THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale I2S

Unité de recherche LGI2P

Gestion de l'incertitude et de l'imprécision dans un processus d'extraction de connaissances à partir des textes

Présentée par Pierre-Antoine JEAN

Le 23 Novembre 2017

Sous la direction de Jacky Montmain et Patrice Bellot

Devant le jury composé de

Mme Béatrice Daille, Professeur à l'Université de Nantes	Président du jury
Mme Catherine Berrut, Professeur à l'Université de Grenoble	Rapporteur
M. Pierre Zweigenbaum, Directeur de Recherche CNRS, LIMSI	Rapporteur
M. Mathieu Roche, Chercheur, HDR, CIRAD	Examinateur
M. Jacky Montmain, Professeur au LGI2P, IMT Mines Alès	Directeur de thèse
M. Patrice Bellot, Professeur à l'Université Aix-Marseille	Co-Directeur de thèse
Mme Sylvie Ranwez, Professeur au LGI2P, IMT Mines Alès	Encadrement
M. Sébastien Harispe, Maître-assistant, IMT Mines Alès	Encadrement

UNIVERSITÉ DE MONTPELLIER

« *Do. Or do not. There is no try.* »

S.W. V

Remerciements

La thèse est un long processus cognitif dans lequel de nombreux acteurs interviennent. Cette page leur est dédiée.

Je commence ces remerciement en m'adressant à mes encadrants : Jacky, Patrice, Sylvie et Sébastien. Je tiens à les remercier pour tout le savoir qu'ils ont réussi à me transmettre tout le long de cette thèse. Grâce à eux, je me suis formé à de nombreux domaines scientifiques : la représentation des connaissances, la recherche et l'extraction d'information, le domaine du questions-réponses et le domaine de l'apprentissage automatique. Chacun d'entre eux m'a accordé du temps pour approfondir ces aspects particuliers liés à la recherche scientifique et je leur en suis reconnaissant.

Je remercie les membres du jury : Mme Catherine Berrut, M. Pierre Zweigenbaum, Mme Béatrice Daille et M. Mathieu Roche. Je les remercie chaleureusement d'avoir accepté de rapporter et examiner cette thèse. Un remerciement tout particulier à M. Pierre Zweigenbaum et à Mme Catherine Berrut pour leurs remarques pertinentes qui ont permis d'améliorer le manuscrit.

Bien entendu que serait une page de remerciement sans remercier tous mes collègues de travail que j'ai côtoyés et appris à connaître pendant ces trois ans. Je vous remercie pour tous ces moments partagés allant des discussions autour de la machine à café aux moments de détente devant un babyfoot. J'ai d'agréables souvenirs en votre compagnie. Un *big up* particulier à Ivan avec qui j'ai énormément partagé durant la thèse.

Ces trois ans, je les dédicace à ma famille : Osmoon, Petite Feuille de Salade, Mikette, Jooooooooël et Papou. Ils ont été d'une aide précieuse et m'ont toujours soutenu et encouragé pour aller de l'avant. Je vous dois énormément et je voulais vous témoigner tout l'amour que je vous porte. Je remercie tout particulièrement ma compagne Marie qui m'a apporté tout son amour et sa bienveillance durant cette thèse, je suis heureux que tu partages ma vie.

Enfin, une pensée à toutes les personnes dont je n'ai pas mentionné le nom et qui m'ont un jour encouragé à persévérer afin d'atteindre mes objectifs.

Résumé

Les concepts de découverte et d'extraction de connaissances ainsi que d'inférence sont abordés sous différents angles au sein de la littérature scientifique. En effet, de nombreux domaines s'y intéressent allant de la recherche d'information, à l'implication textuelle en passant par les modèles d'enrichissement automatique des bases de connaissances. Ces concepts suscitent de plus en plus d'intérêt à la fois dans le monde académique et industriel favorisant le développement de nouvelles méthodes.

Cette thèse propose une approche automatisée pour l'inférence et l'évaluation de connaissances basée sur l'analyse de relations extraites automatiquement à partir de textes. L'originalité de cette approche repose sur la définition d'un cadre tenant compte (i) de l'incertitude linguistique et de sa détection dans le langage naturel réalisée au travers d'une méthode d'apprentissage tenant compte d'une représentation vectorielle spécifique des phrases, (ii) d'une structuration des objets étudiés (*e.g.* syntagmes nominaux) sous la forme d'un ordre partiel tenant compte à la fois des implications syntaxiques et d'une connaissance *a priori* formalisée dans un modèle de connaissances de type taxonomique (iii) d'une évaluation des relations extraites et inférées grâce à des modèles de sélection exploitant une organisation hiérarchique des relations considérées. Cette organisation hiérarchique permet de distinguer différents critères en mettant en œuvre des règles de propagation de l'information permettant ainsi d'évaluer la croyance qu'on peut accorder à une relation en tenant compte de l'incertitude linguistique véhiculée. Bien qu'à portée plus large, notre approche est ici illustrée et évaluée au travers de la définition d'un système de réponse à un questionnaire, généré de manière automatique, exploitant des textes issus du Web. Nous montrons notamment le gain informationnel apporté par la connaissance *a priori*, l'impact des modèles de sélection établis et le rôle joué par l'incertitude linguistique au sein d'une telle chaîne de traitement. Les travaux sur la détection de l'incertitude linguistique et la mise en place de la chaîne de traitement ont été validés par plusieurs publications et communications nationales et internationales. Les travaux développés sur la détection de l'incertitude et la mise en place de la chaîne de traitement sont disponibles au téléchargement à l'adresse suivante : *https ://github.com/PAJEAN/*.

Abstract

Knowledge discovery and inference are concepts tackled in different ways in the scientific literature. Indeed, a large number of domains are interested such as : information retrieval, textual inference or knowledge base population. These concepts are arousing increasing interest in both academic and industrial fields, promoting development of new methods.

This manuscript proposes an automated approach to infer and evaluate knowledge from extracted relations in unstructured texts. Its originality is based on a novel framework making it possible to exploit (i) the linguistic uncertainty thanks to an uncertainty detection method described in this manuscript (ii) a generated partial ordering of studied objects (*e.g.* noun phrases) taking into account of syntactic implications and a priori knowledge defined into taxonomies, and (iii) an evaluation step of extracted and inferred relations by selection models exploiting a specific partial ordering of relations. This partial ordering allows to compute some criteria in using information propagation rules in order to evaluate the belief associated to a relation in taking into account of the linguistic uncertainty. The proposed approach is illustrated and evaluated through the definition of a system performing question answering by analysing texts available on the Web. This case study shows the benefits of structuring processed information (*e.g.* using prior knowledge), the impact of selection models and the role of the linguistic uncertainty for inferring and discovering new knowledge. These contributions have been validated by several international and national publications and our pipeline can be downloaded at *https ://github.com/PAJEAN/*.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Sommaire

1.1 Contexte de la thèse

La naissance de l'expression "*intelligence artificielle*" provient de la conférence de Dartmouth College aux États-Unis en 1956. Cette conférence a attiré de nombreuses communautés comprenant des chercheurs en informatique, des mathématiciens, des économistes et des psychologues tel que Herbert Alexander Simon, Marvin Minsky, John McCarthy, Nathaniel Rochester et Claude Elwood Shannon considérés désormais comme les pères fondateurs de ce domaine (McCarthy et al., 2006). Le nombre et la renommée des acteurs à cette conférence montrait l'engouement autour de ce domaine en émergence. Depuis cette époque, le concept d'intelligence artificielle a bien évolué : des raisonnements logiques et processus d'inférence heuristique, le concept englobe désormais l'apprentissage automatique et l'adaptation des différents modèles à des jeux de données inconnus (Rialle, 1996). Les domaines d'application se sont également largement étendus, de l'économie où elle est utilisée pour la prédiction des indices boursiers à la prédiction d'épidémies et de maladies dans le

domaine de la santé en passant par le secteur du jeu vidéo (ONTANÓN et al., 2013). Cette thèse s'intéresse à l'analyse et à la compréhension du langage naturel par des systèmes automatisés, et à sa sémantique. Dans ce domaine, l'intelligence artificielle doit faire face à de nombreuses problématiques algorithmiques, physiques (puissance de calcul), conceptuelles et d'intelligibilité, rendant cette tâche ardue pour un ordinateur.

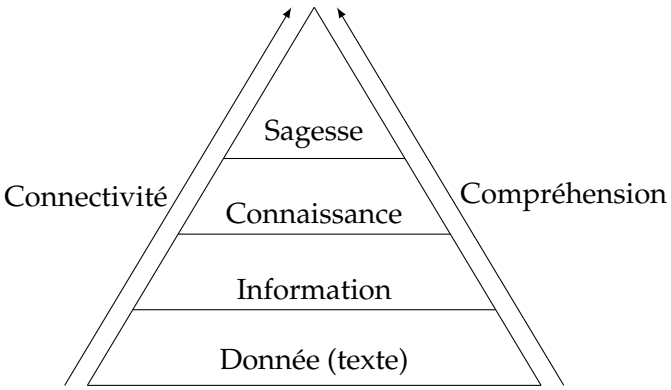
Morphologie, syntaxe, sémantique et pragmatique sont des concepts linguistiques que l'humain manipule quotidiennement et pour lesquels il sait mobiliser de nombreuses capacités cognitives, et ce depuis le plus jeune âge. *A contrario* pour la machine, le texte demeure une donnée comme une autre qui devra pour être traitée efficacement, faire appel à de nombreuses analyses. Comment la machine peut-elle donner du sens à des mots ? Cette question est abordée dans plusieurs domaines de l'informatique. Dans notre cas, nous avons choisi de l'appréhender du point de vue de la représentation des connaissances. Ce domaine, au-delà des mots, modélise et représente la connaissance par des liens sémantiques entre des concepts et/ou des instances. Il constitue un axe privilégié de l'équipe KID (*Knowledge Images Decision making*) du LGI2P (Laboratoire de Génie Informatique et d'Ingénierie de Production) dépendant de l'école des Mines d'Alès et c'est dans cette équipe de recherche que cette thèse a été réalisée. L'ambition de l'équipe est de bénéficier d'un grand nombre de données textuelles pour constituer des bases de connaissances qui pourront servir de support dans un processus décisionnel. Par ailleurs, la collaboration avec le LSIS (Laboratoire des Sciences de l'Information et des Systèmes) à Marseille a permis d'élargir les perspectives initiales de la thèse grâce à l'apport de son savoir-faire dans le domaine de l'extraction d'information et du traitement automatique des langues.

Ainsi, ce manuscrit s'intéresse à diverses difficultés liées à l'exploitation des textes en passant par l'extraction d'information, à son évaluation et à la manière d'en tirer de la connaissance en considération des tournures énonciatives des phrases *e.g.* la présence d'incertitude dans la proposition de l'auteur. Le lecteur y trouvera une réflexion sur le traitement qui permet à partir de la donnée textuelle d'exprimer et de raisonner sur une forme de connaissance ainsi que des propositions pour pallier certains écueils inhérents au langage naturel. Les exemples présentés dans la suite traitent de plusieurs domaines car nous souhaitons développer une approche générique apte à traiter aussi bien des textes techniques (*e.g.* publications) que des textes de vulgarisation (*e.g.* blogs, réseaux sociaux).

1.2 Introduction aux problématiques de l'étude

Disposer de vastes corpus de textes de natures diverses constitue une véritable opportunité pour le domaine de l'Intelligence Artificielle (IA). Cela laisse notamment

FIGURE 1.1 – Représentation des relations entre les concepts de Donnée, d'Information, de Connaissance et de Sagesse.



entrevoir la possibilité d'exploiter de façon systématisée et automatisée les différentes connaissances et éléments d'information explicités dans ces textes, par le couplage d'approches issues du Traitement Automatique des Langues (TAL) et des techniques de raisonnement adaptées. Une variété de traitements et processus pourraient alors bénéficier de la richesse exprimée dans ces corpus (*e.g.* aide à la prise de décision, système de questions-réponses). Cette ligne directrice d'acquisition de la connaissance à partir des textes est le *leitmotiv* de cette thèse. Si nous réalisons l'analogie avec les travaux de ACKOFF, 1989, nous manipulons les différents niveaux d'appréhension de l'esprit humain vis-à-vis du monde, au travers de cette chaîne de traitement. Ces niveaux hiérarchisent quatre concepts fondamentaux : donnée, information, connaissance et sagesse (cf. figure 1.1).

La transposition de ce modèle à notre problématique revient à considérer les textes comme la source de nos données. Ces textes écrits en langage naturel sont à l'origine de nombreuses interrogations (section 1.2.1 dans MANNING et SCHÜTZE, 1999). Notamment :

— Quel genre de choses les personnes disent-elles ?
— Que disent ces choses à propos du monde ?

Ces questions sont à la base du traitement du langage et couvrent à la fois tous les aspects de la structure du langage et toute la sémantique du discours. Pour tenter d'y répondre, le domaine du TAL propose des méthodologies permettant de transformer les données textuelles en information. L'information a été définie par BELLINGER, CASTRO et MILLS, 2004 comme une donnée ayant une signification par le biais d'une connexion relationnelle. Dans notre cas, ces connexions correspondent aux *déclarations* extraites sous la forme de triplets au format RDF[1] <sujet, prédicat, objet> qui peut aussi s'écrire prédicat(sujet, objet) si l'on se conforme à la notation formelle classique en logique du premier ordre[2] (cf. figure 1.2).

1. RDF : *Resource Description Framework*
2. Dans la suite du manuscrit les deux notations pourront être utilisées.

FIGURE 1.2 – De la donnée textuelle à l'information.

"Barrack Obama est né à Hawaï."

↓

<Barrack Obama, naître_à, Hawaï>

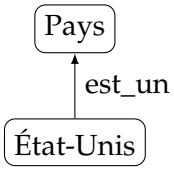
Le terme *déclaration* fait écho aux termes *claim* ou *statement* utilisés dans la littérature anglophone. Il s'agit d'une proposition mettant en jeu deux entités (le sujet et l'objet) par l'intermédiaire d'une relation, la plupart du temps exprimée par un verbe. Extraire une déclaration revient donc à identifier une relation entre deux entités dans un texte. Aussi, par abus de langage, il est possible que nous y fassions référence par la suite en utilisant le terme générique de *relation*.

L'information acquise, l'objectif est désormais d'en extraire de la connaissance. Celle-ci émerge de la compréhension des motifs sous-jacents liés à une collection d'informations. Autrement dit, on parle de connaissance chaque fois qu'une information (ou plusieurs) sert de base à un raisonnement. Il peut s'agir de propriétés particulières sur les relations, d'identifier certaines règles, etc. Par exemple, en présence de la relation <États-Unis, est_un, pays>, nous comprenons les implications liées à l'ordre qui incombe entre ces trois mots et notamment de la relation de subsomption entre *pays* et *États-Unis* (cf. figure 1.3).

FIGURE 1.3 – De l'information à la connaissance.

<États-Unis, est_un, pays>

↓



La connaissance englobe également l'utilisation des mécanismes d'interprétation des informations et d'inférences disponibles afin d'accéder à des relations implicites (cf. figure 1.4).

Finalement, la sagesse « possède la connaissance de toutes les choses, dans la mesure où cela est possible », comme l'a écrit Aristote. Cette citation transparaît au sein de notre problématique comme l'utilisation de ressources de connaissances externes aux données exploitées, permettant notamment un raisonnement pragmatique sur la connaissance (cf. figure 1.5).

Toutefois, la mise en place d'une telle chaîne d'exploitation des textes et de la connaissance est colossale de par sa transversalité dans de nombreux domaines de recherche. Gestion de la complexité linguistique, compréhension sémantique du discours, structuration des connaissances et inférence sont autant de difficultés que doit affronter

FIGURE 1.4 – Les mécanismes d'inférence de connaissances. La connaissance est retranscrite comme les mécanismes d'organisation de l'information et d'inférence de nouvelles connaissances.

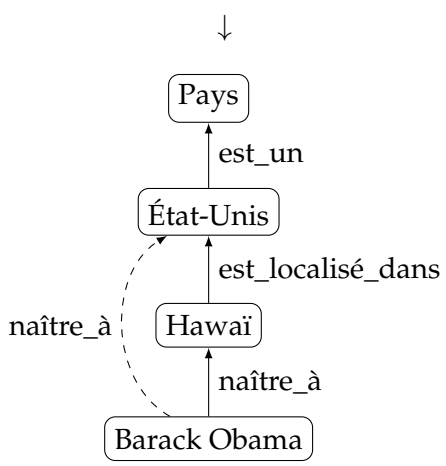
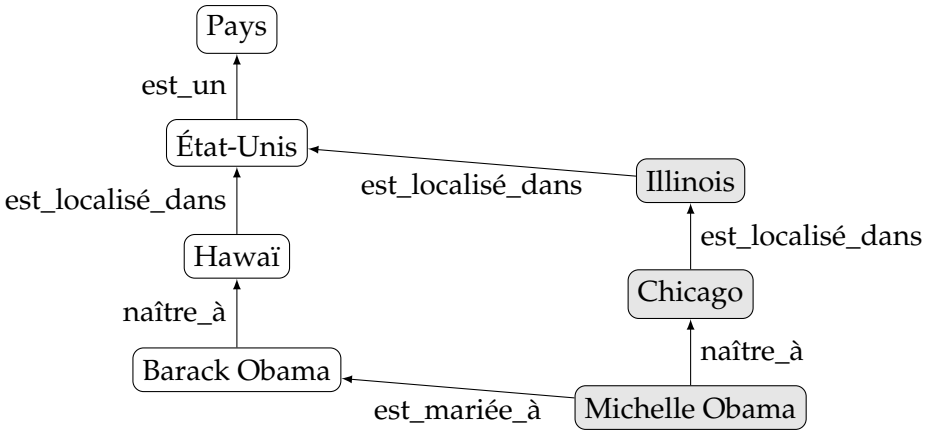


FIGURE 1.5 – Illustration du concept de sagesse dans le contexte de l'extraction de connaissances à partir de textes. La sagesse est modélisée au travers des concepts grisés, extérieurs aux informations initiales.



le système. Ces difficultés sont généralement abordées de manière séparée au sein des communautés du TAL, de l'apprentissage automatique, de la représentation et de l'ingénierie des connaissances.

