aux méthodes d'EI portant sur l'extraction d'entités et des relations qui peuvent les relier.

## 1.3. Problématique

Nous considérons ici qu'une expression est un mot ou une séquence de mots non-nécessairement contigus qui peut désigner une entité, c'est-à-dire un objet réel (De Saussure, 1989), une unité significative (Martinet, 1956) (Figure 1). Cette expression représente alors une mention de cette entité dans un texte. Un couple entité/mention représente donc un signe linguistique. Une entité peut également être vue comme une

instance d'un concept, c'est-à-dire comme un objet (l'instance) qui hérite de toutes les propriétés d'un autre objet plus abstrait (un concept). La tâche d'extraction d'entités est définie comme le processus d'association de la mention d'une entité dans un texte à une ou plusieurs références sémantiques faisant office de signifiant. Ces références peuvent être les entités associées, si elles existent parmi les références, mais sont principalement des concepts, c'est-à-dire des objets plus abstraits que les entités. Selon les objectifs de la tâche, les références disponibles peuvent en effet être plus ou moins abstraites : par exemple, l'expression "le chien de mon voisin" peut tout aussi bien être associée au concept <chien>, qu'à tous les concepts subsumés tels que les concepts <mammifère>, <animal>, <être vivant>, etc. Enfin, la tâche d'extraction de relation (voir Figure 2) consiste à détecter une relation sémantique exprimée dans le texte et reliant une, deux ou plus de mentions d'entités.

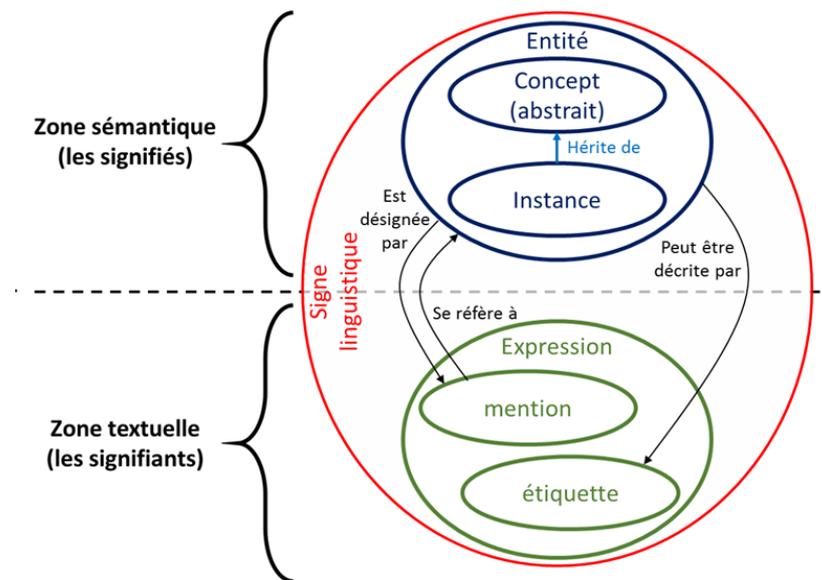


Figure 1 : Schéma explicatif des différents concepts abordés

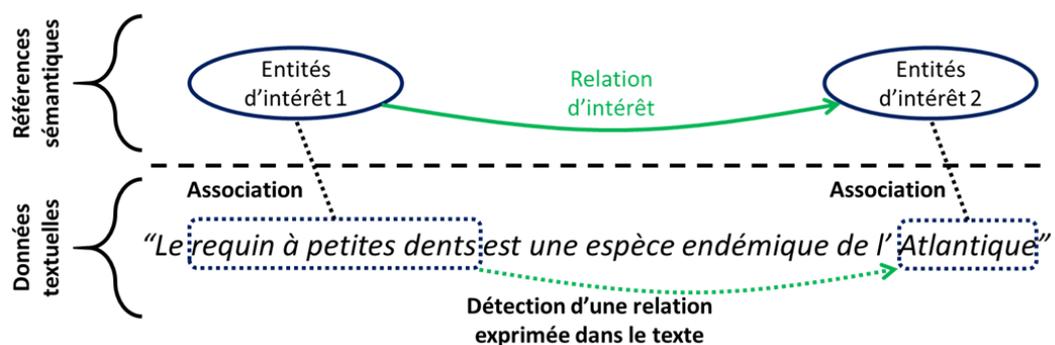


Figure 2 : Exemple d'une extraction d'entité et de relation.

La plupart des approches s'appuie sur des ressources externes (Friedman et al., 2001; Kazama and Torisawa, 2007; Ratinov and Roth, 2009), c'est-à-dire des ressources autres que le corpus de texte duquel les informations sont à extraire. Ces ressources peuvent prendre des formes diverses, dont celle d'une représentation formelle de connaissances telle qu'une ontologie. Ces représentations peuvent contenir de nombreuses autres

connaissances et informations d'intérêt pour la tâche d'Extraction d'Information, et ont de plus une grande utilité lorsqu'il s'agit de définir formellement les types d'informations visées par une tâche, par exemple, les types d'entités (Nédellec et al., 2009). Dans ce cas, une tâche d'EI peut être vue comme la mise en relation entre des données textuelles et une représentation de connaissances (voir Figure 3).

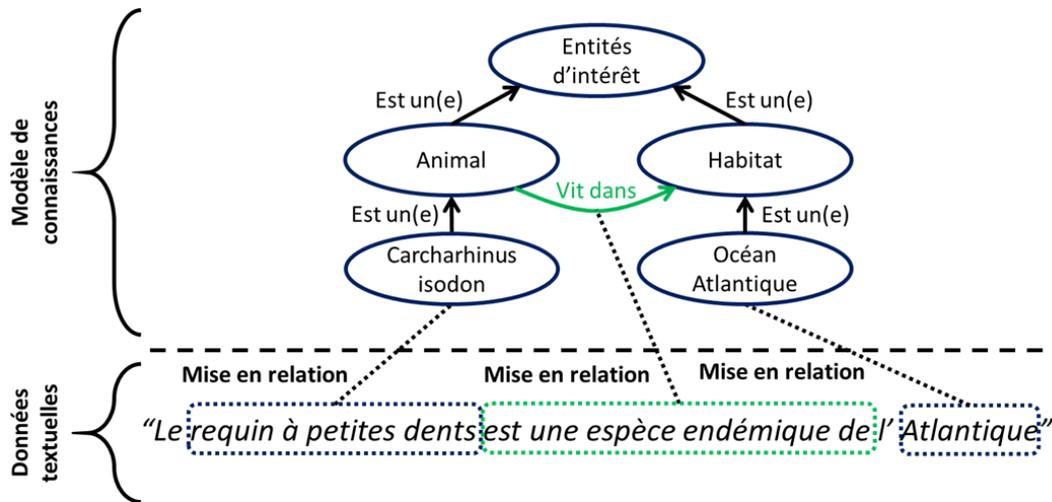


Figure 3 : Exemple d'une mise en relation entre des données textuelles et un modèle de connaissances

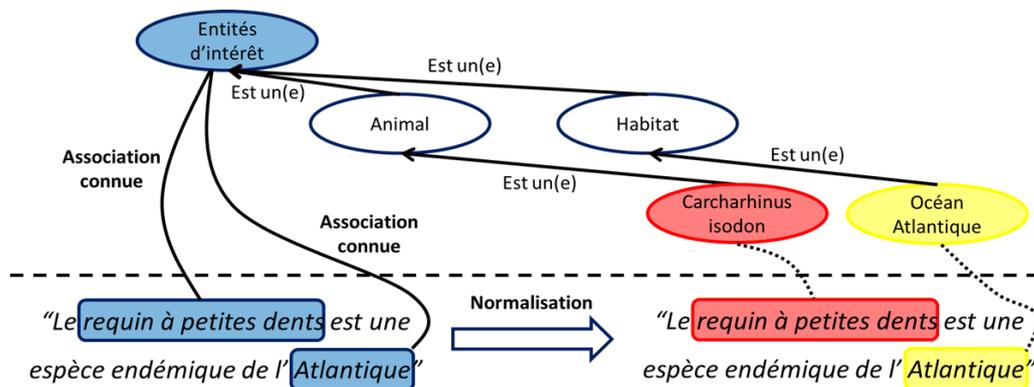


Figure 4 : Exemple de normalisation par les concepts d'une ontologie. La normalisation s'appuie sur une première classification : les mentions d'entités d'intérêts à normaliser sont déjà connues.

Parmi les différentes étapes possibles d'un système d'EI, l'étape de normalisation est celle qui classe des mentions d'entités d'intérêt extraites d'un texte dans des classes définies dans un ensemble de références sémantiques. Lorsque cet ensemble est défini dans une ontologie, ces références prennent la forme de concepts. La normalisation par les concepts d'une ontologie est donc une tâche d'association d'expressions, que nous limiterons ici à des expressions préalablement identifiées comme étant d'intérêt pour la tâche, à des concepts qui les représentent le plus finement possible (voir Figure 4).

Une des approches classiques en normalisation d'entités est d'étudier les similarités morphologiques et/ou syntaxiques entre les mentions d'entités et les étiquettes des

concepts du modèle de connaissances. Des mentions possédant une forte similarité de forme avec une étiquette d'un concept seront associées à celui-ci. Néanmoins, au-delà des possibles flexions et dérivations morphosyntaxiques ou lexicales, il est impossible d'expliciter l'ensemble des expressions qui peuvent être mentions d'une entité. Par exemple, il n'y a pas de similarité de forme entre l'étiquette de concept "requin" et l'expression désignant une des espèces de requin "Carcharhinus isodon". L'étude de la similarité de forme ne peut donc suffire à elle-seule à résoudre le problème de la normalisation d'entités (voir Figure 5).

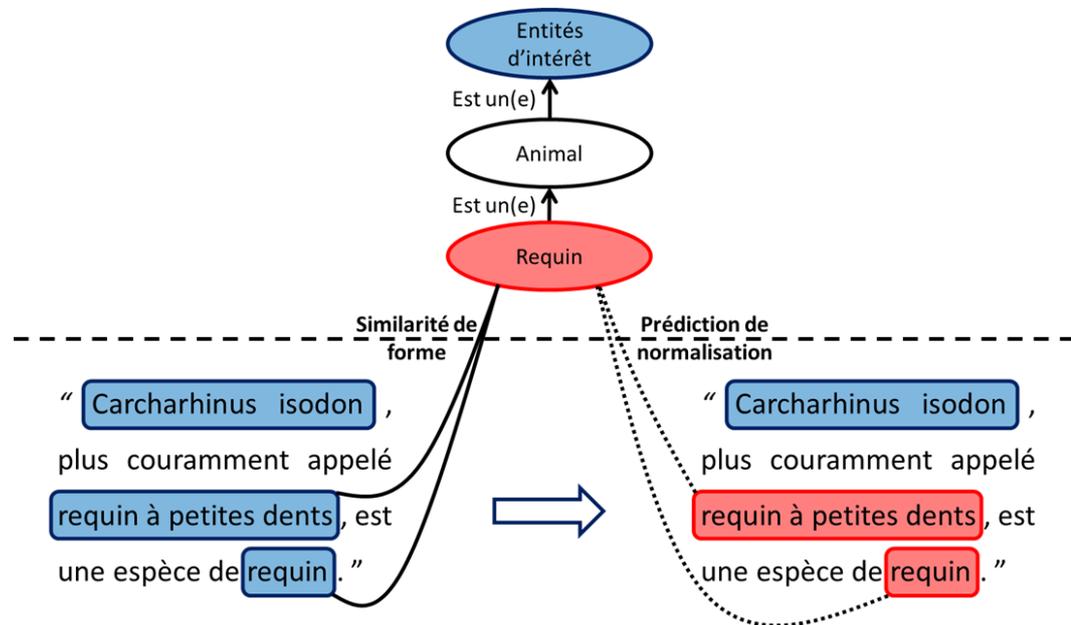


Figure 5 : Schéma d'une normalisation par une méthode qui s'appuierait sur un calcul de similarité de forme entre mentions et étiquettes de concept pour prédire leur association.

Une autre approche consiste à exploiter une partie des informations internes à un corpus pour former des représentations capables de capturer une partie du sens des données textuelles, notamment de mentions d'entités. En particulier, à partir de l'hypothèse de sémantique distributionnelle (Harris, 1954; Firth, 1935), des méthodes utilisent uniquement la distribution des mots ou le nombre d'occurrences, dans un corpus non-annoté pour former de telles représentations. Selon le corpus d'entraînement utilisé, ce genre de méthode peut réussir par exemple à induire, sans aucune autre connaissance préalable, que les mots "dog", "rottweiler" et "beagle" sont des mentions d'entités relativement similaires, car leurs vecteurs se retrouvent les plus proches dans l'espace vectoriel produit. Ainsi :

- Si une méthode obtient l'information qu'une expression donnée est à associer à un type d'entité donnée (par exemple en lui donnant des données annotées, c'est-à-dire des associations expression/concept),
- Si cette expression et toutes celles considérées possèdent une représentation vectorielle par ce genre de méthode,

- Alors cette méthode pourra également prédire que les expressions dont le vecteur est proche du vecteur de l'expression donnée sont à associer au même type d'entités (voir Figure 6).

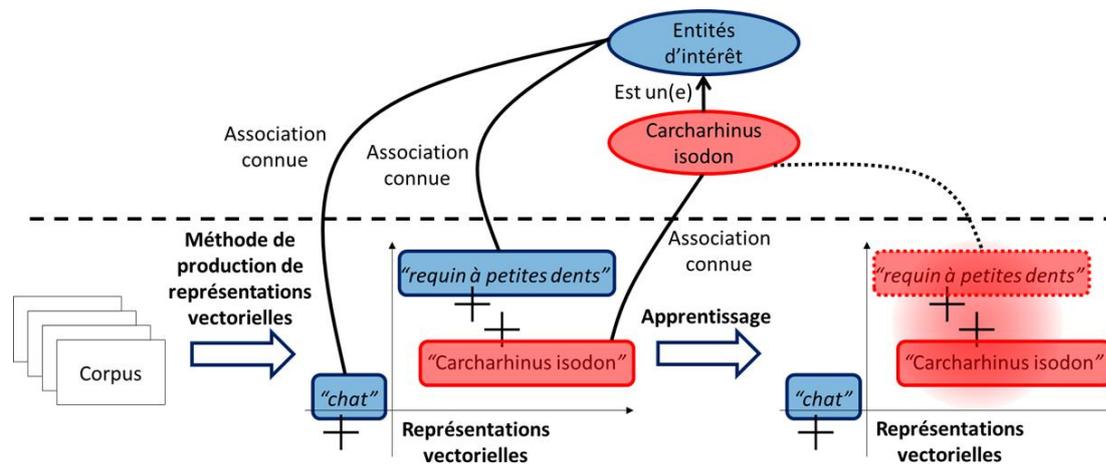


Figure 6 : Schéma des approches de normalisation fondée sur des représentations vectorielles de séquences de mots.

Ce genre de méthode s'appuie principalement sur de l'apprentissage automatique, dont l'une des principales difficultés reste le problème de la généralisation : comment réussir à associer des expressions et concepts pour des concepts absents des données d'entraînement ? Dans les tâches où le nombre de données d'entraînement est relativement faible, notamment comparé au nombre de concepts d'intérêt, il est plus difficile de proposer des méthodes par apprentissage avec une bonne capacité de prédiction (Larochelle et al., 2008).

Les espaces vectoriels permettant de capturer le sens de données textuelles sont qualifiés d'espaces sémantiques (ES) (Baroni and Lenci, 2010), et ont permis de produire des méthodes relativement performantes dans diverses tâches de TALN (Turian et al., ; Collobert et al., ; Socher et al., ; Al-Rfou et al., 2013). Cependant, elles sous-exploitent les bases de connaissances disponibles, dont l'intégration a pourtant montré son intérêt, comme en recherche d'information (Rada et al., 1989; Cohen and Widdows, 2009) ou en classification de documents (Lin et al., 2007), et a pu faciliter la découverte de nouvelles connaissances à partir de la littérature (Agarwal and Yu, 2009; Doshi-Velez et al., 2009).

Cela soulève plusieurs questions :

- 1) Est-ce que ces connaissances pourraient effectivement enrichir les méthodes distributionnelles pour améliorer la tâche de normalisation sur laquelle nous nous concentrons ?
- 2) Comment intégrer ces connaissances aux représentations distributionnelles ?
- 3) Les réponses aux questions précédentes permettent-elles de traiter le problème de l'insuffisance des données disponibles dans certaines tâches ?

Le domaine des sciences du vivant ou du biomédical présente de nombreux avantages comme domaine de développement pour répondre à ces trois questions, notamment :

- Il possède une communauté dynamique et intéressée par les questions de représentation de connaissances et de récupération d'informations dans la littérature de leur domaine,
- Il produit de nombreuses publications scientifiques, riches en informations et facilement accessibles,
- Il produit des connaissances qui mettent au défi nos capacités de modélisation,
- Ses experts produisent de nombreuses représentations de connaissances, dont des ontologies.

Ces ontologies permettent principalement de rassembler une communauté d'un domaine de spécialité autour d'une description formelle des connaissances expertes du domaine. En domaine de spécialité, le nombre de documents textuels produits reste relativement limité par rapport au domaine général et le nombre potentiel d'annotateurs humains compétents l'est encore davantage, ce qui limite considérablement le nombre de documents annotés. Ce contexte représente donc un défi pour les approches par apprentissage qui représentent actuellement l'état de l'art dans de nombreuses tâches de TALN en domaine général, et qui sont fortement dépendantes du nombre de données annotées à disposition.

## 1.4. Contributions

Pour répondre aux questions posées précédemment, nous avons développé une approche originale qui permet d'intégrer une représentation distributionnelle produite à partir d'un corpus de textes et des connaissances externes d'une ontologie.

Pour résoudre la tâche de normalisation, notre approche s'appuie sur un alignement entre l'espace distributionnel des expressions (EDE), fondé sur la distribution des mots dans un corpus externe non-annoté, et un espace sémantique ontologique (ESO). La mise en relation entre ces deux espaces permet de déterminer pour une mention donnée quel est le concept associé dans l'ontologie. Pour que les deux espaces EDE et l'ESO soient de même nature et comparables, nous proposons plusieurs méthodes de construction d'un ESO à partir d'ontologies.

Toutes nos méthodes ont été évaluées sur la tâche de normalisation de la compétition Bacteria Biotope (Deléger et al., 2016). Nos résultats dépassent les meilleurs résultats obtenus par des travaux précédents. Cette performance permet d'argumenter en faveur d'une complémentarité des connaissances externes vis-à-vis de l'information distributionnelle, en permettant par la même occasion de proposer une manière de les intégrer l'une dans l'autre.

Nos méthodes ont également montré une excellente capacité de généralisation à partir de peu de données d'entraînement. Elles ouvrent donc également des pistes intéressantes quant au dépassement des limitations des algorithmes d'apprentissage supervisé dues à

l'insuffisance des données d'entraînement, tout au moins dans leurs applications à des représentations sémantiques.

Nous avons développé plusieurs méthodes autour de cette approche principale, avec des caractéristiques distinctes, notamment vis-à-vis de leur dépendance à des données annotées. Toutes ces méthodes ont été implémentées et partagées en logiciel libre. Elles ont été intégrées à la suite logicielle Alvis<sup>1</sup>, également libre, permettant ainsi le partage avec l'ensemble de la communauté de chercheurs en traitement automatique des langues, leur réutilisation, ainsi que la reproduction des résultats publiés.

## 1.5. Plan

Ce manuscrit s'articule autour de cinq chapitres. Voici une description de leur contenu :

- Le chapitre 2 introduit le contexte général des travaux réalisés durant cette thèse. Elle définit la majeure partie des concepts utilisés pour décrire les travaux, et montre leur positionnement scientifique vis-à-vis des domaines de l'extraction d'information et de la représentation des connaissances. Enfin, une présentation plus approfondie de la tâche de normalisation et des difficultés associées y est faite.
- Le chapitre 3 décrit principalement les méthodes de l'état de l'art pour la tâche de normalisation d'entité. Deux grandes catégories de méthodes y sont présentées : les méthodes par étude de la similarité de forme entre mentions du texte et étiquettes de concepts, et les méthodes fondées sur des représentations sémantiques, fondées principalement sur l'apprentissage supervisé. Les représentations sémantiques sont des représentations vectorielles permettant de capturer une partie du sens des expressions qu'elles représentent. Un état de l'art sur la construction de ces représentations y est également présenté.
- Les chapitres 4 et 5 présentent les contributions principales de ces travaux de thèse. La nouvelle approche proposée se positionne comme une méthode par apprentissage supervisé fondée sur des représentations sémantiques. La méthode CONTES, ainsi que plusieurs méthodes dérivées pour aborder certaines limitations initiales, y sont présentées.
  - Le chapitre 4 permet de détailler les méthodes de construction des représentations sémantiques : celles des mentions d'entités et celles des concepts d'une ontologie.
  - Le chapitre 5 présente comment la méthode CONTES utilise les représentations sémantiques présentées au chapitre 4 pour apprendre une projection permettant de prédire un concept pour une mention. En s'appuyant sur cette méthode initiale, plusieurs dérivées sont également présentées.

---

<sup>1</sup> <https://bibliome.github.io/alvisnlp/>

- Le chapitre 6 nous permettra de conclure et d'énoncer plusieurs perspectives envisageables à l'issue des travaux réalisés durant cette thèse.



# **Chapitre 2**

-

# **Contexte**



## 2.1. Introduction

Pour introduire la normalisation d'entité, tâche particulière d'extraction d'information, il nous faut tout d'abord introduire ce qu'est une information dans notre contexte. Nous verrons que ce concept d'information est intimement lié à celui de donnée et celui de connaissance. Les connaissances ont d'ailleurs une place d'importance particulière pour le travail présenté ici, puisqu'elles peuvent être modélisées par des ontologies, lesquelles seront exploitées par les méthodes présentées dans les chapitres suivants.

Nous présenterons ensuite le domaine de l'extraction d'information par ses tâches les plus représentatives : la préparation de corpus, la reconnaissance d'entité, l'extraction de relation et enfin la normalisation d'entité. Cela nous permettra d'articuler la tâche de normalisation avec le reste du domaine.

Nous pourrions alors introduire les principales difficultés faisant obstacle à la normalisation d'un texte. Nous abordons notamment dans nos travaux deux difficultés :

- Les différences de forme entre les mentions et les étiquettes des concepts qui devraient les normaliser.
- Le nombre trop faible d'exemples annotés par un concept, c'est-à-dire des ensembles de mentions dont le concept qui le normalise est connu.

Enfin, nous finirons ce chapitre par la présentation de la tâche de normalisation d'habitats bactériens Bacteria Biotope du challenge BioNLP Shared-Task 2016 (Deléger et al., 2016). Les données fournies par cette tâche nous permettront d'évaluer les méthodes que nous avons développées et de les comparer avec celles dont les résultats ont été publiés.

## 2.2. Représentations sémantiques pour l'extraction d'information

### 2.2.1. Introduction

Qu'est-ce qu'une connaissance ? Depuis les premières réponses par Platon dans *Theaetetus* (Stocks, 1935) au IV<sup>e</sup> siècle av. J.-C., il ne semble toujours pas y avoir de réponse consensuelle. De même, la définition d'une donnée et d'une information reste une tâche complexe. Avant de pouvoir aborder le domaine de l'extraction d'information, il est avant toute chose nécessaire de définir les contours de ces concepts fondamentaux. Nous faisons le choix ici de nous appuyer sur certaines des définitions existantes et de les adapter à nos besoins en extraction d'information, particulièrement pour la normalisation d'entité. Notamment, nous nous servirons de ces définitions pour introduire les différents

formats structurés utilisés pour représenter des connaissances et/ou des informations, dont les deux sur lesquels se fondent nos travaux : les ontologies et les espaces sémantiques.

## 2.2.2. De la donnée, à l'information, à la connaissance

Nous faisons le choix ici de nous appuyer sur le modèle DIKW (*Data, Information, Knowledge, Wisdom*) (Rowley, 2007), qui a été proposé pour faciliter le traitement d'information à partir de données brutes dans le cadre de la discipline du management de données (voir Figure 7). Ce modèle propose de délimiter la connaissance en s'appuyant sur une définition de l'information, elle-même s'appuyant sur une définition de la donnée. Ce modèle a donc l'intérêt d'introduire et d'articuler simultanément trois concepts fondamentaux en EI.

Une donnée (Ermine et al., 2012) est un élément brut, unitaire, une description élémentaire d'une réalité, qui n'a pas encore été interprétée. Dans le cas de l'EI et des travaux présentés ici, nous considérons une donnée textuelle comme une séquence ordonnée de séquences de caractères, issue d'étapes de préparation d'un corpus brut. Par exemple, les séquences ("*bcat-1*") et ("*bcat-1*", "*is*", "*d*", "*gene*") seront considérées comme des données textuelles. Des séquences de données textuelles sont considérées également comme des données textuelles. Par exemple, ("*bcat-1*", "*is*", "*d*", "*gene*", ("*bcat-1*", "*is*", "*regulated*", "*by*", "*cholestanol*") représentent une donnée textuelle qui pourrait être considérée comme une représentation de deux phrases segmentées. Une donnée textuelle pourra donc représenter des n-grammes, des mots, des phrases, des paragraphes et même des corpus entiers.

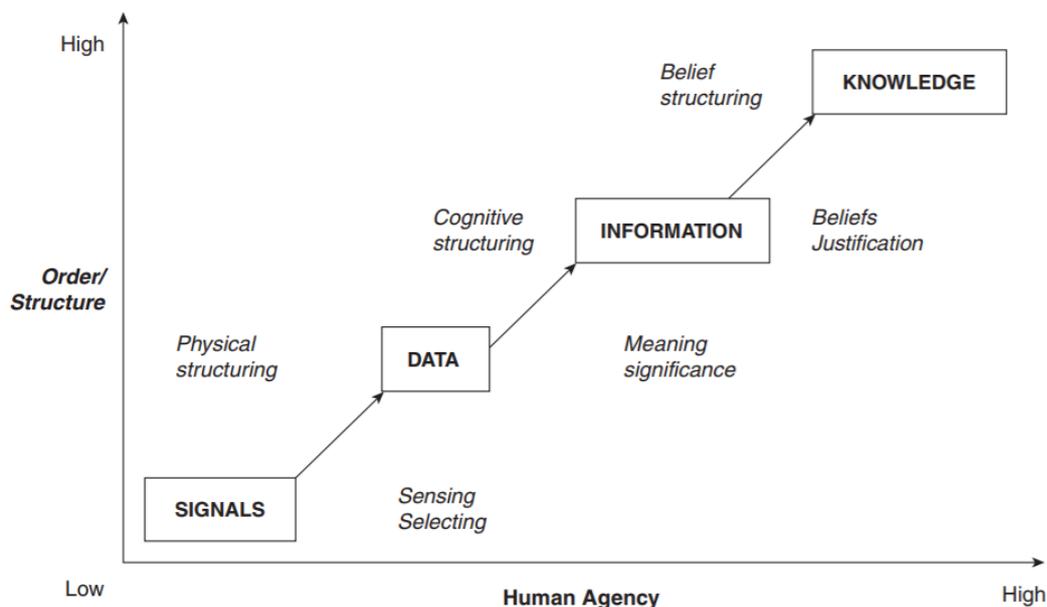


Figure 7 : Schéma du modèle DIKW de (Rowley, 2007), d'après (Choo, 1996)

L'information est une donnée à laquelle les connexions relationnelles ont donné un sens (Floridi, 2010). C'est le résultat de l'interprétation de données, c'est-à-dire une réponse à la question : qu'est-ce que cette donnée ? Dans le cas de l'EI et des travaux présentés ici, nous limiterons une information à une donnée textuelle associée à un objet sémantique servant de référence. Par exemple, si la référence est un concept représentant un gène, associer la donnée "*bcat-1*" au concept <gène> crée une information. On exprimera alors une information structurée sous la forme d'un couple (données, référence).

Une difficulté supplémentaire en français pour réussir à définir la connaissance provient de la confusion courante entre "*savoir*" et "*connaissance*", deux concepts traduits en anglais par le même terme "*knowledge*" (Pesqueux, 2008). La connaissance, au sens de "*knowledge*", est parfois définie comme l'interprétation experte d'une combinaison d'informations (Heisig, 2002). La connaissance, notamment en français, est souvent définie comme une notion opérationnelle et appliquée : elle permet la transformation d'informations en instructions (Ackoff, 1974). A la différence du "*savoir*", qui est plutôt le résultat final de l'acquisition et de la mise en relation de plusieurs connaissances, et souvent sans aucune visée opérationnelle. Dans le cas de l'EI et des travaux présentés ici, nous nous rapprocherons de la définition issue de l'empirisme logique et des travaux en représentation des connaissances et intelligence artificielle : une connaissance est un ensemble de règles logiques connectées entre elles, s'appliquant sur des données et des informations (Russell and Norvig, 2016). En particulier, la règle  $[\forall x \in A \Rightarrow x \in B]$  est une connaissance fondamentale qui exprime la subsomption entre deux concepts A et B.

### 2.2.3. Ontologie : un modèle formel et structuré de connaissances

Un modèle est un système représentant les structures essentielles d'une réalité et capable à son niveau d'en expliquer le fonctionnement (Birou, 1966). Un modèle de connaissances est donc une représentation essentialisée de connaissances réelles d'un domaine de connaissance. Si un modèle est décrit dans un langage formel, c'est-à-dire un langage qui ne permet que de produire des propositions vérifiables, c'est un modèle dit formel. Une ontologie est en particulier un modèle formel interprétable par la machine, définie initialement comme une "*spécification explicite et partagée d'une conceptualisation*" (Gruber, 1993). En EI, ce genre de modèle peut être à la fois utilisé comme ensemble de références sémantiques pour une tâche (Friedman et al., 2001) et comme ressource contenant des connaissances *a priori* d'un domaine. En particulier, une tâche de normalisation peut alors être vue comme une tâche de création automatique d'une base de connaissance à partir de texte et d'ontologie, c'est-à-dire comme une tâche de peuplement d'une ontologie à partir de mentions d'entités extraites de texte (Baader et al., 2003).

Les modèles formels et structurés de connaissances jouent principalement un rôle important pour la représentation de connaissances dans de nombreux domaines (Ashburner et al., 2000; Navigli and Ponzetto, 2010; Miller, 1995; Kim et al., 2013).

Notamment, ils ont l'avantage d'être facilement partageables et réutilisables, et permettent de réunir des communautés spécifiques autour de concepts formellement définis représentant leur domaine. Plus spécifiquement, les méthodes usuelles de construction d'une ontologie peuvent s'appuyer sur des corpus de textes, lesquels représentent une source importante de connaissances à modéliser (Nobécourt, 2000). La première étape est alors d'extraire des termes (c'est-à-dire des expressions canoniques) pouvant représenter des concepts spécifiques du domaine, c'est-à-dire de construire une terminologie, puis de les organiser en thésaurus (Aussenac-Gilles et al., 2000; Lame, 2000) via des relations de subsomption. Cette relation de hiérarchisation sémantique est la relation principale d'une ontologie (Ashburner et al., 2000; Miller, 1995; Suchanek and Weikum, 2008).

Selon le cadre formel, on peut représenter une ontologie sous une forme graphique (voir Figure 8) ou sous la forme d'un ensemble de règles logiques. Une ontologie est principalement composée :

- D'un ensemble de concepts étiquetés formant également une terminologie du domaine représenté. Un concept permet d'unifier différentes étiquettes sous un même objet, malgré les variations des formes de ces étiquettes.
- D'instances (ou individus), c'est-à-dire d'objet héritant de concept(s), même s'il n'est pas requis d'avoir des instances pour constituer une ontologie. Lorsqu'une ontologie est instanciée, on parle d'une base de connaissances.
- De relations sémantiques qui peuvent relier des instances, voire des concepts dans le cas particulier de la relation fondamentale de subsomption entre concepts (nommée "*is a*"), qui est en particulier une relation d'ordre. Cette relation entre deux concepts A et B (noté  $A \text{ is\_a } B$  ou  $A \subset B$ ) permet d'inférer que toute instance de A est aussi une instance de B :  $\forall x \in A \Rightarrow x \in B$ .

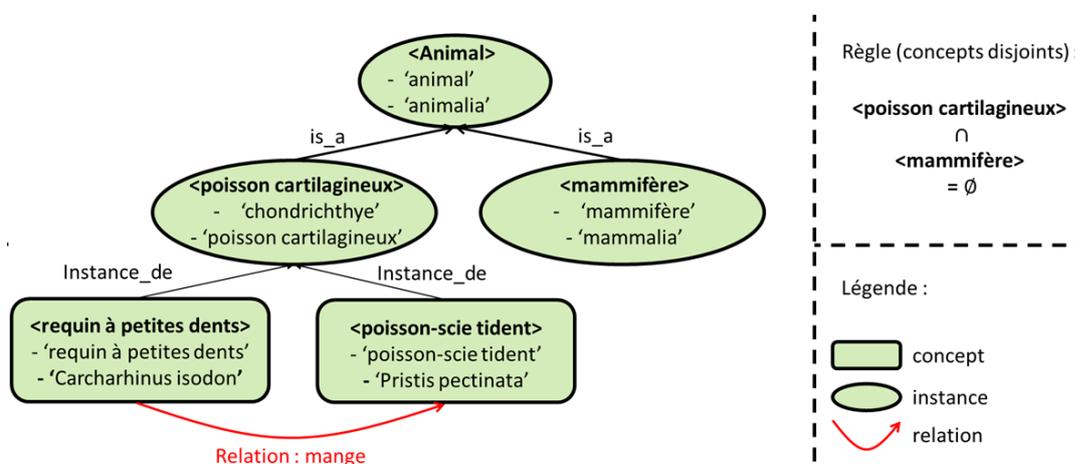


Figure 8 : Exemple d'une base de connaissances avec ses concepts, ses instances de concepts, ses relations entre instances ou entre concepts et ses attributs.

Une ontologie permet donc de représenter des connaissances et d'organiser des informations. Même si elles restent peu utilisées en EI, l'utilisation de connaissances

externes comme ressources, dont l'ontologie reste le modèle le plus riche, a déjà permis d'obtenir des meilleurs résultats en recherche d'information (Cohen and Widdows, 2009) ou en classification de documents (Lin et al., 2007), et a pu permettre des découvertes de nouvelles connaissances à partir de la littérature (Agarwal and Yu, 2009; Doshi-Velez et al., 2009). La modélisation et l'utilisation de connaissances externes est donc un enjeu d'intérêt pour l'EI, notamment pour la littérature de domaines de spécialité (Nédellec et al., 2009).

#### 2.2.4. Espaces sémantiques

Il existe de multiples manières de représenter le sens d'entités, notamment les modèles fondés sur des caractéristiques ("*feature-based model*"), les espaces sémantiques ou encore les ontologies/réseaux sémantiques (voir Figure 9). Avec les ontologies ou les réseaux sémantiques, le sens d'une entité est principalement exprimé en fonction des relations qu'elle entretient avec les autres entités du réseau (Steyvers and Tenenbaum, 2005). Les modèles fondés sur des caractéristiques essaient, quant à eux, de définir une liste de caractéristiques quantitatives permettant de décrire chaque entité (Smith and Medin, 1981). Enfin, un espace sémantique (ES), parfois nommé "*modèle fondé sur un espace vectoriel*" (en anglais : "*Vector Space Model*") est un espace vectoriel dans lequel un ensemble de vecteurs est définis. Chaque vecteur est une représentation d'un objet qui permet de capturer une partie de son sens. Alors que les ontologies et les modèles fondés sur des caractéristiques sont principalement produit manuellement (Hinton and Shallice, 1991), les ES sont principalement construits automatiquement en analysant les contextes des entités (principalement en analysant des contextes de mots dans un corpus).

En TAL, les ES sont majoritairement les espaces vectoriels ( $\mathbb{R}^n$ , +, .) ( $n \in \mathbb{N}^*$ ). Un avantage de tels vecteurs composés de nombres réels est qu'ils peuvent être utilisés directement par des algorithmes d'apprentissage automatique (Riesen et al., 2007). La principale caractéristique dans un ES est que deux objets ayant un sens similaire ont des vecteurs spatialement à proximité (selon une certaine distance), et inversement, moins des objets seront similaires, plus ils seront distants. En conséquence, une distance entre deux objets permet d'estimer la similarité sémantique entre ces deux objets.

Une autre caractéristique couramment observée dans des ES est de représenter des relations sémantiques binaires entre deux entités par un vecteur constant (Mikolov et al., 2013c). Un vecteur de relation peut en effet être déterminé en observant la différence entre deux vecteurs d'entité dans l'ES :  $\langle \textit{Royauté} \rangle = \langle \textit{Roi} \rangle - \langle \textit{Homme} \rangle$  (voir Figure 10). Ce vecteur de relation, une fois appliqué à une autre entité, permet alors de retrouver l'entité-cible :  $\langle \textit{Femme} \rangle + \langle \textit{Royauté} \rangle \approx \langle \textit{Reine} \rangle$ . Cela permet d'effectuer des analogies :  $\langle a \rangle$  est à  $\langle b \rangle$ , ce que  $\langle c \rangle$  est à  $\langle d \rangle$ . L'observation de cette caractéristique est actuellement une des façons courantes d'évaluer la qualité d'un ES (Drozd et al., 2016). Ce genre d'évaluation fondée sur la disposition de représentations d'entités est appelé évaluation intrinsèque. Les évaluations extrinsèques quant à elles sont

des évaluations qui utilisent les scores obtenus en utilisant des représentations vectorielles sur d'autres tâches de TAL (par exemple, reconnaissance d'entité ou extraction de relation).