La section suivante propose d'introduire les concepts fondamentaux pour la compréhension des méthodologies exploitées au sein de cette thèse. Ces derniers appartiennent aux domaines de la représentation des connaissances, du raisonnement, du questions-réponses et de l'enrichissement des bases de connaissances.

1.3 Description des domaines étudiés et concepts manipulés

L'approche décrite dans cette thèse exploite une chaîne de traitement similaire aux travaux concernant l'enrichissement des bases de connaissances à partir de textes non structurés. La caractérisation des rouages de ce domaine passe avant tout par la gestion de la complexité linguistique au travers des méthodes d'extraction d'information et de la compréhension des systèmes à base de connaissances dont l'ontologie fournit l'armature. Cette section est divisée en cinq sous-sections. La première est dédiée à la description des modalités d'extraction d'information à partir de données textuelles. La seconde concerne la définition des systèmes de représentation de la connaissance. La troisième décrit différentes approches pour la construction des bases de connaissances. La quatrième s'intéresse à la définition de la notion de raisonnement. Enfin, la dernière est consacrée au domaine du questions-réponses.

1.3.1 Fouille de textes et gestion de la complexité linguistique

Extraire de la connaissance à partir de données textuelles est une tâche complexe animant la communauté du TALN. Cette sous-section fait figure d'introduction aux différents domaines liés à cette tâche qui seront plus spécifiquement détaillés dans le chapitre 2.

La Recherche d'Information

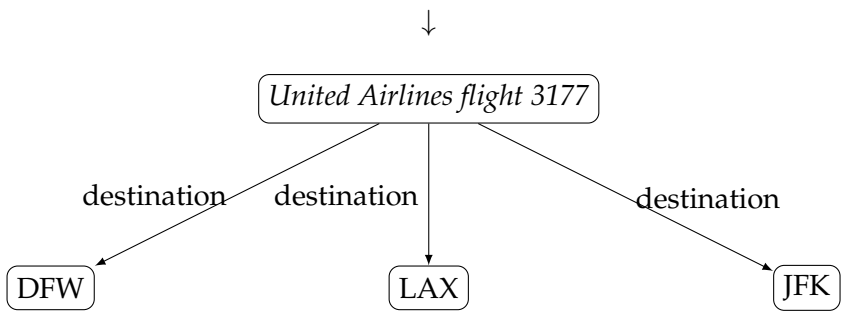
Pour introduire cette sous-section, il est intéressant de définir le domaine de la Recherche d'Information (RI). En effet, ce domaine, bien qu'il ne soit pas dédié à la seule recherche textuelle, est pour nous un point d'entrée des méthodes appartenant au TAL. MANNING, RAGHAVAN et SCHÜTZE, 2008 définissent la RI comme les moyens de trouver une ressource (usuellement des documents) de nature non-structurée (généralement des textes) qui satisfait un besoin d'information à partir d'une large collection de documents. Au cours de cette thèse, ce domaine a été abordé à la fois pour la récupération de données spécifiques afin d'alimenter notre chaîne de traitement et au travers de la compétition COLIEE (*Competition on Legal*

Information Extraction/Entailment). Lors de cette dernière, nous avons eu pour objectif de soumettre un système capable de retrouver un ensemble d'articles pertinents issus du code civil japonais (traduit en anglais) permettant de répondre par *oui* ou par *non* à un cas juridique précis. Dans un but de clarté, les détails de cette tâche, la méthode développée et les résultats obtenus sont plus amplement détaillés dans l'annexe A.

L'extraction d'information

La longueur et la diversité des documents, retournés par un système de RI, peuvent rendre leur exploitation directe difficile par un opérateur humain. Il est donc parfois nécessaire d'en extraire de l'information de manière automatisée. Pour cela revenons sur la notion de triplet <sujet, prédicat, objet> que nous avons introduite précédemment dans ce chapitre. ETZIONI et al., 2011 caractérisent ces relations en tant que relations binaires, qu'ils distinguent des relations n-aires. Cette subtilité crée une première distinction dans les outils d'extraction d'information. En effet, ce domaine est traditionnellement subdivisé en plusieurs catégories parmi lesquelles l'extraction de relations est associée à l'extraction de relations binaires et l'extraction d'événements à l'extraction de relations n-aires. Toutefois, cette dernière catégorie peut dépendre de la première dans la mesure où une relation n-aire peut être perçue comme un ensemble de relations binaires (cf. figure 1.6).

FIGURE 1.6 – Exemple d'une relation n-aires [3]. Cette relation peut être décomposée comme un sous-ensemble de relations binaires. L'extraction d'événements inclut également l'extraction de caractéristiques diverses (date, chiffre, etc.) associées à une entité donnée.



Le domaine de l'extraction d'information a été abordé dans la littérature sous de nombreux aspects regroupant : le développement manuel de motifs textuels (HEARST, 1992), des méthodes supervisées (CULOTTA et SORENSEN, 2004), semi-supervisées (BRIN, 1998) et non-supervisées (WANG et al., 2013). Dans cette thèse, la méthodologie employée pour l'extraction de relations est une approche hybride conciliant à la fois une définition d'un ensemble de motifs morpho-syntaxiques et une approche

3. https ://www.w3.org/TR/swbp-n-aryRelations/

par apprentissage. Elle sera détaillée plus précisément dans le chapitre 2 ainsi que le paradigme de l'extraction d'information.

L'imprécision et l'incertitude inhérentes au langage naturel

Si on s'abstrait des problématiques techniques liées à l'extraction de relations (prise en compte de la syntaxe, délimitation des entités, etc.), de nombreuses autres problématiques subsistent quant à l'exploitation de l'information évoquée dans les textes. Pour les illustrer, un parallèle peut être réalisé avec une discussion orale. En effet, une discussion engage au moins deux personnes et chacune d'entre elle véhicule un ensemble d'informations. Indirectement, chaque personne interprète cette information selon la personne avec qui elle discute et les termes qu'elle emploie. Ces deux aspects appartenant à l'évaluation de l'information peuvent être représentés par les domaines de l'évaluation des sources des documents et l'analyse sémantique des phrases (imprécision, incertitude). Chacun de ces aspects engendre une interprétation différente de l'information et les méta-données connues sur la source constituent des informations essentielles pour notre jugement (cf. tableau 1.1).

TABLEAU 1.1 – Déclarations selon différentes sources. Ce cas pratique s'attarde sur une source de confiance (médecin) et un fait certain, une source de confiance (famille de Jeff) et un fait incertain, une source inconnue et un fait incertain et enfin une source à laquelle on accorde peu de confiance (personne suspicieuse) et un fait certain (et faux). Il souligne toute la complexité d'interprétation et parfois même la subjectivité d'interprétation que chaque personne accorde à une déclaration selon une source donnée et sa manière de l'énoncer.

Source	Déclaration	Confiance
Médecin	Jeff est en bonne santé.	Élevée (+)
Famille de Jeff	Jeff serait en bonne santé.	Élevée (-)
Personne λ	Jeff est probablement en bonne santé.	Faible (+)
Personne suspicieuse	Jeff est en mauvaise santé.	Faible (-)

Le tableau 1.1 ne représente qu'une sous-partie de la complexité linguistique. En effet, il n'aborde pas certaines analyses sémantiques telles que l'imprécision liée aux termes employés et l'incertitude associée à la polysémie du langage. Dans un premier temps, le concept d'imprécision tient compte des notions évoquées d'une part par la logique floue *e.g.* "cette marchandise a un coût élevé" où "coût élevé" est une donnée vague pour décrire la marchandise et d'autre part par le contenu informationnel (*Information Content* – IC) d'une déclaration. Entre les déclarations suivantes : "J'ai vu un animal traverser la route" et "J'ai vu une biche traverser la route", l'IC est plus important dans la seconde phrase car "une biche" est une information plus précise que "un animal". Dans un second temps, la polysémie est généralement traitée par les méthodes de désambiguïsation en considérant le contexte dans lequel elle se situe.

Ainsi, ces différents aspects liés au langage naturel et aux méthodologies d'extraction sont des points clés pour obtenir des extractions de confiance à partir des textes permettant d'assurer la fiabilité d'un processus décisionnel à venir. Le chapitre 2 leur sera consacré.

1.3.2 Représentation de la connaissance

Les systèmes à base de connaissances permettent d'emmagasiner et structurer de la connaissance afin de lui associer une sémantique interprétable par les machines ; formalisme initial du Web sémantique (BERNERS-LEE, HENDLER et LASSILA, 2001). La mise en place d'un tel procédé s'appuie sur les caractéristiques des langages de représentation de la connaissance auxquels les logiques descriptives appartiennent (chapitre 1 dans BAADER, 2003). Les **logiques descriptives** sont les héritières des réseaux sémantiques et des langages de *frames* (MINSKY, 1974). Elles permettent la conceptualisation d'un domaine d'application de façon structurée en définissant une sémantique claire au travers de constructeurs logiques assurant une expressivité relativement riche (COULET, 2008). Ce faisant, la connaissance formalisée au sein des systèmes est considérée comme déclarative, explicite et monotone *i.e.* la remise en cause des connaissances acquises est interdite (BUITELAAR et CIMIANO, 2008). Les éléments fondamentaux constituant ces logiques sont les suivants, les exemples utilisés proviennent de la figure 1.8 :

- Concepts (classes) : ensemble des choses partageant des propriétés communes, *e.g.* Sportif, Sport.
- Instances : Objets appartenant à une classe, *e.g.* Tennis.
- Rôles, englobant :
  - Prédicats : Types de relations définissant les relations sémantiques entre les instances ou les classes, *e.g. sous_classe_de*, *pratique*.
  - Relations : Liens concrets entre les classes et les instances, *e.g. est_un*.
  - Attributs : Propriétés des instances, *e.g.* nbVictoires.

Ces éléments sont formalisés au sein d'une base de connaissances au travers d'un ensemble de constructeurs utilisant une syntaxe particulière (cf. tableau 1.2). Des exemples d'utilisation de cette syntaxe sont fournis par la suite.

Ces constructeurs permettent de générer les règles axiomatiques qui régissent la logique sous-jacente de la base de connaissances (cf. tableau 1.3).

DE GIACOMO et LENZERINI, 1996 ont organisé ces concepts, instances et axiomes au sein d'un cadre théorique constitué de deux grandes parties : la *T-Box* (*Terminological Box*) et la *A-Box* (*Assertional Box*). La *T-Box* permet de représenter la connaissance générale et abstraite d'un domaine. Elle englobe les concepts, les prédicats et les axiomes terminologiques représentés par la relation de subsomption et d'équivalence. La *A-Box*, quant à elle, représente un état particulier du domaine décrit

TABLEAU 1.2 – Principaux symboles de la syntaxe de la logique descriptive. Avec C, C_1 et C_2 des concepts, R un rôle et n un entier non nul.

Syntaxe du constructeur	Sémantique
$\top$	Concept universel
$\bot$	Ensemble vide
$C_1 \sqcap C_2$	Intersection
$C_1 \sqcup C_2$	Union
$\neg C$	Négation
$\forall R.C$	Quantificateur universel
$\exists R.C$	Quantificateur existentiel
($\geq$ n R)	Restriction de cardinalité

TABLEAU 1.3 – Syntaxe et sémantique des axiomes terminologiques et assertionnels. Avec C, C_1 et C_2 des concepts, R, R_1 et R_2 des rôles et a et b des instances de concepts, section 2.2 dans COULET, 2008.

	Type d'axiome	Syntaxe	Sémantique
T-Box	Définition de concept	$C_1 \equiv C_2$	Équivalence
	Définition de rôle	$R_1 \equiv R_2$	Équivalence
	Inclusion de concept	$C_1 \sqsubseteq C_2$	Subsomption
	Inclusion de rôle	$R_1 \sqsubseteq R_2$	Subsomption
A-Box	Assertion de concept	$C(a)$	Instanciation
	Assertion de rôle	$R(a, b)$	Rôle entre deux instances

FIGURE 1.7 – Exemple de constructeurs logiques définissant les règles axiomatiques de la *T-Box* et de la *A-Box*.

T-Box

— Personne $\sqsubseteq \top$
— Sport $\sqsubseteq \top$
— Sportif $\equiv$ Personne $\sqcap \exists$ pratique.Sport
— $\exists$ nbVictoires.$\top \sqsubseteq$ Sportif
— SportifCélèbre $\equiv$ Sportif $\sqcap (\geq 1000$ nbVictoires)

A-Box

— Personne(Pierre-Antoine)
— Sport(Tennis)
— Sport(Badminton)
— pratique(Pierre-Antoine, Tennis)
— nbVictoires(Pierre-Antoine, 150)

par la *T-Box*. Elle est constituée d'axiomes assertionnels permettant d'instancier les concepts et d'assertions de rôles représentées par les relations entre deux instances de concepts (cf. figure 1.7).

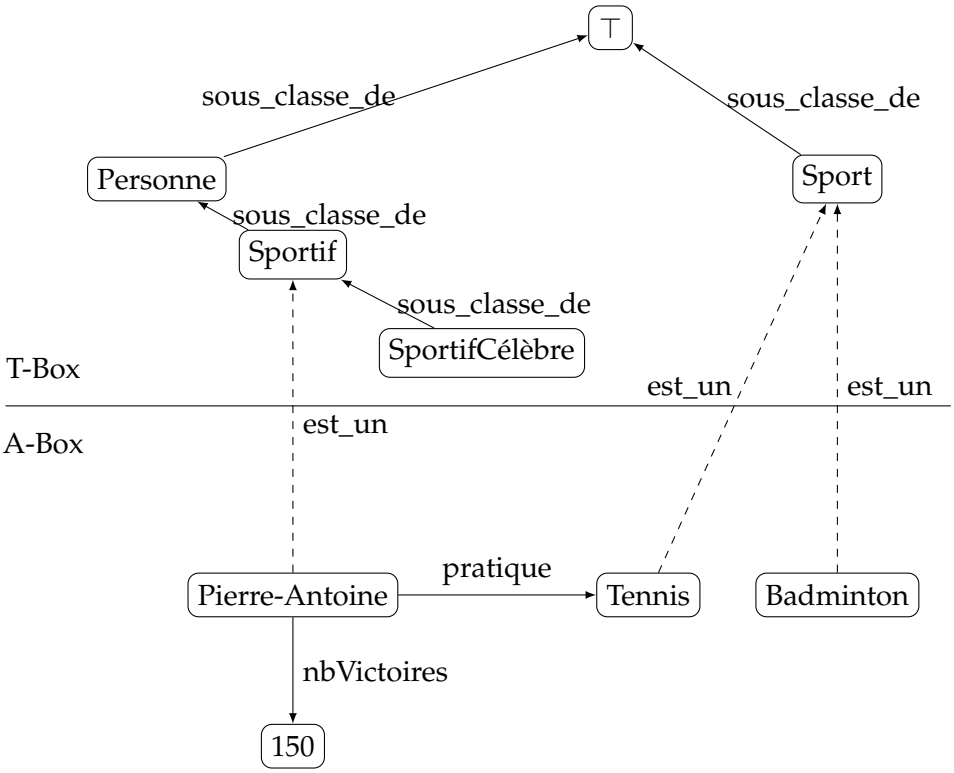
Bien que certaines sémantiques des bases de connaissances ne puissent être retranscrites au sein de graphes orientés acycliques (*e.g.* négation ou disjonction en monde ouvert), une large partie de la connaissance modélisée peut généralement l'être. La figure 1.8 représente une schématisation graphique des constructeurs logiques et axiomes proposés par la figure 1.7. Cette représentation graphique est suffisante dans l'application de nos travaux, et c'est donc celle qui sera utilisée dans le reste du manuscrit.

Il est à noter qu'une distinction peut être faite entre ontologie et base de connaissances. MAEDCHE et al., 2001 associent la *T-Box* et tous ces composants à l'ontologie et la *A-Box* à la base de connaissances. Certains articles adoptent également l'expression de graphe de connaissances en référence à cette seconde partie (NICKEL et al., 2016). Par conséquent, nous envisageons l'enrichissement automatique des bases de connaissances comme un processus guidé par une structure ontologique déjà établie, même s'il existe un cas exceptionnel (cf. sous-section 1.3.3). La section suivante présente les bases des méthodes permettant cet enrichissement.

1.3.3 Enrichissement des bases de connaissances

Les approches les plus reconnues dans le domaine de l'enrichissement automatique de bases de connaissances font référence à des bases très répandues en Web sémantique : YAGO (SUCHANEK, KASNECI et WEIKUM, 2007), DBPEDIA (AUER et al., 2007) ou bien FREEBASE (BOLLACKER et al., 2008). L'enrichissement automatique des bases

FIGURE 1.8 – Schématisation graphique d'une base de connaissances. La *T-Box* est représentée comme une taxonomie de concepts et la *A-Box* contient les informations et relations liées aux instances de ces concepts. On remarque la propriété de transitivité de la relation *est_un*/*sous_classe_de*.



de connaissances, appelé en anglais *knowledge base population* ou *ontology population* (DRUMOND et GIRARDI, 2008), a été popularisé en 2009 avec la *Text Analysis Conference* (MCNAMEE et DANG, 2009). Lors de cet événement, trois tâches ont été définies à partir de l'analyse de textes non structurés :

— *Slot filling* : l'objectif de cette tâche est de compléter toutes les informations connues étant donnée une entité. Un exemple classique est la récupération des informations factuelles à propos d'une personne, d'un lieu ou d'une organisation, *e.g.* étant donnée l'entité *Barack Obama* recueillir à partir d'un corpus de textes : son lieu de naissance, sa date d'anniversaire, son épouse, etc. Ainsi, cette tâche exploite les méthodes d'extraction d'information et plus précisément d'extraction de relations et d'événements.