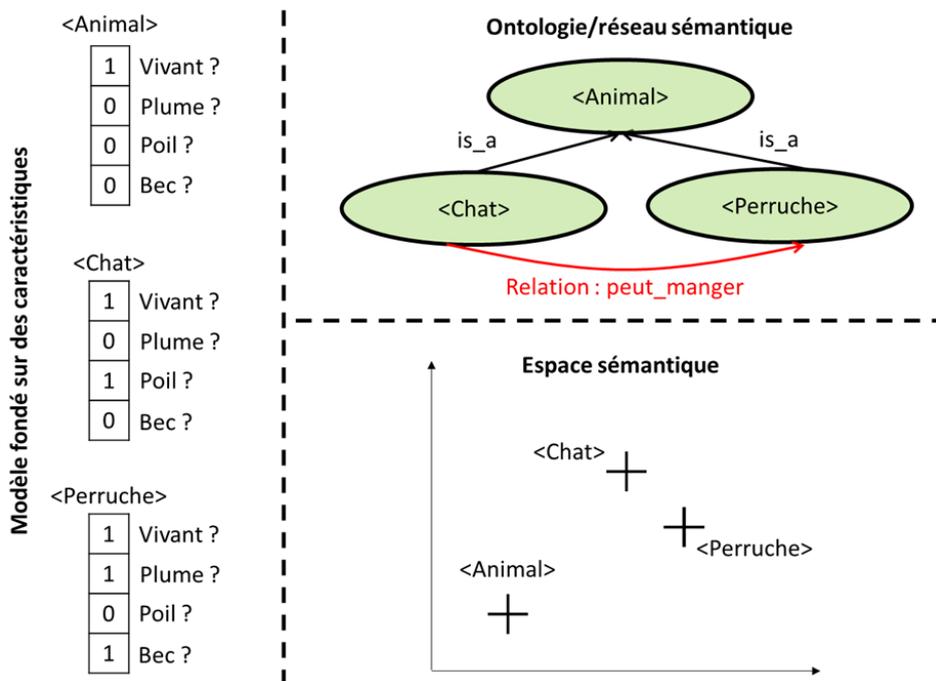


Figure 9 : Exemples de représentations d'entités avec un modèle fondé sur des caractéristiques (à gauche), un graphe de connaissances/ontologie (en haut à droite) et un espace sémantique (en bas à droite).

Néanmoins, cette propriété n'est vraie qu'avec les relations un-à-un, et ne peut pas permettre de représenter d'autres relations telles que :

- Les relations symétriques : par exemple, la relation frère/sœur, car si A est frère de B, alors B est frère de A. En effet, la seule relation symétrique possible serait représentée par un vecteur nul, et l'entité source aurait le même vecteur que l'entité cible.
- Les relations 1:n : par exemple, la relation qui relie un gène aux protéines qu'il code. Un gène peut en effet coder plusieurs protéines différentes. Dans ce contexte, toutes les protéines devraient être les mêmes et de manière générale, toutes les entités ciblées par une relation 1 : n devraient être les mêmes entités.
- Les relations n:1 : par exemple, l'inverse de la relation précédent, à savoir que plusieurs protéines peuvent être codées par un même gène, pour les mêmes raisons que les relations précédentes excepté que ce sont alors les entités sources qui devraient être identiques.
- Les relations multivaleurs ou n:n : par exemple, plusieurs personnes peuvent être auteurs d'un même document (plusieurs-à-un), mais plusieurs documents peuvent avoir également les mêmes auteurs. La relation de subsomption est également une relation multivaleurs. Dans ce cas, toutes les entités sources devraient être identiques, tout comme toutes les entités cibles.

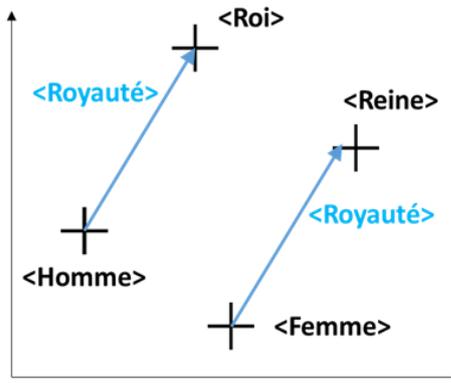


Figure 10 : Exemple de la représentation de relation sémantique dans un ES

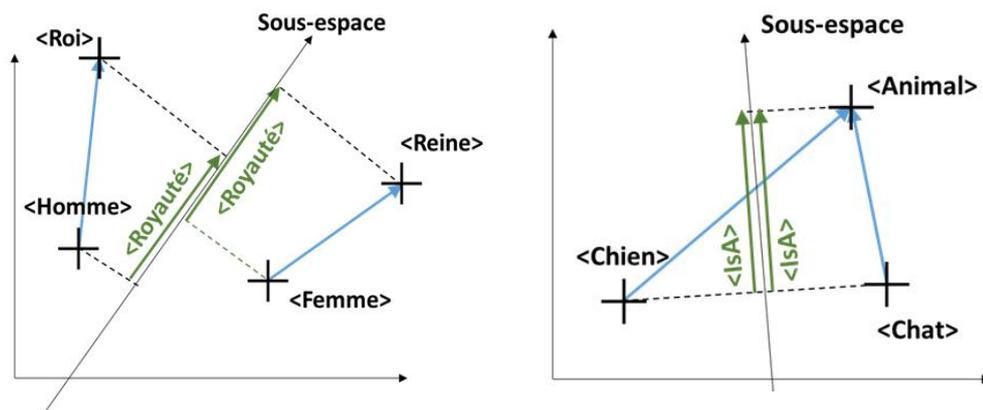


Figure 11 : Représentation de l'approche des méthodes de représentation de relations binaires par projection sur un sous-espace. Dans l'espace sémantique initial, la relation de <Royaute> (à gauche) ou la relation de subsumption <IsA> (à droite) ne sont pas approximées par un vecteur constant dans l'espace. Comme la relation <IsA> est  $n:1$ , si cette propriété était vérifiée, elle impliquerait nécessairement que <Chien> et <Chat>, qui héritent de <Animal>, soient les mêmes entités.

Pour pallier au manque d'expressivité de ce modèle, certains travaux proposent de représenter des relations binaires par des vecteurs constants une fois projeté dans un sous-espace de l'ES considéré (Wang et al., 2014). Autrement dit, pour chaque type de relation binaire, il existe un sous-espace de l'ES initial dans lequel chaque relation de ce type peut être approximée par un vecteur constant entre les vecteurs des entités projetés de l'ES dans ce sous-espace (voir Figure 11).

Néanmoins, pour construire un ES avec ces propriétés, il est nécessaire d'avoir à disposition un réseau sémantique ou une ontologie, représentant l'ensemble des entités d'intérêt mises en relation. Or, l'objectif des méthodes de construction d'espace sémantique est justement de faire émerger ce genre de propriété.

## 2.3. L'extraction d'information

### 2.3.1. Introduction

Historiquement, l'EI a émergé comme un domaine de recherche à la suite du programme DARPA MUC (*Message Understanding Conference*) (Lewis, 1991; Chinchor et al., 1993). Depuis ses débuts, l'extraction d'information ne vise pas à l'extraction de l'ensemble exhaustif des informations contenues dans des textes : l'EI vise à extraire des mentions d'objets particuliers ainsi que les occurrences de types de relations sémantiques qui peuvent relier des mentions entre elles (Russell and Norvig, 2016). Ces objets et ces types de relation peuvent alors être formalisés par un modèle, tel qu'une ontologie (voir Figure 12). Autrement dit, l'EI ne cherche pas à comprendre les textes dans leur ensemble, mais seulement à structurer une partie des informations contenues dans le texte.

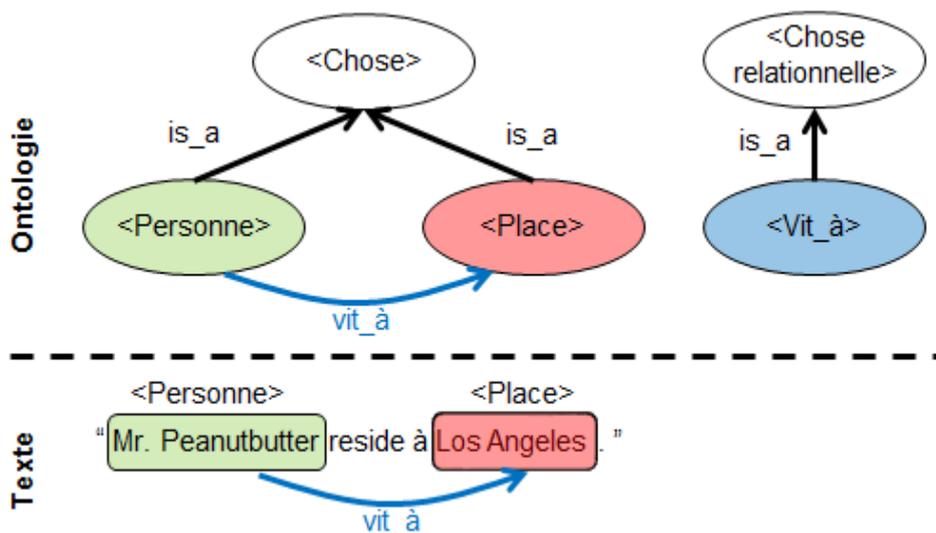


Figure 12 : Un exemple d'EI sur une phrase avec une ontologie comme modèle d'extraction.

Cette section se clôture avec une présentation de la tâche de normalisation d'entité, laquelle est principalement traitée par les travaux présentés dans ce manuscrit.

### 2.3.2. La préparation de corpus

Outre les étapes de "haut niveau" spécifiques à l'extraction d'informations variées, de nombreuses étapes de préparation ainsi que des étapes intermédiaires sont nécessaires pour atteindre les objectifs.

En effet, les formats des textes ne sont pas forcément lisibles pour un programme (ex : images de textes en écriture manuscrite, fichiers PDF, etc.). Une autre étape de préparation est la sélection de documents d'intérêt pour la tâche. Par exemple, si l'objectif est d'extraire des informations en lien avec la recherche en sciences du vivant, il n'est sans doute pas pertinent de sélectionner comme documents des articles de magazines de mode grand public. Cette sélection a donc pour objectif de constituer un corpus de documents en lien avec le domaine de connaissances étudié, contenant certainement plus d'informations que des documents hors-domaine. En conséquence, les textes sélectionnés contiendront par exemple moins d'ambiguïtés. Par exemple, si le domaine d'intérêt est juridique, et que cette sélection de documents est correctement effectuée, on trouvera très majoritairement le mot "avocat" employé dans le sens d'auxiliaire juridique, et certainement très peu ou pas dans le sens de fruit.

De plus, avant d'effectuer une tâche d'EI, il est nécessaire de préparer un corpus de textes en le segmentant en mots, phrases, etc. pour qu'ils soient manipulables par un programme informatique.

Enfin, en plus de ces étapes non-triviales de préparation de corpus, une grande partie des méthodes d'IE nécessite des traitements intermédiaires. Par exemple, l'extraction terminologique correspond à l'extraction de termes d'un corpus. Toutes ces étapes sont parfois nécessaires, et leur qualité peut avoir des répercussions importantes sur les étapes suivantes (Smith, 2011).

### 2.3.3. La reconnaissance d'entités

Le terme "*entité nommée*" est apparu durant le programme MUC au début des années 90. Une entité nommée est une entité dont les mentions sont principalement des noms propres (par exemple, des noms de personnes, de lieux, d'organisations, etc.) (Finkel et al., 2005). Aujourd'hui, la tâche de reconnaissance d'entités vise l'extraction d'entités quelconques et pas seulement les entités nommées.

Dans le texte, une entité est mentionnée principalement sous la forme d'un groupe nominal, étendu ou non (par exemple : "*le petit chat noir du voisin que j'adore*" est un groupe nominal étendu, notamment par une proposition relative). L'ensemble des entités forme une classe ouverte, c'est-à-dire qu'il est impossible d'en faire une liste complète, notamment parce que de nouvelles entités ou de nouvelles formes d'entités sont continuellement créées. De plus, de nombreuses entités similaires sont également susceptibles d'apparaître dans du texte avec des variations morphosyntaxiques importantes.

Une tâche de reconnaissance d'entités est la tâche d'EI d'identification et d'extraction des mentions de ces entités dans un texte vis-à-vis d'un modèle de connaissances de références (*a minima*, un ensemble de classes de références non-relées). Cela correspond

à l'identification d'une mention d'entité, à la détection de ces limites textuelles et à sa classification dans la référence correcte. De façon générale, n'importe quel concept pourrait être utilisé comme référence. Néanmoins, à la différence d'une tâche de normalisation d'entités, cette tâche considère principalement peu de concepts de référence avec un haut niveau d'abstraction, généralement appelés types (voir Figure 13). Par exemple, pour la tâche Bacteria Biotope, la reconnaissance d'entité utilise comme type de référence le concept <Bacterial Habitat>, alors que la normalisation utilisera majoritairement des concepts moins abstraits tels que <Raw Ham> ou <Monkey>. Naturellement, selon les objectifs de la tâche, un concept pourra être considéré comme ayant un haut niveau d'abstraction ou non.

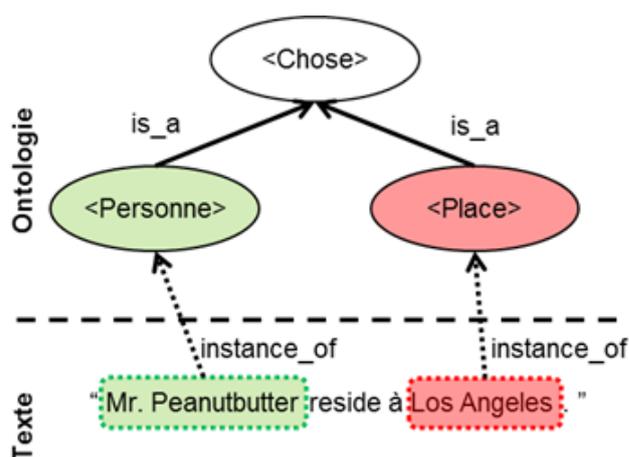


Figure 13 : illustration d'une tâche de reconnaissance d'entités

La reconnaissance d'entités nécessite couramment l'utilisation de terminologies volumineuses représentant une quantité importante d'exemples d'entité à extraire. La tâche est en effet plus difficile si l'on cherche à extraire des entités absentes de la terminologie. Dans les domaines de spécialité, elle peut être légèrement simplifiée car les terminologies pourront être plus exhaustives. Par exemple, dans le cas d'une tâche d'extraction d'entités représentant des espèces vivantes, des dictionnaires d'organismes relativement exhaustifs sont disponibles et donc couramment utilisés (Hirschman et al., 2002). Néanmoins, selon les tâches, des problèmes d'ambiguïtés apparaissent régulièrement. Par exemple, (Hanisch et al., 2005) ont remarqué (en anglais du moins) que les mentions de gènes chez la mouche contenaient souvent des mots appartenant au vocabulaire courant.

Même dans des domaines de spécialité, cette étape reste difficile. Par exemple, en sciences du vivant, de nouveaux termes apparaissent régulièrement. Ceux-ci sont créés par des communautés différentes (Ananiadou et al., 2004), ce qui entraîne également un nombre important de synonymes parfois ambigus. Une ambiguïté fréquente peut être due à des termes employés de façon interchangeable avec des significations pourtant rigoureusement différentes (métonymie). Par exemple, il est courant d'utiliser le nom d'une protéine pour parler également du gène qui l'exprime.

### 2.3.4. L'extraction de relations

Une relation est généralement définie comme un n-uplet :  $r = R(e_1, e_2, \dots, e_n)$ , où les  $e_i$  sont des mentions d'entités et  $R$  est un type de relation de référence (Zhou et al., 2014). Des relations plus complexes peuvent également être utilisées, du type :  $r = R(s_1, s_2, \dots, s_n)$ , où les  $s_i$  sont des mentions d'entités nommées ou des relations extraites, et  $R$  est un type de relation prédéfinie. Lorsque  $n=1$ , la relation est dite unaire. Lorsque  $n=2$ , elle est dite binaire. Enfin, lorsque  $n>2$ , la relation est dite n-aire ou qualifiée d'événement. Les relations binaires entre entités, sont exprimées sous la forme suivante :  $relation = ((e_{tête}, e_{queue}), R)$ , où  $e_{tête}$  et  $e_{queue}$  sont des mentions d'entités respectivement appelées "tête" et "queue" de la relation. La plupart des systèmes d'extraction de relation se concentrent sur le cas particulier des relations binaires (Zhang et al., 2017). De plus, dans la plupart des langages de représentation d'expression d'ontologies, seules des relations sémantiques orientées et binaires sont utilisées.

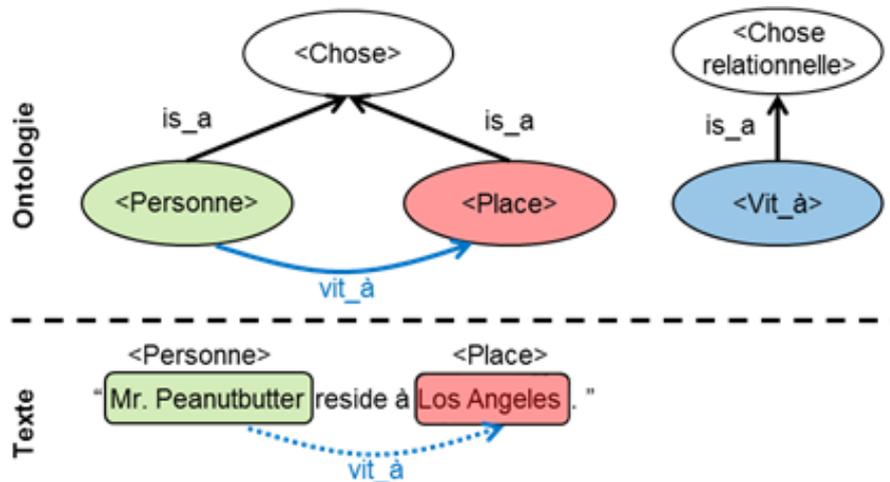


Figure 14 : illustration d'une extraction de relation

En pratique, des relations orientées n-aires sont à considérer pour représenter certaines informations, c'est-à-dire des relations avec plusieurs arguments, ou autrement dit reliant plusieurs mentions d'entité. Dans ce cas, le terme utilisé est plutôt celui d'événement que celui de relation. Un événement peut être défini de manière structurée comme tel :  $événement = ((arg_1, \dots, arg_N), référence)$ , où les  $arg_i$  sont des mentions d'entités et  $référence$  est un type de relation de référence. Néanmoins, il est possible de réduire un événement à plusieurs relations binaires (Xu et al., 2006) au prix d'une augmentation importante de la taille de la représentation et par la création de nombreuses relations artificielles. Les relations binaires restent majoritairement étudiées et utilisées en extraction de relations et en représentation de connaissances.

L'extraction de relation (ER) est définie comme la tâche de détection de relations sémantiques entre des entités, préalablement extraites par une annotation manuelle, une

reconnaissance ou une normalisation d'entités (voir Figure 14). La détection de mentions de relation est également envisageable, mais semble encore rarement abordée (Jiang, 2012). Dans la plupart des cas, la qualité de l'extraction d'entités a donc directement un impact sur la qualité de l'ER.

### 2.3.5. La normalisation d'entités nommées

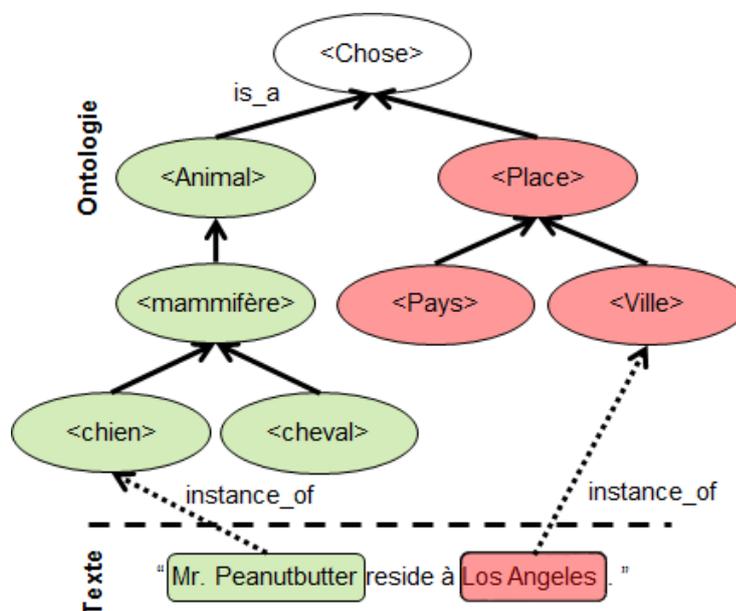


Figure 15 : Illustration d'une normalisation d'entités

La Normalisation d'Entités (NE), appelée de manière équivalente "*liage d'entité*" ("*entity linking*" en anglais) dans le domaine général, est principalement présentée comme l'étape suivant la reconnaissance d'entité. Là où la reconnaissance d'entité vise à extraire des mentions d'entités, puis à les relier à des références sémantiques de haut niveau d'abstraction (type), la NE va relier les mentions d'entités d'intérêt identifiées aux références sémantiques les plus proches des entités, parmi les références disponibles. Par exemple, si une tâche de reconnaissance d'entité extrait les entités qui sont des personnes, alors la tâche de NE reliera chaque entité identifiée comme telle à une référence qui représente cette personne précise (par exemple : "*président de la France*" ou "*M. Macron*" à **<Emmanuel Macron>**). L'ensemble des références sémantiques peut aller d'un modèle non-organisé (par exemple : une liste de personnalités avec identifiants uniques) à une ontologie (par exemple : une ontologie des espèces animales). Selon la liste de concepts de référence pour une tâche, un concept peut être plus ou moins abstrait. Par exemple, "*Mr. Peanutbutter*", un labrador anthropomorphe d'une série télévisée, doit être relié à **<chien>**, car **<chien>** ne possède aucun concept plus spécifique tel que **<labrador>** (voir Figure 15). L'extraction d'une entité par normalisation au moyen d'une ontologie consiste à produire une association (*donnée, concept*) telle que l'entité mentionnée par la *donnée* soit associée au *concept* de l'ontologie le plus précis.

Historiquement, les méthodes automatiques de NE ont émergé grâce aux besoins spécifiques d'automatisation de la maintenance de base de données de termes et d'indexation de document dans le domaine des sciences du vivant (Hirschman et al., 2005b; Lu et al., 2011). Cette maintenance est souvent désignée par terme anglais “*curation*”. La NE a reçu globalement moins d'attention que la reconnaissance d'entités depuis son apparition (Cohen, 2005), mais semble tout de même s'inscrire de plus en plus dans les défis internationaux en EI (Arighi et al., 2017; Ji et al., 2017). Néanmoins, elles sont souvent adossées à une tâche de reconnaissance d'entité, rendant parfois la détermination de la performance stricte des méthodes pour la normalisation moins évidente. De nombreuses méthodes, notamment celles de l'approche fondée sur l'appariement entre expressions d'un dictionnaire et expressions textuelles, effectuent conjointement les deux tâches.

Plus qu'une étape d'uniformisation des mentions d'une même entité, cette étape peut permettre de générer des informations supplémentaires. En effet, lorsque les concepts de référence appartiennent à une ressource plus organisée telle qu'une ontologie, l'association entité-concept donne accès à une possible définition du concept de référence, à des synonymes, à des relations avec d'autres concepts, etc. (Kang et al., 2013). Dans le cas de l'utilisation d'une ontologie, la NE se rapproche aussi de la tâche de Population de Base de Connaissances (PBC, ou “*Knowledge Base Population*”), car on peut aborder la tâche comme la création d'instances à partir de texte à classer dans le ou les concept(s) correct(s) de l'ontologie.

## 2.4. Difficultés de la normalisation d'entités

### 2.4.1. Introduction

Les chapitres précédents ont présenté la tâche de normalisation d'entité (NE). Celui-ci introduit à présent les principales difficultés rencontrées pour résoudre cette tâche.

Une grande partie des approches actuelles de NE, tout particulièrement dans le domaine biomédical (Aronson and Lang, 2010), s'appuient sur une similarité de forme, c'est-à-dire la comparaison entre les deux séquences de caractères que sont les mentions d'entités dans le texte et les expressions désignant les références sémantiques. Cette démarche semble en effet intuitive pour associer deux objets. Néanmoins, une entité peut être mentionnée de nombreuses façons différentes, entraînant fréquemment une différence de forme entre étiquettes et mention (Nenadic et al., 2005). Cette variabilité des expressions représente la principale difficulté de la NE.

A l'inverse, des expressions identiques ou similaires peuvent se référer à des entités différentes. Lorsque deux expressions identiques partagent cette propriété, elles sont

appelées homonymes. Les homonymes, par leur ambiguïté, représentent une difficulté supplémentaire à la NE. Si deux expressions homonymiques sont à normaliser par des concepts distincts, dans certains cas, une expression spécifique dans un texte peut devoir être normalisée également par des concepts distincts. C'est alors un problème de normalisation multiple. Dans certaines tâches de NE, notamment dans le domaine biomédical, les homonymes et les besoins en normalisation multiple sont plutôt rares. Néanmoins, dans le domaine général, ce problème est plus fréquent et donc plus abordé, notamment sous le nom de désambiguïsation du sens de mot ("*Word Sense Disambiguation*") (Ranjan Pal and Saha, 2015).

Enfin, de nouvelles approches de NE sont apparues plus récemment pour répondre à la difficulté que représente la variabilité des expressions. Elles sont principalement fondées sur des méthodes d'apprentissage supervisé, méthodes nécessitant des exemples annotés de mentions normalisées. Or, la quantité de ces données reste faible en NE et pour les domaines de spécialité.

## 2.4.2. Variabilité des formes

Cette section présente plusieurs phénomènes courants introduisant des variations de forme d'expressions.

Premièrement, dans les langues flexionnelles telles que le français ou l'anglais, chaque unité lexicale libre peut posséder des flexions selon son genre, sa nature, son mode, etc. Par exemple, les mots "*chevaux*" et "*chienné*" sont respectivement des flexions des unités lexicales libres "*cheval*" et "*chien*". Une flexion modifie la séquence de caractères d'un mot, et empêche donc une comparaison triviale entre la mention d'entité d'intérêt sous forme de flexion, et l'étiquette qui serait le même mot, sous forme canonique. Dans les langues dotées de lemmatiseurs, cette difficulté semble aujourd'hui dépassée : une étape de lemmatisation des expressions permet alors de réduire les flexions à leur forme canonique.

Les synonymes sont des expressions possédant le même sens, ou un sens relativement proche, mais qui sont décrites par des séquences de caractères différentes. Par exemple, les expressions "*laverie*", "*station de lavage*" et "*buanderie*" sont des synonymes et doivent être normalisés par la même référence. La NE pour les synonymes peut être améliorée si les références possèdent plusieurs étiquettes, idéalement un pour chaque forme canonique utilisée pour mentionner une entité associée à ces références. La construction d'un ensemble de références définies de manière plus exhaustive nécessite néanmoins un effort manuel plus important de la part d'experts d'un domaine (Ushold and King, 1995). De plus, la disparité du vocabulaire utilisé par différents experts pour exprimer les mêmes entités peut être importante (Furnas et al., 1987), rendant la conception d'un ensemble d'expressions de références plus complexe. Enfin, de nouvelles expressions apparaissent régulièrement dans une langue naturelle (Gerlach and Altmann, 2013). Tout cela a pour

conséquence que les lexiques d'un domaine sont incomplets (Klavans and Muresan, 2001).

Au-delà des problèmes de synonymie, une référence peut représenter un concept instanciable par plusieurs entités différentes et dont les mentions ne seront donc pas des synonymes. Par exemple, si la référence est <chien>, les mots hyponymes et non-synonymes tels que “*labrador*” ou “*doberman*”, ne possèdent pas de similarité de forme qui permettrait de les associer. Tenter de référencer tous les hyponymes pour une référence d'intérêt demande alors un effort important, effort à réitérer pour chaque référence, et ne fait que se heurter aux difficultés précédentes.

Enfin, les expressions d'intérêt en NE, ainsi que les étiquettes des références, sont fréquemment des expressions multi-mots, régulièrement étendues par des adjectifs, adverbes, prépositions, propositions relatives, etc. En plus des possibles variations morphosyntaxiques (telle que l'inversion) que peut subir une expression, cela complexifie la similarité entre expressions d'intérêt et expressions de référence. Par exemple, l'expression “le petit chien de mon voisin qui aboie tout le temps” devrait être normalisée par <chien>. Cela complique la comparaison entre deux expressions par rapport à une comparaison entre deux mots simples. L'étude de la structure syntaxique de ces expressions peut alors avoir une importance sémantique. Par exemple, pour différencier “*chocolat au lait*” et “*lait au chocolat*”, relier “*pancreatic cancer*” à “*cancer of the pancreas*”, détecter les têtes sémantiques des expressions, etc. Les expressions figées et les mots composés soulèvent une difficulté supplémentaire car l'étude des unités lexicales libres qui les composent ne permettent pas toujours d'exprimer le sens de l'expression entière (ex : “*pomme de terre*”).

Pour ces raisons, il est fréquent qu'une partie des mentions d'entité dans un texte ne possède aucune forme similaire à une étiquette de référence. Les approches de NE par similarité de forme se limitent donc fréquemment à la normalisation d'une sous-partie des données textuelles et ne peuvent à elles seules proposer des normalisations pertinentes pour le reste.

### 2.4.3. Normalisation multiple

Selon les objectifs d'une tâche de NE, il est nécessaire qu'une expression soit normalisée par plusieurs entités. Par exemple, une tâche pourrait consister à normaliser les expressions se référant à la fois aux concepts <mammifère> et <animal domestique>. Dans ce cas, l'expression “*le labrador de mon voisin*” devrait être normalisée par ces deux concepts. Pour ne pas avoir à considérer cette difficulté, l'ontologie de référence pourrait être complétée en créant des nouveaux concepts pour toutes les conjonctions possibles de concepts. En reprenant l'exemple précédent, cela signifierait de créer le concept <mammifère domestique>. Néanmoins, avec des ontologies volumineuses contenant peu de concepts disjoints, cela augmente considérablement leur taille et rend difficile leur

manipulation. Les approches classiques produisent souvent un score par possible association mention/concept, laissant la possibilité de sélectionner plusieurs concepts pour une même mention (Aronson, 2001; Leaman et al., 2013). Néanmoins, ces méthodes choisissent seulement le concept rapportant le plus haut score pour une mention donnée. Ce problème semble encore peu abordé dans le cas de la normalisation d'entité, notamment parce que ce phénomène reste peu fréquent.

#### 2.4.4. Expression homonymique

Deux homonymes peuvent être vus comme deux expressions à la forme identique mais se référant à des entités distinctes. Par exemple, les expressions des informations (“*avocat*”, <fruit>) et (“*avocat*”, <métier>) sont des homonymes. Des expressions homonymiques peuvent donc se référer à plusieurs entités distinctes, ce qui crée une ambiguïté. Lever une telle ambiguïté est rendu possible en prenant en compte le contexte de l'expression. Les approches de NE qui n'utilisent pas le contexte des expressions qu'elles cherchent à normaliser ont donc des difficultés à aborder ce problème d'ambiguïté.

Tout ou partie de plusieurs étiquettes de concepts peuvent également être des homonymes. Par exemple, le mot anglais “*furrow*” peut à la fois désigner un gène et une partie anatomique d'une mouche (“*morphogenetic furrow*”) (Hanisch et al., 2005). Dans l'ontologie OntoBiotope (Bossy et al., 2016), on trouve le mot “*plant*” dans plusieurs étiquettes : (“*plant*”, <living organism>) et (“*sewage plant*”, <industrial building>). Dans ce cas, c'est la position du concept étiqueté dans le graphe ontologique qui permet de désambigüiser le sens. Les approches de NE qui ne s'appuient que sur les similarités de forme des étiquettes ne peuvent donc résoudre ce problème.

#### 2.4.5. Insuffisance de données annotées

Produire des exemples annotés manuellement demande des efforts importants (Uschold and King, 1995), tout particulièrement pour les domaines de spécialité où le niveau d'expertise des annotateurs doit être élevé. De plus, les tâches de NE possèdent un nombre de concepts de références important par rapport aux tâches de reconnaissance d'entité. Par exemple, l'ontologie de référence pour la tâche de NE de Bacteria Biotope (Deléger et al., 2016), OntoBiotope, contient plus de 2320 concepts. D'autres ensembles de concepts bien plus volumineux sont utilisés, tels que les plus de 275 000 concepts définis dans le “*Medical Subject Headings*” (Lipscomb, 2000) ou les quelques 5 millions du méta-thésaurus biomédical “*Unified Medical Language System*” (UMLS)<sup>2</sup>. En conséquence, l'ensemble des exemples annotés ne porte que sur une sous-partie des concepts de référence pour la tâche. Par exemple, les 747 exemples annotés d'habitats bactériens pour la tâche de NE de Bacteria Biotope n'utilisent que 15% des concepts de l'ontologie. En comparaison avec le domaine général, la tâche de reconnaissance d'entité nommée de

---

<sup>2</sup> <https://www.nlm.nih.gov/research/umls/>

CoNLL en (Sang and De Meulder, 2003) proposait au minimum 3400 exemples annotés dans un jeu de données d’entraînement pour chacun des quatre types d’entités recherchées.

Cette insuffisance des exemples annotés ne présente pas une réelle difficulté pour les approches de NE fondées sur la similarité de forme entre mention et étiquette. À l’inverse, elle en représente une importante pour les approches par apprentissage dont les performances augmentent avec le nombre et la qualité des données d’entraînement. Le problème de classification par apprentissage supervisé pour lequel certaines classes n’ont aucun exemple annoté représente un défi important (Larochelle et al., 2008). Nous présenterons plus en détail ces deux familles d’approches, par étude de la similarité de forme et par apprentissage, dans le prochain chapitre.

## 2.5. Tâche de normalisation Bacteria Biotope

### 2.5.1. Introduction

Pour évaluer la performance d’une méthode de normalisation, il est aujourd’hui courant d’évaluer directement ses prédictions sur une ou plusieurs tâches de normalisation comportant leurs propres objectifs et jeux de données. De plus, cela permet de comparer rapidement plusieurs méthodes distinctes, et de dégager ainsi un état de l’art pour une tâche particulière. Nous présentons dans cette section la tâche de normalisation que nous avons utilisée pour évaluer et comparer les méthodes des travaux de cette thèse : la tâche de normalisation d’habitat bactérien Bacteria Biotope du challenge BioLP Shared Task 2016.

Pour acquérir une meilleure vision des difficultés linguistiques de cette tâche, ainsi que pour participer à la mise en place d’un challenge d’extraction d’information, j’ai participé à l’annotation des documents. Le travail de réalisation du challenge BioNLP Shared Task a notamment conduit à une publication (Deléger et al., 2016) pour le workshop BioNLP Shared Task 2016 adossé à la conférence “*Association for Computational Linguistics*”.

## 2.5.2. Présentation de Bacteria Biotope 2016

La première édition du challenge BioNLP Shared Task qui a introduit la tâche Bacteria Biotope (Deléger et al., 2016) s'est déroulée en 2011, suivie par deux autres éditions en 2013 et 2016. La tâche Bacteria Biotope est une tâche d'extraction d'information visant à extraire les mentions de bactéries et d'habitats bactériens (biotopes et lieux géographiques), ainsi que les possibles relations qui expriment la présence d'une bactérie vivante dans un habitat bactérien (relation normée nommée "*Lives\_In*"). Les documents sont des titres et des résumés d'articles scientifiques disponibles sur PubMed<sup>3</sup>, ou des pages web grand public sur les bactéries, tels que MicrobeWiki<sup>4</sup>. Les références sémantiques pour la normalisation d'entités sont issues de sources de connaissances : NCBI Taxonomy<sup>5</sup> pour les mentions de bactéries et l'ontologie OntoBiotope (Bossy et al., 2015; Bossy et al., 2016) les habitats bactériens.

En 2016, la tâche Bacteria Biotope est divisée en trois sous-tâches, chacune déclinée en deux versions alternatives, avec ou sans extraction d'entité d'intérêt. Les règles d'annotation sont publiquement consultables<sup>6</sup>. Pour toutes les sous-tâches, deux jeux de données annotées sont fournis (un corpus d'entraînement et un corpus de développement). Un jeu de données non-annotées est également fourni, et sert à évaluer la performance des méthodes grâce à un système d'évaluation en ligne<sup>7</sup>. Les trois sous-tâches sont :

- Tâche de normalisation d'entités : extraire les mentions de bactéries et d'habitats bactériens dans le texte en les reliant à des ontologies du domaine.
- Tâche d'extraction de la relation : relier les mentions de bactéries à des mentions d'habitats bactériens (par la relation <Lives\_In>) et de lieux géographiques lorsque le texte exprime explicitement que les bactéries mentionnées étaient vivantes dans l'habitat mentionné.
- Tâche d'extraction de base de connaissance : extraire les mentions de bactéries et d'habitats, les associer aux concepts corrects dans les ressources fournies et extraire les possibles relations <Lives\_In> entre mentions. L'objectif est de créer une base de connaissances répertoriant la présence de tel taxon bactérien dans tel habitat normé. Cette dernière sous-tâche peut être vue comme l'union des deux premières, et est une réponse directe à un besoin de la communauté de biologistes.

---

3 <https://www.nlm.nih.gov/bsd/pubmed.html>

4 <https://microbewiki.kenyon.edu/index.php/MicrobeWiki>

5 <https://www.ncbi.nlm.nih.gov/books/NBK21100/>

6 <http://2016.bionlp-st.org/tasks/bb2/guidelines>

7 <http://bibliome.jouy.inra.fr/demo/BioNLP-ST-2016-Evaluation/index.html>

### 2.5.3. Intérêt scientifique de la tâche

Les informations décrivant des bactéries et leurs habitats sont d'un grand intérêt applicatif dans l'agro-alimentaire, les sciences de la santé, le traitement des déchets, etc. Ces informations sont également importantes pour la recherche fondamentale en microbiologie (métagénomique, phylogéographie, phylogéographie, phyloécologie, etc.). Néanmoins, il n'y a actuellement aucune ressource permettant de centraliser les informations relatives aux relations entre les bactéries et leurs habitats dans un vocabulaire standardisé et structuré. Une grande partie de ces connaissances est dispersée dans des articles scientifiques ou des bases de données contenant des champs avec des données textuelles (par exemple, GenBank<sup>8</sup>, GOLD<sup>9</sup>, DSMZ<sup>10</sup>, GBIF<sup>11</sup>). La plupart des informations sur les biotopes bactériens ne sont donc décrites qu'en langue naturelle. De plus, les biotopes sont des entités très diverses et pouvant être décrites par des mentions variées.

Utiliser à grande échelle, une extraction d'informations performante effectuée sur des documents scientifiques et champs textuels de base de données pourrait alors permettre de collecter ces informations et de remplir automatiquement des bases de connaissances (Bossy et al., 2012). Néanmoins, les besoins vont au-delà de la reconnaissance d'entités et de leurs possibles relations : pour permettre une intégration et une comparaison des informations d'intérêt, il est important de pouvoir utiliser un vocabulaire commun et donc de normaliser les entités décrites dans les données textuelles (Ivanova et al., 2010; Pignatelli et al., 2009; Buttigieg et al., 2013). De telles méthodes d'EI permettraient donc de répondre aux besoins importants des communautés mentionnées précédemment.

Enfin, du point de vue des organisateurs de la tâche, la mise à disposition libre<sup>12</sup>, de toutes les ressources nécessaires permet de mettre en place un environnement adapté à l'évaluation et à la comparaison des méthodes en EI.

### 2.5.4. Tâche de normalisation d'habitat bactérien

L'objectif de la tâche de normalisation d'entités de Bacteria Biotope est de normaliser des mentions d'habitats et de bactéries respectivement avec les concepts de l'ontologie OntoBiotope et les entrées de la base de données NCBI Taxonomy. Nous nous intéressons ici plus particulièrement à la sous-tâche "*BB-cal*" pour laquelle les exemples donnés sont des mentions déjà identifiées comme étant des habitats bactériens.

---

8 Site de GenBank : <https://www.ncbi.nlm.nih.gov/genbank/>

9 Site de GOLD : <https://gold.jgi.doe.gov/>

10 Site de DSMZ : <https://www.dsmz.de/>

11 Site de GBIF : <https://www.gbif.org/>

12 <http://2016.bionlp-st.org/tasks/bb2/>

Pour cette tâche, trois corpus distincts sont fournis (voir Tableau 1) :

- Un corpus pour l'entraînement des méthodes contenant des exemples annotés.
- Un corpus dit de développement, contenant des exemples annotés, et permettant principalement de paramétrer sa méthode et d'analyser les erreurs qu'elle produit.
- Un corpus de test ne contenant que les mentions d'habitats. L'évaluation finale est faite sur une prédiction sur ce corpus grâce à un site d'évaluation en ligne<sup>13</sup>.

	entraînement	développement	test	total
<b>documents</b>	71	36	54	161
<b>mots</b>	16295	8890	13797	38982
<b>mentions d'habitats</b>	747	454	720	1921
<b>mentions d'habitats distinctes</b>	476	267	478	1125
<b>concepts d'habitat mentionnés</b>	825	535	861	2221
<b>concepts d'habitat mentionnés distincts</b>	210	122	177	329

Tableau 1 : Description des données fournies pour la tâche de normalisation d'habitats bactériens "BB-cat"

La précédente édition de *BioNLP Shared Task* précisait que les mentions d'habitats à normaliser sont peu concernées par les problèmes de normalisation multiple et d'homonyme (Bossy et al., 2013), de l'ordre de moins de 3% des cas. En conséquence, les méthodes qui n'abordent pas ces problèmes sont peu pénalisées.

La tâche est évaluée avec une mesure non-stricte : il y a bien un (ou plusieurs) concept(s) correct(s) à prédire pour chaque mention, mais un concept prédit différent ne donnera pas pour autant un score nul. Ce score est en effet proportionnel à une distance sémantique entre le concept correct et celui prédit. L'utilisation de ce score plutôt qu'un score strict est justifiée par les objectifs applicatifs de la tâche : les méthodes d'EI font des erreurs et la mise à jour manuelle des informations extraites est moins coûteuse si la correction est locale. Par exemple, il est plus simple de corriger ("*labrador*", <animal>) par ("*labrador*", <chien>), que de corriger ("*labrador*", <bacterial habitat>). De plus, d'un point de vue sémantique, il semble abrupt de pénaliser de la même façon des erreurs de différents degrés, ce qui est le cas avec les évaluations strictes.

<sup>13</sup> <http://bibliome.jouy.inra.fr/demo/BioNLP-ST-2016-Evaluation/index.html>

La mesure utilisée est celle proposée par (Wang et al., 2007). Pour un concept  $A$ , soit  $T_A$  l'ensemble des concepts-ancêtres de  $A$  ( $A$  inclus), soit  $c$  un concept appartenant à  $T_A$ ,  $d$  le nombre minimal d'arêtes séparant  $A$  et  $c$  dans le graphe ontologique (relation  $\langle is\_a \rangle$ ), et  $w$  un paramètre appelé facteur de contribution sémantique, alors on calcule la valeur suivante :

$$S_A(c) = w^d$$

Pour calculer un score entre un concept prédit  $A$  et le concept correct  $B$ , l'intersection des ensembles  $T_A$  et  $T_B$  est calculée. Enfin, la mesure finale est calculée par :

$$score(A, B) = \frac{\sum_{c \in T_A \cap T_B} (S_A(c) + S_B(c))}{(\sum_{c \in T_A} S_A(c)) + (\sum_{c \in T_B} S_B(c))} \quad (1)$$

Si le concept prédit est le concept correct, alors la valeur du score est 1. Dans les autres cas, la valeur est non-nulle et strictement inférieure à 1. Pour essayer de donner une idée de l'effet de ce score : plus il faut remonter vers la racine pour trouver un concept-ancêtre commun à deux concepts, plus ce score est bas. Le choix d'un facteur de contribution sémantique sert à pénaliser plus ou moins certains concepts. Par exemple, une valeur de 0,8 semble favoriser les prédictions de concepts-frères par rapport aux prédictions de concepts-ancêtres ou descendants. Des valeurs trop faibles ont tendance à donner un poids important à la prédiction du concept-racine. Pour la tâche Bacteria Biotope, le choix a été fait de favoriser les concepts-ancêtres, car cela reste une prédiction vraie (un "labrador" est un  $\langle chien \rangle$ , mais reste aussi un  $\langle animal \rangle$ ), tout en évitant de donner trop d'importance à la racine. Pour cela, la valeur de 0,65 a été expérimentalement choisie (Bossy et al., 2013).

Un score global est calculé en faisant la moyenne de tous les scores pour chaque mention à évaluer, rapportant une valeur strictement supérieure à 0 et inférieure ou égale à 1. Avec le paramètre de contribution sémantique à 0,65, pour la tâche de normalisation d'habitat bactérien, le score naïf obtenu en prédisant le concept racine pour chaque mention à évaluer est alors 0,32.

Cette tâche de normalisation d'habitats bactériens et cette mesure d'évaluation sont celles principalement utilisées pour évaluer les méthodes présentées dans ce manuscrit.

## 2.6. Bilan

Nous avons vu dans ce chapitre que la tâche de normalisation consistait à relier une donnée textuelle (des mentions d'entités) à des références sémantiques, telles que les concepts d'une ontologie. Plus qu'une simple liste de concepts étiquetés, l'ontologie est une représentation formelle de connaissances, qui, par sa structure, enrichit chacun de ses concepts d'informations supplémentaires utiles à exploiter.

En reliant des expressions, mentions d'entité, à des concepts précis, la normalisation ancre les informations d'un texte dans une représentation formelle externe. Ces informations sont d'un grand intérêt pour certaines communautés. Néanmoins, la tâche de normalisation nécessite que les mentions d'entités d'intérêt pour la tâche soient préalablement reconnues, c'est-à-dire que les séquences de mots composant ces mentions aient été détectées et annotées comme étant d'intérêt. Ces données sont fournies sous la forme d'exemples annotés. Ces pré-annotations à la normalisation peuvent être produites par :

- Des annotateurs humains, ce qui permet d'obtenir des annotations de qualité, mais au prix d'un effort important.
- L'application d'une méthode de reconnaissance d'entité, ce qui fournit des annotations de plus faible qualité, mais pour un coût restreint de préparation des données.

Cette tâche présente de nombreuses difficultés, dont deux principales :

- Les différences de forme entre les mentions et les étiquettes des concepts qui devraient les normaliser. Cette différence est fréquente, ce qui empêche la résolution de la normalisation par la seule utilisation d'information morphologique et morphosyntaxique.
- L'insuffisance due au nombre d'exemples annotés manuellement, c'est-à-dire des ensembles de mentions dont le concept qui le normalise est déjà identifié. Cette insuffisance est une difficulté particulièrement sensible pour une famille d'approches de normalisation : les approches par apprentissage automatique.

Enfin, nous avons présenté la tâche de normalisation d'habitats bactériens Bateria Biotope. Cette tâche de normalisation nous donne un cadre pour évaluer nos méthodes et les comparer à l'état de l'art.

## **Chapitre 3**

-

### **Normalisation d'entités fondée sur une ontologie : état de l'art**



### 3.1. Introduction

L'extraction d'information fondée sur une ontologie ou une base de connaissances ('*Ontology-Based Information Extraction*' - OBIE) semble émerger dès la fin des années 90 des questions de construction automatique d'ontologie à partir de texte (Faure and Nédellec, 1998; Hwang, 1999). La différence principale entre des systèmes OBIE et d'autres systèmes d'extraction d'information réside dans le fait que pour les systèmes OBIE, la même ontologie est utilisée à la fois comme ressource d'entrée et comme référence ciblée (Li and Bontcheva, 2007). Une tâche d'OBIE peut aussi être vue comme une tâche de peuplement d'ontologie ou base de connaissances (Declerck et al., 2008; Vargas-Vera et al., 2001) et est équivalente à la tâche de liaison d'entité ("*entity linking*") (Hachey et al., 2013; Derczynski et al., 2015). Ces deux derniers termes sont plutôt utilisés pour les tâches en domaine général. Dans (Wimalasuriya and Dejing Dou, 2010), la définition suivante d'un système OBIE est donnée : un système OBIE est un système qui traite un texte en langage naturel à l'aide d'un mécanisme guidé par des ontologies pour extraire certains types d'informations et présenter le résultat en utilisant des ontologies.

La tâche de normalisation peut être vue comme un problème de classification : l'objectif est de prédire en sortie une ou plusieurs classes en fonction d'observations en entrée. Dans le cas de la normalisation, les entrées sont des expressions textuelles représentant des mentions d'entités, possiblement contextualisées. Une étape préalable consiste à déterminer les expressions textuelles candidates, par exemple, les groupes nominaux, ou les noms simples. De plus, dans le cadre de la normalisation par les concepts d'une ontologie, les classes à prédire sont ces concepts.

L'ensemble des méthodes de normalisation passent par un calcul, implicite ou explicite, d'un score de similarité pour toutes les paires de mentions d'entité à normaliser et de concepts de référence. Les mentions peuvent être représentées par leur séquence de caractères ou par des vecteurs. Les concepts peuvent être directement représentés par des vecteurs, ou plus communément par les séquences de caractères ou les vecteurs de leurs étiquettes. Soit  $S = \{(x_i, y_j)\}_{i,j}$  l'ensemble des associations possibles de représentations de mentions et de représentations de concepts. Soit la fonction *score* définie par :

$$score: S \rightarrow \mathbb{R} \\ (x, y) \mapsto score(x, y) \quad (2)$$

Pour chaque mention à normaliser, les concepts prédits seront ceux dont les représentations obtiendront les meilleurs scores avec la représentation de la mention.

Ce chapitre présente un état de l'art sur les deux principales familles d'approches, qui se différencient selon les représentations de mentions et de concepts utilisées :

- Les approches fondées sur la similarité de forme entre expressions d'un dictionnaire et expressions textuelles : ces approches utilisent directement les

formes des mentions et des étiquettes de concepts, c'est-à-dire leur séquence de caractères.

- Les approches fondées sur des représentations sémantiques. Une représentation sémantique est une représentation vectorielle d'une expression ou d'un concept qui intègre de l'information relative à leur sens.

## **3.2. Normalisation fondée sur la similarité de forme entre expressions d'un dictionnaire et expressions textuelles**

### **3.2.1. Introduction**

Les concepts de l'ontologie possèdent une ou plusieurs étiquettes. L'ensemble des étiquettes forme un dictionnaire. Ces étiquettes sont généralement des expressions textuelles désignant des termes du domaine. Les mentions d'entités dans un texte sont également des expressions textuelles. Une première approche de la normalisation consiste alors à rechercher directement les mentions de même forme que les étiquettes de concepts.

Néanmoins, cette approche par appariement exact est souvent très limitée. Pour autant, des similarités de formes sont fréquentes entre mentions et étiquettes de concepts. Cette section présente un état de l'art de ces méthodes fondées sur la similarité de forme.

### **3.2.2. Limitations des appariements exacts**

Les méthodes par appariement exact cherchent à normaliser des mentions figées par des étiquettes figées, c'est-à-dire que la moindre variation de caractères entre mention et expression suffit à empêcher une association. On peut modéliser ce type d'approche par un score binaire, qui est égal à 1 si les séquences de caractères d'une mention et d'une étiquette sont les mêmes, et à 0 sinon.

Les mentions, comme celles de gènes ou de protéines, sont souvent composées de plusieurs mots. Or, la présence, l'ordre et la spécificité de ces mots au sein d'une mention varient fréquemment, notamment avec les noms contenus dans les bases de données de référence du domaine (Hanisch et al., 2002). Dans ce cas, mention et étiquette ne seront pas identiques.

Pour la tâche de normalisation de BioNLP Shared Task 2013, après une lemmatisation des mentions et des étiquettes de concepts, 60% des mentions d'entité avaient une forme différente de toutes les étiquettes des concepts qui devaient les normaliser (Bossy et al., 2015). Pourtant, l'ontologie de référence OntoBiotope a été développée manuellement à partir de mentions rencontrées dans des textes du même type que ceux de la tâche, ce qui n'est pas le cas de toutes les ontologies. Cela illustre l'inefficacité des méthodes par appariements exacts.

Dans le cas d'expressions multimots, une première amélioration possible est alors de considérer chaque expression comme un sac-de-mots (Wei and Kao, 2011), c'est-à-dire comme un ensemble de plusieurs mots plutôt qu'une séquence ordonnée. A la place de chercher des appariements exacts entre deux séquences, un appariement partiel peut alors être recherché en étudiant les mots identiques et en commun entre ces deux séquences. Par exemple, la méthode FiGO (Couto et al., 2005) se fonde sur l'analyse des mots composant les étiquettes de concepts et les mentions d'entités. La méthode commence par décomposer les étiquettes de concepts en mots, puis à étudier la distribution de chaque mot dans l'ensemble des étiquettes. L'hypothèse est qu'un mot fréquent dans plusieurs concepts est moins important, moins clivant, pour représenter un concept spécifique. Un score est alors calculé à partir des appariements exacts entre les mots des mentions et des étiquettes, en prenant en compte l'importance du mot. Les résultats de la méthode sur plusieurs tâches de normalisation ont néanmoins montré une précision relativement faible (inférieure à 30%).

### 3.2.3. Présentation générale

Les approches fondées sur la similarité de forme s'appuient fréquemment sur des règles linguistiques et des appariements partiels entre mention et étiquette. On trouve fréquemment les expressions anglaises "*rule-based*" et "*dictionary look-up*" pour désigner ces approches. La majorité de ces méthodes peuvent être modélisées par quatre étapes principales (voir Figure 16) :

- 1) Enrichir le dictionnaire de la référence : L'objectif est d'enrichir chaque concept de nouvelles étiquettes. Pour cela, des expressions synonymes, co-hyponymes, acronymes/formes longues, etc. des étiquettes ou des concepts existants sont ajoutés. Cela se fait principalement à partir d'autres ressources terminologiques. Des expressions alternatives peuvent également être ajoutées en faisant varier les formes des étiquettes existantes. Des règles peuvent par exemple être appliquées pour produire des variations d'étiquettes existantes limitées à quelques insertions/délétions de caractères. Dans (Ghiasvand and Kate, 2014), l'observation a été faite que de nombreuses mentions ne différaient en effet d'une étiquette existante qu'à quelques variations de caractères près (d'ordre typographique notamment). En augmentant ainsi le nombre d'étiquettes pour chaque concept, cela permet d'augmenter la probabilité d'un appariement partiel, éventuellement pondéré, entre variations d'étiquettes et variations de mentions. Une approche alternative est de représenter chaque expression par une forme

canonique, par exemple par sa forme lemmatisée ou racinisée. Cela permet de représenter plusieurs expressions initialement différentes, mais supposées partager un sens proche, par une seule. Par exemple, en lemmatisant les mots “*chienné*” et “*chiens*”, ils seront directement appariés avec le mot “*chien*”, étiquette d’un concept <chien>. Nous faisons le choix de classer cette approche dans celle par enrichissement du dictionnaire, car une forme canonique représente à elle seule plusieurs variations. Par exemple, au lieu d’enrichir <chien> par “*chiens*”, lemmatiser toutes les mentions permettra d’apparier directement “*chiens*”, mais également “*chienné*”, etc. Cela augmente donc également les probabilités d’obtenir des appariements exacts.

- 2) Appariement partiel : Un appariement partiel est ensuite opéré. Toutes les associations mention/étiquette qui ne possèdent aucun mot en commun sont exclues (c’est-à-dire qu’elles obtiennent un score nul).
- 3) Filtrer les appariements partiels ambigus : L’étape d’appariement rapporte souvent plusieurs étiquettes pour chaque mention, dont une part importante reste ambiguë. Il est donc souvent nécessaire de procéder à une étape de désambiguïsation, principalement effectuée en supprimant des appariements fondés sur des mots estimés comme étant ambigus par des règles manuelles ou automatiques. L’objectif de cette étape est d’augmenter la précision en limitant les faux positifs.
- 4) Calculer un score d’alignement partiel : Un score d’alignement est alors calculé pour chaque association d’une variation de mention avec une variation d’étiquette, selon différentes métriques selon la méthode. Par exemple, un simple score calculé à partir du nombre de mots en commun entre deux expressions peut être utilisées, permettant ainsi de sélectionner correctement pour la mention “*Human immunodeficiency virus 1*” les étiquettes de concepts différents “*Human*” et “*Human immunodeficiency virus*”. L’association possédant alors le meilleur score permet de déterminer le concept prédit pour une mention.

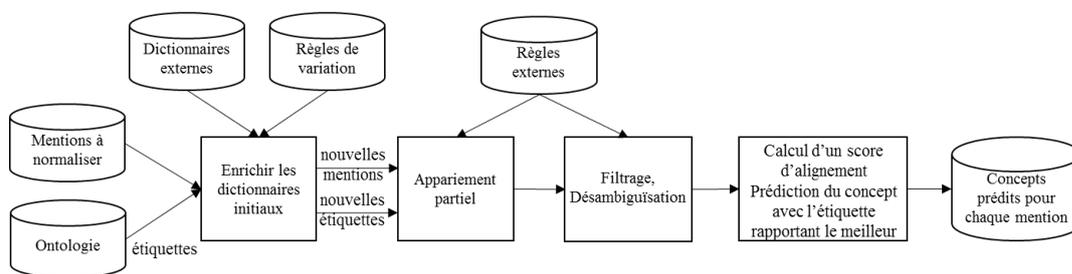


Figure 16 : Schéma général des méthodes fondées sur l’étude de la similarité de forme entre étiquettes de concepts et mentions d’entité extraite.

### 3.2.4. État de l'art

Pour la première étape, plusieurs méthodes commencent donc par augmenter le nombre d'étiquettes pour chaque concept de la référence. Par exemple, la méthode ProMiner (Hanisch et al., 2005; Fluck et al., 2007) commence par générer des nouvelles expressions en :

- Développant certains mots (par exemple, “*a*” est développé en “*alpha*”)
- Développant tous les acronymes sous leur forme longue (par exemple, “*IL*” est développé en “*Interleukin*”), lorsque cela est disponible dans le dictionnaire de référence
- Développant tous les mots contenant des nombres en enlevant ou rajoutant un espace entre séquence de lettres et séquence de chiffres (par exemple, “*Igf1*” est développé en “*Igf 1*”)

La méthode décrite dans (Grouin, 2016) commence par une tentative d'appariement exact entre mention et étiquette, puis, si aucun appariement n'est rapporté, essaye un appariement partiel en générant des flexions pour chaque expression, ainsi que d'autres variations linguistiques générées manuellement.

La méthode Peregrine (Schuemie et al., 2007), après une étape de segmentation en mots des mentions et des étiquettes, ainsi qu'une suppression des mots-outils, effectue une racinisation sur toutes les expressions (au moins à partir d'un certain nombre de caractères). La racinisation ne génère pas plusieurs variations pour une même expression, mais permet de représenter plusieurs variations possibles sous la forme d'une seule forme canonique. Des variations sont alors produites sur chaque séquence racinisée avec les deux règles suivantes :

- Les nombres romains sont remplacés par leur équivalent arabe, et inversement.
- Si un mot se termine par un nombre, un caractère-délimiteur est ajouté (par exemple un espace ou un tiret), et inversement.

La méthode ToMap (Golik et al., 2011) représente chaque expression par sa tête syntaxique. L'hypothèse étant que la tête syntaxique est fréquemment la partie la plus informative d'une expression (au moins dans le cas des habitats bactériens). Une liste de mots évalués comme non-informatifs est néanmoins produite manuellement pour ne pas prendre en compte certaines têtes syntaxiques (par exemple : “*sample*”, “*environment*”, ...).

La méthode de (Tsuruoka et al., 2007) ainsi que UWM (Ghiasvand and Kate, 2014) sont entraînées à prédire de nouvelles règles permettant de produire des variations à partir d'une expression. Notamment, UWM s'appuie sur des suites d'opérations permettant de modéliser le passage d'une chaîne de caractères à une autre, s'appuyant sur une distance de Levenshtein (Levenshtein, 1966). Ces opérations peuvent être des insertions, des substitutions, des délétions ou des conservations. A partir de données d'entraînement (c'est-à-dire associations expressions / étiquettes) et des étiquettes synonymes des concepts de l'ontologie, un ensemble de suites d'opérations sont produites. Ces suites

d'opérations sont chacune spécifiques à une unique association d'expression. La méthode les généralise donc afin d'apprendre les opérations permettant d'obtenir les variations générales à partir d'un mot. Par exemple, à partir des suites permettant de passer de "cyanotic" à "cyanosis" et de "thrombotic" à "thrombosis", la méthode apprend d'une certaine façon le modèle de passage d'une forme adjectivale à une forme nominale des adjectifs se terminant par la chaîne de caractères "otic". Le modèle est dans ce cas de détecter une chaîne de caractères terminale "otic" dans un mot, puis remplacer le 't' et le 'c' par un 's' pour produire de nouvelles variations.

Au cœur de cette étape d'augmentation du nombre de variations pour chaque étiquette de la référence, pouvant être réalisée par des méthodes dédiées (Golik et al., 2013; Aubin and Hamon, 2006), de nombreuses méthodes se concentrent parfois prioritairement sur l'intégration de nouvelles étiquettes en provenance d'autres ressources du domaine. La méthode MetaMap, qui est un des premiers systèmes de normalisation, génère plusieurs variations, pour les mentions seulement, en remplaçant successivement chaque mot de l'expression initiale par un autre fourni par le lexique orienté vers le domaine biomédical SPECIALIST (Browne et al., 2000). Par exemple, l'adjectif "ocular" dans le groupe nominal "ocular complications" aurait pu être remplacé par "eye", "eyes", "optic", "ophthalmic", "oculus", etc. Ce qui produirait alors plusieurs nouvelles séquences de mots telles que "eye complications", "optic complications", etc.

La méthode ProMiner suit une démarche similaire en intégrant les étiquettes issues des bases de données de gènes Entrez Gene (Maglott, 2004) et UniProt (Bairoch, 2004) pour la tâche de normalisation de mentions de gènes et de protéines du challenge Biocreative (Hirschman et al., 2005a). Les résultats de la méthode montrent directement une amélioration des résultats grâce à l'utilisation de dictionnaires externes.

L'outil BIOADI (Kuo et al., 2009) permet de prédire des associations abréviation/forme longue à partir d'un ensemble d'associations d'entraînement. C'est une méthode par apprentissage automatique supervisé, qui s'appuie sur des représentations vectorielles fondées sur des caractéristiques (voir 2.2.4). Elle a été utilisée notamment sur PubMed pour produire un dictionnaire d'associations abréviation/forme longue pour le domaine biomédical. Lors du challenge Biocreative V (Wei et al., 2015), (Lee et al., 2015) ont proposé la méthode avec les meilleurs résultats en s'appuyant principalement sur ce dictionnaire. Cette méthode a utilisé également un filtrage des mots-outils.

LINNAEUS (Gerner et al., 2010) est une méthode spécialisée dans la normalisation de mentions d'organismes vivants, qui s'appuie sur la taxonomie du NCBI (Federhen, 2011) et sur un dictionnaire construit spécifiquement pour cette tâche.

La méthode de (Lee et al., 2015) s'appuie de même sur MEDIC (Davis et al., 2012) et sur le NCBI *disease corpus* (Doğan et al., 2014) pour la normalisation de mentions de maladies lors du challenge Biocreative V.

Toutes les variations produites permettent d'augmenter la probabilité d'une similarité de forme entre les mentions d'entité et les étiquettes de la référence. Néanmoins, cette

augmentation introduit de nouvelles ambiguïtés, notamment des étiquettes similaires mais de concepts différents. Par exemple, dans les améliorations récentes du système MetaMap, la recherche de variations n'est pas effectuée pour les mots d'un ou deux caractères, car elle menait presque toujours à de mauvaises prédictions (Aronson and Lang, 2010). Pour chaque variation d'une mention, MetaMap calcule un score d'alignement entre la variation et toutes les étiquettes partageant au moins un mot exactement apparié. Cela fournit donc également une ensemble d'étiquettes par mention, ordonné selon leur score. Le ou les concepts ayant une étiquette obtenant le meilleur score représente(nt) alors la ou les prédiction(s) de normalisation d'une mention.

(Hanisch et al., 2005) ont développé leur méthode ProMiner en s'attaquant prioritairement au problème des ambiguïtés. En effet, les mentions des gènes d'organismes différents sont parfois identiques ou partiellement identiques, tout comme certaines mentions de gènes différents du même organisme. Certaines autres mentions ne diffèrent que par leur casse. Un premier tri est effectué pour supprimer de la référence des mots connus comme ambigus. À l'inverse, des mots peuvent y être ajoutés s'ils sont connus comme étant d'intérêt. Une seconde étape consiste à classer chaque mot du dictionnaire obtenu dans des catégories sémantiques prédéfinies qui sont censées refléter le degré d'ambiguïté d'un mot. Par exemple, en s'appuyant sur la fréquence d'apparition des mots dans un corpus du domaine et appartenant au dictionnaire, les mots les plus fréquents se verront classer dans une catégorie "synonyme non-spécifique". Les mots de cette catégorie auront moins de poids lors du calcul du score d'alignement final. L'hypothèse sous-jacente est qu'il est peu probable qu'un mot fréquent ne soit utilisé que pour définir uniquement un nom de protéine. Lors du challenge BioCreative II (Morgan et al., 2008), les 250 étiquettes les plus fréquemment rencontrées dans un corpus externe ont été manuellement étudiées, et selon l'appréciation de leur ambiguïté par les auteurs, supprimées, afin d'améliorer les performances de la méthode Peregrine. Toutes les étiquettes produites lors de la première étape sont également filtrées automatiquement selon trois règles :

- Si une étiquette contient moins de trois caractères,
- Si une étiquette ne contient que des nombres,
- Si une étiquette ne contient que des mots-outils.

Enfin, si une étiquette se réfère à plusieurs concepts, elle est considérée comme ambiguë. Lors de l'appariement, si l'étiquette appariée est ambiguë, il faut qu'il y ait une autre mention du même concept dans le même document pour comptabiliser cette prédiction.

La méthode de (Lee et al., 2015) s'appuie quant à elle sur une hiérarchisation des dictionnaires (ceux fournis par la tâche ou intégrés manuellement par les participants) pour désambiguïser. S'il y a plusieurs appariements d'une mention avec des étiquettes de plusieurs dictionnaires, celui avec la plus haute priorité décidera du concept à retenir.

Lors d'un appariement entre têtes sémantiques d'une mention et de plusieurs étiquettes, la méthode ToMap utilisera prioritairement l'appariement des têtes syntaxiques secondaires des expressions pour calculer un score d'alignement. Le concept ayant l'étiquette avec le plus de têtes secondaires en commun avec la mention sera celui prédit.

Enfin, pour traiter les ambiguïtés, (D'Souza and Ng, 2015) organisent plusieurs groupes de règles en fonction de leur niveau de précision estimée. Par exemple, le premier groupe recherche un appariement exact entre mention et étiquette. Le second groupe commence par développer les abréviations constituant une mention, puis recherche un appariement exact entre mention développée et étiquette. Le huitième groupe racinise les mentions avant de rechercher un appariement exact avec une étiquette. L'hypothèse est que les appariements rapportés par les règles du  $i$ -ème groupe sont plus précis que ceux rapportés par le  $(i+1)$ -ème groupe. La méthode commence donc par essayer d'apparier des mentions et des étiquettes avec le premier groupe de règles, puis avec le second, etc. Lors de l'analyse par un groupe, si une mention est appariée de façon non-ambigüe, c'est-à-dire s'il n'y qu'un unique appariement de la mention avec une étiquette d'un seul concept, alors ce concept est choisi pour prédire la normalisation de cette mention. Si une mention ne peut être normalisée de façon non-ambigüe (c'est-à-dire si le groupe propose plus d'un concept pour normaliser la mention), alors aucune prédiction n'est faite. Dès qu'une mention est normalisée, elle est ajoutée comme étiquette de son concept, permettant de rendre cette méthode itérative.

### 3.2.5. Limitations

Bien que ces approches fondées sur l'appariement avec un dictionnaire et des règles peut permettre d'atteindre une excellente précision sur un domaine restreint et des entités à annotées spécifiques, elles possèdent certaines limitations.

S'appuyant fréquemment sur l'utilisation de ressources externes pour augmenter le nombre de variations des mentions et/ou des étiquettes de concepts, elles sont notamment particulièrement spécialisées dans la normalisation d'entités spécifiques d'un domaine (principalement, la normalisation de mentions de gènes/protéines, d'organismes ou de maladies). Plusieurs autres règles viennent parfois accentuer cette spécialisation. Par exemple, pour MetaMap, certains mots des étiquettes de l'UMLS ont été manuellement listées pour ne pas être pris en compte et ainsi limiter les ambiguïtés. De même pour ToMap, dont l'adaptation au domaine et à la tâche (via sa liste des mots non-informatifs pour la reconnaissance d'habitat bactérien) limitent le potentiel d'adaptation de la méthode. Ce problème de sur-spécialisation apparaît donc comme une faiblesse de ce genre d'approche.

Enfin, la plupart des échecs de ces méthodes, c'est-à-dire les mentions qui ne finissent pas normalisées, sont dus à l'absence d'étiquettes avec une nouvelle forme, telles que des synonymes ou des hyponymes (Pratt and Yetisgen-Yildiz, 2003). Ces limitations ont rendu rapidement d'intérêt les approches par apprentissage automatique mettant en avant leur adaptation. Au point où certains travaux relativement récents inscrivent en perspective cette volonté de changement d'approche (Lee et al., 2015).

### 3.2.6. Évaluation de méthodes

Méthode	Score de similarité
ToMap (avec règles spécifiques au domaine)	0,66
ToMap	0,61
Méthode de référence	0,54
(Grouin, 2016)	0,44

Tableau 2 : Résultats des plusieurs méthodes fondées sur la similarité de forme sur la tâche de normalisation d'habitats bactériens de *Bacteria Biotope*. Méthode de référence = lemmatisation et appariement exact.

Lors du challenge BioNLP Shared Task 2016 (Delèger et al., 2016), la seule participation d'une méthode par étude de la similarité de forme, qui ne se fondait pas sur des représentations sémantiques, était celle de (Grouin, 2016). La méthode a obtenu des résultats légèrement en dessous de la méthode de référence, qui effectue une recherche des appariements exacts après lemmatisation. Nous avons pu évaluer deux versions différentes de ToMap sur la tâche de normalisation également. La méthode ToMap possède deux versions dont une plus récente et adaptée à la tâche par intégration manuelle de règles supplémentaires. Les résultats sont présentés ci-dessous dans le Tableau 2.

Des méthodes fondées sur de l'apprentissage supervisé et s'attaquant simultanément à la reconnaissance et à la normalisation émergent ces dernières années (Leaman et al., 2015; Lee et al., 2015), mais le problème de la normalisation reste pour la majorité des cas approché par une approche fondée sur la similarité de forme entre mentions et étiquettes.

## 3.3. Construction d'espaces sémantiques

### 3.3.1. Introduction

Nous venons de présenter un état de l'art des méthodes de NE fondées sur l'étude des formes des expressions, ainsi que leurs limitations, notamment dues à la variabilité importante des formes des mentions par rapport aux étiquettes de concepts. De nouvelles méthodes sont apparues plus récemment pour dépasser ces limitations en s'appuyant sur des représentations sémantiques d'expressions simples ou multi-mots. Ces représentations sont des vecteurs représentant chacun une des expressions considérées dans un même espace vectoriel réel référentiel. L'objectif de ces vecteurs d'expression est de capturer une partie du sens des entités auxquelles elles se réfèrent. Notamment, des

expressions de forme différentes mais se référant à des entités sémantiquement proches doivent posséder des vecteurs spatialement proches. Cette propriété permet alors de dépasser la simple représentation morphologique des expressions. Nous présentons ici un état de l'art sur les méthodes permettant de construire ces espaces.

### 3.3.2. Représentations creuses et non-sémantiques : les représentations 1-parmi-N

Les représentations 1-parmi-N (*“One-Hot”*) ou représentent des objets dans un espace vectoriel réel de la taille du nombre d'objets (par exemple, la taille du vocabulaire). Chaque objet est associé à un vecteur dans une dimension unique de l'espace. Pour représenter un objet particulier, tous les coefficients du vecteur sont à zéro, sauf celui de sa propre dimension. Si l'on considère  $T = \{t_i\}_{i=1}^{|T|}$ , l'ensemble des objets considérés, alors chaque composante  $(t_i)_j$  de la représentation vectorielle de l'objet est définie dans l'espace  $\mathbb{R}^{|T|}$  telle que :

$$(t_i)_j = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases} \quad (3)$$

Les représentations 1-parmi-N ne forment pas d'espaces sémantiques. Au contraire, ce sont des représentations vectorielles d'objets (mots, expressions, concept, etc.) qui n'introduisent aucune similarité privilégiée entre les objets initiaux considérés. Autrement dit, les objets représentés sont tous à la même distance les uns des autres (selon les distances usuelles).

En TAL, les représentations 1-parmi-N sont souvent d'une taille importante, par exemple pour représenter des mots elles sont de la taille du vocabulaire. Elles sont également très creuses, puisque tous les coefficients des vecteurs sont à 0 sauf un. En conséquence, elles peuvent facilement poser des difficultés d'utilisation pour des calculs trop complexes. Néanmoins, leur représentation numérique facilite leur exploitation par des algorithmes d'apprentissage. Elles sont notamment utilisées comme représentations initiales dans certains calculs de représentations sémantiques (Mikolov et al., 2013a).