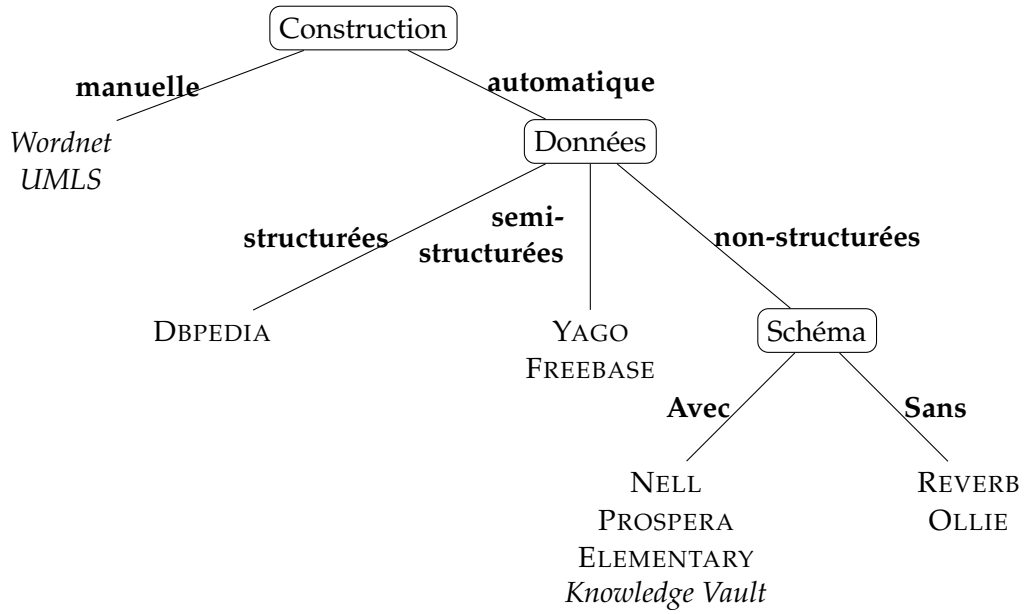
— *Entity linking* : une grande partie de la complexité du langage naturel réside dans son ambiguïté *e.g. Python* se réfère t-il au langage de programmation ou au serpent. L'*entity linking* a pour but de résoudre la polysémie des termes en réalisant une correspondance avec les concepts d'une ontologie. Cette tâche exploite les méthodes mises au point pour la tâche de désambiguïsation.

— *Cold start knowledge base population* : cette tâche démarre avec un schéma ontologique initial décrivant les types d'entités et les relations qui vont composer la base de connaissances. Initialement vide, les méthodes proposées doivent

alors extraire les entités et les relations d'intérêt à partir de documents textuels pour constituer la base de connaissances [4].

Le *slot filling* et l'*entity linking* sont des composants clés de nombreux modèles d'enrichissement. Ils exploitent les techniques du traitement automatique des langues pour convertir des données textuelles en informations exploitables, étapes indispensables aux méthodes automatiques exploitant les textes non structurés. Toutefois, toutes les approches d'enrichissement n'exploitent pas nécessairement des textes non structurés. DONG et al., 2014 proposent une classification des méthodologies en fonction du format de données exploité et de leur mode de fonctionnement. La figure 1.9 résume cette classification sous la forme d'un arbre de décision et propose différents exemples de systèmes.

FIGURE 1.9 – Classification des représentations de la connaissance en fonction de leur construction, des données exploitées et de l'utilisation d'un schéma ontologique initial.



La figure 1.9 propose une schématisation des principales caractéristiques définissant la construction d'une représentation de la connaissance en allant de la taxonomie conçue manuellement à la base de faits générée automatiquement. La première caractéristique discriminante est le mode de construction. Une élaboration manuelle est la plus coûteuse car elle fait appel à un groupe d'experts (ou de volontaires), *e.g.* WordNet (MILLER, 1995) ou bien UMLS (BODENREIDER, 2004) dans le domaine biomédical. La seconde caractéristique est le type des données exploitées : DBPEDIA utilise les données structurées contenues dans les *infobox* de Wikipedia tandis que

4. https ://tac.nist.gov/2017/KBP/ColdStart/guidelines/TAC_KBP_2017_ColdStartTaskDescription_1.0.pdf

YAGO combine les noms des catégories de Wikipedia, qualifiés de données semi-structurées car nécessitant la mise en place de moyens d'extraction, avec la taxonomie de *synsets* de WordNet. Enfin, la dernière caractéristique discriminante est l'utilisation d'un schéma ontologique initial. Lorsqu'un schéma est présent les entités et les relations sont représentées par un identifiant unique. Dans ces méthodes, nous pouvons citer : NELL (CARLSON et al., 2010), PROSPERA (NAKASHOLE, THEOBALD et WEIKUM, 2011), ELEMENTARY (NIU et al., 2012), FRED (PRESUTTI, DRAICCHIO et GANGEMI, 2012) ou bien *Knowledge Vault* (DONG et al., 2014). Tandis que l'absence d'un schéma initial implique que les entités et les relations soient normalisées, mais non désambiguïsées. Il est alors possible que la base de faits possède les triplets <Obama, né_à, Hawaï> et <Barack Obama, lieu_de_naissance, Honolulu>. Cette catégorie contient les méthodes exploitant l'extraction d'information en domaine ouvert telles que REVERB (ETZIONI et al., 2011) et OLLIE (SCHMITZ et al., 2012).

Cette classification (cf. figure 1.9) ne tient pas compte des méthodologies appartenant à l'*ontology learning* dont les principales étapes sont décrites dans le tableau 1.4. Ces méthodes sont orientées sur la construction automatique d'ontologies à partir de textes non-structurés. Ainsi, elles prennent en compte l'extraction des termes d'un domaine et les relations entre ces concepts, voire la génération d'axiomes régissant les concepts et les rôles de l'ontologie (CIMIANO et VÖLKER, 2005 ; VÖLKER, HITZLER et CIMIANO, 2007 ; WU et al., 2012).

TABLEAU 1.4 – Représentation des étapes de l'*ontology learning* réadaptée de l'exemple de BUITELAAR, CIMIANO et MAGNINI, 2005. Les mots *dom* et *range* se réfèrent au type des concepts attendus pour le sujet (domaine) et l'objet (co-domaine) de la relation.

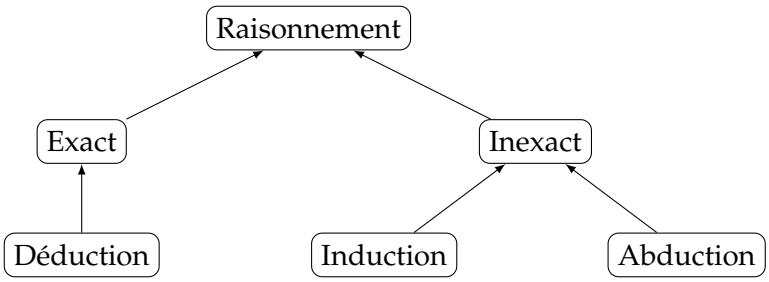
	Étapes	Exemples
1	Mots	matière, cours, école
2	Synonymes	{matière, cours}
3	Concepts	Matière
4	Hiérarchie de concepts	sous_classe_de(Professeur,Personne)
5	Relations	enseigner(*dom* :Professeur,*range* :Matière)
6	Règles	$\forall x, y$(enseigner$(x, y) \rightarrow$ a_étudié(x, y))

L'ensemble des méthodes citées précédemment s'inscrit dans l'hypothèse d'un monde ouvert. Il est important de souligner l'importance de ce paradigme. En effet, la lecture et l'interprétation de la base de connaissances sont conditionnées selon l'hypothèse de départ dans laquelle l'interprétation des triplets non-existants est différente (NICKEL et al., 2016). Contrairement à un raisonnement en monde fermé (*e.g.* dans les bases de données relationnelles) où toute donnée non présente est considérée comme étant fausse, sous l'hypothèse d'un monde ouvert un triplet non-existant est seulement interprété comme *inconnu*. Ainsi, la relation peut être soit vraie soit fausse. Par exemple, sur la figure 1.8, la non-existence d'un lien entre *Pierre-Antoine* et *Badminton* ne signifie pas que *Pierre-Antoine* ne pratique pas le *Badminton*.

1.3.4 Raisonnement

Il est important d'introduire les notions philosophiques liées au raisonnement pour comprendre les mécanismes d'inférence de notre chaîne de traitement. Le raisonnement correspond au procédé d'utilisation de la connaissance existante pour tirer des conclusions, faire des prédictions ou construire une explication. En logique aristotélicienne trois principaux types de raisonnement basés sur des syllogismes [5] ont été introduits : la déduction, l'induction et l'abduction (ANDREEWSKY et BOURCIER, 2000). La déduction permet d'aboutir à une conclusion logique et certaine tandis que l'induction et l'abduction sont qualifiées de raisonnement hypothétique ou inexact (cf. figure 1.10).

FIGURE 1.10 – Taxonomie des différents types de raisonnement.



Le raisonnement déductif commence avec l'assertion d'une règle générale vraie puis se déplace vers une application spécifique (HOEK et al., 2005). La figure 1.11 présente un exemple de raisonnement déductif. Au sein de ce dernier, la prémisse majeure indique que tous les objets appartenant à la classe Homme ont pour attribut "mortel" et la prémisse mineure indique que l'objet Socrate appartient à la classe Homme. Par conséquent, on déduit que Socrate doit être mortel puisqu'il hérite de l'attribut appartenant au concept parent Homme.

FIGURE 1.11 – Exemple de raisonnement déductif.

Prémisse 1	Tous les hommes sont mortels
Prémisse 2	Socrate est un homme
Conclusion	Socrate est mortel

La raisonnement inductif est basé sur la généralisation (KETOKIVI et MANTERE, 2010). Ce type de raisonnement exploite des observations spécifiques afin d'inférer des règles générales à propos d'un domaine. Ces dernières sont probables au vu des observations mais non certaines (cf. figure 1.12).

Le raisonnement abductif désigne une forme de raisonnement qui permet d'expliquer un phénomène ou une observation à partir de certains faits. C'est la recherche

5. Raisonnement logique à deux propositions (prémisses) conduisant à une conclusion. La première prémisse s'appelle la majeure et la seconde la mineure.
6. https ://fr.wikipedia.org/wiki/Induction_(logique).

FIGURE 1.12 – Induction simple réalisée par la généralisation des observations [6].

La proportion Q de l'échantillon a l'attribut A.
Par conséquent :

La proportion Q de la population a l'attribut A.

des causes, ou d'une hypothèse explicative (CATELLIN, 2004). Ce type de raisonnement débute traditionnellement avec un ensemble d'observations incomplètes puis propose une explication la plus probable pour cet ensemble. Pour illustrer ce type de raisonnement, WOUTERS, 1998 a réalisé un parallèle avec le mode d'inférence du personnage de fiction policière Maigret. En effet ce dernier, sur la base des indices qu'il rassemble au cours de son enquête, tente de reconstituer l'histoire du crime. L'auteur qualifie ce procédé de pensée heuristique d'abductif. On peut également réaliser l'analogie entre ce type de raisonnement et la façon de penser d'un médecin lorsqu'il est confronté à un ensemble de symptômes et qu'il doit inférer une maladie permettant de les expliquer.

Par conséquent, la construction d'un raisonnement revient à appliquer l'un de ces trois types de raisonnement sur une problématique initiale dans l'objectif d'inférer des propositions (règles, explications, causes) répondant à cette dernière. Dans le cas de notre chaîne de traitement, la phase d'inférence est inductive. À partir des observations extraites dans les textes, nous tentons d'inférer des généralisations de ces dernières. Ce mode d'inférence inexact s'introduit bien dans le contexte d'un monde ouvert dans lequel notre connaissance du monde est incomplète.

1.3.5 Questions-réponses

La découverte de connaissances, qui est l'objet de ce travail de recherche, peut aboutir par extension à une problématique de questions-réponses permettant d'appuyer ou non d'éventuelles intuitions/questions de la part d'un utilisateur. Le domaine du questions-réponses sera plus amplement détaillé au chapitre 5 dans le cadre de la validation de notre chaîne de traitement. Cette sous-section présente essentiellement les architectures traditionnellement employées dans ce domaine afin de positionner plus précisément notre approche. JURAFSKY et MARTIN, 2014 proposent une décomposition de ces architectures (principalement rattachées aux questions factuelles) en trois principaux types.

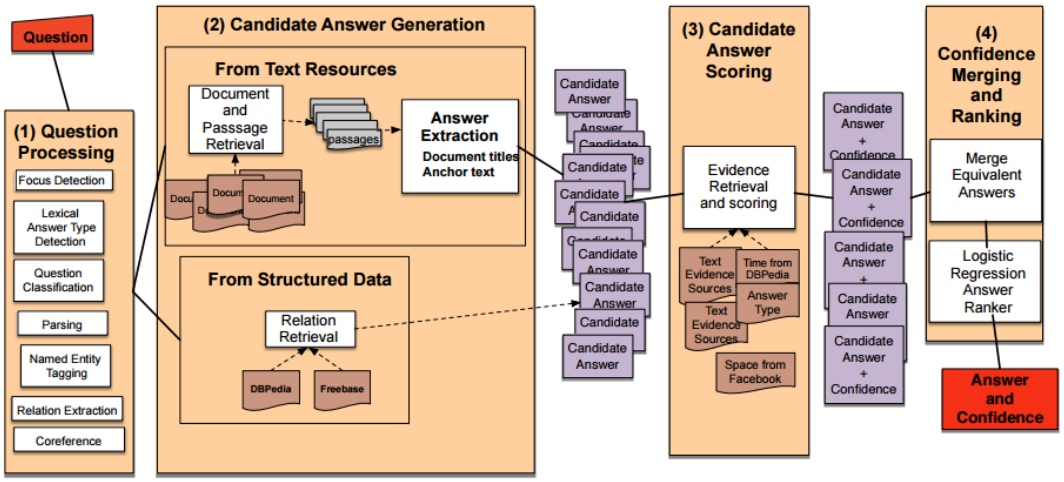
— Systèmes basés sur le traitement automatique des textes.
— Systèmes basés sur des connaissances structurées.
— Systèmes hybrides mixant les approches précédentes.

Les systèmes basés sur le traitement automatique des textes se composent communément de trois modules : le traitement de la question, la récupération des passages

pertinents dans une base de documents et le renvoi de la réponse nécessitant l'extraction d'une information précise. Par conséquent, outre le traitement de la question dont l'objectif est d'extraire le type d'entité recherchée et les mots clés de la requête (LI et ROTH, 2006), les domaines de la recherche et de l'extraction d'information sont le cœur de ces systèmes.

Concernant les systèmes basés sur la connaissance, ils sont fondés sur l'idée de répondre à une question énoncée en langage naturel en réalisant une correspondance entre cette question et une source de connaissances structurées (JURAFSKY et MARTIN, 2014). L'une des premières approches à exploiter ce paradigme fut le système BASEBALL exploitant une base de données stockant diverses informations sur les compétitions de *baseball* (GREEN JR et al., 1961). Ces systèmes sont basés sur des approches permettant de convertir sous une forme logique une phrase écrite en langage naturel. Par la suite, une forme logique peut aisément être traduite au format SQL ou SPARQL selon les structures employées (base de données, ontologie). Pour finir, les systèmes hybrides considèrent à la fois les informations contenues dans les textes et dans les données structurées. La figure 1.13 présente une architecture type d'un de ces systèmes de questions-réponses hybrides.

FIGURE 1.13 – Architecture de l'approche Watson présentée lors de l'émission télévisuelle *Jeopardy !* en 2011 (FERRUCCI, 2012). Figure provenant de l'article de JURAFSKY et MARTIN, 2014.



Notre approche a une architecture comparable à celle présentée lors de cette dernière catégorie. En effet, nous retrouvons les trois principaux blocs contenus au sein de cette architecture : l'extraction d'information, le calcul d'un score pour chaque extraction et la sélection des extractions pertinentes. Toutefois, notre chaîne de traitement aborde chacun d'entre eux d'une manière différente. Dans notre cas, l'extraction d'information est couplée à une phase d'inférence permettant d'acquérir de la connaissance supplémentaire, éventuellement non exprimée dans les textes, et le calcul d'un score est réalisé au travers de l'attribution d'un ensemble de critères exploité lors de la phase de sélection.

Les concepts fondamentaux de la représentation des connaissances et de l'enrichissement de bases de connaissances étant définis, la section suivante présente la ligne directrice de la chaîne de traitement exposée au sein de cette thèse et les solutions proposées pour réaliser un lien entre le texte et la connaissance.

1.4 Intuition de l'approche proposée

La chaîne de traitement décrite dans cette thèse permet d'extraire de l'information textuelle en monde ouvert, d'inférer de la connaissance et de l'évaluer. Ce processus d'inférence tient compte à la fois de l'ensemble des informations délivrées par les relations extraites et d'une source de connaissances externe (si l'utilisateur désire étendre la capacité d'inférence). Sa portée a pour ambition de couvrir plusieurs domaines dont la découverte de connaissances, les systèmes de questions-réponses ou l'enrichissement des bases de connaissances. Compte tenu de ces objectifs, la chaîne de traitement se situe au regard de la figure 1.9 au niveau des méthodes de construction sans schéma ontologique initial. Toutefois, son originalité repose sur un module de raisonnement constitué de deux principales parties. La première correspond à une étape d'inférence de connaissances réalisée sur les données extraites par l'intermédiaire d'un processus d'induction. Ce dernier permet de découvrir de la connaissance par la généralisation des observations. Ainsi, toute déclaration induite lors de ce processus et qui n'est pas explicitée dans les textes est dite découverte[7]. La seconde partie, quant à elle, représente la procédure d'évaluation de la pertinence des connaissances réalisée au travers d'une phase de sélection des déclarations. Pour cela, elle exploite différents modèles de sélection s'appuyant sur des critères spécifiques, calculés pour chaque déclaration. Enfin, une autre originalité de notre approche est de considérer l'incertitude linguistique au niveau du module de raisonnement.

Les sous-sections suivantes présentent une vue globale du module de raisonnement allant de la phase d'inférence à l'étape de sélection en passant par les moyens de détecter et considérer l'incertitude linguistique.

1.4.1 Inférence de connaissances

Le processus d'inférence est réalisé au travers de la structuration des sujets et des objets des déclarations étudiées au sein d'un ordre partiel. Ce dernier permet d'exploiter une propriété d'inclusion lexicale appliquée aux différents syntagmes nominaux. Prenons l'exemple de deux phrases issues de Wikipedia et des relations qu'elles explicitent (cf. figure 1.14).

7. Dans le manuscrit, ce terme de découverte est généralement remplacé par "génération de déclarations" pour le distinguer du domaine scientifique de la découverte de connaissances.

FIGURE 1.14 – Exemples de phrases issues de Wikipedia et des relations <sujet, prédicat, objet> exprimées. Le prédicat est_en_rapport_avec est considéré synonyme du prédicat est_liée_à.

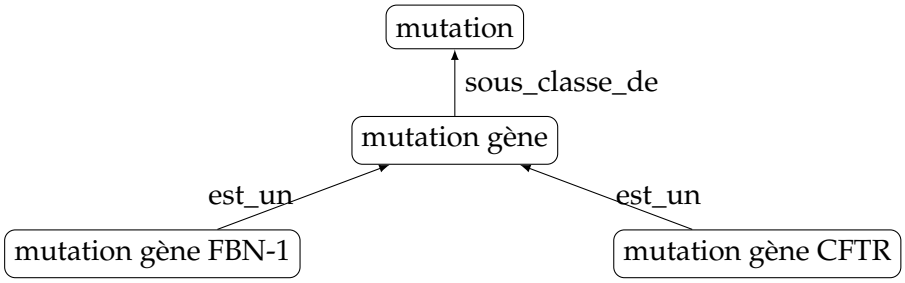
Le syndrome de Marfan est en rapport avec une mutation du gène FBN-1.
La mucoviscidose est liée à des mutations du gène CFTR.

↓

<syndrome de Marfan, est_liée_à, mutation gène FBN-1>
<mucoviscidose, est_liée_à, mutation gène CFTR>

L'observation des syntagmes nominaux "mutation gène FBN-1" et "mutation gène CFTR" évoquent un concept plus général "mutation gène", qui lui même est une spécialisation de "mutation". Ainsi, en tenant compte d'une structuration taxonomique nous pouvons mettre en évidence une hiérarchie entre ces concepts (cf. figure 1.15).

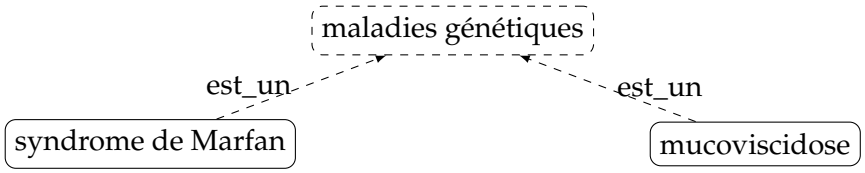
FIGURE 1.15 – Liens d'implication syntaxique tenant compte de la décomposition de deux syntagmes nominaux.



Des premiers éléments de connaissances émergent de cette structuration par rapport aux relations extraites. En effet dans notre exemple, nous pouvons maintenant générer les relations : <syndrome de Marfan, est_liée_à, mutation gène> et <mucoviscidose, est_liée_à, mutation gène>.

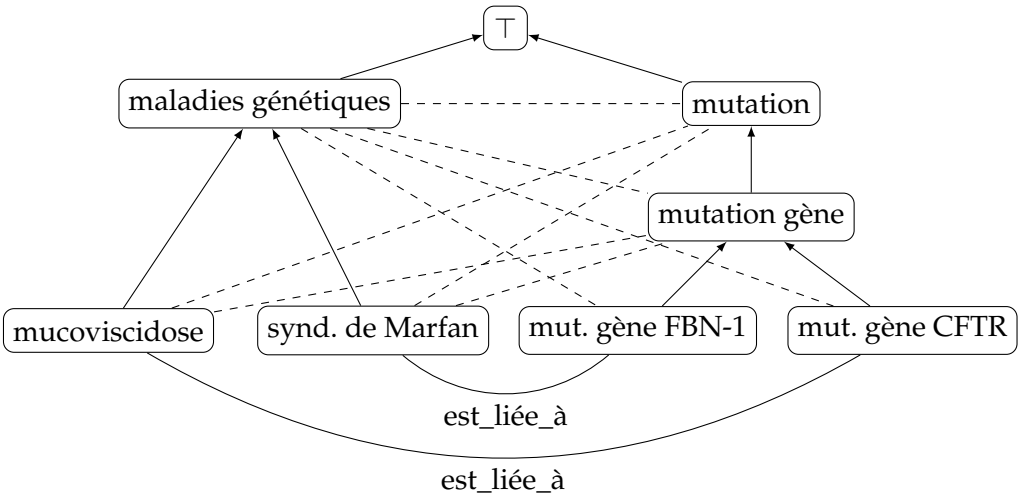
L'ordre partiel établi sur les syntagmes permet également de tirer parti d'une connaissance *a priori* exprimée au sein d'une taxonomie (*T-box*) dans le but d'intégrer par exemple que toute observation des termes "mucoviscidose" et "syndrome de Marfan" correspondent à une évocation de "maladies génétiques" (cf. figure 1.16). Toutefois cela suggère la mise en place au sein de la chaîne de traitement d'un moyen de mise en correspondance entre les termes et les concepts d'une taxonomie, abordant ainsi les problématiques de désambiguïsation.

FIGURE 1.16 – Enrichissement des sujets des relations à partir d'une structuration de la connaissance externe. Les pointillés représentent la connaissance externe.



Ces implications directes (syntaxiques) et indirectes (source externe) renferment des informations supplémentaires par rapport aux relations extraites. Cette connaissance se traduit par la génération de nouvelles relations. En effet, cette structuration permet de mettre en correspondance des concepts faisant référence de façon implicite et non triviale à une même évocation (cf. figure 1.17).

FIGURE 1.17 – Génération de nouvelles relations. Les flèches représentent les relations de subsomption, les lignes pleines les relations extraites et les tirets les relations générées.

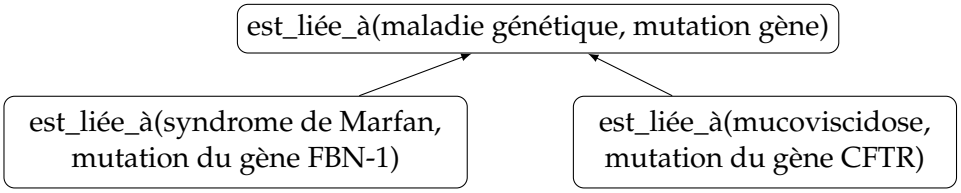


Ce processus de génération permet de faire émerger de nouvelles informations *e.g.* la structuration des déclarations <mucoviscidose, est liée à, mutation gène CFTR> et <syndrome de Marfan, est liée à, mutation gène FBN-1> font implicitement référence à la déclaration éventuellement non observée <maladie génétique, est liée à, mutation gène> puisqu'il *existe* au moins une maladie génétique liée à au moins une mutation génétique.

1.4.2 Évaluation de la pertinence des déclarations

La seconde partie du module de raisonnement porte sur les modalités d'évaluation de la pertinence des déclarations extraites et générées. Pour cela, nous définissons deux critères permettant de discriminer les relations : la croyance et la spécificité. Chacune de ces caractéristiques est estimée à partir de la construction d'un nouveau graphe hiérarchisant l'ensemble des relations extraites et générées en fonction des relations de subsomption entre les sujets et objets des déclarations (cf. figure 1.18).

FIGURE 1.18 – Structuration des relations extraites et générées.



Au travers de ce graphe structurant les déclarations, le critère de spécificité correspond à la profondeur d'une déclaration et la croyance au nombre d'observations d'une déclaration augmenté du nombre d'observations des déclarations plus spécifiques. Ainsi, la valeur de croyance découle d'un processus de propagation ascendante et monotone des observations.

Par la suite, les critères de spécificité et de croyance permettent d'évaluer la pertinence de chaque déclaration au travers de différents modèles de sélection. Ces derniers matérialisent diverses façons d'exploiter ces critères dans l'objectif d'estimer la valeur de vérité des déclarations et de les filtrer.

1.4.3 L'incertitude linguistique dans le module de raisonnement

La chaîne de traitement propose de considérer la prise en compte de l'incertitude linguistique dans le processus d'inférence. En effet, une information contenant un marqueur d'incertitude ne peut être considérée au même niveau qu'une information certaine *e.g.* "Je crois que Barack Obama est américain." et "Barack Obama est américain" représentent deux niveaux d'information différents tels que le poids accordé à la seconde phrase est plus fort que celui de la première phrase. Dans un contexte de découverte de connaissances, nous nous intéressons notamment à l'agrégation de signaux faibles provenant des extractions. Cette agrégation retranscrit des intuitions sur des connaissances à partir de faits peu exprimés dans les textes. Ainsi, nous pouvons supposer que la prise en compte de l'incertitude linguistique peut jouer un rôle non négligeable sur la façon d'exploiter les extractions lors de l'inférence de connaissances. Toutefois, il est probable que ce rôle soit amoindri dans le cadre d'observations nombreuses à propos d'une relation déjà connue.

La prise en compte de l'incertitude linguistique dans le module de raisonnement est réalisée par l'intermédiaire du processus de propagation des observations. En effet, le poids accordé aux observations peut être modifié en fonction de la qualité des déclarations et notamment de l'incertitude qu'elles véhiculent.

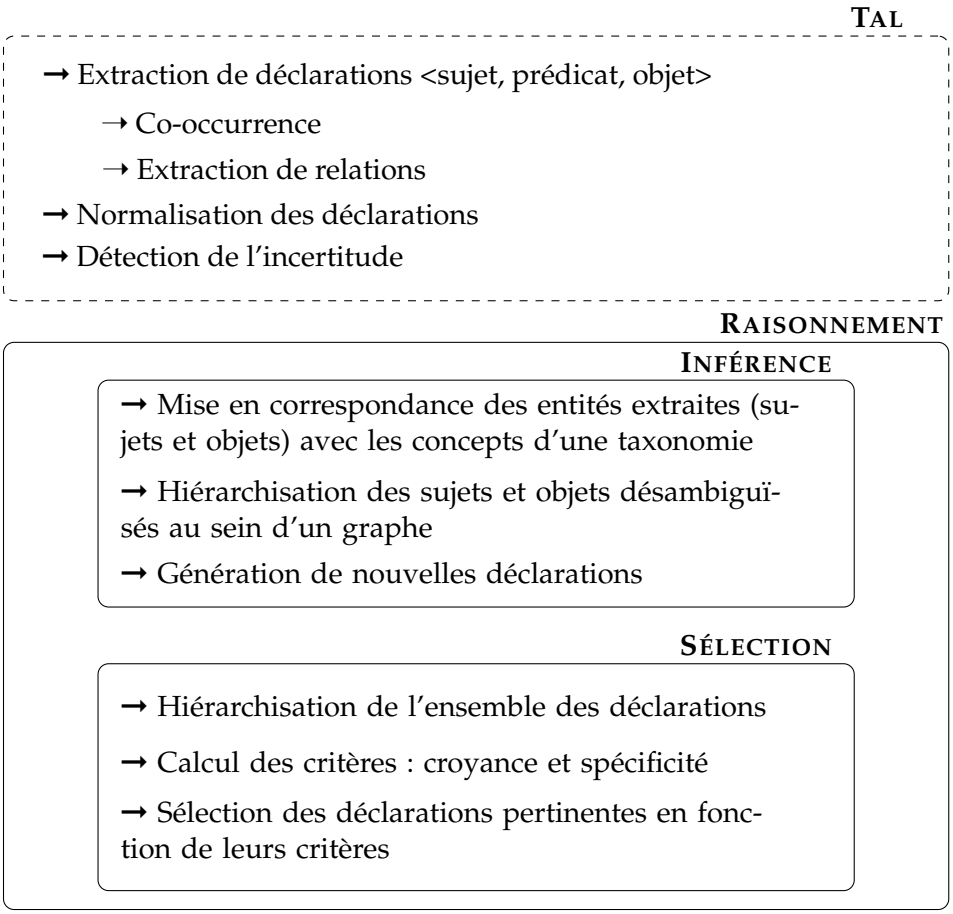
Concernant la détection de l'incertitude à partir du langage naturel, nous avons développé une méthode basée sur de l'apprentissage. Pour cela, nous réalisons une représentation vectorielle des phrases à partir d'une agrégation spécifique des poids associés à l'ensemble des unités qui la composent. Ces unités peuvent être les unigrammes, bi-grammes, etc., tandis que les poids représentent une probabilité conditionnelle d'appartenir à une classe donnée de l'incertitude. Cette approche a été évaluée sur des jeux d'évaluation standards.

La prochaine section expose une vue générale du manuscrit avec une courte présentation des chapitres à suivre.

1.5 Organisation du manuscrit

La structure principale du manuscrit suit la ligne directrice de la chaîne de traitement. La figure 1.19 propose un schéma architectural récapitulant les étapes et les principaux blocs de celle-ci.

FIGURE 1.19 – Architecture de la chaîne de traitement au regard des modules d'extraction d'information et de raisonnement.



Le manuscrit se décompose en six chapitres (en comptant ce premier chapitre introductif) allant du traitement automatique des textes à l'inférence et la sélection des déclarations en terminant par la validation de la méthode.

Le chapitre 2 est axé sur la description des défis posés par la complexité du langage naturel et les moyens d'y répondre. De l'imprécision à l'incertitude en passant par les problématiques posées par la polysémie des termes, ce chapitre propose une vue globale des méthodologies du traitement automatique du langage naturel permettant d'aborder les problématiques générales d'*entity linking* (désambiguïsation) et de *slot filling* (extraction de relations).

Dans le chapitre 3, la première contribution de la thèse est présentée. Celle-ci concerne la détection de l'incertitude linguistique, étape primordiale dans cette étude

visant à découvrir l'impact de ce type d'incertitude au sein d'un processus de raisonnement. Cette méthode de détection s'appuie sur une approche à base d'apprentissage supervisé dans laquelle chaque phrase est représentée comme un vecteur de caractéristiques. Chaque dimension de ce vecteur est caractérisée par une valeur numérique issue d'une agrégation spécifique des scores associés à chaque unité de la phrase (uni-gramme, bi-gramme, etc.). Cette méthode est évaluée au travers des critères de la conférence CoNLL 2010.

Le chapitre 4 présente le module de raisonnement. Ce dernier comprend la phase d'inférence de connaissances et l'étape d'évaluation de la pertinence et de sélection des déclarations. La phase d'inférence est réalisée au travers de la construction d'un ordre partiel structurant les syntagmes nominaux du sujet et de l'objet des déclarations extraites. Cet ordre, pouvant être enrichi par l'intermédiaire d'une taxonomie de concepts, permet de générer de nouvelles déclarations et de guider la phase d'évaluation. En effet, cette seconde phase exploite la structuration de ce dernier graphe pour hiérarchiser l'ensemble des déclarations extraites et générées. Le graphe résultant de cette hiérarchisation permet de calculer différents critères servant à évaluer la pertinence des déclarations au travers d'un ensemble de modèles de sélection.

Le chapitre 5 expose les conditions expérimentales établies pour l'évaluation de la chaîne de traitement. Cette évaluation est réalisée au travers du domaine du questions-réponses et de la modification d'un algorithme de génération automatique de questionnaires à choix multiples (QCM). La modification proposée permet de générer des questionnaires à partir d'une structure taxonomique et d'un ensemble établi de relations. Dans l'optique de répondre à ces questionnaires générés, un jeu de données a été constitué au terme d'une phase d'extraction automatique de phrases issues du Web. Cette évaluation a permis de montrer l'efficacité du module de raisonnement, de comparer les différents modèles de sélections et d'observer l'impact de la prise en compte de l'incertitude.

Le chapitre 6 conclut le manuscrit. Il présente un résumé des recherches menées au sein de cette thèse et les contributions réalisées.

Enfin, le manuscrit comporte deux annexes :

— L'annexe A présente une contribution à la compétition COLIEE (*Competition on Legal Information Extraction/Entailment*). Cette compétition internationale concerne les domaines de la recherche d'information et de l'implication textuelle et elle est appliqués aux textes de loi. L'objectif était de concevoir un modèle capable de retrouver les articles du code civil permettant de répondre à un cas juridique donné. Le système soumis a obtenu le 4e meilleur résultat sur 17 soumissions.

— L'annexe B réalise un lien entre le critère de croyance défini pour caractériser les déclarations au sein des modèles de sélection et la notion de croyance définie par le cadre théorique des fonctions de croyance.