### 3.3.3. Des représentations creuses et sémantiques : les sacs-de-mots TF-IDF

En recherche d'information, TF-IDF (*term frequency-inverse document frequency*) est une mesure de l'importance d'un mot dans un document compris dans un ensemble de documents. Une mesure de TF est une représentation du nombre d'occurrences d'un mot : plus un même mot est fréquent dans un document, plus ce mot devrait être important. Néanmoins, si ce mot est également fréquent dans l'ensemble des documents (par exemple, les mots-

outils), son importance est moindre, car il ne permet pas de discriminer le document. Une mesure IDF est une représentation de cette fréquence d'apparition globale d'un mot dans un corpus (Jones, 1972). La pondération d'une mesure TF par une mesure IDF forme une mesure TF-IDF d'un mot dans un document appartenant à un corpus. Il existe plusieurs variantes de mesures TF et IDF. Classiquement, le TF binaire est égal au nombre d'occurrences de mots dans le document considéré. De même, pour un ensemble  $D$  de documents, et son sous-ensemble  $\{d_j: t_i \in d_j\}$  de documents dans lequel le terme  $t_i$  apparaît, la mesure classique d'un IDF est donnée par la formule suivante :

$$idf(t_i)_j = \log\left(\frac{|D|}{|\{d_j: t_i \in d_j\}|}\right) \quad (4)$$

Enfin, la mesure TF-IDF s'obtient alors en faisant le produit :  $tfidf(t_i)_j = tf(t_i) \cdot idf(t_i)_j$

Cette mesure TF-IDF initiale peut être complexifiée et certaines variations ont été adaptées à la tâche de normalisation en considérant l'étiquette d'un concept comme un document. (Claveau, 2013) a notamment proposé une mesure qui différencie deux mesures TF différentes :

- Une qui est corrélée au nombre d'occurrences d'un mot dans une mention
- Une qui est corrélée au nombre d'occurrences d'un mot dans une étiquette

Enfin, une mesure IDF est calculée, laquelle est corrélée au nombre d'occurrences d'un mot dans l'ensemble des étiquettes de tous les concepts de l'ontologie. Un score de normalisation d'une mention est alors calculé en sommant les produits des trois mesures précédentes pour tous les mots de la mention avec l'étiquette d'un concept candidat. Pour chaque mention, un score est alors calculé pour chaque étiquette de l'ensemble des concepts de l'ontologie. Le concept ayant l'étiquette qui obtient le score le plus élevé est le concept prédit par la méthode. La mesure proposée, fondée sur les formes des expressions et sur leur nombre d'occurrences ne permet pas de dépasser le problème de la haute variabilité de forme entre mention et étiquette.

La mesure TF-IDF peut également être utilisée pour construire des représentations vectorielles d'expressions (Manning et al., 2009). Plutôt que de calculer un poids binaire pour chaque dimension d'un vecteur, une mesure TF-IDF est utilisée en considérant l'expression comme un document : pour chaque dimension d'un vecteur d'une expression, un TF est calculé pour le mot courant par rapport à l'expression considérée, puis celui-ci est pondéré par un IDF du mot par rapport à l'ensemble des expressions considérées.

Soit  $T$ , l'ensemble des mots du vocabulaire. Les représentations vectorielles des expressions de  $T$  sont alors initialisées dans l'espace  $\mathbb{R}^{|T|}$ , chaque dimension étant associée à un élément unique de  $T$ . Par exemple, si les expressions sont "*le petit chien de mon voisin*" et "*le petit chat*". Le vocabulaire est alors :

$$\{\text{"le"}, \text{"petit"}, \text{"chien"}, \text{"de"}, \text{"mon"}, \text{"voisin"}, \text{"chat"}\}$$

Le vocabulaire est constitué de sept mots. Les vecteurs d'expression appartiennent donc à  $\mathbb{R}^7$ . Chaque dimension de ces vecteurs est associée à l'un des mots du vocabulaire. Alors leurs représentations vectorielles en TF-IDF seraient :

$$\begin{aligned} \text{“le petit chien de mon voisin”} &= [\log(\frac{7}{2}), \log(\frac{7}{2}), \log(7), \log(7), \log(7), \log(7), 0] \\ \text{“le petit chat”} &= [\log(\frac{7}{2}), \log(\frac{7}{2}), 0, 0, 0, 0, \log(7)] \end{aligned} \quad (5)$$

On observe alors que les mots “le” et “petit”, bien qu'ils soient communs aux deux expressions, ont un poids plus faible. Dans l'espace sémantique construit, plus des vecteurs d'expressions partagent des mots en commun et discriminants (c'est-à-dire des mots qui n'apparaissent pas fréquemment dans les expressions), plus ils seront proches. C'est donc une représentation vectorielle qui à elle seule ne permet pas de dépasser le problème de la haute variabilité de forme entre mention et étiquette.

### 3.3.4. Des représentations creuses et non-fondées sur la forme : les sacs-de-mots distributionnels

Les représentations 1-parmi-N ne permettent pas de représenter des similarités entre des objets. Les représentations TF-IDF ne permettent d'en représenter que lorsque des expressions partagent des mots en commun. Les méthodes de construction de représentations distributionnelles de mots ne sont pas récentes (Hinton et al., 1986; Pollack, 1990; Deerwester et al., 1990; Elman, 1991) et permettent de représenter des proximités sémantiques entre des mots de forme différente. Ce sont des représentations vectorielles de mots construites à partir de l'étude de la distribution des mots dans des corpus. Ces méthodes s'appuient sur l'hypothèse de sémantique distributionnelle (Harris, 1954) : les objets qui présentent des distributions similaires partagent des similarités sémantiques.

Ces méthodes se basent sur une matrice de cooccurrences dans un corpus, c'est-à-dire une matrice contenant le nombre d'occurrences d'expressions dans des contextes définis. Chaque ligne représente une expression et chaque colonne un contexte. Les premières utilisations de cette approche utilisait des documents entiers comme contexte de mot (Deerwester et al., 1990). Aujourd'hui, dans la majorité des cas, les contextes utilisés sont les mots du vocabulaire (Zweigenbaum and Habert, 2006), mais peuvent être des expressions multi-mots, des phrases, des paragraphes, des documents, etc.

Par exemple, considérons les phrases suivantes : “Le petit chien de mon voisin aboie.” et “Le petit chat miaule.”. En ne considérant pas les mots-outils, le vocabulaire est alors : {“petit”, “chien”, “voisin”, “aboie”, “chat”, “miaule”}. Si nous considérons les mots comme objet à représenter, et que les contextes d'un mot sont définis par les autres mots dans une phrase, nous obtenons la matrice de cooccurrences carrée de la Figure 17.

contextes : - mots :	“petit”	“chien”	“voisin”	“aboie”	“chat”	“miaulé”
“petit”	0	1	1	1	1	0
“chien”	1	0	1	1	0	0
“voisin”	1	1	0	1	0	0
“aboie”	1	1	1	0	0	0
“chat”	1	0	0	0	0	1
“miaulé”	1	0	0	0	1	0

Figure 17 : Exemple d'une matrice de cooccurrences à partir du corpus : "Le petit chien de mon voisin aboie. Le petit chat miaule.", filtré de ses mots-outils et en prenant en compte l'ensemble de la phrase comme contexte. Chaque ligne de cette matrice permet alors de représenter un mot par un vecteur.

Si une matrice de cooccurrences est construite sur un corpus de grande taille, l'hypothèse de sémantique distributionnelle impliquant que des mots partageant un sens proche auront une distribution en contexte proche, alors les vecteurs obtenus pour ces mots devraient être à une distance relativement proche.

Les principaux inconvénients de cette approche initiale :

- La taille des vecteurs reste importante (de l'ordre de la taille du vocabulaire).
- Les vecteurs restent creux, car de nombreux mots n'apparaissent jamais ensemble.
- En essayant d'utiliser directement des expressions multi-mots comme cibles et/ou comme contextes, cela tend à aggraver les deux inconvénients précédents, car dans un même corpus, des expressions multi-mots ont globalement moins d'occurrences que les mots simples.

Une des premières approches pour dépasser ces inconvénients est de réduire le nombre d'objets considérés tout en augmentant leur signal : le nombre de dimensions est alors réduit et les cooccurrences plus nombreuses (la matrice est donc un peu moins creuse). Par exemple, le corpus peut être filtré de ses mots-outils, puis lemmatisé.

Enfin, des représentations continues et de faible dimension peuvent être construites à partir de cette matrice de cooccurrences. L'approche principale étant de réduire la matrice initiale. La solution la plus classique est d'appliquer une décomposition en valeurs singulières (Jurafsky and Martin, 2014; Manning et al., 2009). Cette méthode décompose une matrice de cooccurrences en un produit de trois matrices particulières, dont une matrice centrale creuse contenant des valeurs particulières appelées valeurs singulières de la matrice de cooccurrences. En supprimant des valeurs singulières, et en multipliant la matrice centrale respectivement par celle de gauche ou celle de droite, on obtient une

nouvelle matrice contenant des représentations denses et de plus faible dimension, respectivement des mots ou des contextes. Le défaut principal de cette approche étant que les représentations produites, calculées uniquement sur les statistiques globales du corpus, ne réussissent pas à capturer certains sens tels que les analogies (Pennington et al., 2014). La méthode GloVe (Pennington et al., 2014) produit des représentations différentes appelées plongements lexicaux, mais utilise la matrice de cooccurrences. L'idée de GloVe était de prendre le meilleur des informations globales (telles que celles utilisées pour les sacs-de-mots distributionnels) et locales (telles que celles utilisées pour les plongements lexicaux). Néanmoins, la qualité des représentations a souvent été observée comme moindre par rapport à celle des plongements lexicaux (Muneeb et al., 2015).

### 3.3.5. Des représentations continues et de faible dimension : les plongements lexicaux

De nouvelles méthodes que l'on nomme plongements lexicaux ("*word embeddings*") se sont plus récemment popularisées avec l'intégration de réseaux neuronaux computationnellement plus utilisables (Mikolov et al., 2013a). Ces dernières méthodes permettent de produire des représentations distributionnelles denses et de dimension relativement faible (majoritairement entre 100 et 1000) sans passer par le calcul d'une matrice de cooccurrences.

Tout comme les représentations distributionnelles en sac-de-mots, la qualité principale des plongements lexicaux réside dans leur capacité à représenter des vecteurs de mots de formes différentes (notamment des synonymes) spatialement proches. Ces représentations permettent également d'intégrer de nombreuses autres caractéristiques linguistiques. Notamment, certaines relations entre entités sont parfois exprimées dans l'espace sémantique produit par des translations linéaires (Mikolov et al., 2013c). Par exemple, il suffit de prendre le vecteur résultant du calcul suivant :  $\text{vecteur}(\text{"Londres"}) - \text{vecteur}(\text{"Royaume - Uni"})$ , et d'appliquer le vecteur obtenu à  $\text{vecteur}(\text{"France"})$ . Le vecteur final se retrouve alors à proximité du vecteur du mot "*Paris*", ce qui indique qu'une certaine relation sémantique <a pour capitale> est représentée par une translation linéaire constante dans l'espace.

(Mikolov et al., 2013a) propose initialement deux méthodes par apprentissage non-supervisé pour calculer des plongements lexicaux : CBOV (pour "*Continuous Bag-Of-Words*") et Skip-Gram. CBOV s'appuie sur un réseau à une seule couche cachée, lequel est entraîné à prédire un mot en sortie à partir des mots de son contexte en entrée. Chaque mot est utilisé sous sa version 1-parmi-N. Le contexte utilisé est représenté par les mots contenus dans une fenêtre limitée à la phrase et symétrique autour du mot que l'on cherche à représenter. À l'inverse, Skip-Gram est entraîné à prédire les mots du contexte en fonction du mot à représenter. En entraînant le réseau au fur-et-à-mesure, les deux systèmes vont apprendre une matrice de paramètres permettant d'optimiser l'ensemble des prédictions. Cette matrice possède une colonne par mot du vocabulaire et son nombre

de lignes est égale au nombre de neurones dans la couche cachée, qui est un paramètre libre. La valeur optimale de ce paramètre varie selon les expériences, mais son ordre de grandeur reste principalement entre 100 et 1000 (Mikolov et al., 2013a), ce qui reste bien inférieur à l'ordre de grandeur des vocabulaires.

Une amélioration de l'architecture Skip-Gram a été proposée pour diminuer la complexité des calculs : l'échantillonnage négatif ("*negative sampling*"). Skip-Gram essaye d'optimiser les probabilités de prédire des mots de contexte à partir d'un mot à représenter. Pour chaque mot à représenter durant l'entraînement, le système doit recalculer un score de similarité entre le mot cible et un mot de contexte, et cela pour tous les mots du contexte. La version avec échantillonnage négatif, Skip-Gram NS (Mikolov et al., 2013b) choisit seulement quelques mots de contexte semi-aléatoirement, ce qui permet de diminuer le temps de calcul.

Quelle que soit l'architecture de Word2Vec utilisée, (Chiu et al., 2016) a montré que plusieurs paramètres peuvent avoir un impact significatif sur la qualité des représentations produites. De plus, l'architecture Skip-Gram avec échantillonnage négatif semble également produire de meilleures représentations que l'architecture CBOW pour le domaine biomédical.

D'autres méthodes ont depuis fait leur apparition, telles que FastText (Bojanowski et al., 2016). Cette méthode s'appuie sur la même architecture que Skip-Gram mais en calculant des représentations pour des n-grammes de caractères plutôt que pour des mots. Cette approche permet notamment de répondre au problème des mots hors-vocabulaire qui n'avaient initialement pas de représentation. Pour cela, une représentation d'un mot hors-vocabulaire peut être calculée à partir des représentations des n-grammes qui le compose et qui eux ont certainement été rencontrés lors de l'entraînement.

Enfin, des méthodes récentes, ELMo (Peters et al., 2018) ou BERT (Devlin et al., 2018), ont permis non plus de représenter des mots, mais des mots dans leur contexte par des plongements lexicaux. Selon le mot et son contexte, à partir d'un espace sémantique initial, ces méthodes produisent dynamiquement une représentation dense et de faible dimension. Cette utilisation du contexte permet de répondre notamment aux problèmes d'homonymes (voir 2.4.4). Nous ne considérerons pas dans ces travaux ces dernières méthodes, car celles-ci ne construisent pas d'espaces sémantiques dans lequel chaque expression est représentée par un vecteur. Or, cette propriété est importante pour notre approche.

### 3.3.6. Compositionnalité et représentation vectorielle d'expression

Les méthodes classiques de construction de plongements lexicaux utilisent les fréquences des mots. La question de la construction de plongements lexicaux pour des expressions

composées de plusieurs mots n'est donc pas triviale. Le problème de compositionnalité a été modélisé par (Mitchell and Lapata, 2010). Soient  $v_{m1}$  et  $v_{m2}$ , des vecteurs représentant respectivement les mots  $m1$  et  $m2$  composant une expression. Une fonction  $f$  de compositionnalité peut alors être définie telle que :

$$f: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$v_{m1}, v_{m2} \mapsto f(v_{m1}, v_{m2}) \quad (6)$$

Quel que soient les mots  $m1$  et  $m2$ , on attend de  $f(v_{m1}, v_{m2})$  qu'elle représente l'expression composée de  $m1$  et  $m2$  (c'est-à-dire " $m1 m2$ "). Les fonctions  $f$  les plus communément utilisées sont la somme et la multiplication, avec une certaine efficacité (Mitchell and Lapata, 2010), ainsi que la moyenne :

$$f: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$v_{m1}, v_{m2} \mapsto f(v_{m1}, v_{m2}) = \frac{v_{m1} + v_{m2}}{2} \quad (7)$$

Laquelle peut être généralisée pour n'importe quelle expression ( $m_1, \dots, m_N$ ) ( $N \geq 2$ ) composées de  $N$  mots par la formule suivante :

$$f: (\mathbb{R}^n)^N \rightarrow \mathbb{R}^n$$

$$v_{m1}, \dots, v_{mN} \mapsto f(v_{m1}, \dots, v_{mN}) = \frac{1}{N} * \sum_{k=1}^N v_{m_k} \quad (8)$$

Lorsque la similarité cosinus est utilisée, il n'y a pas de différence entre l'utilisation de l'addition ou de la moyenne, puisque :

$$\cos\left(v_1, \frac{v_{m1} + v_{m2}}{2}\right) = \frac{2}{|v_1| * |v_{m1} + v_{m2}|} * v_1 \cdot \left(\frac{1}{2} * (v_{m1} + v_{m2})\right) \quad (9)$$

$$\Rightarrow \cos(v_1, v_2) = \cos(v_1, v_{m1} + v_{m2})$$

Cette méthode reste néanmoins insensible à l'ordre des mots, et plus généralement à la structure syntaxique, donnant la même représentation à toutes les constructions qui partagent le même vocabulaire. Elle a néanmoins montré son efficacité sur plusieurs tâches de TAL (Landauer and Dumais, 1997; Foltz et al., 1998; Kintsch, 2001).

### 3.3.7. Adaptation de plongements lexicaux à des connaissances externes

Historiquement, les méthodes de construction de plongements lexicaux généraux les plus performantes et les plus utilisées à ce jour sont apparues à partir de 2013 (Mikolov et al., 2013a; Pennington et al., 2014). Les premiers travaux à se pencher sur l'intérêt d'adapter

des plongements lexicaux à des connaissances externes ont ensuite émergé à partir de 2014 avec (Yu and Dredze, 2014).

La construction de plongements lexicaux étant fondée sur l'hypothèse de sémantique distributionnelle, ceux-ci sont sujets aux limitations de cette approche. Par exemple, en raison de la similarité des contextes, les vecteurs d'antonymes (“*weak*” et “*strong*”) se retrouvent fréquemment à proximité (Ono et al., 2015). Or, la discrimination d'antonymes est un besoin fréquent. L'apport de connaissances spécifiques à la tâche et au domaine, telles que les expressions représentant des antonymes, pourrait donc permettre de modifier les plongements lexicaux de manière à augmenter ou diminuer certaines similarités et ainsi améliorer les plongements lexicaux pour cette tâche ou ce domaine en les adaptant au besoin.

Une autre limitation des méthodes de production des plongements lexicaux est qu'elles nécessitent d'être entraînées sur de grandes quantités de textes. Pour un domaine de spécialité et une tâche précise, il est difficile d'obtenir un corpus composé uniquement de textes du même domaine et d'un même style de langue. Un compromis est donc souvent à faire entre la taille et l'homogénéité, vis-à-vis de la tâche et du corpus d'entraînement choisi (Chiu et al., 2016). Les plongements vont alors encoder des informations hors-domaine qui n'ont potentiellement aucun intérêt pour le contexte applicatif. Autrement dit, il y aura des dimensions sémantiques qui seront utilisées pour représenter des informations peu pertinentes. Par exemple, dans des textes du domaine général, les mots « fraise » et « tomate » seront représentés de façon relativement éloignée, car leurs contextes sont plutôt différents : l'un est plutôt assimilé aux fruits sucrés, aux confiseries, etc., alors que l'autre, bien qu'étant un fruit, est plutôt assimilé aux légumes, à la cuisine salée, etc. Pour une tâche de détection des entités représentant des fruits, cette distinction fondée principalement sur l'utilisation culinaire, ne semble pas pertinente. L'utilisation de connaissances sur le domaine étudié peut non seulement permettre d'éliminer les dimensions non-pertinentes, mais aussi permettre d'utiliser l'espace vectoriel à disposition pour améliorer l'expressivité des plongements lexicaux.

Enfin, les plongements lexicaux sont identifiés par les formes de surface des mots du vocabulaire. Un mot donné ne possède donc qu'une seule représentation vectorielle. Ils sont inaptes à différencier des mots polysémiques et auront donc des difficultés à les désambiguïser. Par exemple, le mot français « culture » peut avoir le sens d'une exploitation agricole, ou d'un élevage de cellules biologiques, ou encore d'un ensemble de traits sociologiques partagé par un groupe d'individus. Si le corpus initial est relativement général, voire hors-domaine, un mot polysémique ne sera majoritairement pas ou peu utilisé pour son ou ses sens d'intérêt pour le domaine d'intérêt. La connaissance du contexte applicatif, notamment en domaine de spécialité, peut donc permettre de corriger certains plongements lexicaux de mots polysémiques.

Pour toutes ces raisons, des plongements lexicaux adaptés à des connaissances spécifiques à un domaine et à une tâche ont un réel potentiel pour améliorer les performances de méthodes d'extraction d'informations.

Il existe deux grandes catégories de méthodes de construction de plongements lexicaux exploitant des connaissances externes : les méthodes dites “conjointes” (Yu and Dredze, 2014; Kiela et al., 2015; Ono et al., 2015; Nguyen et al., 2016; Liu et al., 2015) et celles dites “post-traitement” (Faruqui et al., 2014; Wieting et al., 2015; Mrkšić et al., 2016; Mrkšić et al., 2017). Les connaissances externes principalement utilisées dans les évaluations sont des relations de similarité entre des mots (synonymes, co-hyponymes, ...), ou de dissimilarité (antonymes, ...) telles que celles que l'on peut trouver dans WordNet dans le graphe de mots dont les arêtes expriment la synonymie (Miller, 1995). Malgré cette catégorisation, ces méthodes possèdent des principes relativement proches.

Les méthodes conjointes prennent en compte des connaissances externes en même temps qu'elles construisent les plongements. La plupart de ces méthodes s'appuie sur une adaptation des méthodes neuronales initiales, à quelques exceptions près (Osborne et al., 2016). Dans ce cas majoritaire, l'intégration des connaissances se fait au niveau de la fonction objectif du réseau neuronal. Au terme distributionnel classique (noté  $terme_{distributionnel}$ ), la méthode ajoute un terme supplémentaire permettant de prendre en compte ces connaissances (noté  $terme_{connaissance}$ ). Elles possèdent donc toutes une fonction objectif  $L$  de la forme suivante :

$$L = terme_{distributionnel} + terme_{connaissance} \quad (10)$$

Ces méthodes permettent de calculer directement des plongements lexicaux adaptés à des connaissances externes, mais ne permettent pas d'adapter des plongements existants.

Les méthodes par post-traitement prennent un ensemble de plongements lexicaux, appartenant au même espace vectoriel en entrée, et le modifient en prenant en compte des connaissances externes. La majorité des méthodes existantes se fondent sur une fonction objectif possédant un terme (noté  $terme_{connaissance}$ ) visant à rapprocher ou éloigner les plongements initiaux selon des connaissances externes (par exemple : rapprocher des plongements de synonymes, ou éloigner des antonymes), ainsi qu'un terme visant à préserver les plongements initiaux (noté  $terme_{préservation}$ ). Elles possèdent donc toutes une fonction objectif de la forme suivante :

$$L = terme_{préservation} + terme_{connaissance} \quad (11)$$

L'avantage principal des méthodes post-traitement est qu'elles sont applicables à tout type de plongements lexicaux initiaux en entrée.

## 3.4. Normalisation fondée sur des représentations sémantiques

### 3.4.1. Introduction

Nous avons présenté dans la section précédente un état de l'art des méthodes de construction de représentations sémantiques, c'est-à-dire des représentations vectorielles d'expressions simples ou multi-mots capables de capturer une partie du sens des entités désignées par ces expressions. Selon le sens capturé, ces représentations permettent de manipuler des expressions sans être totalement dépendantes de leur forme. Leur utilisation en normalisation laisse donc envisager un dépassement des limitations des approches fondées sur la seule étude de la similarité de forme. Cette section présente un état de l'art des méthodes fondées sur ce type de représentation.

### 3.4.2. Approche par calcul de scores entre représentations sémantiques

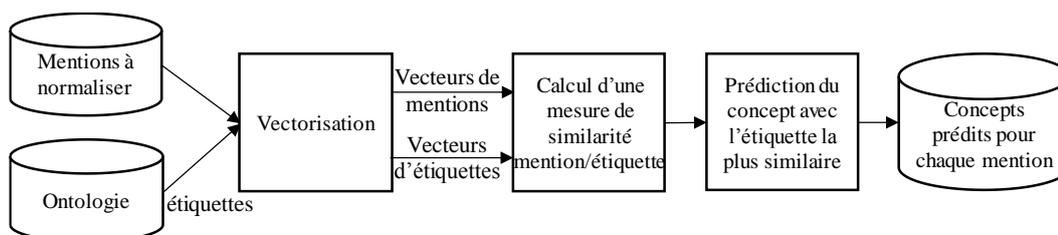


Figure 18 : Schéma de synthèse des méthodes de normalisation fondée sur des espaces sémantiques

Les méthodes de normalisation fondées sur des représentations vectorielles peuvent également s'appuyer sur la similarité de forme entre expressions d'un dictionnaire et expressions textuelles, à laquelle elles ajoutent de nouvelles caractéristiques sémantiques. Notamment, les représentations TF-IDF sont une façon de représenter sous forme vectorielle des informations par rapport aux formes de surface des mots qui les composent, mais pondérées par un poids que l'on peut considérer comme sémantique. Ces méthodes commencent par construire des représentations vectorielles pour chaque mention et pour chaque étiquette de concept. Dans l'espace sémantique produit, un score, fondée sur une mesure de similarité, est calculé entre chaque association d'une représentation de mention et d'étiquette. Le concept prédit est celui dont l'étiquette obtient le meilleur score pour une mention (voir Figure 18).

La méthode développée par l'équipe BOUN (Tiftikci et al., 2016) commence par essayer d'apparier exactement chaque mention avec une étiquette de l'ontologie. Si aucun concept

n'est trouvé de cette manière, la méthode s'appuie alors sur des représentations TF-IDF pour les mentions et les étiquettes de concepts. Soit  $T = \{mot_i\}_{i=1}^{|T|}$ , l'ensemble des mots du vocabulaire. Un score est alors calculé par similarité cosinus entre toutes les mentions et toutes les étiquettes, définit tel que :

$$score: \mathbb{R}^{|T|} \times \mathbb{R}^{|T|} \rightarrow \mathbb{R} \\ (x, y) \mapsto score(x, y) = \cos(x, y) = \frac{x^\top \cdot y}{\|x\| \cdot \|y\|} \quad (12)$$

Méthode	Score de similarité
Turku (Mehryary et al., 2017)	0,63
BOUN (Tiftikci et al., 2016)	0,62
Méthode de référence	0,54

Tableau 3 : Résultats de méthodes fondées sur un calcul de score entre représentations sémantiques de mentions et d'étiquettes. La méthode de référence effectue une lemmatisation, puis recherche les appariements exacts.

Le concept possédant l'étiquette rapportant le plus haut score est choisi pour normaliser chaque mention. La méthode développée par (Mehryary et al., 2017) se fonde également sur des représentations TF-IDF et n'intègre pas d'étape d'appariement exact. A la place d'utiliser des représentations TF-IDF fondées sur les mots, elle utilise des représentations fondées sur des segmentations en n-grammes de caractères (de taille un, deux et trois). Avant de produire ces représentations, une racinisation est effectuée sur les mots des mentions et des concepts. Enfin, le même score fondé sur la similarité cosinus est utilisé, et le concept possédant l'étiquette rapportant le plus haut score est également choisi pour normaliser chaque mention. L'équipe BOUN avait obtenu les meilleurs résultats durant l'édition 2016 de BioNLP Shared Task pour la tâche Bacteria Biotope. L'année suivante, la méthode de (Mehryary et al., 2017) avait dépassé légèrement ces résultats. Les résultats de ces deux méthodes sont précisés dans le Tableau 1.

### 3.4.3. Approche par apprentissage supervisé

Pour qu'une fonction de score fondée sur une distance mathématique soit pertinente pour la normalisation, il est nécessaire que les représentations sémantiques soient d'une qualité suffisamment importante pour que chaque représentation de mention soit à faible distance de la représentation de leur concept associé. Autrement dit, il faudrait que ces représentations soient adaptées spécifiquement à une tâche précise de normalisation. Or, même si certaines méthodes tendent à adapter des espaces sémantiques à des domaines (voir 3.3.7), ce n'est pas ce que des méthodes telles que GloVe ou Word2Vec produisent (Faruqui et al., 2016). Une solution est alors d'utiliser une fonction de score qui elle, soit

optimisée pour une tâche précise. Une telle fonction de score, adaptée à une tâche et à des représentations de mentions et d'entités disponibles, peut être estimée par apprentissage supervisé, et entraînée sur des exemples annotés disponibles.

L'apprentissage supervisé a en effet pour objectif d'apprendre une fonction de prédiction, dépendante d'un ensemble de paramètres, à partir d'exemples annotés. Dans le cas d'une méthode de normalisation par des concepts d'une ontologie et par apprentissage supervisé fondées sur des espaces sémantiques :

- Les expressions textuelles sont représentées par des vecteurs sémantiques.
- Les classes à prédire sont :
  - Soient des représentations vectorielles sémantiques d'étiquettes des concepts
  - Soient des représentations vectorielles de concepts, principalement des variables catégorielles représentées par des représentations 1-parmi-N
- Les exemples annotés sont des couples composés d'une représentation d'expression et d'une représentation de concept.
- La fonction de prédiction se fonde principalement sur une fonction de score de normalisation permettant de calculer un score entre n'importe quelle représentation d'expression et n'importe quelle représentation d'étiquette de concept.

Soit  $L$  une fonction objectif, c'est-à-dire une fonction ayant pour objectif d'approcher les paramètres optimaux  $\theta_f$  pour la fonction de score vis-à-vis de la tâche de normalisation et des exemples annotés. L'apprentissage consiste donc à calculer :  $\theta_f = \operatorname{argmin}_{\theta}(L(\theta))$ . Une des méthodes les plus utilisées pour résoudre ce problème est l'algorithme du gradient stochastique (Burgess et al., 2005). Cette méthode permet d'approcher la matrice optimale  $\theta_f$  en itérant sur des valeurs intermédiaires en fonction d'un paramètre d'apprentissage  $\lambda$ , telle que :

$$\theta_{n+1} = \theta_n - \lambda \cdot \nabla L(\theta_n) \quad (13)$$

Du fait de la possible difficulté à calculer le gradient d'une fonction, la fonction objectif est souvent choisie sous la forme de sommes de termes simples.

Les méthodes fondées sur des représentations vectorielles et par apprentissage d'un score d'association sont définies par deux étapes :

- Une étape d'entraînement (voir Figure 19) pour apprendre les paramètres optimaux pour la fonction de score,
- Une étape de prédiction (voir Figure 20) qui applique la fonction de score à toutes les associations possibles de mentions et de concepts. Le concept (ou son étiquette) qui obtient le meilleur score est alors choisi pour normaliser une mention.

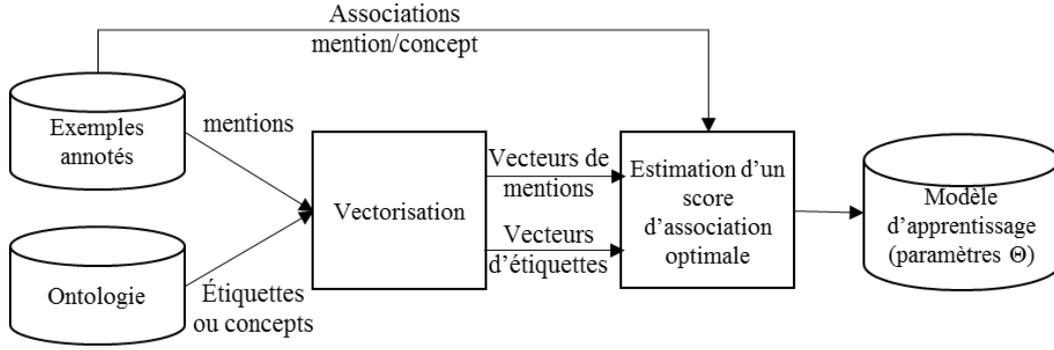


Figure 19 : Description schématique de l'étape d'entraînement des méthodes fondées sur des représentations vectorielles et par apprentissage d'un score d'association

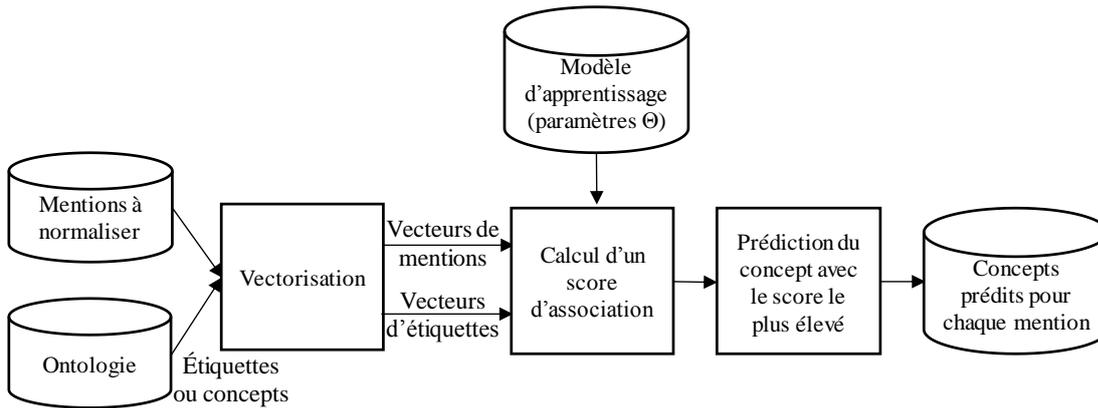


Figure 20 : Description schématique de l'étape de prédiction des méthodes fondées sur des représentations vectorielles et par apprentissage d'un score d'association

### 3.4.4. État de l'art

La méthode Dnorm (Leaman et al., 2013) est historiquement la première méthode de normalisation basée sur l'apprentissage supervisé et des représentations vectorielles sémantiques d'expressions. Elle s'appuie sur des représentations TF-IDF pour les mentions d'entités ainsi que pour les étiquettes de concepts. L'objectif de la méthode est d'apprendre une fonction de score s'appliquant à un couple de vecteurs TF-IDF d'expressions. Soit  $T = \{mot_i\}_{i=1}^{|T|}$ , l'ensemble des mots du vocabulaire. L'espace sémantique défini par les représentations TF-IDF est donc  $\mathbb{R}^{|T|}$ . La fonction inspirée de (Bai et al., 2010), dépendante d'une matrice de paramètre  $\theta \in \mathcal{M}_{|T|}(\mathbb{R})$ , est définie pour deux représentations TF-IDF, telle que :

$$score_{\theta}: \mathbb{R}^{|T|} \times \mathbb{R}^{|T|} \rightarrow \mathbb{R} \quad (x, y) \mapsto score_{\theta}(x, y) = x^T \cdot \theta \cdot y \quad (14)$$

Cette fonction doit retourner une valeur haute lorsque la mention est à normaliser par le concept dont l'étiquette est utilisée en argument. Inversement, elle doit retourner une valeur faible lorsque la mention et l'étiquette ne sont pas à associer.

Soit  $X = \{x_1, \dots, x_N\}$ , un ensemble de  $N$  représentations vectorielles sémantiques de mentions d'entités telle que  $\forall i \in \llbracket 1, N \rrbracket, x_i \in \mathbb{R}^n$  ( $n \in \mathbb{N}^*$ ). Soit  $\Omega = \{\text{concept}_i\}_{i=1}^M$ , un ensemble de  $M$  concepts d'une ontologie et  $E(\mathcal{C}) = \{\text{etq}_i\}_{i=1}^P$ , l'ensemble des  $P$  étiquettes des concepts de  $\mathcal{C}$ . Soit  $\mathcal{C} = \{c_1, \dots, c_N\}$ , un ensemble de  $N$  concepts appartenant à  $\mathcal{C}$ . Soit  $A^+ = \{(x_1, c_1), \dots, (x_N, c_N)\}$ , l'ensemble des couples de représentations TF-IDF de mentions et d'étiquettes de concepts représentant les exemples annotés pour l'entraînement de la méthode. Soit  $g$  une fonction qui permet d'obtenir l'ensemble des représentations vectorielles des étiquettes d'un concept. Soit  $g^-$  une fonction qui permet d'obtenir les représentations vectorielles de toutes les étiquettes de concepts qui n'appartiennent pas aux étiquettes d'un concept spécifique :  $\forall i \in \llbracket 1, M \rrbracket, g^-(c) = \{\text{etq} \in g(c') / c' \in \mathcal{C} - \{c\}\}$ . La fonction objectif suivante, appelée "*margin ranking loss*" (Herbrich, 2000), est alors définie :

$$L(\theta) = \sum_{(x,c)}^{A^+} \sum_{y^+}^{g(c)} \sum_{y^-}^{g^-(c)} \max(0, 1 - \text{score}_\theta(x, y^+) + \text{score}_\theta(x, y^-)) \quad (15)$$

Pour déterminer une matrice  $\theta_f$  qui minimise cette fonction objectif, un algorithme de gradient stochastique est utilisé en parcourant aléatoirement l'ensemble des exemples annotés. A chaque itération, chaque score calculé entre une mention et l'étiquette de concept qui la normalise correctement doit être supérieur au score obtenu avec une étiquette d'un autre concept. Si  $x^\top \cdot \theta \cdot y^+ - x^\top \cdot \theta \cdot y^- < 1$ , une mise à jour est effectuée avec le pas d'itération  $\lambda$  suivant :

$$\theta_{n+1} = \theta_n + \lambda \cdot (x \cdot (y^+)^\top - x \cdot (y^-)^\top) \quad (16)$$

Enfin, une amélioration de la méthode consiste à ne prendre en compte pour chaque mention qu'une seule étiquette à chaque itération parmi celles du concept associé (c'est-à-dire  $\text{card}(g(c)) = 1$ ). L'étiquette utilisée est celle qui rapporte le plus haut score avec la mention utilisée. De-même, il est attendu que le meilleur score pour des étiquettes d'autres concepts reste inférieur au score précédent.

La méthode TaggerOne (Leaman and Lu, 2016) est une autre méthode apprenant une fonction de score entre des représentations TF-IDF. Elle effectue en plus de façon simultanée la reconnaissance et la normalisation d'entités. Elle s'inspire des approches pour la reconnaissance d'entité fondées sur des modèles semi-markoviens (Cohen and Sarawagi, 2004) en calculant des scores de normalisation directement pour des séquences continues de mots plutôt que mot par mot. Néanmoins, la prise en compte des prédictions précédentes dans les séquences de mots (qui représentent une phrase, un corpus, etc. dans lequel il faudrait reconnaître et normaliser des expressions) pour effectuer une nouvelle

prédiction, bien qu'étant au centre des modèles markoviens, n'est pas utilisée, car elle réduisait systématiquement les performances de la méthode.

Dans les modèles de Markov cachés appliqués à la tâche de reconnaissance d'entité, le problème est formalisé par la recherche d'une séquence de catégories sémantiques alignée avec une séquence de mots. Les catégories sémantiques sont celles spécifiées par la tâche, ainsi qu'une catégorie <vide> en plus, qui est à aligner avec tous les mots qui n'appartiennent à aucune mention à identifier par les autres catégories. Par exemple, considérons la phrase suivante segmentée en 7 mots : *['le', 'petit', 'chien', 'aboie', 'contre', 'un', 'chat']*.

Alors l'objectif est de déterminer et d'aligner une séquence de 7 catégories sémantiques sur cette séquence de mots. Si les catégories d'intérêt pour la tâche sont <chien> et <chat>, alors la séquence solution devrait être :

*[< chien >, < chien >, < chien >, < vide >, < vide >, < chat >, < chat >]*

Les modèles semi-markoviens introduisent la notion de segment de texte pour un texte préalablement segmenté en mots. Ces segments représentent des sous-séquences de mots. Pour qu'un texte segmenté en mots soit décomposé en segments valides, il faut que tous les mots du texte initial appartiennent à un unique segment et que tous les segments soit continus. Par exemple, la séquence suivante est une séquence de segments valides de la phrase précédente et qui représente certainement la meilleure détection des bornes des mentions d'entités d'intérêt :

*[['le', 'petit', 'chien'], ['aboie'], ['contre'], ['un', 'chat']]*

A la différence d'un modèle markovien, l'objectif d'un modèle semi-markovien est alors de déterminer une décomposition en segments valides et d'aligner directement chaque segment avec une catégorie. Pour l'exemple précédent, la séquence de catégories sémantiques solution devrait alors être :

*[< chien >, < vide >, < vide >, < chat >]*

Notons que la catégorie additionnelle <vide> ne doit être alignée qu'avec des segments à un mot.

La méthode TaggerOne va alors chercher une fonction de score qui va s'appliquer à une décomposition en segment et une séquence de catégories de même taille. En calculant un score pour toutes les décompositions en segments valides possibles et tous les alignements de catégories de la même taille, la méthode effectue simultanément une reconnaissance et une normalisation d'entité. Notons  $S = [s_1, \dots, s_M]$  une séquence de  $M$  segments telle que  $S$  soit une représentation d'une décomposition en segments valide d'une séquence de mots. Notons  $Y = [y_1, \dots, y_M]$  une séquence de  $M$  étiquettes de concepts de l'ontologie. Soit  $T$ , l'ensemble des mots du vocabulaire. Une fois l'ensemble de paramètres  $\theta$

définissant la fonction de score optimale déterminé, la prédiction d'une décomposition en segments  $S_{pred}$  et d'une séquence de catégories alignée  $Y_{pred}$  peut être déterminée telle que :

$$(S_{pred}, Y_{pred}) = \operatorname{argmax}_{S, Y} (\operatorname{score}_{\theta}(S, Y)) \quad (17)$$

L'objectif de la méthode est donc d'apprendre une fonction de score qui permette de maximiser la valeur obtenue sur un ensemble d'exemples annotés. La fonction de score  $\operatorname{score}'_{\theta}$  entre une représentation TF-IDF de segment et une représentation TF-IDF d'étiquette est définie comme la somme de deux scores :

$$\operatorname{score}'_{\theta}: \mathbb{R}^{|T|} \times \mathbb{R}^{|T|} \rightarrow \mathbb{R} \\ (s, y) \mapsto \operatorname{scoreNER}_{\theta}(s, y) + \operatorname{scoreNorm}_{\theta}(s, y) \quad (18)$$

Le score  $\operatorname{scoreNorm}_{\theta}$  évalue seulement la normalisation et s'inspire de celui de DNorm et utilise un ensemble de paramètre  $\theta$  représenté par une matrice  $W \in \mathcal{M}_{|T|}(\mathbb{R})$  et un scalaire  $\alpha$ . Le score est alors défini tel que :

$$\operatorname{scoreNorm}_{\theta}(s, y) = \alpha \cdot (s^{\top} \cdot y) + s^{\top} \cdot W \cdot y \quad (19)$$

Le score global est alors calculé en sommant tous les scores  $\operatorname{score}'_{\theta}$  entre chaque segment et catégorie.

La phase d'entraînement est effectuée par l'algorithme MIRA (Crammer and Singer, 2001), directement sur le score global pour la reconnaissance et la normalisation d'entité. En conséquence, la méthode peut apprendre à utiliser l'information de l'une des deux tâches pour améliorer sa prédiction sur l'autre, là où la plupart des méthodes actuelles les réalisent l'une à la suite de l'autre. Pour être utilisée seulement pour la normalisation, il suffit d'utiliser une décomposition en segment qui représente chaque mention d'entité d'intérêt comme un segment, et tous les autres mots du texte comme un segment à un mot, préalablement aligné avec la catégorie <vide>.

Malgré l'utilisation de représentations TF-IDF ne permettant directement que de comparer des expressions possédant des mots en commun, DNorm et TaggerOne, de par leur approche par apprentissage, semblent pouvoir permettre d'apprendre à reconnaître des mentions de formes différentes d'une même entité.

(Limsopatham and Collier, 2016) proposent une méthode neuronale fondée sur un réseau neuronal convolutif ("*convolutional neural networks*" - CNN). Cette méthode a systématiquement obtenu de meilleurs résultats que son alternative avec un réseau récurrent ("*recurrent neural network*" - RNN). Elle s'inspire de (Limsopatham and Collier, 2015), qui fixait manuellement la fonction de score à partir des plongements lexicaux. Cette nouvelle méthode est spécialisée dans la normalisation d'expression multi-mots ou de phrases, et a été développée pour la normalisation de messages issus de médias sociaux par des concepts biomédicaux. Par exemple, la phrase "*I don't hunger or thirst*" pourrait être

normalisée par le concept <Loss of Appetite>. L’objectif du réseau est de prendre en entier une séquence de représentations vectorielles de mots, représentant une expression ou une phrase, et de la classer dans l’une des classes définies par les concepts de l’ontologie. Elle s’inspire de réseaux appliqués à la traduction automatique, en formulant la tâche comme une tâche de traduction automatique de phrase vers un nom de concept parmi ceux disponibles.

Pour le CNN, une mention est représentée par une séquence ordonnée de représentations vectorielles de chaque mot composant la mention. Soit  $S = (m_1, \dots, m_l)$ , une représentation définie par une séquence de  $l$  représentations vectorielles de mots dans  $\mathbb{R}^d$  ( $d \in \mathbb{N}^*$ ) telles que :  $\forall i \in \llbracket 1, l \rrbracket, m_i \in \mathbb{R}^d$ . Les représentations vectorielles testées sont des plongements lexicaux calculés par Word2Vec (Mikolov et al., 2013a) et pré-entraînés sur Google News<sup>14</sup>. Soit  $h$ , un entier représentant la taille d’une fenêtre de mots. L’ensemble des paramètres d’apprentissage  $\Theta$  est composé d’un ensemble de vecteurs filtres  $w(h) \in \mathbb{R}^{d \times h}$ , et d’un paramètre de biais  $b \in \mathbb{R}$ . Soit  $m_{i:i+h-1} \in \mathbb{R}^{d \times h}$ , la concaténation de  $h$  représentations vectorielles de mots consécutifs d’une expression. Soit  $f$  une fonction d’activation telle que la fonction ReLu (Nair and Hinton, 2010). On définit la valeur  $c_i(h)$  telle que :

$$\forall i \in \llbracket 1, l \rrbracket, c_i(h) = f(w(h) \cdot m_{i:i+h-1} + b) \quad (20)$$

Pour toutes les variations de fenêtres de taille  $h$ , on peut donc calculer  $l$  valeurs dont on voudrait qu’elles représentent chacune un score de dépendance entre les mots composant la séquence de mots commençant à la position  $i$  dans l’expression. Pour capturer la caractéristique la plus importante de l’expression, vis-à-vis de la taille de la fenêtre commençant à la position  $i$ , une étape de “max-pooling” est appliquée pour obtenir un vecteur de caractéristique  $c_{max}$  de l’expression :

$$c_{max} = [\max_h(c_1(h)), \dots, \max_h(c_l(h))] \quad (21)$$

Ce vecteur  $c_{max}$  est finalement utilisé en entrée d’une couche “softmax” pour obtenir une probabilité de classification pour chaque concept de l’ontologie.

L’entraînement est effectué par descente de gradient stochastique avec une mise à jour de type Adadelta (Zeiler, 2012). En utilisant des plongements lexicaux plutôt que des représentations TF-IDF, la méthode peut aborder le problème de la variabilité des formes de mentions d’un même concept. En utilisant de l’apprentissage, elle peut s’adapter à une tâche particulière grâce à une étape d’entraînement. Néanmoins, l’architecture CNN utilisée reste dépendante d’un nombre important d’exemples annotés.

<sup>14</sup> <https://code.google.com/archive/p/word2vec/>

## 3.5. Bilan

Nous venons de voir que les méthodes de normalisation d'entité fondées sur la similarité de forme entre mention et étiquette de concept avaient de fortes limitations, notamment leur faible capacité de généralisation et de traitement des variations de forme. Les approches par apprentissage apportent une réponse au problème de généralisation, et les approches fondées sur des plongements lexicaux apportent une réponse au problème de variation de forme. De plus, à partir de représentations vectorielles de mention et d'étiquette (ou de concept), ces méthodes peuvent produire un score pour l'association mention/étiquette ou mention/concept. En pratique, elles prédisent un seul concept en choisissant celui qui donne le meilleur score. Néanmoins, le classement des concepts fourni par ces méthodes pourrait permettre une étape de prédiction plus complexe, en choisissant par exemple plusieurs concepts dans certains cas. La normalisation multiple reste néanmoins un problème peu exploré.

Pour les méthodes par apprentissage et représentation sémantique, les deux voies pour réaliser tout le potentiel pour la normalisation, sont :

- Soit une augmentation des données fournies pour une tâche, c'est-à-dire principalement de l'augmentation du nombre des données annotées et/ou de l'augmentation du nombre d'étiquettes pour chaque concept,
- Soit de nouveaux progrès dans les méthodes même d'apprentissage pour des données étiquetées en quantité limitée (Ching et al., 2018).



## **Chapitre 4**

-

## **Construction d'espaces sémantiques**



## 4.1. Introduction

Dans la section précédente, nous avons présenté le potentiel des méthodes fondées sur des plongements lexicaux et sur l'apprentissage supervisé. Néanmoins, rappelons que les tâches de normalisation en domaine spécialisé sont contraintes par le petit nombre d'exemples annotés manuellement par rapport au nombre de concepts considérés qui est souvent très important (au moins de plusieurs milliers). De plus, l'annotation manuelle demande un effort supplémentaire car il faut avoir recours à des experts du domaine. Or, les approches par apprentissage sont fortement dépendantes de la quantité et de la qualité des exemples annotés.

L'approche développée durant cette thèse prend donc inspiration dans l'approche par apprentissage automatique supervisé fondée sur des espaces sémantiques et apporte une réponse au problème d'apprentissage avec peu ou pas d'exemples annotés. En particulier, nous avons fait les choix suivants :

- Conserver l'utilisation des plongements lexicaux pour leurs propriétés sémantiques
- Représenter chaque concept de l'ontologie directement par une représentation vectorielle plutôt que de représenter indépendamment leurs étiquettes de concepts.

Son originalité vient de la "sémantisation" des représentations de concepts, au contraire des représentations 1-parmi-N : on représente un maximum d'informations contenues dans une ontologie par un espace sémantique, dans lequel deux concepts ayant un sens proche possèdent des vecteurs spatialement proches. Et pour cela, les connaissances externes utilisées sont celles contenues dans l'ontologie de référence.

L'objectif est de donner une représentation vectorielle à l'ontologie en créant un espace sémantique de référence pour le domaine et plus particulièrement pour la tâche. (Mikolov et al., 2013c) ont montré que l'espace des plongements lexicaux possède des propriétés structurelles, et notre objectif est de construire un espace sémantique ontologique (ESO) possédant ce genre de propriétés. De plus, notre objectif est de créer un espace de référence plus proche de la structure de l'ontologie que les solutions proposées jusqu'ici. Cet espace devrait avoir l'avantage de ne pas présenter d'ambiguïté, c'est-à-dire de permettre de discriminer ce qu'il est nécessaire de différencier en fonction de la tâche. Les mentions d'entités à représenter dans cet espace, représentées par des vecteurs, devraient être situées spatialement à proximité de leurs concepts de rattachement.

Nous avons vu que les méthodes de construction de plongements lexicaux sont justement capables de représenter des mots dans un espace sémantique où les mots avec un sens proche auront des représentations vectorielles à proximité. Des méthodes de composition de vecteurs de mots permettent également de transférer cette propriété à des représentations d'expressions multi-mots, formant ce que nous appellerons un espace

distributionnel des expressions (EDE). Comme les mentions d'un même concept désignent des entités avec un sens proche, l'EDE devrait être structuré en plusieurs regroupements de représentations de mentions désignées par un même concept.

Néanmoins, la qualité d'un EDE pour représenter un domaine et une tâche particulière est dépendante de nombreux paramètres. Notamment, il faut faire le choix :

- D'un corpus d'entraînement pour les méthodes de construction de plongements lexicaux, ainsi que les pré-traitements à appliquer sur celui-ci ;
- De tous les hyper-paramètres de ces méthodes ;
- Et d'une méthode de composition de représentations de mots.

Notre hypothèse est que si l'EDE est structuré de façon à bien représenter le domaine, c'est-à-dire s'il y a une similarité structurelle globale entre l'EDE et l'ESO, alors il est possible de construire un alignement entre les deux espaces qui aura de bonnes propriétés prédictives.

Nous avons développé la méthode CONTES (“*CONcept-TErm System*”) qui recherche cet alignement, par apprentissage supervisé, sous la forme d'une fonction de projection de l'EDE vers l'ESO. Cette méthode s'appuie sur l'utilisation de la méthode Word2Vec-SkipGram (Mikolov et al., 2013a) pour construire un EDE, et sur la méthode Ancestry (introduite dans la section 0) pour calculer un ESO. Nous détaillerons les choix de paramètres et de ressources utilisées pour produire des EDE (voir section 0). Pour guider ces choix, nous avons évalué CONTES sur la tâche de normalisation d'habitat bactérien Bacteria Biotope (voir 2.5.4) et avons utilisé l'ontologie OntoBiotope. Les exemples annotés utilisés sont ceux du corpus d'entraînement. Les scores présentés ont principalement été obtenus sur le corpus de développement.

La méthode CONTES initiale, intégrant les parties sur l'EDE et l'ESO, a été publiée dans les actes du workshop BioNLP en 2017, adossé à la conférence “*Association for Computational Linguistics*”. Elle a donné lieu à une présentation orale. Nous avons collaboré avec Robert Bossy sur les questions d'optimisation de l'implémentation de CONTES et de son intégration dans la plateforme AlvisNLP, ce qui nous a permis de mener à bien toutes les expérimentations de façon reproductible.

## 4.2. Méthode de représentation vectorielle de concepts d'une ontologie hiérarchique : Ancestry

### 4.2.1. Introduction

L'objectif de notre approche est la représentation des concepts d'une ontologie par des vecteurs d'un même espace vectoriel. Cet ensemble de vecteurs doit également intégrer une partie des informations contenues dans l'ontologie, dont les informations relationnelles reliant les concepts de l'ontologie. L'espace sémantique produit à partir d'une ontologie est appelé espace sémantique de l'ontologie (ESO) (voir Figure 21). Le graphe acyclique dirigé formé par la relation de subsumption ( $\langle \text{is\_a} \rangle$ ) étant communément le seul graphe connexe contenu dans une ontologie, nous avons commencé prioritairement à l'exploiter. Notons que ce graphe est également toujours acyclique. Nous retenons comme critère la possibilité de reconstruction du graphe à partir de l'ensemble des vecteurs. Pour cela, nous pouvons par exemple nous appuyer sur des calculs de distances entre les vecteurs de concepts : les vecteurs les plus proches spatialement doivent représenter les concepts connectés entre eux dans le graphe. Des concepts connectés entre eux par des relations de subsumption partageant des similarités sémantiques (ex :  $\langle \text{chien} \rangle \langle \text{is\_a} \rangle \langle \text{mammifère} \rangle$ ), cette propriété permet d'établir un score de similarité sémantique entre des concepts en calculant une distance spatiale ou un score entre les vecteurs associés. En conséquence, le choix d'une mesure de score/distance aura une influence sur les scores estimés. Un score fondée sur la similarité cosinus a été principalement utilisé dans nos expériences, car celui-ci est fréquemment une bonne solution pour mesurer des similarités sémantiques entre représentations vectorielles d'espaces distributionnels et/ou de grande dimension (Mikolov et al., 2013b; Pennington et al., 2014; Kamper et al., 2015).

Même si la représentation des vecteurs de concept de type 1-parmi-N ne suffit naturellement pas à répondre à notre objectif d'intégration de l'information structurelle de l'ontologie, nous partons de ce modèle de représentation et essayons de l'améliorer pour qu'il se rapproche de notre objectif. En effet, les représentations 1-parmi-N de concepts représentent des concepts de façon indépendante, or, ils ne le sont pas par définition d'une ontologie. De plus, conserver une dimension distincte pour chaque classe nous permet de limiter l'impact des mentions normalisées par plusieurs concepts dans les exemples annotés. Soit  $C = \{c_k\}_{k=1}^n$ , l'ensemble des  $n$  ( $n \in \mathbb{N}^*$ ) concepts de l'ontologie tel que le concept  $c_k$  est associé à la dimension  $k$ . Soit  $V = \{v_{c_k}\}_{k=1}^n$ , l'ensemble des  $n$  représentations vectorielles de ces concepts, tel que le vecteur  $v_{c_k}$  représente le concept  $c_k$ . Soit  $w_{c_k}^i$ , le  $i$ -ème composant du vecteur du concept  $c_k$ . Alors on peut représenter chaque vecteur de concept par la formule suivante :

$$v_{c_k} = (w_{c_k}^1, \dots, w_{c_k}^n) \quad (22)$$

Pour exploiter la relation de subsumption, les coefficients associés aux dimensions de concepts qui appartiennent aux mêmes lignées (c'est-à-dire tous les enfants jusqu'aux feuilles et tous les parents jusqu'à la racine) que celles du concept courant peuvent se voir attribuer une valeur non-nulle. Pour chaque composante d'un vecteur de concept défini en (22), on a donc :

$$w_{c_k}^i = \begin{cases} 1 & \text{si } i = k \\ a_{c_k}^{c_i} & \text{si } c_i \text{ est un ancêtre de } c_k \\ a'_{c_k}^{c_i} & \text{si } c_i \text{ est un descendant de } c_k \\ 0 & \text{sinon} \end{cases} \quad (23)$$

Notre méthode Ancestry représente chaque vecteur de concept en ne prenant en compte que les concepts ancêtres dans les lignées du concept courant, mais pas les concepts de sa descendance. Pour cela, toutes les valeurs associées à des ancêtres du concept courant auront une valeur de 1, des premiers parents du concept courant jusqu'à la racine.

En utilisant l'équation (23), on a donc :

$$a_{c_k}^{c_i} = \begin{cases} 1 & \text{si } c_i \text{ est un ancêtre de } c_k \\ 0 & \text{sinon} \end{cases} \quad (24)$$

$$a'_{c_k}^{c_i} = 0 \quad (25)$$

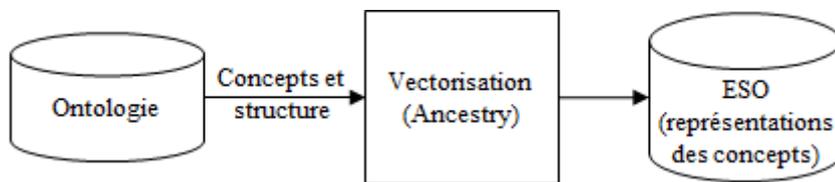


Figure 21 : Schéma de la méthode de construction de représentations vectorielle de concepts Ancestry. La méthode utilise les concepts ainsi que la structure globale de l'ontologie (via ses relations de subsumption  $\langle IsA \rangle$ ) pour construire l'espace sémantique de l'ontologie (ESO).

## 4.2.2. Étude du modèle

On peut étudier ce modèle par un exemple d'ontologie minimaliste, comme celle représentée dans la Figure 22. Avec cette ontologie et les formules (22), (23), (24) et (25), nous calculons les représentations vectorielles représentées dans la Figure 23.

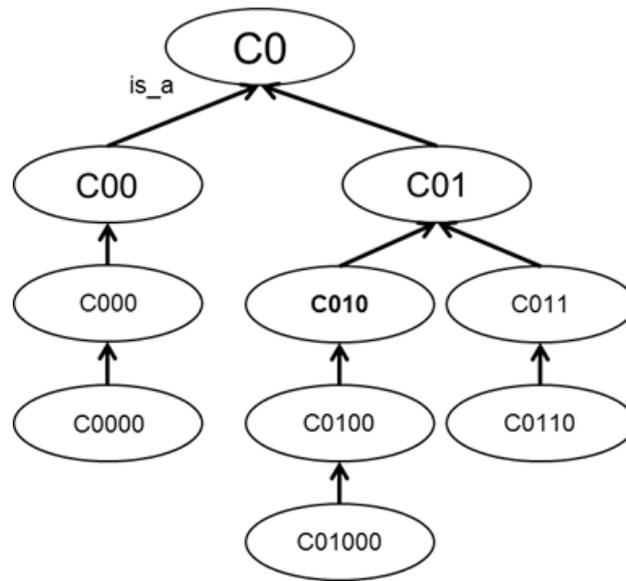


Figure 22 : Exemple d'ontologie minimaliste

Et grâce à ces représentations, nous pouvons calculer la similarité cosinus et la distance euclidienne entre vecteurs de concepts. Le Tableau 4 montre les mesures de similarité cosinus entre le concept C010 et les autres concepts de l'ontologie. On peut constater que le classement global des concepts ne change pas selon la mesure utilisée, bien qu'il y ait des différences au niveau des concepts à égale distance. Notamment, la distance euclidienne introduit beaucoup d'égalité, et ne permet donc pas de discriminer certains concepts. La Figure 24 montre une visualisation de l'ESO après une réduction de dimension par ACP, laquelle conserve les propriétés de similarité sémantique de l'espace non-réduit.

dimensions vecteurs	C0	C00	C01	C000	C010	C011	C0000	C0100	C0110	C01000
C0	1	0	0	0	0	0	0	0	0	0
C00	1	1	0	0	0	0	0	0	0	0
C01	1	0	1	0	0	0	0	0	0	0
C000	1	1	0	1	0	0	0	0	0	0
C010	1	0	1	0	1	0	0	0	0	0
C011	1	0	1	0	0	1	0	0	0	0
C0000	1	1	0	1	0	0	1	0	0	0
C0100	1	0	1	0	1	0	0	1	0	0
C0110	1	0	1	0	0	1	0	0	1	0
C01000	1	0	1	0	1	0	0	1	0	1

Figure 23 : Représentations vectorielles des concepts de l'ontologie selon la méthode Ancestry

Concept	Similarité cosinus	Distance euclidienne
C010	1,00	0,00
C01000	0,87	1,00
C01	0,82	1,00
C01000	0,77	1,41
C011	0,67	1,41
C0	0,58	1,41
C0110	0,58	1,73
C00	0,41	1,73
C000	0,33	2,00
C0000	0,29	2,24

Tableau 4 : Similarité cosinus et distance euclidienne entre le concept C010 et les autres concepts de l'ontologie (arrondi à deux chiffres après la virgule)

En calculant les similarités cosinus entre chaque paire de vecteurs de concepts de l'ontologie, on peut alors représenter sa structure : en partant de n'importe quel vecteur, on peut déterminer ses concepts-fils (les plus similaires) et ses concepts-parents (les seconds plus similaires), et ainsi remonter jusqu'au concept-racine de l'ontologie, ce qui est une propriété que nous recherchons.

Dans l'exemple de la Figure 25, l'ontologie présente un héritage et cette propriété de représentation persiste. Même si ces exemples n'ont pas valeur de preuve, ils illustrent comment le modèle Ancestry répond à nos objectifs. Les résultats de l'utilisation de ces représentations seront comparés à ceux de l'utilisation de représentations 1-Parmi-N (voir 0 - Tableau 11).

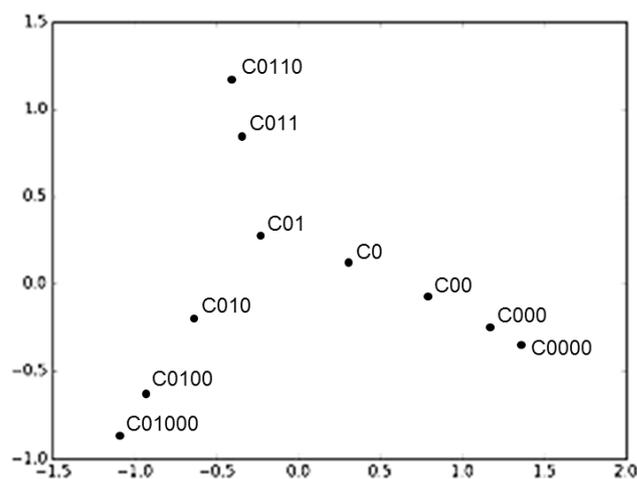


Figure 24 : Réduction de dimension par ACP (avec distance euclidienne)

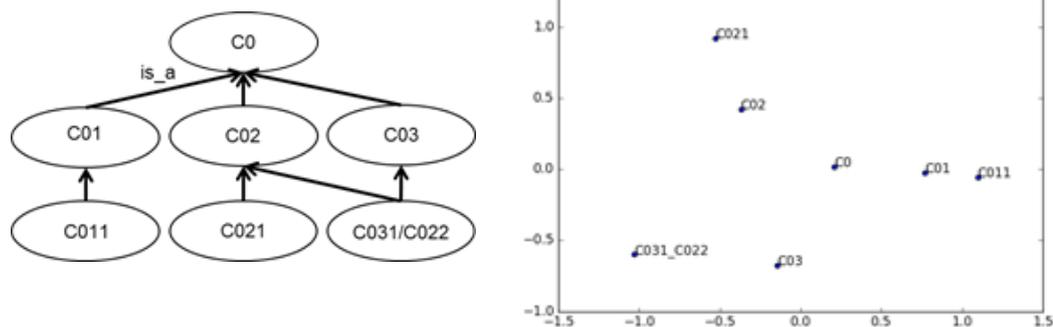


Figure 25 : Exemple d'une ontologie contenant un héritage multiple (à gauche) et la réduction par ACP correspondante (à droite).

### 4.2.3. Limites du modèle Ancestry

Tout comme dans un modèle 1-Parmi-N (voir 3.3.2), la taille de chaque vecteur du modèle Ancestry est égale au nombre de concepts différents dans l'ontologie. Cette taille devient donc rapidement importante pour certaines ontologies comme le méta-thésaurus UMLS qui contient environ 5 millions de concepts. Avec des ontologies volumineuses, l'utilisation pratique des représentations de ce modèle devient difficile en terme de complexité algorithmique, puisque la complexité de l'espace évolue en  $O(n^2)$ .

De plus, les représentations produites sont très creuses. En effet, le maximum de coefficients à 1 dans un vecteur est obtenu pour les concepts les plus profonds dans le graphe hiérarchique. Or, dans de nombreuses ontologies, la profondeur maximale reste relativement faible par rapport au nombre de concepts. Par exemple, dans l'ontologie OntoBiotope, la profondeur maximale est de 13 pour plus de 2320 concepts. Les représentations vectorielles sont donc toutes remplies à au moins 99,4% de valeurs nulles.

L'héritage multiple peut altérer la qualité des représentations, même si la proximité spatiale entre vecteurs de concept ne semble pas devenir aberrante pour autant. La figure suivante montre en effet que des concepts-parents peuvent être à des distances différentes de leur concept-enfant, prévenant la possibilité de reconstruire le graphe initial à partir des seules informations de distances entre les vecteurs de concepts. Néanmoins sur l'exemple de la Figure 26, cela n'altère pas le classement des concepts et n'atteint donc pas les propriétés sémantiques de la représentation.

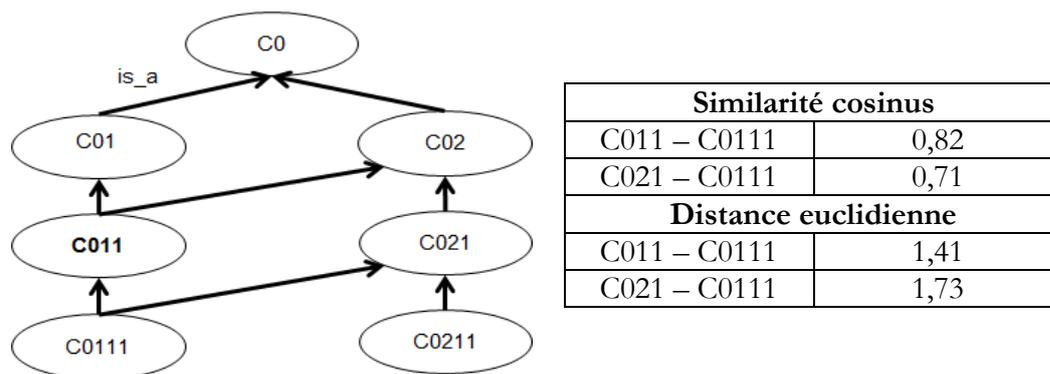


Figure 26 (à gauche) : Ontologie possédant des héritages multiples

Tableau 5 (à droite) : Calculs de distance en le concept C0111 et ses deux concepts parents C011 et C021. La mesure de similarité entre un concept-enfant et ses deux concepts-parents est différentes.

Enfin, avec la similarité cosinus, des concepts parents et enfants seront d'autant plus similaires qu'ils seront profonds dans l'arborescence. C'est également le cas pour des concepts-frères. Ces propriétés sont sémantiquement pertinentes : par exemple, il y a bien moins de similarité entre les mammifères et les oiseaux, qu'entre les chiens et les chats. Pour conséquence sur la prédiction de concepts pour la normalisation, on s'attend à ce que cela favorise des erreurs de prédiction d'un concept frère, parent ou fils au lieu du concept attendu. Parmi les différents cas de prédiction d'un concept autre que celui attendu, la détection d'un concept parent direct est d'ailleurs la moins pénalisante dans la tâche Bacteria Biotope 3. De plus, la prédiction d'un parent direct n'entraîne aucune erreur logique : une mention d'un concept est également une mention d'un concept parent.

#### 4.2.4. Rapprochement entre représentations de concepts fils et parents : Ancestry+

Pour essayer de favoriser des rapprochements entre concepts fils et parents, tout en conservant les propriétés sémantiques visées, nous avons développé la méthode Ancestry+. Dans le cas d'erreur de prédiction pour la tâche de normalisation, l'idée est de favoriser des prédictions de concepts parents du concept correct, plutôt que des prédictions des concepts frères ou fils. Autrement dit, nous cherchons à faciliter la prédiction de concepts plus généraux qu'un concept exact, ce qui reste logiquement vrai. Cette méthode vise donc à éloigner chaque vecteur de concept des vecteurs de ses descendants, et à l'inverse de le rapprocher des vecteurs de ses concepts parents. Si *sim* est une mesure de similarité, alors la propriété globale suivante était donc recherchée :

$$sim(\langle \text{concept} \rangle, \langle \text{enfant} \rangle) < sim(\langle \text{concept} \rangle, \langle \text{parent} \rangle) \quad (26)$$

dimensions vecteurs	C0	C00	C01	C000	C010	C011	C0000	C0100	C0110	C01000
C0	0	0	0	0	0	0	0	0	0	0
C00	1	1	0	0	0	0	0	0	0	0
C01	1	0	1	0	0	0	0	0	0	0
C000	1	1	0	2	0	0	0	0	0	0
C010	1	0	1	0	2	0	0	0	0	0
C011	1	0	1	0	0	2	0	0	0	0
C0000	1	1	0	1	0	0	3	0	0	0
C0100	1	0	1	0	1	0	0	3	0	0
C0110	1	0	1	0	0	1	0	0	3	0
C01000	1	0	1	0	1	0	0	1	0	4

Figure 27 : Représentations Ancestry+ pour l'ontologie de la Figure 22.

Pour cela, au lieu d'affecter la valeur associée au concept à 1, la méthode lui affecte la valeur égale à la distance du concept de la racine. Soit la fonction *dist* qui calcule une distance entre deux concepts  $c_i$  et  $c_j$  définie telle que :

$$dist(c_i, c_j) = \text{plus court chemin entre } c_i \text{ et } c_j \quad (27)$$

Notons qu'en particulier, nous avons :

$$dist(c_k, c_k) = 0 \text{ et } dist(c_k, \text{parent direct de } c_k) = 1 \quad (28)$$

Alors d'après la formule (22), on définit la méthode Ancestry+ par la formule suivante :

$$w_{c_k}^i \begin{cases} dist(c_k, \text{racine}) & \text{si } i = k \\ 1 & \text{si } c_i \text{ ancêtre de } c_k \\ 0 & \text{sinon} \end{cases} \quad (29)$$

En calculant des représentations pour ce modèle avec l'ontologie utilisée en Figure 22, on obtient les représentations de la Figure 27. On peut observer qu'il y a peu de différence avec les représentations construites par la méthode Ancestry (voir Figure 23).

On peut observer que ce modèle, tout en répondant aux objectifs initiaux, répond à la propriété souhaitée. On peut néanmoins observer une diminution générale des valeurs des scores de similarité. Cela indique une dispersion importante de l'ensemble des concepts dans l'espace considéré.

	Similarité cosinus	Distance euclidienne
C0–C02	0,45	2,00
C02–C021	0,23	3,16
C021–C0211	0,15	4,47
C0211–C02111	0,11	5,83
C01–C02	0,20	2,83
C021–C022	0,18	4,24

Tableau 6 : Calculs de distance euclidienne et de similarité cosinus entre certains concepts de l'ontologie pour la représentation non-réduite.

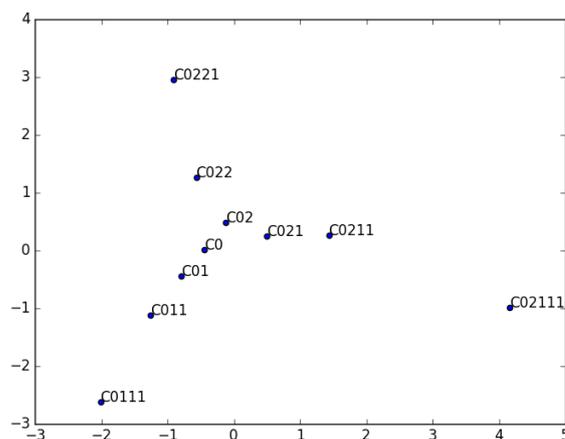


Figure 28 : Représentation de l'ESO après ACP pour le modèle Ancestry+ et appliqué sur l'ontologie de la Figure 22.

#### 4.2.5. Tentative de densification et de diminution du nombre de dimensions

Les deux autres limites du modèle Ancestry mentionnées précédemment sont la taille importante des vecteurs produits et leur caractère très creux. Pour améliorer les représentations, une méthode de réduction de dimension peut être appliquée. Nous avons testé plusieurs analyses en composantes principales (ACP). Pour évaluer la qualité des espaces réduits, et les comparer à l'espace initial, nous avons évalué la méthode CONTES sur la tâche de normalisation Bacteria Biotope de BioNLP Shared Task, en utilisant des espaces ontologiques avec différentes réductions par ACP. Les différents résultats sont présentés dans la Figure 29.

On observe qu'une réduction de dimension par ACP dégrade le score final de la tâche de normalisation. Néanmoins, cette dégradation reste relativement faible jusqu'à une taille de vecteur de dimension 600, soit aux alentours de 75% de taux de réduction. Nous avons émis trois hypothèses, non validées, quant à cette diminution de performance :

- Il n'y a plus assez de dimensions pour permettre d'exprimer correctement l'information structurelle de l'ontologie

- Il faut un nombre de dimensions suffisant pour représenter les héritages multiples. À partir d'un certain seuil de réduction de dimension, ce nombre n'est plus suffisant et la représentation perd trop d'informations sur la structure initiale de l'ontologie. L'ontologie OntoBiotope contient environ 20% de concepts qui héritent d'au moins deux parents. Un taux de réduction de plus de 80% ne peut alors que détériorer l'espace ontologique initial. Ce taux se rapproche justement du taux de réduction observé à partir duquel les performances de la méthode CONTES chutent drastiquement.
- L'ACP va optimiser la répartition des vecteurs de concepts dans l'espace disponible pour maximiser leur variance. Cette répartition pourrait éloigner les vecteurs les uns des autres par rapport à l'espace initial. En conséquence, leurs scores de similarité sémantique peuvent globalement diminuer de façon importante, et cela pourrait avoir un impact négatif sur les résultats de la méthode CONTES.