Chapitre 2

Traitement de l'information textuelle

Sommaire

Ce chapitre est dédié à la description des problématiques linguistiques auxquelles notre module d'extraction de connaissances est confronté et à la présentation des méthodologies d'analyse qu'il exploite. Il introduit les enjeux de cette étape et présente un état de l'art dans le domaine de l'extraction de relations et de la désambiguïsation. La première partie énumère les subtilités du langage naturel auxquelles nous nous intéressons, tandis que la seconde propose l'état de l'art sur les méthodologies d'extraction d'information. Cette thèse est écrite en français mais les traitements qui y sont proposés se veulent aussi génériques que possible. Cette remarque a toute son importance dans un contexte d'analyse linguistique où justement les traitements devront être adaptés aux spécificités de chaque langue. Nous n'en abordons que deux : le français et l'anglais. Les exemples donnés dans ce chapitre seront donc indifféremment en français ou en anglais. Lorsqu'une particularité liée à la langue impose un traitement différent, celui-ci sera décrit.

2.1 Quelques problématiques majeures du traitement automatique des langues

Le succès d'une chaîne d'extraction de connaissances à partir des textes est principalement conditionné par ses aptitudes à gérer la complexité du langage naturel. Celle-ci fait l'objet d'une recherche avancée depuis le début du XX^e siècle avec l'apparition des premiers modèles de la linguistique moderne (DE SAUSSURE, 1916 ; HARRIS, 1962 ; CHOMSKY, 1964) et des premiers programmes informatiques analysant le langage naturel dans un monde fermé *e.g.* SHRDLU un programme communiquant avec un opérateur humain pour réaliser des tâches rudimentaires (WINOGRAD, 1971) ou LUNAR un programme fournissant des réponses en langage naturel à des questions sur l'analyse géologique de la lune par les missions Apollo (WOODS, KAPLAN et NASH-WEBBER, 1972).

Le traitement automatique des langues (et du signal) propose des solutions linguistiques et algorithmiques pour analyser les différentes branches de la linguistique (cf. figure 2.1). Nous nous intéressons plus particulièrement à trois de ces branches : la morphologie, la syntaxe et la sémantique. Les analyses morphologiques et syntaxiques ont pour objectif d'analyser les caractéristiques intrinsèques du langage. Tandis que l'analyse sémantique permet d'appréhender la signification littérale des phrases.

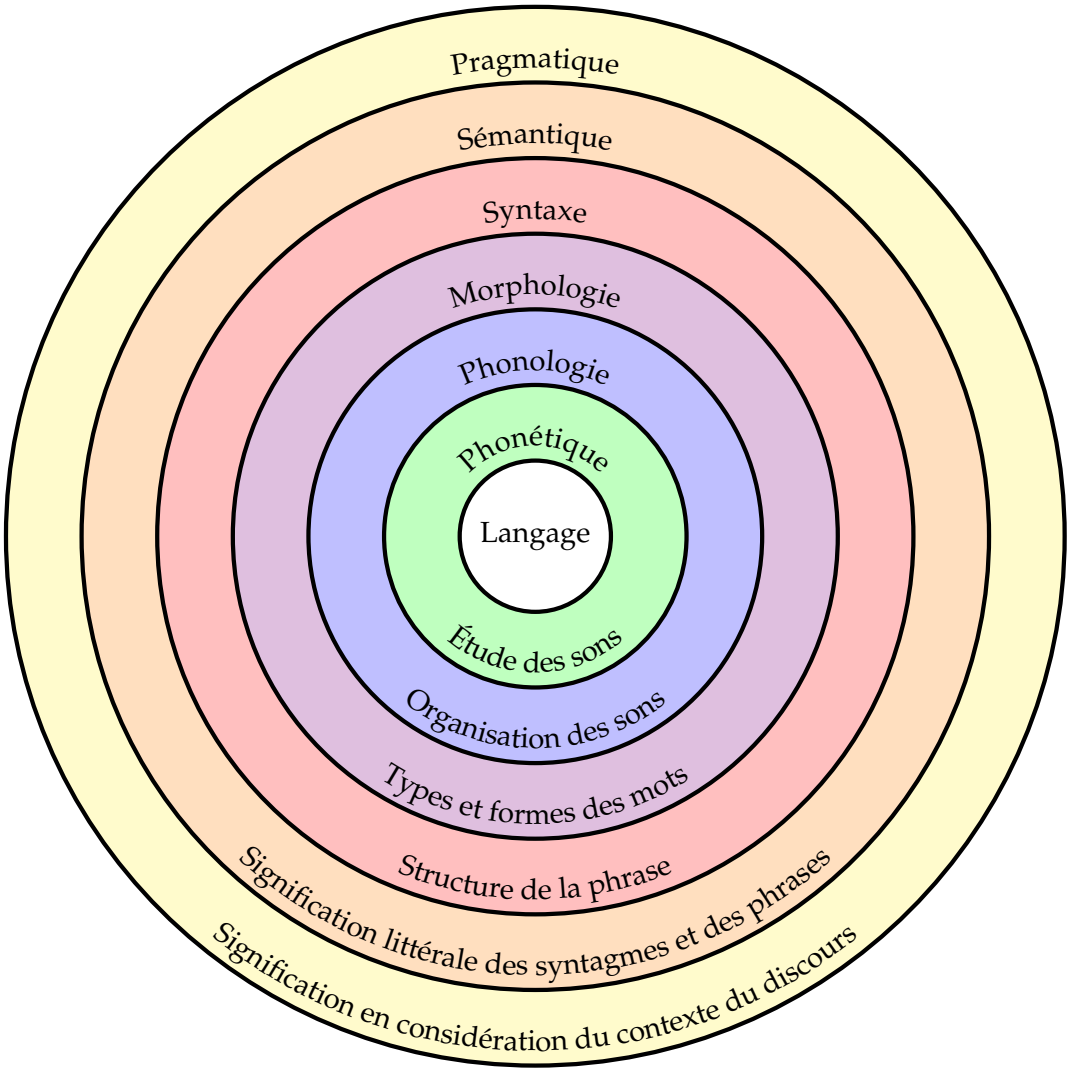
2.1.1 L'analyse morphologique

La morphologie lexicale est abordée au travers des modules de lemmatisation dont l'objectif est de présenter un mot sous sa forme non fléchie, appelée lemme (BRUNET, 2002). Les formes fléchies correspondent aux formes conjuguées ou accordées d'un mot de base, *e.g.* le lemme *jouer* possède plusieurs flexions correspondant à ses formes conjuguées à diverses personnes, temps et modes : *il jouera , nous jouons* [1]... La lemmatisation est distincte de la racinisation d'un mot, pour lequel il s'agit d'obtenir sa racine c'est-à-dire la partie restante une fois le préfixe et le suffixe supprimés (PORTER, 1980) *e.g.* le mot *chercher* a pour radical (ou *stemme*) *cherch* qui ne correspond pas à un mot réel. Selon les contextes d'utilisation, l'intérêt de la lemmatisation est plus ou moins discutable. Dans le cadre de petit corpus, la réduction du nombre de formes à considérer a pour avantage d'augmenter les fréquences des formes. Toutefois, dans un contexte d'extraction de significations ce traitement entraîne une perte d'information qui peut être préjudiciable (LEMAIRE, 2008).

Au sein de notre chaîne de traitement, nous exploitons les formes fléchies des termes explicités dans les relations. En effet, cette opération permet de diminuer le nombre de nœuds à considérer dans le graphe des syntagmes et des déclarations permettant

1. http ://blog.onyme.com/lemmatisation-et-racinisation-en-francais-flexion-lemme-et-racine-dun-mot/

FIGURE 2.1 – L'imbrication des différentes branches appartenant à la linguistique (MOESCHLER et REBOUL, 1998).



d'améliorer le temps et l'espace consommé par la mémoire dans le cadre de corpus volumineux (plusieurs centaines de Go). Nous la réalisons par le biais de la bibliothèque *nltk* [2] sous Python.

2.1.2 L'analyse syntaxique

En ce qui concerne la syntaxe, elle est généralement traitée au travers de l'ordre des mots et des règles de grammaire régissant les phrases. L'application la plus courante est l'étiquetage morpho-syntaxique (aussi appelé étiquetage grammatical ou *POS tagging* (*part-of-speech tagging*) en anglais). Cette méthode permet d'attribuer la catégorie grammaticale correspondant à chaque mot de la phrase (cf. figure 2.2).

Le tableau 2.1 présente l'ensemble des fonctions grammaticales défini lors de la construction du corpus annoté *The Penn Treebank* (MARCUS, MARCINKIEWICZ et

2. www.nltk.org

FIGURE 2.2 – Étiquetage morpho-syntaxique d'une phrase. Avec NNP un nom propre, VBZ un verbe au présent à la troisième personne du singulier, DT un déterminant, JJ un adjectif, NN un nom commun et IN une préposition (SANTORINI, 1990).

John likes the blue house at the end of the street.

↓

John | **NNP** likes | **VBZ** the | **DT** blue | **JJ** house | **NN** at | **IN** the | **DT** end | **NN** of | **IN** the | **DT** street | **NN** . | .

SANTORINI, 1993). Il regroupe 36 étiquettes morpho-syntaxiques et 12 autres étiquettes utilisées principalement pour la ponctuation.

TABLEAU 2.1 – Étiquettes morpho-syntaxiques exploitées dans le projet *The Penn Treebank* de l'Université de Pennsylvanie.

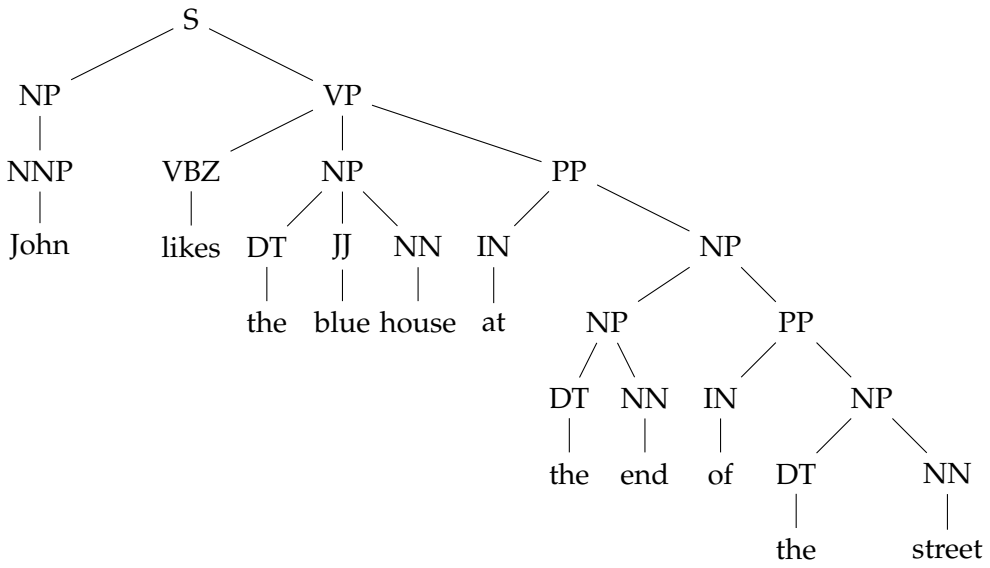
CC	Conj. de coordination	PRP$	Pronom possessif
CD	Nombre cardinal	RB	Adverbe
DT	Déterminant	RBR	Adverbe, comparatif
EX	Clause existentielle (*there*)	RBS	Adverbe, superlatif
FW	Mot étranger	RP	Particule
IN	Prep./conj. de subordination	SYM	Symbole (math./scientifique)
JJ	Adjectif	TO	*to*
JJR	Adjectif, comparatif	UH	Interjection
JJS	Adjectif, superlatif	VB	Verbe, forme de base
LS	Puce	VBD	Verbe, passé
MD	Modal	VBG	Verbe, gérondif/participe présent
NN	Nom, sing.	VBN	Verbe, participe passé
NNS	Nom, pluriel	VBP	Verbe, non-3^e pers., sing. présent
NNP	Nom Propre, sing.	VBZ	Verbe, 3^e pers., sing. présent
NNPS	Nom Propre, pluriel	WDT	Wh-déterminant
PDT	Prédéterminant	WP	Wh-pronom
POS	Forme possessive	WP$	wh-pronom possessif
PRP	Pronom personnel	WRB	Wh-adverbe

Les dépendances grammaticales entre les mots sont fréquemment exploitées pour représenter la structure syntagmatique et grammaticale d'une phrase. Cette structure est souvent représentée sous la forme d'arbre dans lequel les feuilles sont les mots, leurs parents les étiquettes morpho-syntaxiques et les nœuds supérieurs les syntagmes (cf. figure 2.3).

Les caractéristiques syntaxiques citées précédemment sont au cœur des méthodes permettant d'analyser la co-référence, la temporalité, les déclinaison des mots ou bien la grammaire, autant de facteurs déterminants pour la structure d'une phrase et l'analyse précise de sa sémantique. En effet, l'analyse de la structure d'une phrase est primordiale pour comprendre sa sémantique. Une dé-contextualisation des mots peut modifier les informations véhiculées.

La chaîne de traitement présentée dans le manuscrit s'appuie notamment sur des patrons syntaxiques dans le but de normaliser l'information exploitée (cf. chapitre

FIGURE 2.3 – Arbre des dépendances grammaticales pour une phrase donnée. Avec NP l'abréviation pour un syntagme nominal (nominal phrase), VP un syntagme verbal et PP un syntagme prépositionnel.



4). Cette opération est réalisée par la bibliothèque nltk.

2.1.3 L'analyse sémantique

L'analyse sémantique évalue le sens des informations véhiculées en tenant compte des déclarations et de leur contexte. Elle représente une grande difficulté pour la machine de par la richesse du langage naturel. Considérons par exemple ces deux phrases :

— *Turkish authorities have said all the suicide bombers were Turks.*
— *Ankara says all four terrorists were Turkish.*

Un humain réalise facilement une correspondance entre ces deux phrases, sachant qu'Ankara est la capitale de la Turquie et connaissant la forte similarité entre *suicide bombers* et *terrorists*. Tandis que la machine fait uniquement la correspondance avec *all* et *were*, voire avec *said/says* et *Turks/Turkish* si elle est dotée d'un lemmatiseur : soit des mots ne retranscrivant pas l'information véhiculée. Par conséquent, la machine requiert un ensemble d'outils pour lui permettre de calculer la similarité entre des groupes de mots et saisir les connaissances implicites et contextuelles. L'ingénierie des connaissances propose des méthodologies pour tenter de modéliser ces outils : mesures de similarité, ordre partiel sur les éléments de la connaissance, etc. Cette problématique de similarité a fait l'objet de compétitions à l'image de SemEval en 2012[3] dans laquelle les participants devaient évaluer 2000 paires de phrases en attribuant un score entre 0 (faible) et 5 (élevé) pour représenter le degré de similarité entre les phrases (AGIRRE et al., 2012).

3. https ://www.cs.york.ac.uk/semeval-2012/task6.html

Toutefois, la sémantique d'une information est fortement soumise aux variations énonciatives des auteurs et implique une analyse approfondie des termes et marqueurs employés pouvant modifier une information exacte en une information imprécise et/ou incertaine. Par exemple :

— *Some people say all terrorists were Turkish.*

Cette phrase présente ici un marqueur d'incertitude. En effet, nous passons d'une information certaine énoncée par des autorités compétentes à une information véhiculée par quelques personnes non identifiées. Ainsi, nous ne pouvons pas accorder le même crédit à ces deux informations.

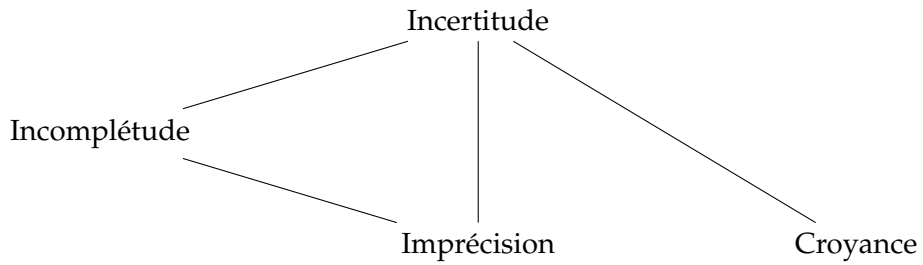
Cette thèse se focalise sur les problématiques d'inférence de connaissances à partir des textes en tenant compte de la manière dont une information est énoncée. L'imprécision et l'incertitude inhérentes au langage naturel peuvent changer fortement la sémantique d'une phrase ; elles constituent donc des données centrales au sein de notre chaîne de traitement. La section suivante présente ces deux notions au travers de différentes classifications et définitions transposées au domaine de la linguistique.

2.2 Imprécision et incertitude dans l'exploitation des textes en domaine ouvert

De nombreuses classifications de l'incertitude et de l'imprécision ont été proposées pour distinguer les différents tenants et aboutissants de ces deux notions. JOUSSELME, MAUPIN et BOSSÉ, 2003 mettent en regard diverses classifications dans un contexte élargi par rapport au langage naturel (théorie des ensembles flous, fusion de l'information). L'analyse de toutes ces classifications démontre une absence de consensus dans la littérature. Toutefois, plusieurs classifications se distinguent. La première est proposée par BOUCHON-MEUNIER et NGUYEN, 1996, elle impute l'incertitude au concept plus général de l'imperfection sur la connaissance (cf. figure 2.4). Cependant, ce modèle dénote l'imprécision comme une part de l'incertitude, compréhensible dans les résultats d'approche automatisée mais difficilement transposable au langage naturel. En effet, une déclaration linguistique peut être précise et incertaine "Je ne suis pas sur que John ait 30 ans" mais également certaine et imprécise "Je suis sur que John a au moins 35 ans". Cette caractérisation de l'incertitude par rapport à l'imprécision est semblable à celle proposée par FUCHS, 2008. En effet, l'auteur qualifie l'imprécision comme un degré de certitude interprétative dans laquelle l'incertitude décroît lorsqu'une expression se précise.

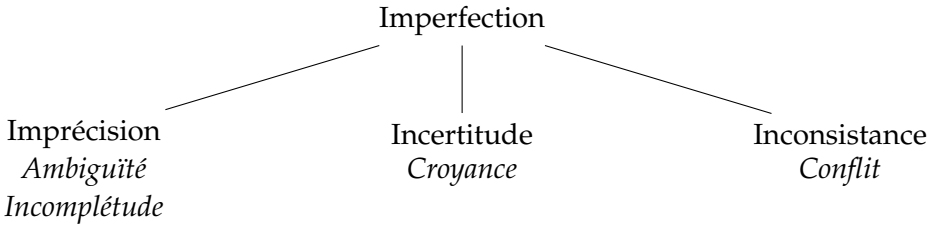
La seconde classification est proposée par SMETS, 1997. Celle-ci est plus polyvalente et facilement transposable au langage naturel. En effet, elle est basée essentiellement sur l'opposition entre l'imprécision et l'incertitude, et considère que ces deux notions

FIGURE 2.4 – Classification proposée par BOUCHON-MEUNIER et NGUYEN, 1996 dans laquelle l'imprécision est un sous-concept de l'incertitude.



sont une spécification d'un concept plus général, l'imperfection de l'information (cf. figure 2.5). Au sein de cette classification, l'imprécision et l'inconsistance sont vues comme des propriétés de l'information elle-même, alors que l'incertitude est considérée comme une propriété de la relation entre l'information et la connaissance d'un agent à propos du monde.

FIGURE 2.5 – Classification proposée par SMETS, 1997 dans laquelle l'imprécision et l'incertitude sont les sous concepts de l'imperfection de l'information.



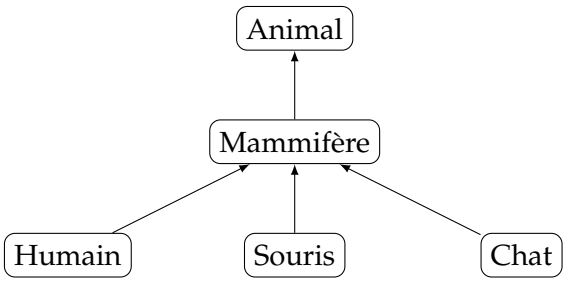
Au travers de cette dernière classification, détaillons plus particulièrement les spécificités de l'imprécision et de l'incertitude. Les sous-sections suivantes sont consacrées à la définition des contours abstraits dans lesquels ces notions évoluent. Nous y abordons également les choix de modélisation réalisés dans cette thèse.

2.2.1 L'imprécision dans le langage naturel

Selon BOUCHON-MEUNIER, 1995 deux raisons peuvent expliquer l'emploi d'informations imprécises. La première est due à notre incapacité à préciser des concepts qui attendent une valeur exacte. Ce niveau peut être subdivisé en deux notions. D'un côté, l'imprécision numérique dans laquelle une description exacte est impossible à produire, *e.g.* "La manifestation à Paris était importante", nombre exact de personnes impossible à déterminer. De l'autre côté, l'imprécision considérée d'un point de vue conceptuel au regard d'une taxonomie de concepts (*T-Box* d'une ontologie) dans laquelle une entité est décrite par un concept plus général (relation d'hyperonymie), *e.g.* "Les chats mangent des souris"/"Les chats mangent des mammifères" (cf. figure 2.6).

La seconde raison est inhérente à la nature même de certains concepts. L'appartenance d'un objet à un concept dont les contours sont mal définis est nécessairement imprécise. Considérons les concepts de chaud et tiède. Existe-t-il un seuil de température pour différencier l'eau chaude de l'eau tiède ? Un bain est dit chaud ou tiède selon les goûts d'une personne, de la saison, de l'âge de la personne, etc. (WEISSENBACHER, 2008). Ainsi, l'imprécision peut être due à un ensemble de propriétés subjectives.

FIGURE 2.6 – Taxonomie de concepts représentée sous la forme d'un graphe. Les flèches représentent une relation de spécification (subsomption ou instanciation).



Dans la modélisation de la chaîne d'extraction de connaissances présentée dans la suite de cette thèse, nous avons fait le choix de considérer l'imprécision au regard d'un ordre partiel sur les syntagmes et les concepts qu'ils représentent. Ce choix découle de la stratégie employée pour obtenir un module permettant d'inférer de la connaissance. Cette stratégie ambitionne de conserver toute la sémantique des phrases et réalise une structuration des syntagmes par le biais de règles linguistiques, éventuellement enrichie par une taxonomie de concepts existants.

2.2.2 Les dimensions de l'incertitude

Cette section tente d'introduire la notion d'incertitude au sens large pouvant être rencontrée dans le traitement automatique des langes et l'extraction d'information. SMETS, 1997 définit l'incertitude comme la distinction entre l'état de connaissance d'un agent (humain ou machine) sur une information et la véritable valeur de vérité de celle-ci. En effet, une information est soit vraie, soit fausse. Cependant, la connaissance de l'agent sur le monde peut ne pas lui permettre de décider de la valeur de cette dernière. Ainsi, la certitude est une pleine connaissance de la valeur de vérité d'une donnée et l'incertitude une connaissance partielle. DRAGOS, 2013 propose une décomposition de l'incertitude dans le cadre des *soft data* [4]. Cette décomposition distingue trois dimensions de l'incertitude :

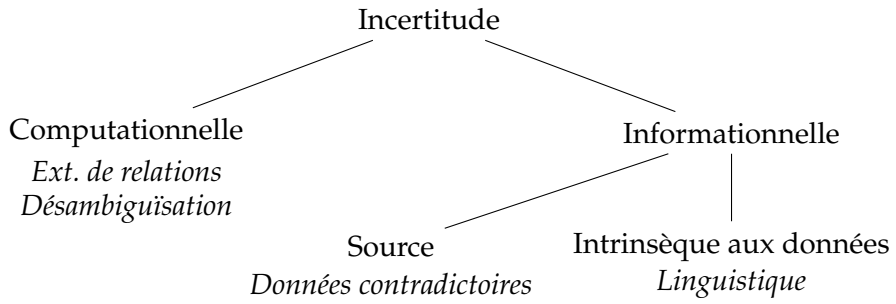
— L'incertitude intrinsèque aux données.
— L'incertitude liée à la source de diffusion.

4. Ces données font référence aux choses difficilement mesurables telles que les opinions ou les ressentiments des gens. Elles s'opposent aux *hard data* provenant par exemple de capteurs.

— L'incertitude relationnelle correspondant à l'inexactitude de l'information au vu de différentes sources (sources conflictuelles).

Ces différentes dimensions génériques s'adaptent à la problématique d'extraction de connaissances. Toutefois, nous pouvons ajouter une autre dimension du fait de la dépendance du domaine à la qualité des résultats des approches exploitées. Cette qualité peut s'apparenter à la gestion des *hard data* impliquant une prise en compte de la certitude des données. Ainsi, nous pouvons concevoir l'incertitude dans le domaine de l'extraction de connaissances comme à la fois liée aux méthodes d'extraction automatisées (inexactitude des extractions) et à la nature de l'information extraite (cf. figure 2.7).

FIGURE 2.7 – Les types d'incertitudes auxquels une approche d'extraction de connaissances est confrontée. L'incertitude computationnelle est due aux potentielles erreurs des modèles d'extraction. Tandis que l'incertitude informationnelle provient des sources des documents et des caractéristiques énonciatives d'une phrase (marqueurs d'incertitude).



Détaillons brièvement les différentes dimensions de l'incertitude présentées en figure 2.7. L'incertitude computationnelle peut être induite à différents niveaux. En effet, une approche d'inférence de connaissances à partir des textes est un module complexe faisant intervenir des techniques en constante évolution telles que l'extraction de relations, la détection de l'incertitude linguistique ou la désambiguïsation des entités. Cette évolution des méthodes entraîne une variabilité dans le temps des résultats et la possibilité de recueillir des résultats incertains. La pertinence des méthodes exploitées est abordée dans DONG et al., 2014. Les auteurs y présentent une approche d'enrichissement des bases de connaissances à partir des textes en tenant compte de l'incertitude engendrée par plusieurs types d'extracteurs linguistiques.

Le second type d'incertitude est lié à l'information elle-même. Il nous ramène à l'exactitude d'une information décrite par un opérateur humain. Dans ce cas, deux critères peuvent être considérés : d'une part la fiabilité de la source, d'autre part l'incertitude qui peut être exprimée dans le texte. Différentes méthodes d'évaluation de la fiabilité des sources ont été proposées dans le domaine de l'analyse de données, et plus particulièrement pour la recherche de vérité (DONG et al., 2015 ; BERETTA et al., 2016). Le principe communément admis est le suivant : une source qui fournit fréquemment des informations pertinentes est qualifiée de fiable. Par extrapolation,

une information soutenue par une source fiable, a plus de chance d'être vraie (LI et al., 2016). Cette estimation peut être réalisée au travers d'une procédure itérative dans laquelle le degré de véracité (pertinence) des informations et le degré de fiabilité d'une source sont calculés alternativement, jusqu'à convergence (DONG, BERTI-EQUILLE et SRIVASTAVA, 2009). L'exactitude d'une information est ainsi estimée par l'intégration et l'analyse de données conflictuelles issues des différentes sources.

Enfin, en ce qui concerne l'incertitude linguistique, elle se traduit dans les textes par un ensemble de marqueurs lexicaux. Ces marqueurs peuvent être de nature très diverse et former dans certains cas des expressions incertaines complexes. Dans la littérature anglo-saxonne, ces marqueurs sont usuellement appelés des *hegdes* [5]. Cette notion provient de la publication de LAKOFF, 1975 dans laquelle il définit les *hedges* comme l'ensemble des mots ayant pour objectif de rendre une proposition plus ou moins floue. Par la suite, cette définition est reprise par HYLAND, 1998 dans laquelle les *hedges* sont définis comme des marqueurs linguistiques permettant d'indiquer à la fois, un manque d'engagement sur la valeur de vérité d'une proposition et un désir de ne pas exprimer un engagement catégorique. Ainsi au travers de ces deux définitions, une décomposition de l'incertitude linguistique portant sur l'objectivité et la subjectivité des faits apportés par un auteur émerge. Par conséquent, ce type d'incertitude est intrinsèque à la phrase et s'attarde sur la sémantique et la valeur de vérité d'une proposition. À notre connaissance, cette forme d'incertitude n'a jamais été traitée au sein d'un module d'extraction de connaissances. Elle se veut au cœur de l'approche présentée au sein de la thèse. Pour capturer ce type d'incertitude, nous proposons un modèle de détection spécifique basé sur la classification de l'incertitude de SZARVAS et al., 2012 (cf. chapitre 3).

Les principales difficultés du langage naturel viennent d'être exposées. Voyons maintenant les moyens à notre disposition pour gérer ces contraintes afin d'extraire de l'information désambiguïsée à partir de textes non structurés. Ces étapes, essentielles au fonctionnement de la chaîne de traitement sont présentées dans la prochaine section.

2.3 Les méthodologies pour l'extraction d'information

Les difficultés pour appréhender le langage naturel abordées dans la section précédente relient de nombreux domaines de recherche plus ou moins récents exploitant les textes à la recherche d'informations pertinentes. De la recherche d'information, dont les premiers modèles remontent aux années 1940 (SHAW, 1948), à l'extraction d'information, popularisée par la conférence MUC en 1992 (SUNDHEIM, 1992) en passant par l'enrichissement des bases de connaissances, chacun se confronte aux

5. La notion de *hedge* peut différer selon la classification de l'incertitude que l'on considère. Selon ces classifications et le contexte, les termes *weasel* ou *peacock* peuvent être employés (VINCZE, 2013).

subtilités du langage naturel. Les approches d'extraction de connaissances nécessitent des modèles appartenant à l'ensemble de ces domaines pour rechercher des documents, extraire de l'information et les relier à une base de connaissances. Cette section considère que la base de documents a été constituée et se concentre plus particulièrement sur l'extraction d'information.

L'extraction d'information est un vaste domaine de recherche. Il est communément divisé en trois principales branches : la reconnaissance d'entités nommées, l'extraction de relations et l'extraction d'événements (JEAN-LOUIS, 2011). Bien que ces domaines soient distincts dans le type d'information recueillie, ils accomplissent leur objectif en employant des méthodes similaires tirées des modèles symboliques et numériques. Les modèles symboliques constituent les premiers travaux dans le domaine de l'extraction d'information. Ils emploient un ensemble de règles définies manuellement par des experts ou une forme d'apprentissage pour extraire des informations (MUSLEA, 1999 ; MINTZ et al., 2009). Dans ce contexte, les règles sont le plus souvent composées de mots et d'autres attributs issus des traitements linguistiques. Ces systèmes sont généralement composés d'un ensemble conséquent de règles, avec un recouvrement possible entre certaines d'entre elles. Par conséquent, un ensemble de contraintes est nécessaire pour leur déclenchement (JEAN-LOUIS, 2011). Concernant les modèles numériques, ils ont pour objectif d'apprendre à associer automatiquement des classes à un ensemble d'éléments (BUNESCU et MOONEY, 2005 ; CHAN et ROTH, 2011). Ces méthodes se distinguent par le degré de supervision qu'elles requièrent selon la disponibilité de données d'entraînement.

Dans le cadre d'un module d'extraction de connaissances à partir des textes, la reconnaissance des entités nommées incluant une désambiguïsation et l'extraction de relations sont les branches indispensables au processus. En effet, elles déterminent les informations (entités, relations) exploitées dans la suite du traitement. C'est pourquoi, les sous-sections suivantes se concentrent sur ces méthodes et présentent les approches expérimentées.

2.3.1 La désambiguïsation des entités d'intérêt

Théorie et méthodologie

La reconnaissance d'entités nommées permet d'identifier et classer des entités d'intérêt à l'intérieur de catégories prédéfinies, *e.g.* dans le domaine du biomédical les entités peuvent être des noms de protéines, de gènes ou bien de maladies (KULICK et al., 2004). Cette tâche s'accompagne généralement d'une normalisation sous forme canonique et non ambiguë des entités, appelée également phase de désambiguïsation (LAFOURCADE et SANDFORD, 1999 ; BAZIZ, AUSSENAC-GILLES et BOUGHANEM, 2003). Un terme est dit ambigu si deux sens ou plus peuvent lui être associés dans des contextes différents (DINH et TAMINE, 2010). Par exemple, le terme "has"

en anglais indique à la fois un verbe et le nom d'une protéine. Ainsi, la désambiguïsation permet d'identifier de manière unique une entité *e.g.* au travers d'une représentation de la connaissance décrivant de manière unique les sens des termes ou concepts qu'elle contient (dictionnaire, thésaurus, ontologie, etc.). Cette sous-section s'emploie principalement à évoquer les problématiques de la reconnaissance des entités nommées au regard de la désambiguïsation des entités.

Les premiers travaux dans ce domaine remontent aux années 80 avec la mesure de LESK, 1986. Cette dernière permet de sélectionner le sens d'un mot au regard du nombre de mots en commun entre sa définition et celle d'un autre mot situé dans un voisinage proche (définition pouvant être récupérée au sein d'un dictionnaire). L'exemple proposé dans cet article repose sur la désambiguïsation de l'expression anglaise *pine cone* pour laquelle deux définitions peuvent être associées à *pine* :

— *King of evergreen tree with needle-shaped leaves [...]*
— *Waste away through sorrow or illness [...]*

ainsi que trois définitions pour *cone* :

— *Solid body which narrows to a point [...]*
— *Something of this shape whether solid or hollow [...]*
— *Fruit of certain evergreen trees [...]*

Par conséquent, *evergreen* et *tree* sont communs à deux définitions, suggérant que si les mots *pine* et *cone* apparaissent ensemble alors le sens le plus probable serait l'arbre et son fruit. Cette mesure de similarité définie, la désambiguïsation d'un texte est réalisée par l'évaluation de toutes les combinaisons possibles de sens en choisissant celle qui maximise la somme des scores des sens choisis. Ce score évalue une combinaison C_i d'un fragment de texte contenant N mots avec $LeskSim(S_u, S_w)$ la similarité de Lesk entre les sens S_u du mot u et les sens S_w du mot w (cf. équation 2.1).

$$score(C_i) = \sum_{u=1}^{N} \sum_{w=u+1}^{N} LeskSim(S_u, S_w) \tag{2.1}$$

La mesure de similarité de Lesk peut être remplacée par des mesures de similarité basées sur des distances taxonomiques (RESNIK, 1999 ; BUDANITSKY et HIRST, 2001 ; TCHECHMEDJIEV, 2012). La problématique liée à cette désambiguïsation est le nombre de combinaisons possibles (GELBUKH, SIDOROV et HAN, 2005). En effet, la complexité de la tâche est exponentielle en N. Pour y répondre, différentes approches d'optimisation ont été proposées, comprenant les algorithmes génétiques (GELBUKH, SIDOROV et HAN, 2003) et les algorithmes de colonies de fourmis (SCHWAB et al., 2012). Outre l'optimisation algorithmique, BANERJEE et PEDERSEN, 2002 ont étendu l'approche initiale de Lesk avec l'incorporation des définitions des concepts possédant une relation taxonomique dans WordNet [6] avec les définitions initiales.