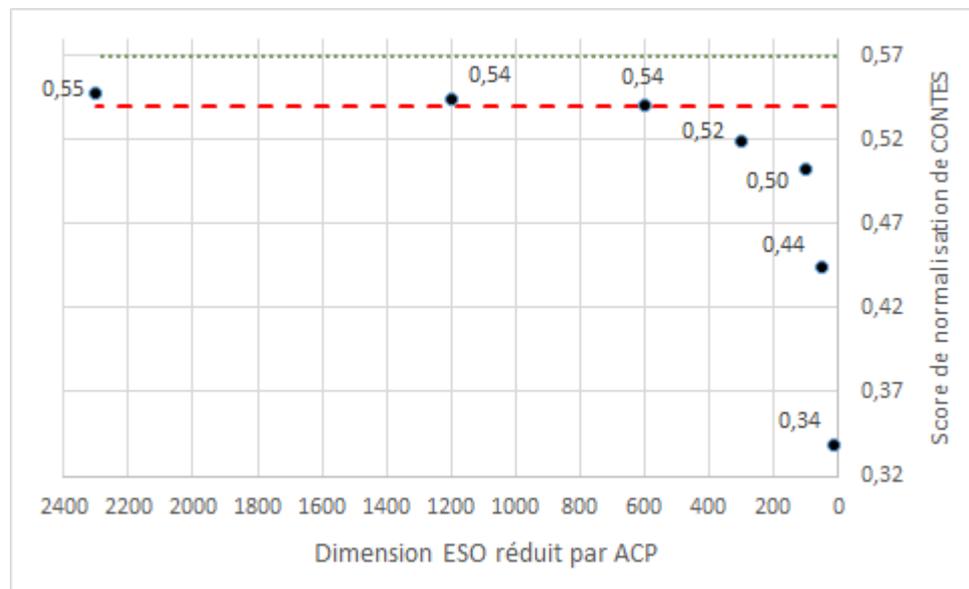


Figure 29 : Influence d'une réduction de dimension de l'espace ontologique par ACP. L'axe des abscisses représente le nombre de dimension de l'ESO après réduction. La droite en tirets rouges représente le score de la méthode de référence.

La droite en points verts représente le score de la méthode CONTES sans réduction de dimension de l'ESO (expérimentation effectuée avec quelques paramètres différents pour le calcul de l'EDE).

En conséquence, nous avons conservé l'ESO construit avec la méthode Ancestry, non-réduit, dans le reste de nos expérimentations.

## 4.3. Construction initiale de l'espace distributionnel des expressions

### 4.3.1. Introduction

La méthode CONTES se fonde sur des plongements lexicaux, qu'elle adapte au domaine et à la tâche pour estimer un score de normalisation. Selon les représentations choisies, et plus largement, selon le paramétrage des méthodes de construction de vecteurs choisies, la similarité structurelle entre les deux espaces sémantiques de mots du texte et de concepts de l'ontologie variera, et la performance de CONTES en sera affectée. Il y a donc des choix à faire pour garantir cette performance. L'étape de construction de l'EDE est présentée sur la Figure 30.

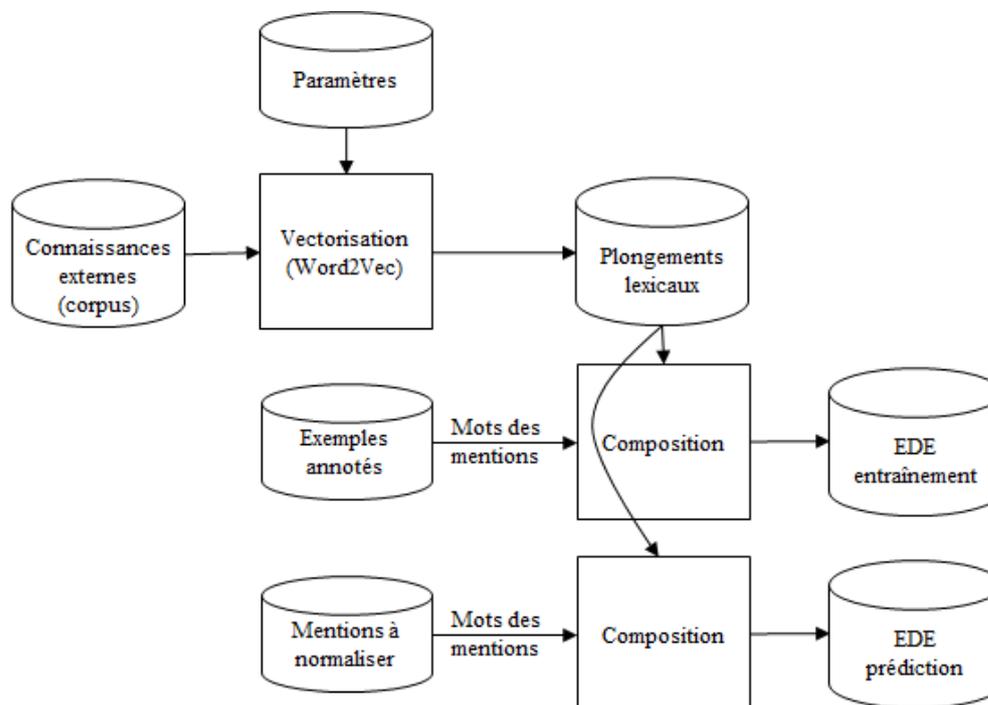


Figure 30 : Schéma de la méthode de construction des espaces distributionnels des expressions utilisées par la méthode CONTES. Un corpus externe est choisi et pré-traité. Un calcul de plongement lexicaux est alors effectué avec certains paramètres pour obtenir des représentations de mots. À partir des mots des mentions (issues des exemples annotés ou d'exemples à normaliser), une méthode va construire des représentations denses et de faible dimension pour chaque mention. En fonction de l'origine des exemples, deux espaces distributionnels des expressions sont construits : celui d'entraînement à partir des exemples annotés, et celui de prédiction pour les exemples à annoter.

Ces choix sont expliqués et justifiés dans les sections suivantes, et leur impact est directement évalué avec le score obtenu par la méthode CONTES sur la tâche de normalisation Bacteria Biotope, principalement sur le jeu de données de développement.

### 4.3.2. Choix de méthode et de corpus d'entraînement

Selon (Muneeb et al., 2015) et (Chiu et al., 2016), le modèle de création de plongements lexicaux le plus performant dans le domaine biomédical serait Word2Vec avec l'architecture Skip-Gram. Dans le domaine général, (Levy et al., 2015) montre également que cette architecture représente un état de l'art sur plusieurs tâches d'évaluation extrinsèques. Pour le confirmer, nous avons comparé cette architecture et l'architecture CBOW de Word2Vec avec un corpus d'articles biologiques et pour la tâche de normalisation Bacteria Biotope (voir 4.3.3 - Figure 31).

Pour éviter le problème de la représentation des mots hors-vocabulaire et pour favoriser l'apparition de chaque mot dans un contexte similaire (domaine et style de langue) à celui de la tâche Bacteria Biotope, nous avons choisi un corpus d'entraînement de Word2Vec composé de titres et résumés d'articles scientifiques du domaine biologique. La base de données bibliographique MEDLINE<sup>15</sup> donne accès à des millions de documents répondant à cet objectif. Pour s'approcher au mieux du domaine de la tâche Bacteria Biotope, plus de deux millions de documents ont été sélectionnés grâce aux termes MeSH de leurs index tels que "*bacterid*". D'autres études ont montré que l'utilisation de textes composés uniquement de titres et de résumés, plutôt que d'articles complets, améliore les résultats sur des tâches d'évaluations intrinsèques et extrinsèques de représentations distributionnelles (Chiu et al., 2016). Dans notre expérience, tous les caractères majuscules du corpus ont été mis en minuscules. L'objectif est de ne pas différencier certains mots sur cet aspect (comme le fait qu'ils soient en début de phrase ou non) et de limiter le nombre de mots possibles. Le Tableau 7 indique quelques caractéristiques de notre corpus d'entraînement des méthodes Word2Vec

Nombre de documents	2 333 943
Nombre de phrases	16 384 331
Nombre de mots	412 240 083
Nombres de mots (mots-outils exclus)	244 856 229
Vocabulaire (forme de surface des mots en minuscule)	1 565 740

Tableau 7 : Pour le filtrage des mots-outils, les mots suivants ont été enlevés du corpus (pour l'anglais) : les déterminants, les prépositions, les conjonctions, le mot "to", les caractères ':', '( et )'. Les catégories grammaticales ont été détectées par l'outil TreeTagger (Schmid, 1999). Pour la segmentation en phrase, l'outil SegMig<sup>16</sup> a été utilisé. Pour la segmentation en mots, l'outil WoSMig<sup>17</sup> a été utilisé.

<sup>15</sup> <https://www.nlm.nih.gov/bsd/medline.html>

<sup>16</sup> <https://bibliome.github.io/alvisnlp/reference/module/ScSMig>

<sup>17</sup> <https://bibliome.github.io/alvisnlp/reference/module/WoSMig>

### 4.3.3. Choix des hyper-paramètres de Word2Vec et de prétraitements

La méthode Word2Vec possède de nombreux paramètres pouvant prendre des valeurs variables. Pour choisir certains paramètres à tester, et fixer les autres, nous nous sommes inspirés de (Chiu et al., 2016), qui montre que plusieurs paramètres n'ont pas vraiment d'influence tant qu'ils restent dans un certain intervalle. Nous avons notamment fixé les paramètres indiqués dans le Tableau 8 dans toutes nos expériences.

Paramètres	Valeurs
Taille de l'échantillonnage négatif	5
Sous-échantillonnage	0,001
Pas d'apprentissage	0,025
Nombre minimum d'occurrences	0

Tableau 8 : Paramètres par défaut conservés dans nos expérimentations

Notamment, le nombre minimum d'occurrences qu'un mot doit avoir pour être pris en compte a été mis à 0. En effet, ce paramètre n'a apparemment qu'un impact limité, mais permet de conserver des mots de faible fréquence d'apparition, potentiellement d'intérêt. Cela permet donc de limiter également l'impact du problème des mots hors-vocabulaire. L'inconvénient principal est que cela augmente la taille du vocabulaire, ce qui rend la manipulation des données plus difficile et augmente les temps de calcul. Pour utiliser Word2Vec sur d'autres corpus d'entraînement plus volumineux, d'autres expérimentations devraient être conduites afin d'observer l'impact de ce paramètre.

Selon ces mêmes travaux, les paramètres les plus importants semblaient être la taille de la fenêtre contextuelle utilisée et la taille des vecteurs finaux. Les autres paramètres possibles, peu influents, ont été laissés à leur valeur par défaut. Nous avons également voulu observer l'impact de prétraitements du corpus :

- Avec ou sans filtrage des mots-outils (ou mots-vides) : certains mots ne sont pas vraiment significatifs, c'est-à-dire ne portent pas un sens important (au moins vis-à-vis de la tâche), tels que les déterminants. Pour tenter d'améliorer les mots contenus dans les fenêtres contextuelles utilisées, nous avons donc testé de filtrer ces mots dans le corpus d'entraînement. Les mots suivants ont été enlevés du corpus (pour l'anglais) : les déterminants, les prépositions, les conjonctions, le mot "to", les caractères ':', '(' et ')'. Les catégories grammaticales ont été détectées par l'outil TreeTagger (Schmid, 1999).
- Forme de surface, lemmatisation ou racinisation : le nombre de mots dans le vocabulaire est important, de l'ordre de plusieurs millions (nous n'avons enlevé aucun mot rare). Ce nombre rend les calculs plus longs et la manipulation des

fichiers plus difficile. Une réduction de la taille du vocabulaire a donc une première utilité. De plus, des mots porteurs d'un sens très proche sont traités différemment dans les contextes utilisés. Représenter différentes variations par un même représentant peut donc permettre d'enrichir les contextes. Remplacer les mots représentés par leur forme de surface par leur lemme ou leur racine pourrait répondre à ces deux enjeux.

- Masquage de certains types d'entités, générales (les nombres) ou spécifiques au domaine (microorganismes et mots-clés permettant de décrire qu'une bactérie vit dans un habitat) : de manière similaire à l'application d'une lemmatisation ou racinisation, normaliser directement certains mots ou expressions pourrait enrichir les contextes. Des normalisations spécifiques à au domaine pourraient notamment permettre d'améliorer la représentation des termes du domaine.

De manière générale, ces prétraitements permettent de diminuer le vocabulaire du corpus tout en limitant la perte d'information, ce qui facilite la manipulation du corpus d'entraînement et diminue les temps de calculs. De plus, ils permettent d'enrichir les contextes utilisés. Les mots-outils ne portent *a priori* que peu d'informations utiles pour caractériser un habitat bactérien, et les supprimer permet d'utiliser des mots qui n'auraient pas été pris en compte sans cela pour une même taille de fenêtre contextuelle. Les deux autres pré-normalisations permettent aussi d'avoir moins de formes dans le vocabulaire, mais surtout d'augmenter le nombre de leurs occurrences dans chaque contexte.

La méthode Word2Vec est une méthode neuronale qui commence par initialiser une matrice pseudo-aléatoire. Dans ce genre de cas, sur des évaluations extrinsèques notamment, on observe des variations importantes des scores en fonction de l'initialisation (Reimers and Gurevych, 2017). En effet, selon la matrice initiale, les résultats globaux sur la tâche de normalisation Bateria Biotope varient, et il est donc important de prendre cette variation en considération. Nous avons donc testé plusieurs jeux de paramètres avec différentes initialisations aléatoires. En faisant varier seulement la taille de la fenêtre contextuelle, le nombre d'itérations et le filtrage des mots-outils sur le corpus brut, nous avons obtenu un écart-type minimum de 0.56% pour un jeu de paramètre fixé et notamment pour une fenêtre de taille 5. Le nombre d'itérations ne semble pas avoir d'impact significatif. Une valeur maximale d'écart-type de 0.011 a été obtenue pour une fenêtre de taille 2 et un filtrage des mots-outils. En conséquence, si un score obtenu pour un jeu de paramètres est 0.03 points au-dessus de celui obtenu pour un autre jeu, il est inutile de démontrer la supériorité significative du premier jeu sur le second.

Le Tableau 9 présente les résultats de l'ensemble des expérimentations effectuées. Deux paramètres ont un impact important sur la performance de la normalisation : l'utilisation de la racine à la place de la forme de surface des mots, et le filtrage des mots-outils avant d'entraîner Word2Vec. L'utilisation de la racine dégrade systématiquement les résultats (en moyenne 3 points de moins). L'utilisation des formes de surface ou des lemmes n'a pas permis de distinguer une nette différence de performance. A l'inverse, le filtrage des

mots-outils semble augmenter significativement les résultats, à l'inverse de certains travaux sur la classification en domaine biomédical (Agarwal and Yu, 2009).

Filtrage mots-outils	Forme utilisée	Précision moyenne
oui	forme de surface	0,60
oui	lemme	0,59
non	forme de surface	0,58
non	lemme	0,58
oui	racine	0,57
non	racine	0,57

Tableau 9 : Impact de l'utilisation des formes de surfaces des mots comparée à l'utilisation de leur lemme ou de leur racine. Chaque expérience est déclinée en version avec filtrage des mots-outils ou non. Ces expériences sont réalisées pour une taille de vecteur de 200 et une fenêtre de taille 2.

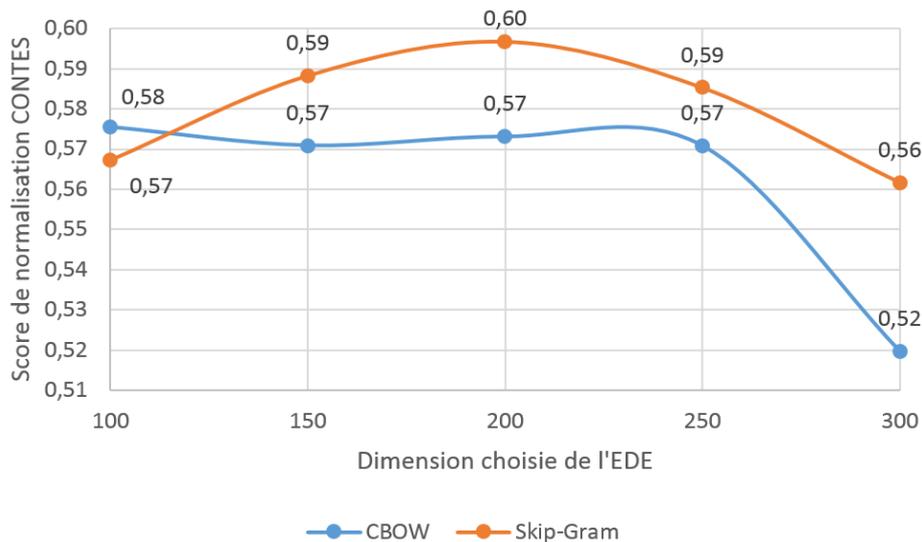


Figure 31 : Comparaison entre les architectures CBOW et Skip-Gram de Word2Vec en fonction de la taille des vecteurs (sur corpus non-filtré, taille fenêtre 2).

La valeur par défaut de la taille des représentations est de 100, et les valeurs conseillées sont souvent entre 100 et 300. Celle de 200 est la meilleure pour la représentation d'expressions biomédicales dans les travaux de (Muneeb et al., 2015), et cela quelle que soit la méthode testée (CBOW, Skip-Gram ou GloVe) ou l'évaluation, intrinsèque ou extrinsèque. Dans les travaux de (Chiu et al., 2016), quelle que soit la taille du corpus utilisé, une valeur de 200 est toujours celle qui obtient les meilleurs résultats, avec un écart vraisemblablement peu significatif néanmoins. Dans nos expérimentations, une valeur de

200 semble être une dimension adaptée à la tâche aussi bien pour l'architecture Skip-Gram que l'architecture CBOW (voir Figure 31). Néanmoins, l'architecture Skip-Gram a produit des résultats globalement meilleurs pour notre tâche de normalisation, et en particulier pour la valeur optimale de 200.

Comme observé dans certains travaux, la taille de la fenêtre contextuelle n'a pas d'impact réellement significatif dans le cas d'évaluation extrinsèque (voir Tableau 10). Une légère amélioration est obtenue avec une taille de fenêtre de 12 mots pour le corpus filtré de ses mots-outils (0,64 au lieu de 0,62), mais la confirmation de l'amélioration nécessitera une étude statistique plus approfondie de la variabilité des résultats. L'absence d'influence de la taille de la fenêtre s'observe également pour un corpus dont les mots-outils n'ont pas été filtrés. Néanmoins, dans ce cas, la taille de 12 mots n'est plus significativement la meilleure (0,60 au lieu de 0,61 avec une fenêtre de 2 ou 3 mots).

Taille fenêtre	Score (corpus filtré)	Score (corpus non-filtré)
1	0,62	0,59
2	0,60	0,61
3	0,61	0,61
5	0,60	0,60
8	0,61	0,59
12	0,64	0,60
20	0,62	0,60

Tableau 10 : Test de l'influence de la taille de la fenêtre contextuelle sur les résultats de la tâche de normalisation de *Bacteria Biotope*, pour une taille de vecteur de 200 et un corpus filtré de ses mots-outils.

En conséquence, une petite taille de fenêtre de 2 mots a été retenue pour la plupart des expérimentations, notamment pour diminuer les temps de calcul. Aucun effet significatif des pré-normalisations (des nombres, des microorganismes et des mots-clés permettant de décrire qu'une bactérie vit dans un habitat), positif ou négatif, n'a pu être observé expérimentalement. Le nombre des modifications apportées au corpus n'est peut-être simplement pas assez important pour avoir un effet notable sur le calcul distributionnel.

#### 4.3.4. Construction de plongements lexicaux pour des expressions multi-mots

Après avoir produit des représentations vectorielles pour chacun des mots du corpus (voir section précédente), il est nécessaire de passer par une étape de production de représentations vectorielles d'expressions multimots, qui sont des mentions à normaliser des corpus d'entraînement et d'évaluation pour la tâche. Pour cela, la première étape est de segmenter ces expressions en différents mots. Ensuite, chaque mot composant une expression et possédant une représentation vectorielle va être utilisé pour calculer une représentation de cette expression.

L'objectif est de représenter les expressions par des vecteurs denses et de faible dimension, tout comme les plongements lexicaux. De plus, ces vecteurs sont à représenter dans le même espace vectoriel que ceux des plongements, notamment pour permettre l'étude de leur localisation par rapport à l'ensemble des mots initialement représentés.

La méthode de compositionnalité de la moyenne (voir formule (8)) a été choisie. Si une expression est composée de mots qui n'ont pas de représentation vectorielle, ce qui est le cas des mots hors-vocabulaire, alors ceux-ci ne sont pas pris en compte et la moyenne est faite avec le reste des mots qui la compose.

## 4.4. Bilan

Nous venons de présenter les méthodes utilisées par la méthode CONTES pour construire :

- Des représentations de mentions denses et de faible dimension dans un espace distributionnel des expressions (EDE),
- Des représentations creuses de concepts dans un espace sémantique de l'ontologie (ESO).

L'objectif principal visé était de construire un EDE avec de fortes similarités structurelles avec l'ESO. Rappelons que notre hypothèse est qu'une similarité structurelle globale entre les deux espaces permettrait de concevoir un alignement des vecteurs sémantiquement interprétable. Pour estimer la qualité de cette similarité, nous avons directement évalué notre méthode sur la tâche de normalisation Bacteria Biotope. Nous avons identifié les paramètres de la méthode qui produisent les meilleurs scores pour la tâche.

Notamment pour l'EDE, nous avons déterminé une influence significativement positive de la conservation des formes de surface des mots du vocabulaire et d'une dimension des vecteurs égale à 200. Pour l'ESO, nous avons pu constater qu'une réduction de dimension par ACP dégrade rapidement les performances de la méthode. Nous avons donc choisi de conserver l'ESO non-réduit. Pour une utilisation de la méthode CONTES avec une ontologie de l'ordre de 2000 concepts, cela ne soulève pas de difficultés. Néanmoins, la méthode pourrait en rencontrer si elle utilisait une ontologie d'un ordre de grandeur plus élevé.

## **Chapitre 5**

-

### **Méthodes de normalisation d'entités**



## 5.1. Introduction

La méthode CONTES (“*CONcept-TErm System*”) est fondée sur la projection d’un espace distributionnel des expressions (EDE) dans un espace de référence pour une tâche précise de normalisation fondée sur une ontologie, l’espace sémantique de l’ontologie (ESO). La similarité structurelle entre les deux espaces dont nous faisons l’hypothèse, nous permet de rechercher une projection linéaire. L’objectif de l’approche est d’exploiter cette similarité globale entre les deux espaces pour compenser la faible quantité de données d’entraînement.

Pour déterminer une projection linéaire pertinente, notre méthode visera à apprendre à projeter toutes les représentations de mentions au plus près des représentations des concepts qui leur sont associés (voir Figure 32). En conséquence, nous avons opté pour une régression linéaire multivariée qui est entraînée sur des exemples annotés.

Une fois la fonction de projection linéaire déterminée à partir des exemples annotés, nous pouvons l’appliquer aux représentations des mentions non-annotées d’un EDE. Dans l’ESO, CONTES recherche alors le concept le plus similaire pour chaque mention. Pour cela, un score de similarité est calculé entre chaque mention et chaque concept. Pour chaque mention, le concept prédit est alors celui qui obtient le score le plus élevé (voir Figure 33).

Ce chapitre présente plus en détail la méthode CONTES, ainsi que les résultats obtenus sur la tâche de normalisation Bacteria Biotope, puis nous énumérerons plusieurs limitations auxquelles nous répondons par plusieurs méthodes alternatives à CONTES présentées ici. L’une des méthodes présentées, HONOR (“Hierarchical Ontological Normalization”), a été publiée dans les actes de la conférence LREC (“The International Conference on Language Resources and Evaluation”). Elle a donné lieu à une présentation orale. Une partie de ces travaux a été réalisée en collaboration avec Louise Deléger (MaIAGE, INRA), qui a effectué les analyses avec l’outil ToMap afin de les combiner à la méthode CONTES (pour la mise en place de la méthode HONOR, notamment). Nous avons également collaboré avec Robert Bossy pour mettre en œuvre les expérimentations que j’ai conçues pour déterminer l’impact de différentes configurations, hyperparamètres et prétraitements du corpus d’apprentissage. Nous avons bénéficié des importants moyens de calcul et de stockage de la plateforme Migale<sup>18</sup>.

---

<sup>18</sup> <http://migale.jouy.inra.fr/>

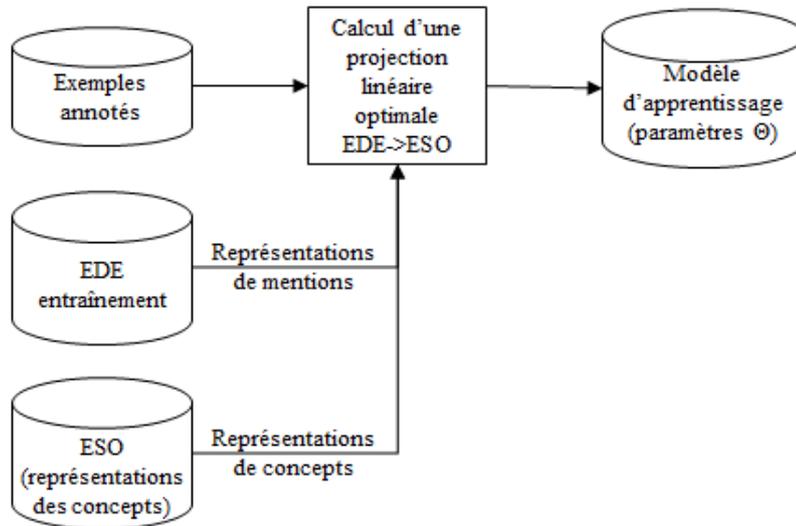


Figure 32 : Schéma de l'étape d'entraînement de la méthode CONTES.

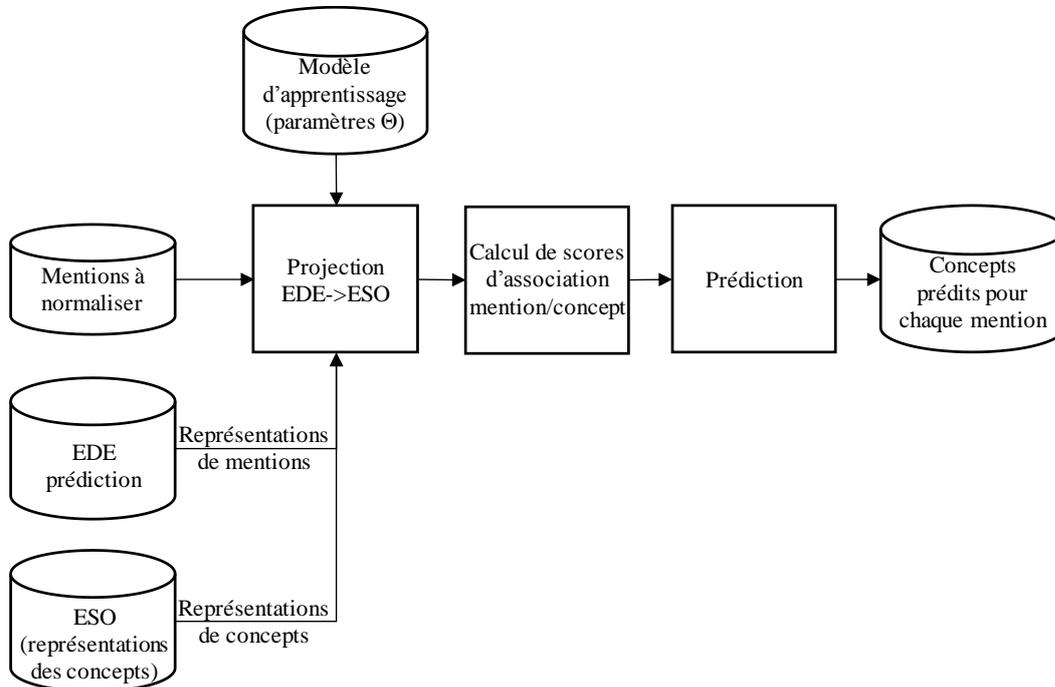


Figure 33 : Schéma de l'étape de prédiction de la méthode CONTES. Après la projection des représentations de mentions dans l'ESO, un score est calculé entre chaque association de mentions et de concepts par une mesure de similarité cosinus. Pour chaque mention, le concept qui a obtenu le score le plus élevé est celui prédit par la méthode.

## 5.2. Apprentissage d'une projection conservant la structure de l'espace source et normalisation

### 5.2.1. Introduction

Le chapitre précédent a détaillé les méthodes de construction des deux espaces sémantiques d'intérêt pour la méthode CONTES, à savoir :

- L'espace distributionnel des expressions (EDE) : espace sémantique contenant des plongements lexicaux pour tous les mots du corpus d'entraînement et pour toutes les mentions d'entités considérées.
- L'espace sémantique ontologique (ESO) : espace sémantique contenant les représentations vectorielles des concepts de l'ontologie et portant l'information hiérarchique de l'ontologie.

Notre objectif est à présent de construire une projection dans l'ESO des vecteurs de mentions issus d'un jeu de données d'entraînement. Pour cela, nous voulons apprendre une fonction telle que l'ensemble des vecteurs de mentions projetés se retrouvent à proximité des vecteurs de leur concept associé. Cette fonction doit ensuite permettre de projeter les représentations de n'importe quelle mention à proximité du concept à prédire. Enfin, une fonction de score est utilisée pour classer l'ensemble des couples mention/concept. Les résultats de la méthode CONTES sont présentés et comparés aux autres méthodes évaluées sur la tâche de normalisation Bacteria Biotope. Ils sont publiés dans (Ferré et al., 2017; Ferré et al., 2018).

### 5.2.2. Détermination d'une projection linéaire entre espaces sémantiques

Nous souhaitons apprendre une fonction de projection telle que l'ensemble des vecteurs de mentions projetés soient à proximité des vecteurs de leur concept associé. Pour cela, nous avons retenu une fonction linéaire. L'intérêt principal étant que les fonctions linéaires de  $\mathbb{R}^n$  ( $n \in \mathbb{N}^*$ ) dans  $\mathbb{R}^m$  ( $m \in \mathbb{N}^*$ ) peuvent facilement être décrites par une matrice  $\theta \in \mathcal{M}_{m,n}(\mathbb{R})$ . C'est donc une transformation relativement simple, qui tendra à conserver certaines similarités structurelles entre l'ensemble des vecteurs initiaux et celui des vecteurs projetés (c'est-à-dire des combinaisons de translations, rotations, homothéties, réduction linéaire de dimension, etc.). Si l'espace distributionnel des expressions (EDE) est représenté par  $\mathbb{R}^n$  et si l'espace sémantique ontologique (ESO) est représenté par  $\mathbb{R}^m$ , alors cette fonction  $f_\theta$  est définie telle que :

$$f_{\theta}: \begin{array}{l} \mathbb{R}^n \rightarrow \mathbb{R}^m \\ x \mapsto f_{\theta}(x) = \theta \cdot x \end{array} \quad (30)$$

On cherche alors à déterminer la fonction permettant de minimiser la dissimilarité entre projetés de vecteurs de mentions et vecteurs de concepts associés. Cela revient à un problème de régression : nous essayons de déterminer la fonction  $f_{\theta}$  qui pour des représentations de mentions prédira globalement le mieux les représentations de concepts associés. Ce type de problème est un problème de régression linéaire multivariée (“*multivariate regression*”) (Borchani et al., 2015). On trouve également les termes anglais de “*multi-target regression*” ou “*multi-response regression*”. Ce type de régression permet en effet de prédire plusieurs variables continues en fonction de plusieurs variables continues explicatives.

Ce problème peut se résoudre en le décomposant pour chaque dimension de l'ESO par la méthode de la cible unique (Spyromitros-Xioufis et al., 2012). On obtient alors  $m$  problèmes de régression linéaire multiple, c'est-à-dire des problèmes de régression où l'on cherche à prédire une variable réelle à partir de plusieurs variables réelles explicatives.

Soit  $X = \{x_1, \dots, x_N\}$ , un ensemble de  $N$  représentations vectorielles sémantiques de mentions d'entités telle que  $\forall i \in \llbracket 1, N \rrbracket, x_i \in \mathbb{R}^n$  ( $n \in \mathbb{N}^*$ ). Soit  $Y = \{y_1, \dots, y_N\}$ , un ensemble de  $N$  représentations vectorielles sémantiques de concepts d'une ontologie telle que  $\forall i \in \llbracket 1, N \rrbracket, y_i = (y_1^i, \dots, y_m^i) \in \mathbb{R}^m$  ( $m \in \mathbb{N}^*$ ). Soit  $A = \{(x_1, y_1), \dots, (x_N, y_N)\}$ , l'ensemble des couples de représentations de mentions et de concepts représentant les exemples annotés pour l'entraînement de la méthode. Alors on peut créer  $m$  jeux de données d'entraînement  $A_q = \{(x_1, y_q^1), \dots, (x_N, y_q^N)\}$  (pour  $q \in \llbracket 1, m \rrbracket$ ). Nous cherchons à déterminer pour chacun de ces  $m$  jeux de données, un paramètre  $\theta \in \mathbb{R}^n$  optimal pour la fonction  $g_{\theta}$  définie telle que :

$$g_{\theta}: \begin{array}{l} \mathbb{R}^n \rightarrow \mathbb{R} \\ x \mapsto g_{\theta}(x) = \theta^{\top} \cdot x \end{array} \quad (31)$$

Pour résoudre chacun de ces problèmes de régressions, nous avons utilisé la méthode des moindres carrés ordinaire (MCO). La méthode MCO cherche à minimiser la fonction objectif  $L_q$  définie telle que :

$$L_q: \begin{array}{l} \mathbb{R}^n \rightarrow \mathbb{R} \\ \theta \mapsto L_q(\theta) = \sum_{i=1}^N |y_q^i - g_{\theta}(x_i)|^2 \end{array} \quad (32)$$

On cherche donc les  $m$  vecteurs  $\theta_q = \operatorname{argmin}_{\theta} (L_q(\theta))$  qui permettront de produire les  $m$  lignes de la matrice optimale  $\theta_f$  du problème de régression linéaire multivariée. Or, pour chaque jeu de données d'entraînement  $A_q$ , la fonction objectif  $L_q$  possède un minimum unique calculable analytiquement. La fonction  $f_{\theta_f}$  (voir (30)) obtenue permet

alors de calculer la projection de n'importe quel vecteur d'expression de l'espace distributionnel dans l'espace des concepts.

Pour effectuer ces calculs, nous utilisons la fonction "*LinearRegression*"<sup>19</sup> de la librairie Scikit-learn<sup>20</sup>.

Une projection inverse pourrait tout aussi bien être recherchée pour répondre à la tâche de normalisation (ESO vers EDE). Cette variante n'a pas été testée expérimentalement, car dans l'EDE, des vecteurs de mentions d'un même concept peuvent être très dispersés. Il nous semble donc moins évident d'interpréter le résultat d'une telle projection. De plus, notre objectif est de produire des représentations distributionnelles améliorées par des connaissances externes, ce qui ne peut être réalisé qu'en projetant les vecteurs à modifier dans l'espace construit à partir de ces connaissances.

### 5.2.3. Effets de la régression linéaire sur la projection

Une similarité structurelle entre l'espace distributionnel et l'espace ontologique est attendue. Néanmoins, des "défauts" sont possibles dans l'espace distributionnel. L'impact de ces défauts sur la qualité de la projection peut être étudié visuellement avec des espaces en deux dimensions. Nous avons construit plusieurs vecteurs fictifs d'expressions et de concepts pour représenter plusieurs cas afin d'illustrer des cas réels. Quelques expérimentations sont présentées en Figure 34, Figure 35 et Figure 36.

Notamment, le cas de la Figure 34 montre que si des vecteurs de mention d'un même concept sont fortement dispersés, alors cela a un impact global sur la qualité de la projection linéaire. Si ce phénomène est fréquent, alors la projection linéaire obtenue sur les données d'entraînement risque donc d'être fortement impactée. Néanmoins, nous supposons que les représentations distributionnelles que nous avons produites tendent à limiter ce phénomène.