6. http ://wordnetweb.princeton.edu/perl/webwn

Par la suite, les modèles de désambiguïsation en domaine spécialisé ont émergé. Par exemple MetaMap dans le domaine biomédical propose de lever les ambiguïtés des entités en choisissant un concept dans le metathesaurus UMLS[7] ayant le type sémantique le plus probable pour un contexte lexical donné (ARONSON, 2001 ; ARONSON et LANG, 2010). Le type sémantique est une information supplémentaire propre à UMLS. Il permet de regrouper les concepts selon une sémantique précise, *e.g.* CHEM pour *Chemicals & Drugs* ou ANAT pour *Anatomy*. Une approche similaire a été entreprise à partir de ressources médicales françaises (PEREIRA et al., 2008). Par ailleurs, ces méthodes attribuent aux termes désambiguïsés les identifiants d'un concept appartenant à une ontologie. Ce procédé est de plus en plus populaire dans la littérature et s'inscrit activement dans la tâche d'*entity linking* (HAN et ZHAO, 2009 ; DAIBER et al., 2013 ; FERRET et LE BORGNE, 2016). Cette thématique constitue toujours un sujet de recherche actuel. Elle est désormais orientée vers la reconnaissance des entités en domaine ouvert, lequel suggère des problématiques de volume et de diversité des données (SACK et al., 2016).

Expérimentations

Dans le cadre de notre chaîne de traitement plusieurs modèles de désambiguïsation ont été expérimentés. En effet, le domaine sémantique des données textuelles conditionne grandement le modèle utilisé. L'évaluation de l'approche a été réalisée sur deux principaux types de textes : domaine général et domaine spécialisé (ici le bio-médical). Pour ce dernier domaine, MetaMap a été le modèle de prédilection. Ses options ont permis de désambiguïser uniquement les concepts appartenant à un type sémantique particulier et de restreindre la désambiguïsation à la taxonomie du MeSH. Ces options de personnalisation ont permis de cibler certaines relations spécifiques. Toutefois, cette restriction est spécifique à cette validation. En effet, l'approche a pour objectif d'être exécutée en domaine ouvert signifiant une portée de désambiguïsation la plus large possible. Concernant le domaine général, deux modèles ont été employés. Le premier est *DBpedia Spotlight* (DAIBER et al., 2013) permettant de relier les entités à DBPEDIA. Ce dernier désambiguïse les entités candidates en utilisant une mesure de similarité cosinus entre le contexte du mot à désambiguïser et les documents Wikipedia appartenant aux entités candidates. Le second modèle employé se base sur la taxonomie de WordNet et une approche de maximum de similarité entre un mot (ou un ensemble de mots) et une phrase comme décrit précédemment avec la mesure de similarité de Lesk. Il emploie la bibliothèque *pywsd* (TAN, 2014) et la mesure de similarité de WU et PALMER, 1994. Elle permet de réaliser le lien entre un mot et un nœud de la taxonomie de WordNet. Dans notre cas, nous autorisons uniquement une désambiguïsation respectivement pour le sujet et l'objet. Cette désambiguïsation considère la plus longue entité pouvant être désambiguïsée. La figure 2.8 montre les étapes conduisant à sa recherche.

7. https ://www.nlm.nih.gov/research/umls/

FIGURE 2.8 – Désambiguïsation de la plus longue entité issue d'un sujet ou d'un objet. Les mots soulignés sont ceux évalués par la méthode de maximum de similarité. La numérotation représente l'ordre des motifs évalués au sein des phrases et le symbole $\surd$ signifie l'entité ayant été désambiguïsée. Dans cet exemple *AIMP2-DX2* est désambiguïsé dans le MeSH par l'identifiant unique C575685.

0 *AIMP2-DX2 increased expression*
↓
1 *AIMP2-DX2 increased expression*
2 *AIMP2-DX2 increased expression*
3 *AIMP2-DX2 increased expression*
4 *AIMP2-DX2 increased expression*
5 *AIMP2-DX2 increased expression*
6 *AIMP2-DX2 increased expression* $\surd$

En conclusion, les méthodes présentées dans cette sous-section permettent d'appréhender la phase d'*entity linking* essentielle dans l'enrichissement des bases de connaissances et des méthodes destinées à lier les textes avec ces représentations de la connaissance. Cette tâche se compose généralement de trois principales parties :

1. Repérage des formes de surface (mots ou ensemble de mots).
2. Correspondance avec les entités candidates possibles.
3. Désambiguïsation (fonction de pondération tenant compte des entités et de leur contexte).

Dans notre chaîne d'extraction de connaissances, cette phase permet de désambiguïser selon une taxonomie donnée le sujet et l'objet d'une relation extraite à partir des modèles présentés dans la sous-section suivante.

2.3.2 L'extraction de relations

Théorie et méthodologie

L'extraction de relations est définie par CRAVEN et KUMLIEN, 1999 de la façon suivante : « Étant donné un ensemble d'entités d'intérêt, des relations entre ces entités et un corpus de documents à traiter, extraire de ces documents des entités et leurs relations respectives ». L'extraction de relations consiste en la découverte de connections sémantiques entre des entités d'intérêt. Usuellement, les méthodes se basent sur un ensemble de paires d'entités dans un document et essayent de déterminer à partir d'indices (syntaxiques, contextuels ou sémantiques) si une relation existe entre les entités de cette paire.

Les méthodes les plus simples reposent sur la co-occurrence entre deux termes d'intérêt dans une même phrase. Cette méthodologie est utilisée dans ZHOU et al., 2014 afin d'extraire les relations entre des maladies et les symptômes qui leur sont associés. Leur approche réalise une correspondance lexicale entre les labels associés aux concepts du MeSH [8] et les mots-clés fournis dans les articles scientifiques extraits de PubMed [9]. Si de bons résultats ont été observés sur ces relations spécifiques en domaine fortement contraint, un bruit trop important est à déplorer dans le cas d'une application en domaine général et d'un prédicat factuel tel que *bornIn* (SURDEANU et al., 2012), pour lequel les types d'entités d'intérêt occasionnent un trop grand nombre d'étiquettes possibles. Dans la continuité des méthodes non supervisées, WANG et al., 2013 proposent de regrouper des relations sémantiquement équivalentes en considérant un ensemble de types d'entités prédéfinis (par rapport au module de reconnaissance des entités nommées qu'ils exploitent). Pour cela, les auteurs proposent une procédure basée sur un *clustering* multi-niveaux. Le premier niveau permet de regrouper des expressions linguistiques similaires telles que *create the* et *who create*, aboutissant à de nombreux petits *clusters* lexicaux précis. Le second niveau de *clustering*, quant à lui, exploite une similarité sémantique afin de regrouper les premiers *clusters* au sein de *clusters* sémantiques plus larges. Par conséquent, ces derniers retranscrivent des expressions lexicales sémantiquement identiques pour des types d'entités similaires *e.g.* entre deux organisations, l'approche obtient un *cluster* regroupant les expressions suivantes : *purchase, buy, acquire, trade*.

Toutefois, les méthodes présentées ci-dessus sont limitées à un champ lexical ou à des types d'entité précis. En réponse à cette limitation, l'université de Washington avec le projet *KnowItAll* a introduit le concept *Open Extraction Relation* (ETZIONI et al., 2004). Ce paradigme a la spécificité de ne pas être contraint par le type des entités, en plus de se confronter aux problématiques liées au volume des données. L'un des modèles les plus célèbres se nomme REVERB et exploite un ensemble de contraintes syntaxiques et lexicales (ETZIONI et al., 2011). Les contraintes syntaxiques se présentent sous la forme de patrons basés sur les étiquettes morpho-syntaxiques des mots (cf. figure 2.9). Tandis que les contraintes lexicales exploitent des patrons lexicaux basés sur la construction d'un large dictionnaire d'entités afin d'éviter l'extraction de déclarations trop spécifiques. Cette première phase permet de récupérer 85% des relations verbales binaires. Enfin, une fois ces relations extraites, la méthode détermine les limites du sujet et de l'objet des déclarations contenues dans la phrase en utilisant une méthode d'apprentissage spécifiquement entraînée.

REVERB est une méthode robuste, rapide et surpassant les modèles TEXTRUNNER (YATES et al., 2007) et WOE (WU et WELD, 2010) conçus par la même équipe de recherche. Cependant, elle est incapable d'extraire certains types de relations telles que les relations nominales. Par exemple, *"Microsoft co-founder Bill Gates spoke at ..."*

8. Medical Subject Heading – https ://www.ncbi.nlm.nih.gov/mesh
9. https ://www.ncbi.nlm.nih.gov/pubmed

FIGURE 2.9 – Expressions rationnelles basées sur les fonctions grammaticales des relations verbales (FADER, SODERLAND et ETZIONI, 2011). Ces expressions extraient soit un verbe simple (*invented*), soit un verbe suivi par une préposition (*located in*), soit un verbe suivi par un syntagme nominal et terminant par une préposition (*has atomic weight of*).

V | V P | V W * P
W = (*noun* | adj | adv | pron | det)
P = (prep | *particle* | *inf. marker*)

doit entraîner l'extraction de <*Bill Gates, be co-founder of, Microsoft*>. Pour répondre à cette limite, SCHMITZ et al., 2012 proposent la méthode OLLIE. Cette dernière exploite les dépendances syntaxiques des phrases par le biais d'un apprentissage sur un corpus de phrases collecté par des relations *graines* issues de REVERB. Ce choix algorithmique permet une meilleure précision des relations extraites et favorise la conservation des relations nominales.

Expérimentations

Dans le cadre de notre chaîne de traitement, nous exploitons les extractions issues de REVERB. Ce choix est motivé d'une part par sa souplesse et sa rapidité d'exécution [10] et d'autre part par une délimitation plus simple des sujets et des objets, facilitant ainsi leur analyse par rapport à OLLIE (cf. figure 2.10).

FIGURE 2.10 – Exemples d'extraction de relations réalisée par REVERB et OLLIE sur une même phrase. Les prédicats ont été soulignés pour une meilleure visibilité. La phrase provient du jeu de données BioASQ de 2015.

The mineralization of the psammoma bodies is induced principally by the collagen fibers synthesized by the meningocytes and that the form of mineralization is spherical and growth is radial, controlled by the tumoral cells.

↓

Reverb	<*the mineralization of the psammoma, induce by, the collagen fibers*>
Ollie	<*the mineralization of the psammoma bodies, is induced principally by, the collagen fibers synthesized by the meningocytes and that the form of mineralization is spherical and growth is radial controlled by the tumoral cells*>

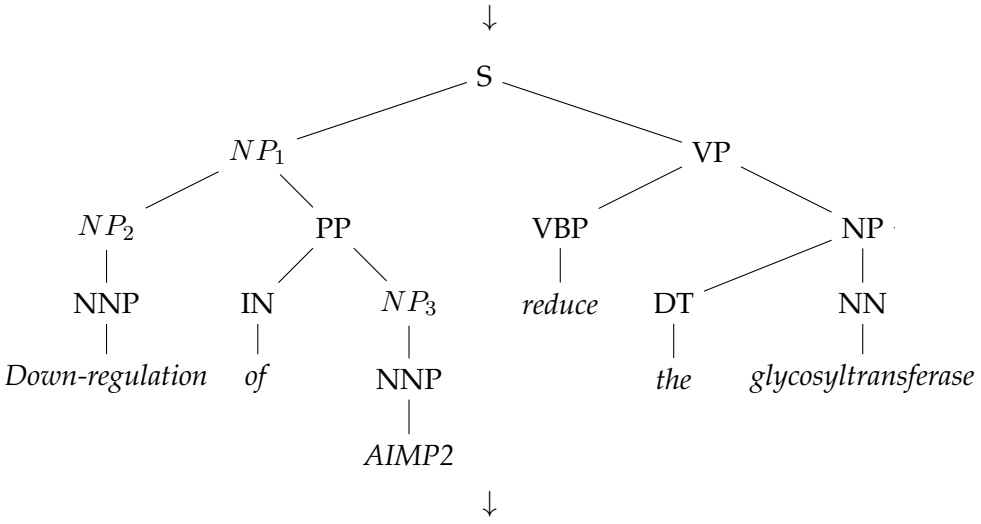
L'identification des limites gauches et droites des sujets et des objets, réalisée par REVERB, exploite de simples heuristiques telles que l'extraction des syntagmes nominaux ou l'identification d'entités provenant de Wikipedia. À cela s'ajoute un module complémentaire nommé ARGLEARNER pour améliorer la précision de certaines extractions. En effet, les heuristiques initialement établies peuvent produire une information erronée sur certaines structures de phrases. Par exemple, *The cost of the*

10. REVERB a été conçu pour être exécuté sur de large corpus de textes : http ://openie.allenai.org/

war against Iraq has risen above 500 billion dollars permet d'extraire la relation <*Iraq, has risen above, 500 billion dollars*>. Ainsi ce module, présent sous la forme d'une option dans REVERB, emploie une méthode de classification spécifiquement entraînée pour une meilleure identification des limites des entités. Par conséquent, les extractions de REVERB facilitent la normalisation des entités tout en conservant d'importantes informations au sein de la phrase. Ces informations sont généralement retranscrites par la conservation des adjectifs et des syntagmes prépositionnels associés au syntagme nominal principal. L'arbre syntaxique en figure 2.11, obtenu avec StanfordNLP [11], illustre l'heuristique de REVERB quant aux délimitations des sujets et objets. Dans le cas présent c'est le syntagme nominal NP_1 qui est conservé.

FIGURE 2.11 – Importance de conserver certaines informations au sein des phrases. L'option sans conservation peut être obtenue en utilisant une méthode de reconnaissance d'entités nommées spécifique au milieu bio-médical. Cette dernière s'intéresse à l'identification des concepts généraux *e.g.* les noms des protéines plutôt que des modificateurs leurs étant appliqués, où en effet, *down-regulation* est une information primordiale au sein de la phrase.

Down-regulation of AIMP2 reduced the glycosyltransferase.



Sans conservation <*AIMP2, reduce, glycosyltransferase*>
Avec conservation <*AIMP2 down-regulation, reduce, glycosyltransferase*>

Toutefois l'inconvénient de cette méthode d'extraction de relations, qui est également sa force, est l'utilisation des motifs syntaxiques pour identifier les relations par rapport à la structure du prédicat. En effet, les prédicats extraits ne sont pas aisés à manipuler selon les structures des phrases. Dans le cas d'une phrase simple, il est facile d'extraire un prédicat unique *e.g. Asbestos causes cancer* permet d'obtenir <*asbestos, cause, cancer*>. Toutefois, dans le cas d'une phrase plus élaborée, il est fréquent d'obtenir des prédicats plus complexes pouvant modifier les idées véhiculées au sein d'une phrase s'ils ne sont pas correctement analysés *e.g. Aspririn fails to cure*

11. http ://nlp.stanford.edu :8080/parser/

acne entraîne l'extraction <*aspririn, fail to cure, acne*>. Dans cet exemple, il est important d'identifier le marqueur de négation du prédicat changeant radicalement la signification de la phrase. Outre l'aspect négatif, il est également possible de rencontrer des relations dont l'identification du prédicat n'a pas été correctement réalisée du fait par exemple à une faute de langage. Ces prédicats peuvent alors contenir un élément modificateur essentiel à la compréhension de la phrase. Par exemple, dans le jeu de données ClueWeb09 [12] on remarque l'extraction du prédicat *cause decrease*. Ce dernier contient un modificateur (*decrease*) devant être séparé du prédicat afin d'éviter d'inférer des informations erronées. Une solution possible est de supprimer ce modificateur du prédicat et de l'ajouter à la fin de l'objet de la relation (cf. figure 2.12).

FIGURE 2.12 – Exemple extrait du jeu de données ClueWeb09. La modification du prédicat est nécessaire pour éviter d'inférer une information erronée autour du prédicat *cause*.

<*bradycardia, cause decrease, blood pressure*>

↓

<*bradycardia, cause, blood pressure decrease*>

Un autre exemple issu de ce jeu de données provient des relations de comparaison entre deux entités faisant intervenir une expression imagée *e.g. Jeff eats like a pig* ne doit pas signifier que Jeff mange du cochon. Enfin, il faut également considérer la voix passive des prédicats *e.g. be VBD by*. Dans ce dernier cas, une inversion du sujet et de l'objet doit être opérée.

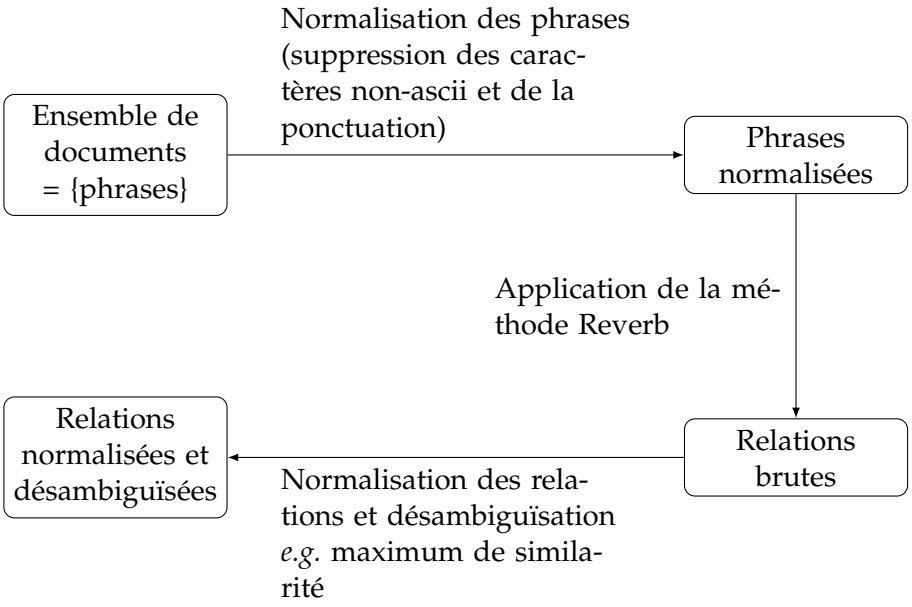
Ainsi, dans le cadre de la construction d'une interface de requêtage sur ces relations, il est important de pré-traiter les données en fonction de leur prédicat. Une fois ces relations extraites, la chaîne de traitement emploie une des approches de désambiguïsation vues précédemment pour tenter de réaliser un lien avec les concepts d'une taxonomie donnée.