---

<sup>19</sup> [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LinearRegression.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html)

<sup>20</sup> <https://scikit-learn.org/stable/index.html>

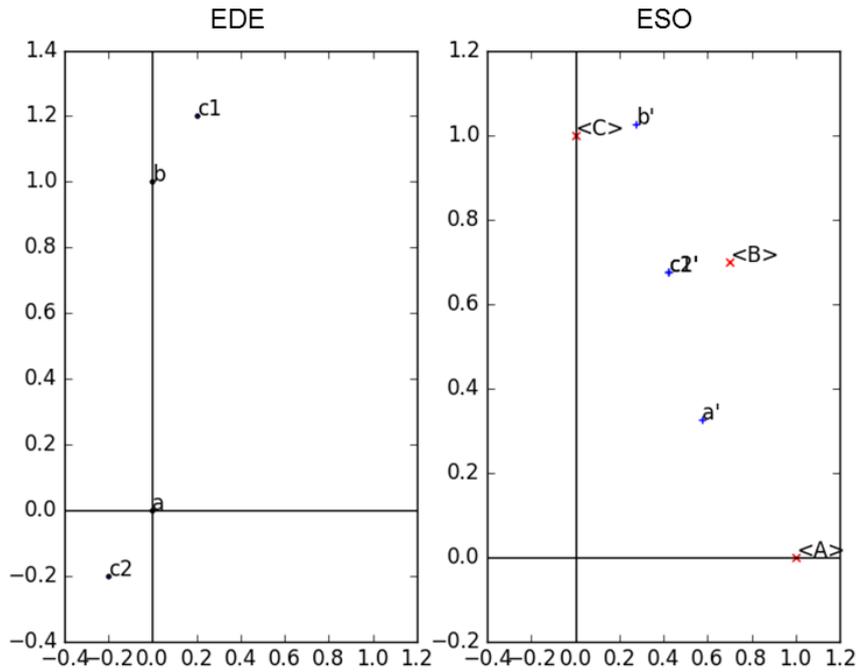


Figure 34 : La représentation de l'EDE n'est pas pertinente : deux points appartenant à la même catégorie sémantique sont séparés par d'autres points appartenant à d'autres catégories. La régression linéaire ne réussit pas à déterminer une projection pertinente.

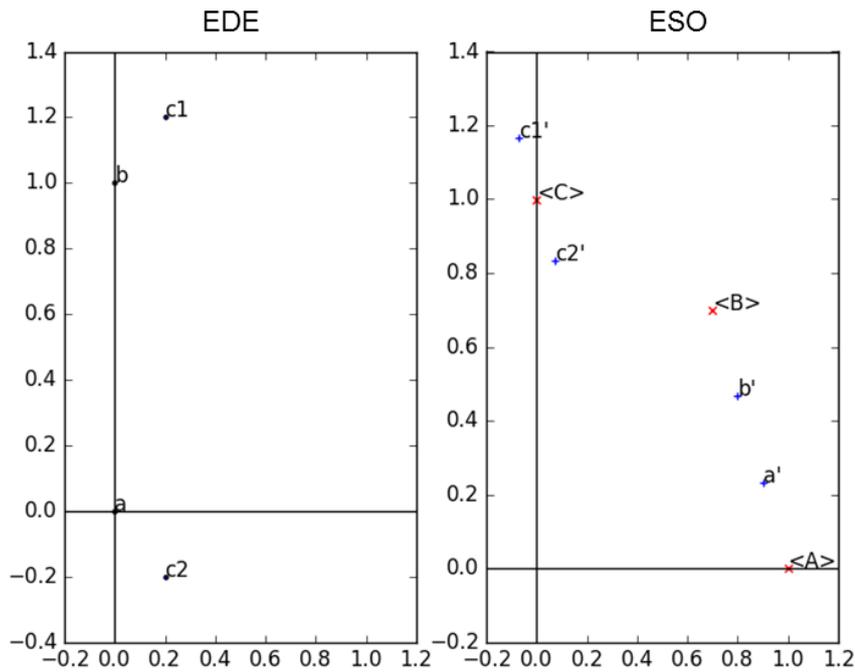


Figure 35 : La représentation de l'EDE est contestable : c2 est plus proche de a que de b ou c1 (qui appartiennent pourtant à des catégories sémantiques plus proches). Néanmoins, tous les points appartenant à une même catégorie, et uniquement eux, peuvent facilement être contenus dans une même forme convexe. La régression linéaire réussit à déterminer une projection pertinente.

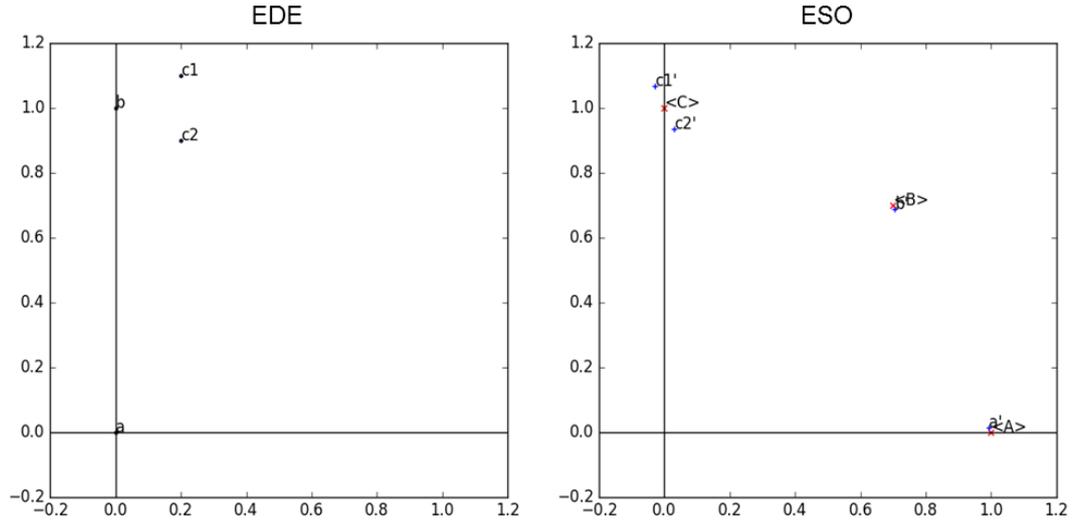


Figure 36 : La représentation de l'EDE est relativement pertinente (même si  $a$  est plus proche de  $c2$  en distance euclidienne que de  $b$ ). La régression linéaire réussit à déterminer une projection pertinente.

#### 5.2.4. Prédiction de normalisation par calcul de distance

La fonction de projection apprise par entraînement peut être ensuite appliquée à n'importe quel vecteur de l'espace sémantique des expressions. Notamment, en calculant une représentation vectorielle pour chaque mention d'intérêt d'un nouveau corpus, on peut alors calculer sa projection dans l'espace ontologique. Pour prédire un concept pour normaliser une mention, la méthode retient celui dont le vecteur est le plus proche du projeté en terme de distance. Pour cela, nous avons retenu la distance cosinus, qui expérimentalement a produit de meilleurs résultats que ceux obtenus avec une distance euclidienne.

Nous pouvons donc définir notre fonction de score telle que :

$$\mathbb{R}^{|T|} \times \mathbb{R}^{|\Omega|} \rightarrow \mathbb{R}$$

$$score_{A,b}: (x, y) \mapsto score_{A,b}(x, y) = \frac{f_{A,b}(x)^\top \cdot y}{\|f_{A,b}(x)\| \cdot \|y\|} \quad (33)$$

Avec cette méthode de prédiction, la méthode CONTES initiale a obtenu des résultats encourageants proches de ceux de l'état de l'art en 2017 sur la tâche de normalisation d'habitat bactérien Bacteria Biotope (voir Tableau 11). Les représentations de concepts construits à partir de la méthode Ancestry produisent une amélioration significative du score par rapport à l'utilisation de représentations 1-Parmi-N (+7 points). L'approche a permis le développement d'une autre méthode, HONOR (*Hierarchical Ontological*

*Normalization*'), qui représente aujourd'hui l'état de l'art sur ce même jeu de données (Ferré et al., 2018).

En se fondant sur l'approche de la méthode CONTES, nous avons développé d'autres méthodes par la suite qui sont décrites dans la section suivante.

Méthode	Score
<b>HONOR</b>	<b>0,74</b>
ToMap (avec règles spécifiques au domaine)	0,66
Turku (Mehryary et al., 2017)	0,63
ToMap	0,62
BOUN (Tiftikci et al., 2016)	0,62
<b>CONTES</b>	<b>0,61</b>
Baseline	0,54
<b>1-Parmi-N/CONTES</b>	<b>0,54</b>
(Grouin, 2016)	0,44

Tableau 11 : Résultats de la méthode CONTES sur Bacteria Biotope pour une taille des vecteurs de mentions de 200. La méthode 1-Parmi-N/CONTES est l'utilisation de représentations 1-Parmi-N à la place des représentations construites par la méthode Ancestry (pour les mêmes paramètres de l'EDE que pour CONTES).

## 5.3. Expérimentations avec différentes sources d'exemples

### 5.3.1. Introduction

La méthode CONTES a été présentée dans les chapitres précédents. Nous avons développé d'autres méthodes complémentaires s'appuyant sur elle pour répondre à certaines limitations. Deux principales limitations ont été identifiées :

- Dépendance à des exemples annotés (même peu nombreux) : la méthode CONTES est une méthode par apprentissage, elle requiert donc des exemples annotés. Dans des domaines de spécialité, produire ces exemples reste toujours un effort important.

- Non-utilisation des étiquettes de concepts : La méthode CONTES n'utilise pas les étiquettes des concepts de l'ontologie de référence, mais seulement la structure de son graphe de connaissances. Or, ces étiquettes forment une information précieuse, utilisée comme principale source d'information dans de nombreuses autres méthodes de normalisation d'entités (Aronson and Lang, 2010; Golik et al., 2011).

Pour répondre à ces limitations, trois approches principales ont été étudiées :

- Approche par apprentissage faible : l'utilisation d'exemples annotés de plus faible qualité pour l'entraînement d'une méthode par apprentissage. Les méthodes suivantes ont été développées :
  - WSOT-CONTES (*“Weak Supervised on Ontology Tag CONcept-TErm System”*) : Utilisation des étiquettes des concepts de l'ontologie comme mentions possibles de concepts.
  - WSEP-CONTES (*“Weak Supervised on External Predictions CONcept-TErm System”*) : Utilisation des prédictions produites par une autre méthode, ToMap (Golik et al., 2011), sur un corpus non-annoté.
  - Full-CONTES (*“Supervised and Weak Supervised on Ontology Tag CONcept-TErm System”*) : La combinaison de CONTES et de WSOT-CONTES, c'est-à-dire l'utilisation à la fois d'exemples annotés manuellement et des étiquettes de concepts de l'ontologie.
- Approche mixte par intégration d'une méthode fondée sur l'étude de similarité de forme. La méthode HONOR (*“Hierarchical Ontological Normalization”*) a été développée en intégrant la méthode ToMap à CONTES.
- Approche intégrant les deux approches précédentes, c'est-à-dire de l'apprentissage faiblement supervisé et l'intégration d'une méthode fondée sur l'étude de similarité de forme. La méthode WSEP-HONOR (*“Weak Supervised on External Predictions Hierarchical Ontological Normalization”*) a été développée, qui est une version de la méthode HONOR en supervision faible grâce à l'utilisation des prédictions de la méthode ToMap, sur un corpus non-annoté.

### 5.3.2. Paramètres expérimentaux

Tous les résultats de cette section ont été obtenus avec les mêmes paramètres pour permettre leur comparaison. Notamment, pour l'entraînement de Word2Vec, les paramètres sont :

Paramètres	Valeur
architecture	skip-gram
alpha	0,05
min-count	0
negative	5
sample	0,001
size	200
window	2

Tableau 12 : Paramètres sélectionnés pour les évaluations comparatives de ce chapitre

De plus, un filtrage des mots-outils a été systématiquement appliqué.

### 5.3.3. Approche mixte par intégration d'une méthode fondée sur l'étude de similarité de forme : HONOR

Les deux versions de ToMap (Golik et al., 2011) seules obtiennent déjà des scores importants sur la tâche de normalisation de Bacteria Biotope, relativement aux autres méthodes évaluées (voir Tableau 2), mais ToMap ne réussit à proposer une prédiction de normalisation que pour environ la moitié des mentions d'habitat bactérien, quel que soit le jeu de données de la tâche Bacteria Biotope testé. Pour les autres mentions, pour lesquelles la méthode ne réussit pas à prédire directement un concept par similarité de forme, le concept racine est choisi. Cela est principalement dû à des variations importantes de formes entre mentions et étiquettes. A l'opposé, les approches distributionnelles sur lesquelles s'appuie CONTES visent à dépasser le problème de la variabilité de forme. Pour prédire, CONTES n'utilise d'ailleurs pas les étiquettes des concepts de l'ontologie. Les deux approches sont donc très différentes et complémentaires, et des performances différentes sont également attendues. ToMap a une bonne précision, c'est-à-dire qu'un concept prédit est souvent le bon concept, mais que dans les autres cas, aucun concept n'est prédit. Une première approche pour exploiter ces différences est de :

- Conserver toutes les prédictions d'une méthode fondée sur la similarité de forme telle que ToMap. On attend de ces prédictions qu'elles soient relativement précises.
- Puis de produire des prédictions avec CONTES pour toutes les mentions qui n'ont pas pu être normalisées par la première méthode.

Méthode	Score
HONOR (ToMap avec règles)	<b>0,74</b>
HONOR (ToMap sans règle)	0,69

Tableau 13 : Résultats de la méthode HONOR avec une taille de vecteur de 200.

La méthode HONOR (“*Hierarchical Ontological Normalization*”) suit cette approche en intégrant les résultats de la méthode ToMap avec ceux de la méthode CONTES.

Les résultats publiés ont placé HONOR, avec la version de ToMap contenant des règles spécifiques au domaine, comme la meilleure méthode à l'état de l'art en 2018 sur la tâche de normalisation Bacteria Biotope (voir Tableau 13). De plus, avec la version initiale de ToMap (voir 3.2.4), HONOR atteint également d'excellentes performances avec un score au-dessus de ToMap et de CONTES.

### 5.3.4. Approche par apprentissage supervisé faible

#### 5.3.4.1. WSOT-CONTES : Utilisation des étiquettes des concepts de l'ontologie

CONTES ne s'appuie que sur la structure de l'ontologie. Néanmoins, les étiquettes de concepts représentent une source d'informations, le plus souvent de qualité, présente dans une ontologie. Notamment, une étiquette représente en pratique une possible mention d'un concept dans un texte. Pris indépendamment, une étiquette devrait donc être normalisée par son concept. Si l'ontologie a été construite avec des étiquettes pertinentes pour chaque concept, alors elle contient au moins un exemple d'association mention/concept par concept.

Ce jeu de données représente donc un intérêt considérable :

- Il fournit un exemple d'association pour chaque concept de l'ontologie. Inversement, pour la tâche Bacteria Biotope, l'ensemble des associations des données d'entraînement du texte ne portait que sur 15% des concepts de l'ontologie OntoBiotope.

- Il fournit un nombre d'associations relativement important. Pour la tâche Bacteria Biotope, le corpus d'entraînement contient environ 1200 associations, alors que l'ontologie Ontobiotope contient 2320 concepts, soit environ le double d'associations en utilisant seulement les étiquettes.

WSOT-CONTES (*“Weak Supervised on Ontology Tag CONcept-TErm System”*) s'appuie uniquement sur les associations étiquette/concept présentes dans l'ontologie pour apprendre une projection entre les deux espaces sémantiques. Les scores de WSOT-CONTES obtenus sur la tâche de normalisation d'habitat bactérien Bacteria Biotope sont présentés dans le Tableau 14.

Taille des vecteurs	CONTES	WSOT-CONTES
100	0,60	0,49
200	<b>0,61</b>	0,55
300	0,59	0,57
500	0,49	0,61
1000	0,40	<b>0,63</b>

Tableau 14 : Scores de la méthode WSOT-CONTES en fonction des associations étiquette/concept utilisées issues des concept d'OntoBiotope. Ces scores sont déclinés par rapport à la taille des vecteurs utilisée. WSOT-CONTES est comparée à CONTES, qui ne s'appuie que sur les associations mention/concept issues du corpus annoté. Expérience réalisé avec une taille de fenêtre de 2.

Si l'on observe les résultats seulement pour une taille de vecteurs de 200, l'utilisation du corpus manuellement annoté reste la source d'une meilleure performance. Néanmoins, avec l'augmentation de la taille des vecteurs, les associations étiquette/concept gagnent en performance, à l'inverse des associations issues du corpus annoté dont les résultats se dégradent nettement. Avec une taille des vecteurs de 1000, WSOT-CONTES a légèrement dépassé CONTES (0.63 contre 0.61). Nous pouvons émettre l'hypothèse qu'à partir d'une taille de vecteurs élevée, l'augmentation du nombre de dimensions nécessite une augmentation du nombre d'associations utilisant des concepts différents de l'ontologie. Avec l'augmentation du nombre de dimensions, il y a plus de possibilités de projection, et donc besoin de plus de contraintes pour guider cette projection. Ces contraintes sont plus satisfaites par les associations étiquette/concept.

#### 5.3.4.2. WSEP-CONTES : Utilisation des prédictions d'une méthode fondée sur la similarité de forme

A la place d'un corpus manuellement annoté, les résultats d'une autre méthode de normalisation peuvent être utilisés pour entraîner une méthode telle que celle présentée précédemment. Ainsi, si l'autre méthode utilisée ne se fonde pas sur de l'apprentissage supervisé, cette approche permet alors de se passer de données manuellement annotées. L'idée est de générer ainsi plus de données d'entraînement que ce qui pourrait être disponible par annotation manuelle, en espérant que cette quantité plus importante de données, de qualité moindre, permettra néanmoins de conserver une performance

similaire qu'avec l'utilisation seule d'annotations manuelles. Le choix d'une méthode non-supervisée initiale est importante : plus la méthode pourra générer des exemples corrects, plus notre approche devrait produire des résultats pertinents.

La méthode WSEP-CONTES (*“Weak Supervised on External Predictions CONcept-TERM System”*) suit cette approche en utilisant la méthode ToMap pour produire des exemples annotés pour l'apprentissage. Les scores de la méthode WSEP-CONTES sont à interpréter comparativement au rappel de la méthode ToMap qui ne réussit à produire une prédiction que pour environ la moitié des mentions d'habitats bactériens du jeu de données de développement. Les scores importants montrent donc une grande précision de la méthode, laissant envisager que ses prédictions de normalisation sont de qualité. La méthode a donc un potentiel à fournir automatiquement un nombre important de données d'une relativement bonne qualité, qui pourront servir à entraîner CONTES.

Nombre de données sélectionnées aléatoirement	Moyenne des scores	Ecart-type
100 000	0,59	0,005
10 000	0,57	0,012
1 000	0,54	0,016

Tableau 15 : Etude de l'impact du nombre de prédictions de ToMap (version sans règle spécifique au domaine) sélectionnées aléatoirement et sur 10 exécutions.

ToMap a été utilisé sur un sous-ensemble de notre corpus d'entraînement pour le calcul de plongements lexicaux, contenant plus de 7 000 000 phrases, et comportant plus d'un million d'expressions normalisées. Des groupes d'associations expression/concept aléatoirement sélectionnées sont alors formés avec des tailles différentes afin de tester l'influence du nombre d'exemples d'entraînement sur la performance de la méthode. Les plus petits jeux de données possèdent 1000 associations sélectionnées aléatoirement, ce qui est de l'ordre de grandeur du jeu de données fourni par l'annotation manuelle. Au-dessus de 100 000 associations, la complexité algorithmique de la régression linéaire par la méthode des moindres carrés rend le calcul très coûteux.

Les résultats de la méthode WSEP-CONTES sont présentés dans le Tableau 15. On peut observer que l'augmentation du nombre de prédictions de ToMap utilisées augmente également la performance des résultats, tout en les stabilisant quel que soit l'échantillon de prédictions. La méthode se rapproche de la performance de la méthode CONTES, tout en se passant donc d'exemples annotés manuellement.

### 5.3.4.3. Full-CONTES : Utilisation des étiquettes et d'exemples manuellement annotés

La méthode Full-CONTES s'appuie à la fois sur des données issues d'une annotation manuelle et sur les associations étiquette/concept de l'ontologie pour déterminer une projection pertinente entre les deux espaces sémantiques. Elle reprend les mêmes principes de la méthode WSOT-CONTES, en incluant également la prise en compte d'associations mention/concept issues d'un corpus manuellement annoté. Les scores de Full-CONTES obtenus sur la tâche de normalisation d'habitat bactérien Bacteria Biotope sont présentés dans le Tableau 16.

Taille des vecteurs	CONTES	WSOT-CONTES	Full-CONTES
100	0,60	0,49	0,58
200	<b>0,61</b>	0,55	0,62
300	0,59	0,57	0,65
500	0,49	0,61	0,67
1000	0,40	<b>0,63</b>	<b>0,72</b>
2000	N.D.	0.38	0.47
3000	N.D.	0.39	0.46

Tableau 16 : Scores de la méthode Full-CONTES en fonction des associations mention/concept utilisées issues des concepts d'OntoBiotope et du corpus annoté. Ces scores sont déclinés par rapport à la taille des vecteurs utilisée. Full-CONTES est comparée à CONTES (entraîné seulement sur des exemples annotés manuellement) et à WSOT-CONTES (entraîné seulement sur l'ontologie). N.D. = non-disponible.

Pour une taille des vecteurs de 200, l'utilisation des deux sources de données d'entraînement augmente légèrement le score global. L'ajout des associations étiquettes/concept semble donc améliorer la pertinence du jeu de données d'entraînement. De même qu'avec WSOT-CONTES, les scores de la méthode augmentent avec l'augmentation de la taille des vecteurs (au moins jusqu'à une taille de vecteur de 1000). De plus, les scores obtenus par Full-CONTES sont au minimum 5% supérieurs à ceux de WSOT-CONTES quelle que soit la taille de vecteur utilisée. Avec une taille de vecteur de 1000, Full-CONTES obtient un score de 0.72, ce qui représente le meilleur score des méthodes actuelles de type CONTES. Cela montre une certaine capacité de la méthode à évoluer avec la qualité des jeux de données avec lesquels elle est entraînée.

### 5.3.5. Approche mixte et par apprentissage faiblement supervisé : WSEP-HONOR

Nombre de données sélectionnées aléatoirement	Moyenne	Ecart-type
100 000	0,66	0,001
10 000	0,66	0,003
1 000	0,64	0,007

Tableau 17 : Scores de la méthode WSEP-HONOR en fonction du nombre de prédictions de la méthode ToMap sélectionnées (version sans règle) sur 10 exécutions.

La méthode WSEP-HONOR (*Weak Supervised on External Predictions Hierarchical Ontological Normalization*) est fondée sur l'approche HONOR, à la différence qu'elle commence par produire un nombre important d'associations mention/concept dans un corpus non-annoté grâce à l'utilisation d'une méthode non-supervisée. Ce sont ces associations qui serviront à entraîner la méthode par alignement d'espaces sémantiques, plutôt que des associations issues d'un corpus manuellement annoté.

WSEP-HONOR intègre également ToMap et CONTES de manière similaire à HONOR. De même qu'avec WSEP-CONTES, ToMap a été utilisé sur un corpus inclus dans le corpus d'entraînement de Word2Vec contenant plus de 7 000 000 de phrases, et rapportant plus d'un million de prédictions d'expressions normalisées. Parmi ces prédictions, des groupes d'associations expression/concept aléatoirement sélectionnées ont été formés avec des tailles différentes. Les résultats obtenus sur la tâche de normalisation de Bacteria Biotope sont présentés dans le Tableau 17 pour l'utilisation de la version de ToMap sans règle, et dans le Tableau 18 pour l'utilisation de celle avec règles.

Nombre de données sélectionnées aléatoirement	Moyenne	Ecart-type
100 000	0,72	000,1
10 000	0,72	0,003
1 000	0,70	non disponible

Tableau 18 : Etude du nombre de données sélectionnées (version avec règles), sur 10 exécutions.

Comme attendu, les résultats avec la version de ToMap adaptée à la tâche sont globalement meilleurs qu'avec la version initiale. Les résultats sont également moins élevés que ceux de la méthode HONOR, qui elle, a bénéficié d'exemples de référence. Le score évolue positivement avec l'augmentation du nombre d'associations expression/concept, mais semble se stabiliser autour d'un palier entre 1% et 3% en dessous des résultats d'HONOR. Un corpus manuellement annoté semble donc rester pour l'instant une meilleure ressource à utiliser quand elle est disponible.

Une diminution importante de l'écart-type est également observée par rapport à la méthode WSEP-CONTES et cela quel que soit la version de ToMap utilisée. Cela peut s'expliquer en partie par le fait que la moitié des mentions à prédire le sont par ToMap, qui ne s'appuie pas sur des traitements stochastiques.

La méthode WSEP-HONOR avec ToMap sans règle produit des résultats équivalents à ceux de la méthode ToMap avec règles seule (0.66), et cela dès 10 000 associations. L'approche WSEP-HONOR sans règle représente donc pour la tâche Bacteria Biotope, l'état de l'art des méthodes *a priori* non spécifiques au domaine et ne nécessitant pas d'exemples annotés manuellement, ce qui la rend prometteuse pour d'autres tâches.

## 5.4. Discussion et explorations

Les méthodes présentées précédemment abordent plusieurs difficultés de la normalisation, telles que le problème de variation de forme entre mentions et étiquettes, l'adaptabilité à de nouvelles tâches de normalisation ou encore la rareté, voire l'absence, d'exemples annotés. D'autres difficultés persistent néanmoins, telles que :

- Le traitement des mots hors-vocabulaire : certains mots appartenant au vocabulaire du corpus à normaliser peuvent ne pas appartenir au corpus d'entraînement de Word2Vec. En conséquence, ces mots n'ont pas de représentations. Si des mentions à normaliser contiennent des mots hors-vocabulaire, leurs représentations seront peu exploitables, voire inexistantes.
- La dimension importante de l'ESO : La taille des représentations dans l'ESO est égale au nombre de concepts dans l'ontologie. Pour l'ontologie OntoBiotope, cela représente une taille de 2320, mais d'autres ontologies classiques du domaine biomédical peuvent avoir une taille d'un ordre de grandeur de 10 à 1000 fois supérieure (par exemple le méta-thésaurus biomédical "*Unified Medical Language System*" et ses plus de 5 millions de concepts). Cela pourrait entraîner des problèmes de calculabilité.
- La faible similarité structurelle entre EDE et ESO : les méthodes présentées dans ce chapitre s'appuient sur l'hypothèse que si les deux espaces sémantiques utilisés ont des structures proches, alors on peut réussir à les aligner de façon pertinente. Même si les résultats semblent indiquer une certaine similarité, celle-ci peut être relativement faible (différence entre informations distributionnelles d'un corpus et informations ontologiques, différence de domaines, limitations des

représentations, ...). En conséquence, une projection linéaire peut éprouver des difficultés à déterminer une projection pertinente.

- Algorithme d'apprentissage non-incrémental : la méthode des moindres carrés ordinaire (MCO) utilisée pour résoudre la régression linéaire multivariée est non-incrémentale. En conséquence, l'ensemble des représentations issues des exemples annotés sont traitées simultanément. Dans le cadre de jeux de données de petites tailles, une MCO a des avantages, mais deviendrait difficile à effectuer sur des jeux de données plus importants.
- Concepts prédits parfois assez généraux : les méthodes ont tendance à prédire pour une mention un concept plus général que le concept le plus précis disponible dans l'ontologie.

Pour répondre à ces difficultés, certaines pistes ont été explorées et sont en cours de test sur des jeux de données complets :

- Le traitement des mots hors-vocabulaire : la méthode de calcul de représentations distributionnelles FastText (Bojanowski et al., 2016) permet de produire des représentations pour des mots hors-vocabulaire à partir de n-grammes de caractères les composant. Cette méthode a été intégrée à CONTES comme méthode possible pour le calcul de représentations distributionnelles et des premiers tests ont été effectués (Ferré et al., 2019).
- La dimension importante de l'ESO : deux solutions ont été envisagées (Ferré et al., 2019). La première consiste en la réduction des représentations initiales. La seconde consiste en la modification de la méthode construction de représentations de concepts. Dans le premier cas, des méthodes de réduction ont été intégrées à CONTES comme options, notamment les méthodes d'Analyses en Composantes Principales (ACD), de Positionnement Multidimensionnel (MDS) et t-SNE (Maaten and Hinton, 2008), avec une perte de la qualité de la normalisation. Dans le second cas, la méthode Node2Vec (Grover and Leskovec, 2016) a été intégrée à la méthode CONTES comme option, et d'autres méthodes pourraient être testées telles que trans-H (Wang et al., 2014) ou Onto2Vec (Smaili et al., 2018).
- La faible similarité structurelle entre EDE et ESO : pour augmenter la similarité entre les deux espaces, des connaissances externes pourraient être intégrées à l'EDE. Des méthodes permettant d'adapter des espaces distributionnelles à des connaissances externes pourraient alors être utilisées telles que Retrofitting (Faruqui et al., 2014) ou Counterfitting (Mrkšić et al., 2016). Ses méthodes s'appuient principalement sur des listes de synonymes vis-à-vis d'un domaine précis et vise à rapprocher leurs représentations dans l'espace. Ces méthodes pourraient être adaptées à nos besoins en considérant les représentations de mentions d'un même concept comme devant être rapprochées.
- Algorithme d'apprentissage non-incrémental : des solutions neuronales ainsi que des méthodes telles que celle de la rétro-propagation du gradient (E. Rumelhart et al., 1986) pourraient être envisagées pour déterminer des projections non-linéaires. Une solution neuronale ouvrirait la voie à l'utilisation d'architectures plus complexes abordant notamment la prise en compte de contexte, telles que les

architectures LSTM (Hochreiter and Schmidhuber, 1997), Bi-LSTM, Graph-LSTM (Peng et al., 2017).

- Concepts prédits parfois assez généraux : les composantes associées aux concepts parents du vecteur d'un concept sont toujours égales à 1. Durant l'apprentissage, cela a pour effet de ne pas trop pénaliser la prédiction d'un concept parent plutôt que le concept le plus précis attendu. Nous avons donc recherché un facteur permettant de diminuer progressivement la valeur d'une composante plus le concept associé est éloigné du concept courant. Sur un jeu de données restreints, une valeur a été déterminée permettant d'augmenter la précision des méthodes (Ferré et al., 2019).

## 5.5. Diffusion des méthodes

Les codes Python<sup>21</sup> de la méthode CONTES sont partagés sous licence libre sur la plateforme GitHub<sup>22</sup> à l'adresse suivante : <https://github.com/ArnaudFerre/CONTES>. La méthode y est régulièrement mise à jour avec les dernières améliorations stables. La méthode est documentée et peut être directement utilisée en ligne de commande. Cette utilisation de la méthode a été rendue possible grâce au travail de Mouhamadou Ba, post-doctorant dans l'unité MaIAGE de l'INRA (financé par le projet Ontobedding, sélectionné lors de l'appel à projets émergents 2018 du département STIC de l'Université Paris-Saclay<sup>23</sup>).

Toutes les méthodes développées durant ces travaux de thèse et présentées dans ce chapitre sont intégrées dans la suite logicielle libre AlvisNLP/ML<sup>24</sup> (Ba and Bossy, 2016). AlvisNLP/ML intègre plusieurs modules de traitement automatique des langues, dont des modules de segmentation en phrases et en mots, d'étiquetage morpho-syntaxique et d'extraction terminologique, etc. L'intégration des méthodes dans la suite logicielle permet de changer chacun des pré-traitements (par exemple pour changer de langue) et permet la connexion à d'autres méthodes d'extraction d'information en post-traitement. Cela facilite également les tests de reproductibilité.

---

21 <https://www.python.org/>

22 <https://github.com/>

23 <https://www.universite-paris-saclay.fr/fr/recherche/appel-projet/appel-a-projets-emergents>

24 <https://bibliome.github.io/alvisnlp/>



Home Taxon lives in Habitat Habitat is inhabited by Taxon Taxon exhibits Phenotype Phenotype is exhibited by Taxon Tutorials About Florilege

Search relations by taxon

150 relations for the taxon Lactobacillus delbrueckii subsp. bulgaricus

Source   
GenBank  
CIRM  
DSMZ

Habitat

SOURCE TEXT	TAXON	RELATION TYPE	HABITAT	SOURCE
26585479	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	aerobic environment	OpenMinTeD
8620186, 25380800	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	agar	OpenMinTeD
10555300	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	Allium	OpenMinTeD
3094303, 11375147	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	animal	OpenMinTeD
26059517, 11375147, 16459231	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	animal probiotic	OpenMinTeD
17985843	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	apple and primary derivative thereof	OpenMinTeD
7440507	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	axial filament	OpenMinTeD
25638445	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	bean and related product	OpenMinTeD
11375147, 26133176, 22806866	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	bile	OpenMinTeD
2998270, 3231712, 23253642	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	blood	OpenMinTeD
11101474	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	blood plasma	OpenMinTeD
2998270, 2516523, 2890386	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	blood serum	OpenMinTeD
23216386	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	body fluid	OpenMinTeD
22782266	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	bovine	OpenMinTeD
22782266	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	bovine milk	OpenMinTeD
17484385	Lactobacillus delbrueckii subsp. bulgaricus 2038	Lives in	caecum	OpenMinTeD
17484385	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	caecum	OpenMinTeD
22968411, 26581248, 20121195	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	carbohydrate	OpenMinTeD
9361445	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	carrier	OpenMinTeD
12377803, 62862, 14763849	Lactobacillus delbrueckii subsp. bulgaricus	Lives in	cell	OpenMinTeD

Figure 37 : Capture d'écran de la plateforme Florilège. Un moteur de recherche permet d'interroger la base de données contenant des taxons bactériens (deuxième colonne) et des habitats dans lesquels les taxons vivent (quatrième colonne). Les documents d'où ces informations ont été extraites sont également référencés (première colonne). Il s'agit principalement de PubMed, GenBank et BacDive.

La suite AlvisNLP/ML est également connectée à la plateforme européenne d'extraction d'information OpenMinted<sup>25</sup>. Cette plateforme propose des services d'extraction d'information, particulièrement adaptés aux publications et contenus scientifiques. A court terme, nos méthodes seront également accessibles à travers cette plateforme.

Enfin, l'unité MaiAGE prévoit à court terme d'utiliser ces méthodes pour construire et maintenir des bases de données en biologie telles que la base de données Florilège (Falentin et al., 2017) dont la tâche Bacteria Biotope représente l'objectif (voir Figure 37). Elles remplaceront alors la méthode ToMap qui occupe actuellement cette fonction. Au-delà des biotopes, l'édition 2019 de la tâche Bacteria Biotope permettra d'évaluer la qualité de la prédiction des informations de phénotypes pour Florilège.

<sup>25</sup> <https://services.openminted.eu/home>

## 5.6. Bilan

Nous avons présenté CONTES qui est une méthode de normalisation qui vise à aligner :

- Des représentations vectorielles de mentions d'entités des textes, formant un espace distributionnel des expressions (EDE) ;
- Et des représentations vectorielles de concepts d'une ontologie, formant un espace sémantique de l'ontologie.

Pour les aligner, la méthode apprend une fonction de projection à partir d'exemples annotés et d'un modèle de régression linéaire multivariée, dont l'objectif est de projeter les représentations des mentions à proximité des représentations des concepts associés. Pour une nouvelle mention, une représentation peut être calculée dans l'EDE, et celle-ci peut alors être projetée dans l'ESO grâce à la fonction apprise précédemment. Un score, fondé sur une mesure de similarité (similarité cosinus), peut alors être calculé pour la projection de la mention et chacun des vecteurs de concept de l'ESO. Le concept prédit est celui dont le vecteur obtient le meilleur score.

Les bons résultats de la méthode CONTES sur la tâche Bacteria Biotope sont un argument en faveur de l'existence d'une similarité structurelle partielle entre EDE et ESO. Dans le cas contraire, nous avons pu voir qu'une projection linéaire ne réussirait pas à projeter convenablement les vecteurs de mentions à proximité des vecteurs de concepts associés.

Ces résultats sur la tâche Bacteria Biotope montrent également que la méthode atteint des performances équivalentes à celles qui sont basées sur des similarités de forme entre mention et étiquette de concept. CONTES étant fondée sur de l'apprentissage supervisé, ces résultats indiquent que la méthode réussit à pallier la faible quantité d'exemples annotés pour son entraînement grâce à l'intégration de l'information distributionnelle et de l'information ontologique.

Pour rendre la méthode encore moins dépendante d'exemples annotés, plusieurs alternatives ont été développées et présentées dans ce chapitre. Notamment, des approches par supervision distante ont été explorées pour obtenir des exemples annotés à moindre coût, mais de qualité moindre, telles que :

- L'utilisation des étiquettes et leurs concepts
- L'utilisation des prédictions d'une autre méthode de normalisation ne nécessitant pas d'exemples annotés

La combinaison dans la méthode HONOR de notre approche et d'une approche fondée sur la similarité de forme, a permis de montrer une forte complémentarité de ces deux approches. Cette complémentarité place HONOR comme la méthode état de l'art sur la tâche Bacteria Biotope.

Enfin, CONTES et toutes les méthodes présentées dans ce chapitre ont été intégrées dans une suite logicielle libre d'extraction d'information. Ces méthodes sont donc accessibles

à l'ensemble de la communauté et l'intégralité des résultats présentés pourront être reproduits. Grâce à cette intégration, HONOR sera utilisée à court terme pour l'extension d'une base de données biologique (voir section 5.4).



## **Chapitre 6**

-

## **Conclusion et perspectives**



Nous avons pu constater l'émergence récente (Leaman et al., 2013) des méthodes par apprentissage pour la résolution de la tâche de normalisation en domaine de spécialité. Cette émergence s'est produite simultanément à celle des méthodes de production de plongements lexicaux de qualité (Mikolov et al., 2013a; Pennington et al., 2014), utilisées pour plusieurs tâches de traitement automatique des langues naturelles. L'avantage principal des approches par apprentissage est qu'elles ont potentiellement une bonne capacité d'adaptation, sous condition que les méthodes soient entraînées avec de nouveaux exemples annotés spécifiques à une tâche. En domaine de spécialité, l'utilisation conjointe de l'apprentissage supervisé et de plongements lexicaux pour la normalisation est encore plus récente (Limsopatham and Collier, 2016). Dans ces domaines, l'insuffisance d'exemples annotés est, en effet, une difficulté supplémentaire pour les méthodes par apprentissage. Nous avons proposé une nouvelle approche qui s'appuie pleinement sur les avantages qu'apportent l'apprentissage supervisé et les plongement lexicaux et qui intègre des connaissances externes décrites dans une ontologie du domaine pour répondre à la tâche de normalisation, et qui s'attaque au problème de l'insuffisance des exemples annotés.

Les représentations vectorielles denses d'objets ont montré leur intérêt pour de nombreuses tâches, notamment lorsqu'elles sont utilisées dans des approches par apprentissage. La performance de notre nouvelle approche indique que sous certaines conditions, des similarités entre deux espaces sémantiques représentant des objets distincts peuvent être exploitées :

- Un espace sémantique distributionnel décrivant des mots et des expressions issus d'un corpus
- Un espace sémantique ontologique décrivant des concepts d'un domaine

L'exploitation de ces similarités permet alors de pallier partiellement la quantité faible de données d'entraînement pour la tâche de normalisation. Les conditions propices à l'émergence de ces similarités résident dans le choix d'un corpus du domaine d'intérêt pour le calcul distributionnel, dans un choix de paramètres pertinents pour exécuter la méthode de calcul distributionnel, et dans le choix d'une ontologie décrivant le domaine d'intérêt. Notre approche se veut donc indépendante du domaine de la tâche de normalisation puisque les seules connaissances nécessaires à son utilisation sont externes. Autrement dit, pour l'adapter à une autre tâche, le choix d'autres paramètres pour le calcul distributionnel (notamment le corpus d'entraînement) et d'une autre ontologie devrait être suffisant.

Les versions de la méthode par supervision distante ou par supervision faible ouvrent également des pistes très prometteuses. Notamment, leurs résultats montrent qu'il est déjà possible de se passer d'exemples annotés manuellement. L'utilisation des étiquettes de concepts de l'ontologie est aussi une nouvelle façon d'approcher le problème par supervision distante. Enfin, l'approche ouvre d'emblée d'intéressantes perspectives d'utilisation de représentations distributionnelles adaptées au domaine pour la résolution d'autres tâches en extraction d'information comme la reconnaissance d'entité et l'extraction de relation notamment.

La méthodes CONTES ainsi que ses versions dérivées sont donc les premières implémentations d'une nouvelle approche visant à traiter la tâche de normalisation. Les résultats et le potentiel théorique sous-jacent de ces méthodes valident à eux seuls l'intérêt de l'approche. De plus, leur nouveauté soulève de nouvelles questions et ouvre de nombreuses pistes. Nous détaillons certaines de ces perspectives dans les sections suivantes.

## 6.1. Adaptabilité à d'autres jeux de données

Une première piste d'amélioration est celle de l'adaptabilité : les méthodes peuvent-elles atteindre les mêmes résultats sur d'autres tâches de normalisation ? Pour les autres jeux de données existants, ce test se heurte fréquemment à un problème de calculabilité. En effet la taille des ontologies de référence dans les autres jeux de données disponibles est d'un ordre de grandeur de 10 à 1000 fois supérieur à celle de l'ontologie OntoBiotope. Par exemple, le méta-thésaurus biomédical "*Unified Medical Language System*" (UMLS)<sup>26</sup> contient environ 5 millions de concepts et est déjà utilisé depuis plusieurs années pour réaliser des annotations de publications scientifiques (Aronson, 2001). Le corpus "*NCBI disease corpus*"<sup>27</sup> est un corpus manuellement annoté par les concepts de maladies issus du "*Medical Subject Headings*" (Lipscomb, 2000) et de la base de données "*Online Mendelian Inheritance in Man*" (Hamosh, 2004). Cela correspond à environ 13 000 concepts. En conséquence, l'espace sémantique de référence au cœur des calculs des méthodes développées possède un nombre de dimensions très élevé pour ces autres jeux de données, limitant ainsi leur utilisation. Deux approches peuvent alors être considérées pour résoudre ce problème :

- Réduire l'espace ontologique initial avant d'y projeter les plongements lexicaux. Les méthodes de réduction de dimension testées jusqu'ici (ACP, MDS) ne sont certainement pas adaptées à ce que l'on souhaiterait conserver, c'est-à-dire au moins le classement des plus proches voisins de chaque vecteur de concept. Certaines méthodes (Maaten and Hinton, 2008) se proposent de conserver ce genre de propriétés, mais semblent se restreindre à de la réduction pour de la visualisation graphique d'espaces à grande dimension. Leur utilisation pour des espaces réduits au-dessus de trois dimensions reste incertaine.
- Trouver une alternative aux méthodes testées et permettant surtout de créer un espace ontologique réduit tout en conservant ses propriétés. Cela revient à suivre une démarche similaire à celles des représentations vectorielles d'expressions pour aboutir à des "plongements ontologiques". Des possibles solutions sont à explorer du côté des méthodes existantes de "*graph embeddings*" (Cai et al., 2017).

## 6.2. Adaptation de l'espace distributionnel et amélioration de la projection

L'hypothèse de l'approche est que si les deux espaces sémantiques utilisés ont des structures proches, alors on peut réussir à les aligner de façon pertinente. Cet alignement est recherché sous la forme d'une projection linéaire, et les paramètres de cette projection sont déterminés à l'aide d'une régression linéaire multivariée. Cela ouvre directement deux axes de travail non-exclusifs :

- Augmenter la similarité structurelle entre les deux espaces considérés afin de faciliter la recherche d'une projection linéaire pertinente. Dans ce cas, une piste intéressante est à

<sup>26</sup> <https://www.nlm.nih.gov/research/umls/>

<sup>27</sup> <https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/>

envisager du côté des nouvelles méthodes d'adaptation de plongements lexicaux à des connaissances externes (Faruqui et al., 2014; Mrkšić et al., 2016), afin de rendre l'espace distributionnel des expressions directement plus similaire à celui de l'ontologie.

- Déterminer une projection plus complexe, permettant de mieux combler les défauts de similarité structurelle entre les deux espaces. La régression linéaire a de nombreux avantages lorsque peu de données d'entraînement sont disponibles, notamment limiter le surapprentissage, mais elle possède deux défauts principaux dans notre cas : elle ne permet pas de déterminer une fonction de projection très expressive et elle n'est pas incrémentale, c'est-à-dire qu'elle doit considérer l'ensemble des données d'entraînement simultanément. Ces deux défauts sont également les principales limitations théoriques à l'utilisation de l'approche en domaine. En effet, dans une tâche en domaine général, une méthode aura potentiellement à utiliser une quantité de données d'entraînement plus importante. Cette quantité nécessitera certainement l'utilisation d'une méthode incrémentale. De plus, cette quantité d'information devrait permettre d'atteindre une projection plus expressive.

### **6.3. Application à d'autres tâches en extraction d'information**

Au cœur du domaine de l'extraction d'information, notre approche permet de définir un espace sémantique intégrant conjointement entités textuelles et entités conceptuelles. Les représentations produites peuvent donc également être vues comme des représentations adaptées au domaine d'intérêt. Dans ce cas, la méthode CONTES pourrait être utilisée pour adapter des plongements lexicaux initiaux telle que (Faruqui et al., 2014). Ces nouveaux plongements pourraient alors être utilisés par d'autres méthodes d'extraction d'information, notamment la reconnaissance d'entité et l'extraction de relation. Notre hypothèse est que, si ces méthodes sont appliquées à des tâches du même domaine que celui sur lequel notre méthode a produit des représentations, alors ces représentations augmenteront leur performance. De plus, et à la différence des méthodes d'adaptation à des connaissances externes, CONTES pourrait alors répondre au problème de manque d'interprétabilité des plongements lexicaux. En effet, chaque représentation d'entité textuelle plongée dans l'espace ontologique peut directement être interprétée grâce à sa proximité avec un vecteur de concept.



---

# Publications durant la thèse

## Publications liées à la thèse

Arnaud Ferré, Mouhamadou Ba, and Robert Bossy. 2019. Improving at BLAH5 the CONTES Method for Normalizing Biomedical Text Entities with Concepts from an Ontology with (almost) no Training Data. *Journal of Genomics & Informatics*.

Arnaud Ferré, Louise Deléger, Pierre Zweigenbaum, and Claire Nédellec. 2018. Combining rule-based and embedding-based approaches to normalize textual entities with an ontology. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Hélène Falentin, Estelle Chaix, Sandra Derozier, Magalie Weber, Solange Buchin, Bedis Dridi, Stéphanie-Marie Deutsch, Florence Valence-Bertel, Serge Casaregola, Pierre Renault, et al. 2017. Florilege: a database gathering microbial phenotypes of food interest. *4th International Conference on Microbial Diversity 2017*.

Arnaud Ferré, Pierre Zweigenbaum, and Claire Nédellec. 2017. Representation of complex terms in a vector space structured by an ontology for a normalization task. *BioNLP 2017*.

Arnaud Ferré. 2017b. Représentation de termes complexes dans un espace vectoriel relié à une ontologie pour une tâche de catégorisation. *Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCLA 2017)*.

Arnaud Ferré. 2017a. Normalisation de termes complexes par sémantique distributionnelle guidée par une ontologie. *19es REcontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*.

Robert Bossy, Estelle Chaix, Louise Deleger, Arnaud Ferré, Mouhamadou Ba, Philippe Bessières, and Claire Nédellec. 2016. OntoBiotope : une ontologie pour croiser les habitats microbiens avec les analyses de génomes. *Les journées Bioinformatique de l'Inra*.

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessières, and Claire Nédellec. 2016. Overview of the Bacteria Biotope task at BioNLP Shared Task 2016. *Proceedings of the 4th BioNLP Shared Task Workshop*.

## Publications non directement liées à la thèse

### Représentation de connaissances en sciences du vivant

Vincent Henry, Anne Goelzer, Arnaud Ferré, Stephan Fischer, Marc Dinh, Valentin Loux, Christine Froidevaux, and Vincent Fromion. 2017. The bacterial interlocked process ONtology (BiPON): a systemic multi-scale unified representation of biological processes in prokaryotes. *Journal of biomedical semantics*.

Vincent Henry, Arnaud Ferré, Christine Froidevaux, Anne Goelzer, Vincent Fromion, Sarah Cohen-Boulakia, Sandra Derozier, Marc Dinh, Ghislain Fiévet, Stephan Fischer, et al. 2016. Représentation systémique multi-échelle des processus biologiques de la bactérie. *IC2016: Ingénierie des Connaissances*.

### Bioinformatique et médiation scientifique

William Briand, Ousmane Dao, Guillaume Garnier, Raphaël Guegan, Britany Marta, Clémence Maupu, Julie Miesch, Kenn Papadopoulo, Arthur Radoux, Julie Rojahn, et al. 2018. Dégradation d'un anticancéreux dans les eaux usées - une médaille d'or pour l'équipe GO Paris-Saclay. *médecine/sciences*.

Nika Abdollahi, Alexandre Albani, Eric Anthony, Agnes Baud, Mélissa Cardon, Robert Clerc, Dariusz Czernecki, Romain Conte, Laurent David, Agathe Delaune, et al. 2018. Meet-U: educating through research immersion. *PLoS computational biology*.

Nicolas Allias et al. 2016. Bioinfo-fr. net: présentation du blog communautaire scientifique francophone par les Geekus biologicus. *Journées Ouvertes de Biologie Informatique & Mathématiques*.

Marguerite Benony, Marianne Cardon, Arnaud Ferré, Jean Coquet, Nathan Foulquier, Florian Thonier, Lucas Le Lann, Henry De Belly, Alexandre Evans, Aakriti Jain, et al. 2016. The smell of us - crowdsourcing human body odor evaluation. *Human Computation-A Transdisciplinary Journal*.

# Bibliographie

- Russell L. Ackoff. 1974. The systems revolution. *Long Range Planning*, 7(6):2–20, December.
- S. Agarwal and H. Yu. 2009. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics*, 25(23):3174–3180, December.
- Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. *arXiv:1307.1662 [cs]*, July. arXiv: 1307.1662.
- Sophia Ananiadou, Carol Freidman, and Jun'ichi Tsujii. 2004. Introduction: named entity recognition in biomedicine. *Journal of Biomedical Informatics*, 37(6):393–395.
- Cecilia Arighi, Lynette Hirschman, Thomas Lemberger, Samuel Bayer, Robin Liechti, Donald Comeau, and Cathy Wu. 2017. Bio-ID track overview. *Cell*, 482(7310):376.
- Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMLA Symposium*, page 17. American Medical Informatics Association.
- Alan R. Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, May.
- Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, et al. 2000. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, May.
- Sophie Aubin and Thierry Hamon. 2006. Improving term extraction with terminological resources. In *International Conference on Natural Language Processing (in Finland)*, pages 380–387. Springer.
- Nathalie Aussenac-Gilles, Brigitte Biébow, and Sylvie Szulman. 2000. *Corpus Analysis for Conceptual Modelling*.
- Mouhamadou Ba and Robert Bossy. 2016. Interoperability of corpus processing workflow engines: the case of AlvisNLP/ML in OpenMinTeD. In *Meeting of working Group Medicago sativa*, page np.
- Franz Baader, Deborah L McGuinness, Daniele Nardi, and Peter F Patel-Schneider. 2003. The description logic handbook: Theory, implementation and applications. :510.
- A. Bairoch. 2004. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(Database issue):D154–D159, December.
- Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.
- Alain Birou. 1966. *Vocabulaire pratique des sciences sociales*. Editions Economie et humanisme edition.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *ACL 2017*, July. arXiv: 1607.04606.

- Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larrañaga. 2015. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233.
- Robert Bossy, Estelle Chaix, Louise Deleger, Arnaud Ferré, Mouhamadou Ba, Philippe Bessières, and Claire Nédellec. 2016. OntoBiotope: une ontologie pour croiser les habitats microbiens avec les analyses de génomes. In *Les journées Bioinformatique de l'Inra*, page 1.
- Robert Bossy, Wiktor Golik, Zorana Ratkovic, Philippe Bessières, and Claire Nédellec. 2013. BioNLP shared Task 2013 – An Overview of the Bacteria Biotope Task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 161–169, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Robert Bossy, Wiktor Golik, Zorana Ratkovic, Dialekti Valsamou, Philippe Bessières, and Claire Nédellec. 2015. Overview of the gene regulation network and the bacteria biotope tasks in bionlp'13 shared task. *BMC bioinformatics*, 16(10):S1.
- Robert Bossy, Julien Jourde, Alain-Pierre Manine, Philippe Veber, Erick Alphonse, Maarten Van De Guchte, Philippe Bessières, and Claire Nédellec. 2012. BioNLP Shared Task-The Bacteria Track. In *BMC bioinformatics*, volume 13, page S3. BioMed Central.
- Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 89–96, Bonn, Germany. ACM Press.
- Pier Luigi Buttigieg, Norman Morrison, Barry Smith, Christopher J Mungall, and Suzanna E Lewis. 2013. The environment ontology: contextualising biological and biomedical entities. *Journal of biomedical semantics*, 4(1):43.
- Hongyun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. 2017. A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications. *arXiv:1709.07604 [cs]*, September. arXiv: 1709.07604.
- Nancy Chinchor, David D Lewis, and Lynette Hirschman. 1993. Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3). *Computational linguistics*, 19(3):409–449.
- Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, and others. 2018. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387.
- Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. *Proceedings of BioNLP16*:166.
- C. W. Choo. 1996. The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions. *International Journal of Information Management*, 16(5):329–340, October.
- Vincent Claveau. 2013. IRISA participation to BioNLP-ST13: lazy-learning and information retrieval for information extraction tasks. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 188–196, Sofia, Bulgaria, August. Association for Computational Linguistics.

- Aaron M Cohen. 2005. Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *Proceedings of the acl-ismb workshop on linking biological literature, ontologies and databases: Mining biological semantics*, pages 17–24. Association for Computational Linguistics.
- Trevor Cohen and Dominic Widdows. 2009. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390–405, April.
- William W Cohen and Sunita Sarawagi. 2004. Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods. :10.
- Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural Language Processing (Almost) from Scratch. *Natural Language Processing*:45.
- Francisco M Couto, Mário J Silva, and Pedro M Coutinho. 2005. Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6(Suppl 1):S21.
- Jim Cowie and Yorick Wilks. 2000. Information extraction. *Handbook of Natural Language Processing*, 56:57.
- Koby Crammer and Yoram Singer. 2001. Ultraconservative Online Algorithms for Multiclass Problems. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, volume 2111, pages 99–115. Springer, Berlin, Heidelberg.
- A. P. Davis, T. C. Wieggers, M. C. Rosenstein, and C. J. Mattingly. 2012. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database*, 2012(0):bar065–bar065, March.
- Ferdinand De Saussure. 1989. *Cours de linguistique générale: Édition critique*. volume 1. Otto Harrassowitz Verlag.
- Thierry Declerck, Christian Federmann, Bernd Kiefer, and Hans-Ulrich Krieger. 2008. Ontology-based information extraction and reasoning for business intelligence applications. In *Annual Conference on Artificial Intelligence*, pages 389–390. Springer.
- Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.
- Louise Delèger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferrè, Philippe Bessières, and Claire Nédellec. 2016. Overview of the Bacteria Biotope task at BioNLP Shared Task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 12–22.
- Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of Named Entity Recognition and Linking for Tweets. *Information Processing & Management*, 51(2):32–49, March. arXiv: 1410.7182.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, October. arXiv: 1810.04805.
- Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, February.

Finale Doshi-Velez, Byron C Wallace, and Ryan P Adams. 2009. Graph-Sparse LDA: A Topic Model with Structured Sparsity. :7.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530.

Jennifer D’Souza and Vincent Ng. 2015. Sieve-Based Entity Linking for the Biomedical Domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 297–302, Beijing, China, July. Association for Computational Linguistics.

Michel Dupont, Jean-Marc Vuillaume, Bernard Victorri, Patrice Enjalbert, Yann Mathet, and Nicolas Malandain. 2002. Nouvelles perspectives en extraction d’information. *Revue des Sciences et Technologies de l’Information-Série TSI: Technique et Science Informatiques*, 1(21):37–63.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. Learning internal representations by back-propagating errors.

Jeffrey L Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7(2–3):195–225.

Jean-Louis Ermine, Mahmoud Moradi, and Stéphane Brunel. 2012. Une chaîne de valeur de la connaissance. *Management international*, 16:29.

Hélène Falentin, Estelle Chaix, Sandra Derozier, Magalie Weber, Solange Buchin, Bedis Dridi, Stéphanie-Marie Deutsch, Florence Valence-Bertel, Serge Casaregola, Pierre Renault, and others. 2017. Florilege: a database gathering microbial phenotypes of food interest. In *4th International Conference on Microbial Diversity 2017*, page np.

Manaal Faruqi, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2014. Retrofitting word vectors to semantic lexicons. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Manaal Faruqi, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. *ACL 2016*, May. arXiv: 1605.02276.

David Faure and Claire Nédellec. 1998. A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition. In *LREC workshop on adapting lexical and corpus resources to sublanguages and applications*, pages 5–12.

Scott Federhen. 2011. The NCBI taxonomy database. *Nucleic acids research*, 40(D1):D136–D143.

Arnaud Ferré, Mouhamadou Ba, and Robert Bossy. 2019. Improving at BLAH5 the CONTES Method for Normalizing Biomedical Text Entities with Concepts from an Ontology with (almost) no Training Data.

Arnaud Ferré, Louise Deléger, Pierre Zweigenbaum, and Claire Nédellec. 2018. Combining rule-based and embedding-based approaches to normalize textual entities with an ontology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

- Arnaud Ferré, Pierre Zweigenbaum, and Claire Nédellec. 2017. Representation of complex terms in a vector space structured by an ontology for a normalization task. *BioNLP 2017*:99–106.
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.
- John Rupert Firth. 1935. The Technique of Semantics. *Transactions of the philological society*, 34(1):36–73.
- Luciano Floridi. 2010. *Information: A Very Short Introduction*. Oxford: Oxford University Press, February.
- Juliane Fluck, Heinz Theodor Mevissen, and Holger Dach. 2007. ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries. :3.
- Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2–3):285–307.
- Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. In *ISMB (supplement of bioinformatics)*, pages 74–82.
- G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, November.
- Howard Gardner. 2011. *Frames of mind: The theory of multiple intelligences*. Hachette UK.
- Martin Gerlach and Eduardo G Altmann. 2013. Stochastic model for the vocabulary growth in natural languages. *Physical Review X*, 3(2):021006.
- Martin Gerner, Goran Nenadic, and Casey M. Bergman. 2010. LINNAEUS: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85.
- Omid Ghiasvand and Rohit Kate. 2014. UWM: Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 828–832, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.
- Wiktorija Golik, Robert Bossy, Zorana Ratkovic, and Claire Nédellec. 2013. Improving term extraction with linguistic analysis in the biomedical domain. *Research in Computing Science*, 70:157–172.
- Wiktorija Golik, Pierre Warnier, and Claire Nédellec. 2011. Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. In *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence*, pages 37–39.
- Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A Brief History. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.
- Cyril Grouin. 2016. Identification of Mentions and Relations between Bacteria and Biotope from PubMed Abstracts. *Proceedings of the 4th BioNLP Shared Task Workshop*:64.

- Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM.
- Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating Entity Linking with Wikipedia. *Artificial Intelligence*, 194:130–150, January.
- A. Hamosh. 2004. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–D517, December.
- Daniel Hanisch, Juliane Fluck, Heinz-Theodor Mevissen, and Ralf Zimmer. 2002. Playing biology’s name game: identifying protein names in scientific text. In *Biocomputing 2003*, pages 403–414, Kauai, Hawaii, December. World Scientific.
- Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. 2005. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S14.
- Zellig S. Harris. 1954. Distributional Structure. *Word*, 10(2–3):146–162, August.
- Peter Heisig. 2002. European guide to good practice in knowledge management. *IPK, Berlin*.
- Ralf Herbrich. 2000. Large margin rank boundaries for ordinal regression. *Advances in large margin classifiers*:115–132.
- G. E. Hinton, J. L. Rumelhart, and James L. McClelland. 1986. Distributed representations.
- Geoffrey E Hinton and Tim Shallice. 1991. Lesioning an attractor network: investigations of acquired dyslexia. *Psychological review*, 98(1):74.
- Lynette Hirschman, Marc Colosimo, Alexander Morgan, and Alexander Yeh. 2005a. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, 6(Suppl 1):S11.
- Lynette Hirschman, Alexander A Morgan, and Alexander S Yeh. 2002. Rutabaga by any other name: extracting biological names. *Journal of Biomedical Informatics*, 35(4):247–259.
- Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005b. *Overview of BioCreAtIvE: critical assessment of information extraction for biology*. BMC Bioinformatics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- John Hutchins. 2005. The history of machine translation in a nutshell. Retrieved December, 20:2009.
- Chung Hee Hwang. 1999. Incompletely and Imprecisely Speaking : Using Dynamic Ontologies for Representing and Retrieving Information. :13.
- Natalia Ivanova, Susannah G Tringe, Konstantinos Liolios, Wen-Tso Liu, Norman Morrison, Philip Hugenholtz, and Nikos C Kyrpides. 2010. A call for standardized classification of metagenome projects. *Environmental microbiology*, 12(7):1803–1805.
- Heng Ji, Xiaoman Pan, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee, Cash Costello, and Sydney Informatics Hub. 2017. Overview of TAC-KBP2017 13 Languages Entity Discovery and Linking. In *TAC*.

- Jing Jiang. 2012. Information extraction from text. In *Mining text data*, pages 11–41. Springer.
- Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.
- Dan Jurafsky and James H Martin. 2014. *Speech and language processing*. volume 3. Pearson London.
- Herman Kamper, Weiran Wang, and Karen Livescu. 2015. Deep convolutional acoustic word embeddings using word-pair side information. *arXiv:1510.01032 [cs]*, October. arXiv: 1510.01032.
- Ning Kang, Bharat Singh, Zubair Afzal, Erik M van Mulligen, and Jan A Kors. 2013. Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association*, 20(5):876–881, September.
- Jun’ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.
- Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2048.
- Soonho Kim, Marta Iglesias-Sucasas, and Virginie Viollier. 2013. The FAO Geopolitical Ontology: A Reference for Country-Based Information. *Journal of Agricultural & Food Information*, 14(1):50–65, January.
- Walter Kintsch. 2001. Predication. *Cognitive science*, 25(2):173–202.
- J. L. Klavans and S. Muresan. 2001. Evaluation of the DEFINDER system for fully automatic glossary construction. *Proceedings of the AMLA Symposium*:324–328.
- Cheng-Ju Kuo, Maurice HT Ling, Kuan-Ting Lin, and Chun-Nan Hsu. 2009. BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. *BMC Bioinformatics*, 10(15):S7, December.
- Guiraudé Lame. 2000. *Knowledge acquisition from texts towards an ontology of French law*.
- Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.
- Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. 2008. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3.
- R. Leaman, R. Islamaj Dogan, and Z. Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, November.
- Robert Leaman and Zhiyong Lu. 2016. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 32(18):2839–2846.
- Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, 7(Suppl 1):S3.

- Hsin-Chun Lee, Yi-Yu Hsu, and Hung-Yu Kao. 2015. An enhanced CRF-based system for disease name entity recognition and normalization on BioCreative V DNER Task. :8.
- Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.
- Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3(0):211–225, May.
- David D. Lewis. 1991. Data Extraction as Text Categorization: An Experiment With the MUC-3 Corpus. In *third message understanding conference (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991*.
- Yaoyong Li and Kalina Bontcheva. 2007. Hierarchical, perceptron-like learning for ontology-based information extraction. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*, page 777, Banff, Alberta, Canada. ACM Press.
- Nut Limsopatham and Nigel Collier. 2015. Adapting Phrase-based Machine Translation to Normalise Medical Terms in Social Media Messages. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, August. arXiv: 1508.02285.
- Nut Limsopatham and Nigel Collier. 2016. Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, Berlin, Germany. Association for Computational Linguistics.
- Yongjing Lin, Wenyan Li, Keke Chen, and Ying Liu. 2007. A Document Clustering and Ranking System for Exploring MEDLINE Citations. *Journal of the American Medical Informatics Association*, 14(5):651–661, September.
- Carolyn E. Lipscomb. 2000. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265–266, July.
- Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1501–1511.
- Zhiyong Lu, Hung-Yu Kao, Chih-Hsuan Wei, Minlie Huang, Jingchen Liu, Cheng-Ju Kuo, Chun-Nan Hsu, Richard Tzong-Han Tsai, Hong-Jie Dai, Naoaki Okazaki, and others. 2011. The gene normalization task in BioCreative III. *BMC bioinformatics*, 12(8):S2.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.
- D. Maglott. 2004. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33(Database issue):D54–D58, December.
- Christopher Manning, Prabhakar Raghavan, and Hinrich Schuetze. 2009. Introduction to Information Retrieval. *Natural Language Engineering*:581.

- André Martinet. 1956. *La description phonologique, avec application au parler franco-provençal d'Hauteville (Savoie)*. volume 56. Librairie Droz.
- Farrokh Mehryary, Kai Hakala, Suwisa Kaewphan, Jari Björne, Tapio Salakoski, and Filip Ginter. 2017. End-to-End System for Bacteria Habitat Extraction. *BioNLP 2017*:80.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June. Association for Computational Linguistics.
- George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.
- Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.
- Alexander A Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, Chengjie Sun, Heng-hui Liu, Rafael Torres, Michael Krauthammer, William W Lau, Hongfang Liu, Chun-Nan Hsu, Martijn Schuemie, K Bretonnel Cohen, et al. 2008. Overview of BioCreative II gene normalization. *Genome Biology*, 9(Suppl 2):S3.
- Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting Word Vectors to Linguistic Constraints. *arXiv:1603.00892 [cs]*, March. arXiv: 1603.00892.
- Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.
- T. H. Muneeb, Sunil Kumar Sahu, and Ashish Anand. 2015. Evaluating distributed word representations for capturing semantics of biomedical concepts. *Proceedings of ACL-IJCNLP*:158.
- Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.
- Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, July. Association for Computational Linguistics.
- Claire Nédellec, Adeline Nazarenko, and Robert Bossy. 2009. Information extraction. In *Handbook on ontologies*, pages 663–685. Springer.

- Goran Nenadic, Irena Spasic, and Sophia Ananiadou. 2005. Mining Biomedical Abstracts: What's in a Term? In Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee, and Oi Yee Kwong, editors, *Natural Language Processing – IJCNLP 2004*, volume 3248, pages 797–806. Springer, Berlin, Heidelberg.
- Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. *arXiv preprint arXiv:1605.07766*.
- Jérôme Nobécourt. 2000. A method to build formal ontologies from texts. In *EKAW-2000 Workshop on ontologies and text, Juan-Les-Pins, France*.
- Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word Embedding-based Antonym Detection using Thesauri and Distributional Information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989, Denver, Colorado, June. Association for Computational Linguistics.
- Dominique Osborne, Shashi Narayan, and Shay B Cohen. 2016. Encoding prior knowledge with eigenword embeddings. *Transactions of the Association for Computational Linguistics*, 4:417–430.
- Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Yvon Pesqueux. 2008. La dualité "savoir-connaissance" en sciences des organisations. In *Séminaire de recherche en anthropologie de l'imaginaire*.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv:1802.05365 [cs]*, February. arXiv: 1802.05365.
- Miguel Pignatelli, Andrés Moya, and Javier Tamames. 2009. EnvDB, a database for describing the environmental distribution of prokaryotic taxa. *Environmental Microbiology Reports*, 1(3):191–197.
- Jordan B. Pollack. 1990. Recursive Distributed Representations. *Artificial Intelligence*, 46:77–105.
- Wanda Pratt and Meliha Yetisgen-Yildiz. 2003. A Study of Biomedical Concept Identification: MetaMap vs. People. *AMIA Annual Symposium Proceedings*, 2003:529–533.
- R. Rada, H. Mili, E. Bicknell, and M. Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, February.
- Alok Ranjan Pal and Diganta Saha. 2015. Word Sense Disambiguation: A Survey. *International Journal of Control Theory and Computer Modeling*, 5(3):1–16, July.
- Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning*, pages 147–155. Association for Computational Linguistics.

- Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. *arXiv preprint arXiv:1707.09861*.
- Kaspar Riesen, Michel Neuhaus, and Horst Bunke. 2007. Graph embedding in vector spaces by means of prototype selection. In *International Workshop on Graph-Based Representations in Pattern Recognition*, pages 383–393. Springer.
- Jennifer Rowley. 2007. The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of information science*, 33(2):163–180.
- Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.
- Philip Russom. 2007. BI search and text analytics. *TDWI Best Practices Report*:9–11.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *arXiv:cs/0306050*, June. arXiv: cs/0306050.
- Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora*, pages 13–25. Springer.
- Martijn J. Schuemie, Rob Jelier, and Jan A. Kors. 2007. Peregrine: Lightweight gene name normalization by dictionary lookup. In *Proc of the Second BioCreative Challenge Evaluation Workshop*, pages 131–133.
- Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.
- Amit Singhal and others. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.
- Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. 2018. Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics*, 34(13):i52–i60, July.
- Edward E Smith and Douglas L Medin. 1981. *Categories and concepts*. volume 9. Harvard University Press Cambridge, MA.
- Noah A Smith. 2011. Linguistic structure prediction. *Synthesis lectures on human language technologies*, 4(2):1–274.
- Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. :11.
- Eleftherios Spyromitros-Xioufis, Grigorios Tsoumakos, William Groves, and Ioannis Vlahavas. 2012. Multi-label classification methods for multi-target regression. *arXiv preprint arXiv:1211.6581*:1159–1168.
- Mark Steyvers and Joshua B. Tenenbaum. 2005. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29(1):41–78, January.
- J. L. Stocks. 1935. CORNFORD, F. M. -Plato’s Theory of Knowledge. *Mind*, 44(n/a):526.

- Mert Tiftikci, Hakan Sahin, Berfu Büyüköz, Alper Yayıkçı, and Arzucan Ozgür. 2016. Ontology-based Categorization of Bacteria and Habitat Entities using Information Retrieval Techniques. *ACL 2016*:56.
- Y. Tsuruoka, J. McNaught, J.;c. Tsujii, and S. Ananiadou. 2007. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*, 23(20):2768–2774, October.
- Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word Representations: A Simple and General Method for Semi-Supervised Learning. *Proceedings of the 48th annual meeting of the association for computational linguistics*:11.
- Mike Uschold and Martin King. 1995. Towards a Methodology for Building Ontologies. :15.
- Maria Vargas-Vera, Enrico Motta, John Domingue, Simon Buckingham Shum, Mattia Lanzoni, Walton Hall, and Milton Keynes. 2001. Knowledge Extraction by using an Ontology-based Annotation Tool. :8.
- James Z. Wang, Zhidian Du, Rapeeporn Payattakool, Philip S. Yu, and Chin-Fu Chen. 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, May.
- Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *AAAI*, pages 1112–1119.
- Warren Weaver. 1955. Translation. *Machine translation of languages*, 14:15–23.
- Chih-Hsuan Wei and Hung-Yu Kao. 2011. Cross-species gene normalization by species inference. *BMC Bioinformatics*, 12(8):S5, October.
- Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wieggers, and Zhiyong Lu. 2015. Overview of the BioCreative V Chemical Disease Relation (CDR) Task. *Proceedings of the fifth BioCreative challenge evaluation workshop*:14.
- John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.
- Daya C. Wimalasuriya and Dejing Dou. 2010. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323, June.
- Feiyu Xu, Hans Uszkoreit, and Hong Li. 2006. Automatic event and relation detection with seeds of varying complexity. In *Proceedings of the AAAI workshop event extraction and synthesis*, pages 12–17.
- Victor H Yngve. 1961. *Random generation of English sentences*. Massachusetts Inst. of Technology.
- Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 545–550.
- Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *arXiv:1212.5701 [cs]*, December. arXiv: 1212.5701.

Qianqian Zhang, Mengdong Chen, and Lianzhong Liu. 2017. A review on entity relation extraction. In *2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, pages 178–183. IEEE.

Deyu Zhou, Dayou Zhong, and Yulan He. 2014. Biomedical Relation Extraction: From Binary to Complex. *Computational and Mathematical Methods in Medicine*, 2014.

Pierre Zweigenbaum and Benoît Habert. 2006. Faire se rencontrer les parallèles : regards croisés sur l’acquisition lexicale monolingue et multilingue. *Revue de sociolinguistique en ligne GLOTTOPOL*, 8:22–44.

**Titre :** Représentations vectorielles et apprentissage automatique pour l'alignement d'entités textuelles et de concepts d'ontologie : application à la biologie

**Mots clés :** extraction d'information, normalisation, plongement lexical, intelligence artificielle, traitement automatique des langues

**Résumé :** L'augmentation considérable de la quantité des données textuelles rend aujourd'hui difficile leur analyse sans l'assistance d'outils. Or, un texte rédigé en langue naturelle est une donnée non-structurée, c'est-à-dire qu'elle n'est interprétable que par un programme informatique spécialisé, sans lequel les informations des textes restent largement sous-exploitées. Parmi les outils d'extraction automatique d'information, nous nous intéressons aux méthodes d'interprétation automatique de texte pour la tâche de normalisation d'entité qui consiste en la mise en correspondance automatique des mentions d'entités de textes avec des concepts d'un référentiel.

Pour réaliser cette tâche, nous proposons une nouvelle approche par alignement de deux types de représentations vectorielles d'entités capturant une partie de leur sens : les plongements lexicaux pour les mentions textuelles et des "plongements ontologiques" pour les concepts, conçus spécifiquement pour ce travail. L'alignement entre les deux se fait par apprentissage supervisé. Les méthodes développées ont été évaluées avec un jeu de données de référence du domaine biologique et elles représentent aujourd'hui l'état de l'art pour ce jeu de données. Ces méthodes sont intégrées dans une suite logicielle de traitement automatique des langues et les codes sont partagés librement.

**Title:** Vector representations and machine learning for alignment of text entities with ontology concepts: application to biology

**Keywords:** information extraction, normalization, word embedding, artificial intelligence, natural language processing

**Abstract:** The impressive increase in the quantity of textual data makes it difficult today to analyze them without the assistance of tools. However, a text written in natural language is unstructured data, i.e. it can be interpreted only by a specialized computer program, without which the information in the texts remains largely under-exploited. Among the tools for automatic extraction of information from text, we are interested in automatic text interpretation methods for the entity normalization task that consists in automatically matching text entity mentions to concepts in a reference terminology.

To accomplish this task, we propose a new approach by aligning two types of vector representations of entities that capture part of their meanings: word embeddings for text mentions and concept embeddings for concepts, designed specifically for this work. The alignment between the two is done through supervised learning. The developed methods have been evaluated on a reference dataset from the biological domain and they now represent the state of the art for this dataset. These methods are integrated into a natural language processing software suite and the codes are freely shared.