2.4 Synthèse

Nous avons mis en évidence au sein de ce chapitre les complexités linguistiques auxquelles une chaîne d'extraction de connaissances doit faire face en domaine ouvert. De l'extraction à la désambiguïsation en passant par les subtilités énonciatives d'un auteur, le langage naturel regorge d'embûches devant être démystifiées afin d'améliorer *l'intelligibilité* du langage par la machine. La figure 2.13 récapitule les différentes tâches accomplies par la chaîne de traitement au regard des informations acquises jusqu'à présent.

12. http ://reverb.cs.washington.edu/

FIGURE 2.13 – Récapitulation de la chaîne de traitement au regard du module d'extraction d'information et de la langue anglaise. La normalisation des relations est abordée au chapitre 4. Lorsqu'il est question des relations désambiguïsées, nous incluons également les relations n'ayant pas eu de correspondance avec un concept de la taxonomie.



Parmi les étapes mentionnées sur cette figure la phase de normalisation des relations sera discutée dans le chapitre 4. Cette phase permet notamment de structurer l'information extraite afin de générer de nouvelles relations et de les évaluer. Cette évaluation est une étape importante au sein de notre chaîne d'extraction de connaissances. Sa particularité est de prendre en compte l'incertitude linguistique qui est un modificateur de la valeur de pertinence d'une information. En effet, les subtilités énonciatives caractéristiques du langage naturel font que l'information extraite peut être accompagnée d'incertitude *e.g. AIMP2-DX2 down-regulation probably reduces glycosyltransferase*. Par conséquent, la chaîne de traitement doit employer des méthodologies d'évaluation de cette information en considération de l'incertitude énoncée. Pour rendre cela possible, elle doit comprendre et repérer les notions liées à l'incertitude linguistique. Le chapitre suivant présente la première contribution de cette thèse, un module de détection de l'incertitude linguistique.

Chapitre 3

Détection de l'incertitude linguistique

Sommaire

Qu'elle soit d'ordre linguistique, numérique ou due à la subjectivité de certains jugements, l'incertitude est omniprésente dans toute situation langagière. En général levée par un *récepteur* humain qui réinterprète la phrase dans un contexte de co-énonciation (FUCHS, 2008), cette incertitude est beaucoup plus difficile à identifier de manière automatique dans des fragments de textes, qui peuvent de surcroît être sortis de leur contexte. Pourtant, ceux-ci peuvent être à la base d'un raisonnement approché ou, de façon plus globale, intégrés dans un processus décisionnel, pour ne citer que quelques-unes des applications possibles du traitement automatique des langues. La détection automatique de l'incertitude dans les textes a suscité un grand nombre de travaux ces dernières années (KERDJOUDJ et CURÉ, 2015) et des événements majeurs comme la *Conference on Natural Language Learning* (CoNLL) en 2010 ont contribué au développement de méthodes dédiées. Leur intégration dans

des applications d'analyse de sentiments, de recherche d'information, de questions-réponses ou encore d'extraction d'information à partir de textes a montré une réelle plus-value. Cependant, les formes diverses d'incertitude détectées ainsi que la forte dépendance de cette détection à la nature des textes analysés laissent largement ouvertes les perspectives de recherche dans ce domaine.

Dans ce chapitre, nous proposons une méthode de détection de l'incertitude linguistique basée sur une analyse statistique de différentes caractéristiques lexicales et syntaxiques. Après une classification des différentes formes d'incertitude et la présentation de notre positionnement, en particulier concernant les modalités d'évaluation, notre modèle est présenté. Celui-ci offre des résultats particulièrement intéressants lorsqu'il est confronté au processus de validation défini dans le cadre de la conférence CoNLL. Cette évaluation nous a permis différentes observations concernant, entre autres, l'influence de la nature des textes analysés et les dimensions de l'incertitude considérées.

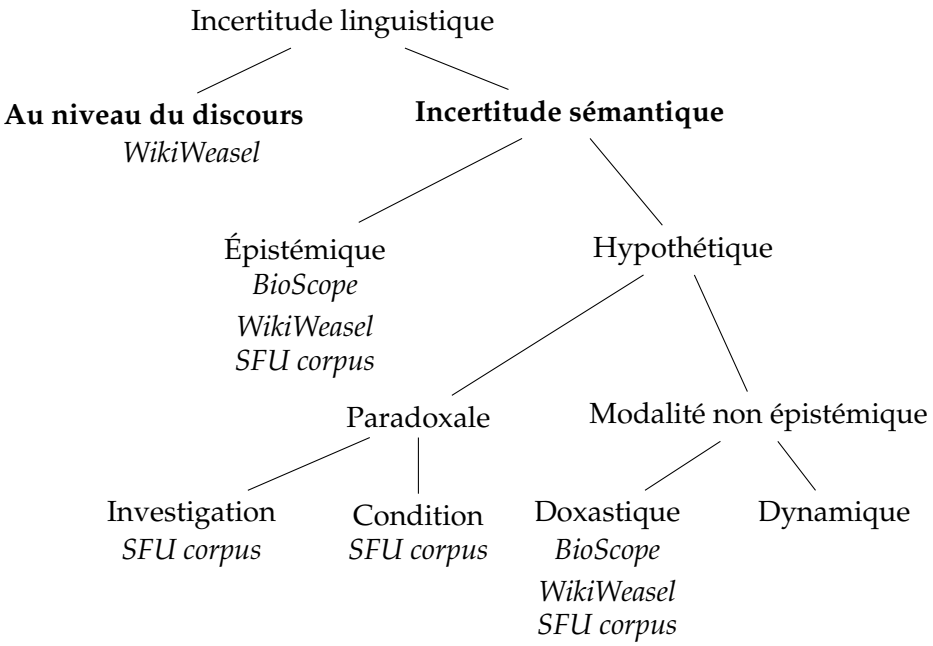
3.1 Classification de l'incertitude linguistique

Les classifications de l'incertitude présentées au chapitre précédent ne sont pas spécifiquement axées sur l'incertitude linguistique et certaines dimensions de cette incertitude n'y sont pas décrites (*e.g.* la *condition*). Pour les compléter, nous introduisons la classification proposée par SZARVAS et al., 2012 dans laquelle les auteurs distinguent deux principales branches de l'incertitude : l'incertitude au niveau du discours et l'incertitude sémantique (cf. figure 3.1). Ainsi, cette classification vient détailler la branche de l'incertitude linguistique présentée dans la figure 2.7. Nous faisons figurer dans la classification ci-dessous les noms des corpus exploités dans la suite de cette étude.

L'**incertitude au niveau du discours** dénote dans la proposition du locuteur un manque d'information, intentionnel ou non. Ainsi la proposition « Certaines personnes ont manifesté [...] » appelle des compléments d'information : quelles personnes, combien étaient-elles ? (FERSON et al., 2015). La subjectivité d'une proposition fait également partie de cette dimension de l'incertitude. Ainsi, l'incertitude au niveau du discours dépend principalement du contexte, du discours et de l'orateur ; en l'absence de connaissance sur ces différentes dimensions, l'incertitude persiste (VINCZE, 2015).

Par ailleurs, on appelle **incertitude sémantique** les propositions dont on ne peut pas déterminer la valeur de vérité étant donné l'état mental actuel du locuteur, soit le degré de confiance qu'il associe à sa proposition. Cette branche de l'incertitude se subdivise en deux catégories, d'une part l'incertitude épistémique et, d'autre part, l'incertitude hypothétique elle-même subdivisée en quatre catégories : dynamique, doxastique, d'investigation et de condition. La principale différence entre ces deux

FIGURE 3.1 – Classification de l'incertitude proposée par SZARVAS et al., 2012. Les trois corpus exploités dans la validation de notre approche sont indiqués dans les différentes formes d'incertitude qu'ils considèrent. BioScope (SZARVAS et al., 2008) est annoté sur l'incertitude sémantique et plus particulièrement sur les formes épistémiques et doxastiques. WikiWeasel (FARKAS et al., 2010) considère, en plus de ces formes, l'incertitude au niveau du discours. Enfin, le corpus SFU considère les mêmes dimensions de l'incertitude que BioScope (KONSTANTINOVA et al., 2012) en y rajoutant les formes paradoxales de l'incertitude sémantique : l'investigation et la condition.



catégories est que les propositions d'incertitude hypothétique peuvent être vraies, fausses ou incertaines *e.g. Il croit que la Terre est plate*. Ce dernier exemple représente une modalité doxastique exprimant les croyances et les hypothèses d'un orateur, lesquelles pouvant être connues comme vraies ou fausses par d'autres dans l'état actuel du monde. Dans ce cas, les connaissances actuelles nous permettent d'infirmer cette proposition. En ce qui concerne les propositions d'incertitude épistémique, elles sont définitivement incertaines *e.g. Il peut pleuvoir*, la factualité de la proposition ne peut être connue. Le tableau 3.1 expose des exemples supplémentaires associés aux différentes catégories de l'incertitude sémantique. Sur ces exemples, on observe que l'incertitude s'applique à la fois sur la proposition principale de la phrase *e.g.* pour l'investigation, l'incertitude porte sur le rôle de NF-kappa B dans l'activation protéique soit la conclusion de l'enquête entreprise, et sur l'incertitude de l'action à venir avec la modalité dynamique et la condition.

La section suivante propose une analyse des différentes méthodes de détection de l'incertitude citées dans la littérature, leurs caractéristiques et leurs performances.

TABLEAU 3.1 – Exemples d'incertitude sémantique. La phrase véhiculant une incertitude épistémique est tirée d'un abstract d'un article scientifique (BioScope) et les autres exemples de la thèse de VINCZE, 2015.

Épistémique		LFA-3 peut jouer un rôle dans la régulation de l'expression du VIH
Hypothétique	Investigation	Nous avons examiné le rôle de NF-kappa B dans l'activation protéique
	Condition	S'il pleut, nous resterons à la maison
	Doxastique	Il croit que la terre est plate
	Dynamique	Je dois y aller

3.2 Les corpus et méthodes pour la détection de l'incertitude

De nombreux travaux, dans différents domaines, ont été consacrés à la détection des différentes formes d'incertitude et à leur prise en compte dans différentes applications de TAL, ce qui a permis d'en améliorer les performances. Par exemple, WU et al., 2011 démontrent que la détection de l'incertitude linguistique permet d'améliorer la précision des informations extraites à partir de rapports radiologiques. Dans le domaine de l'analyse des sentiments, PANG et LEE, 2004 ont montré que la détection de la subjectivité, considérée comme une forme d'incertitude au niveau du discours, aide à améliorer la classification de la polarité des phrases. En ce qui concerne les systèmes questions-réponses, BEN ABACHA, 2012 ; BEN ABACHA et ZWEIGENBAUM, 2015 montre de manière empirique comment la détection de l'incertitude linguistique peut améliorer les performances du système MEANS. L'évaluation de la qualité des données est un domaine de recherche exploitant également l'incertitude dans le langage naturel. Ce domaine tient compte d'un nombre important de caractéristiques afin de qualifier la pertinence des données (MENDES, MÜHLEISEN et BIZER, 2012). Ces caractéristiques considèrent notamment certains aspects dérivés de l'incertitude linguistique tels que la subjectivité.

Cette section se divise en deux sous-sections. La première présente les principaux corpus exploités dans la tâche de détection de l'incertitude et dans la validation de notre approche. Ces corpus ont été conçus spécifiquement pour certaines formes d'incertitude. La seconde partie réalise un tour d'horizon des principales méthodes relatées dans la littérature pour la détection de l'incertitude linguistique.

3.2.1 Description de corpus pour la détection de l'incertitude

L'intérêt pour la détection et la prise en compte de l'incertitude dans les textes est justifié par le fait qu'elle est largement présente dans le langage naturel. Ainsi, LIGHT,

QIU et SRINIVASAN, 2004 estiment que 11 % des phrases dans les résumés des articles de MEDLINE [1] sont incertaines. Ce pourcentage est largement revu à la hausse dans le tableau 3.2, où les pourcentages d'incertitude des phrases issues des trois corpus exploités dans la thèse sont présentés : BioScope, WikiWeasel et SFU corpus.

TABLEAU 3.2 – Statistiques concernant les corpus exploités dans l'article.

	BioScope	WikiWeasel	SFU
Nb de phrases	16874	14726	17263
Nb de phrases incertaines	4218	4262	3912
%Phrases incertaines	25	29	22
Nb moyen des mots/phrase	27,1	29,3	19,0
Nb moyen des marqueurs/phrase incertaine	1,5	3,1	1,8

BioScope (SZARVAS et al., 2008) est un corpus spécifique au domaine biomédical constitué d'articles scientifiques. Il fournit à la fois une annotation des mots clés et une annotation de la portée des marqueurs dans les phrases (pour la tâche de détection de l'incertitude et de négation). BioScope est divisé en deux parties : les résumés des articles bio-médicaux et les articles complets. Les annotations de ces corpus ont été réalisées par deux étudiants en linguistique, supervisés par un linguiste. Les pourcentages d'accords inter-annotateurs sont présentés dans le tableau 3.3.

TABLEAU 3.3 – Pourcentages d'accords inter-annotateurs pour BioScope réalisés par deux étudiants en linguistique et un linguiste. Le premier chiffre correspond au pourcentage d'accord entre les deux étudiants et le deuxième et troisième chiffre au pourcentage d'accord de chaque étudiant avec le chef annotateur.

BioScope	Résumés	Articles complets
Mots clés	79 / 84 / 92	78 / 81 / 91
Portée des marqueurs	77 / 80 / 97	63 / 67 / 90

WikiWeasel (FARKAS et al., 2010) est un corpus généraliste constitué de paragraphes de Wikipedia. Le nom de ce corpus provient de l'appellation anglaise *Weasel word* [2]. Cette expression désigne un mot ou une expression dans Wikipedia, exprimant souvent l'opinion personnelle d'une personne sans aucun retour ni source pour étayer son argumentation. En effet, WikiWeasel est un corpus complexe faisant intervenir des dimensions de l'incertitude difficiles à analyser, notamment au travers de l'incertitude au niveau du discours qui exprime la notion de subjectivité. Celle-ci se reflète dans le pourcentage inter-annotateurs obtenu par deux linguistes qui est seulement de 46% et qui témoigne de cette difficulté y compris pour des opérateurs humains.

1. https ://www.nlm.nih.gov/bsd/pmresources.html
2. *Weasel* désigne une belette en langue anglaise et au sens figuré, une personne sournoise.

Enfin, **SFU** (KONSTANTINOVA et al., 2012) est un corpus généraliste constitué de critiques de divers produits de consommation. La stratégie d'annotation utilisée est la suivante : un premier linguiste a réalisé l'ensemble des annotations et 10% de ces annotations ont été choisies aléatoirement puis soumises à un autre linguiste afin de calculer le pourcentage d'accord inter-annotateurs qui est de 89%.

Des exemples de phrases représentatives tirées de chaque corpus sont présentés dans le tableau 3.4. Ces exemples tiennent compte de la diversité des dimensions de l'incertitude considérées dans les différents corpus.

TABLEAU 3.4 – Exemples de phrases issues des corpus BioScope, WikiWeasel et SFU. Les phrases tirées de BioScope et SFU révèlent deux marqueurs d'incertitude épistémique et celle de WikiWeasel un marqueur d'incertitude épistémique et un autre d'incertitude au niveau du discours (manque d'information au niveau de la source).

BioScope	*We **suggest** that these IL-10 producing effector T cells **may** contribute to clearing malaria infection without-inducing immune-mediated pathology.*
WikiWeasel	*Pedro Chamijo was **probably** born in Spain, but **some sources** say he was born in Quito*[3].
SFU	*It **seems** the back **could** use a little more leg room.*

La principale différence entre ces corpus au regard de la classification de SZARVAS et al., 2012 est le type d'incertitude considéré (cf. figure 3.1). BioScope et SFU prennent en compte uniquement l'incertitude sémantique tandis que WikiWeasel ajoute à celle-ci la prise en compte d'une partie de l'incertitude au niveau du discours, notamment au travers des mots *weasel* qui se rapportent à la notion de source dans le texte : *Qui dit ça ?* et à la part de subjectivité apportée par un contributeur de Wikipedia. L'identification automatique de ces mots *weasel* a été étudiée par GANTER et STRUBE, 2009.

L'analyse détaillée de ces corpus révèle que la nature des textes est généralement caractéristique des différentes dimensions de l'incertitude employées (*cf.* tableau 3.5). Par exemple, l'incertitude doxastique est plus représentée dans les textes issus de Wikipedia par rapport aux articles biomédicaux. Ces dimensions s'expriment sous des formes diverses par l'emploi de verbes spéculatifs (suggérer, présumer), d'adjectifs et adverbes se rapportant naturellement à l'incertitude (probablement, possible), d'auxiliaires modaux permettant d'exprimer une modalité (pouvoir, devoir) ou encore l'emploi de certains temps ou modes de conjugaison (subjonctif, conditionnel).

Toutefois, l'incertitude peut s'exprimer sous la forme d'expressions complexes. Par exemple, la phrase *L'épaississement de la paroi de la vessie soulève la question de la cystite* (FARKAS et al., 2010), évoque un propos incertain alors même qu'aucun mot pris séparément ni le mode de conjugaison n'exprime d'incertitude. Ainsi, la prise en

3. https ://en.wikipedia.org/wiki/Pedro_Bohórquez

TABLEAU 3.5 – Les marqueurs d'incertitude les plus fréquemment utilisés dans BioScope et WikiWeasel pour trois catégories de l'incertitude, sachant que l'incertitude doxastique et conditionnelle sont des sous-catégories de l'incertitude hypothétique. Ces annotations proviennent d'une étude a posteriori de la conférence CoNLL 2010, détaillée dans SZARVAS et al., 2012 ce qui explique la recherche de marqueurs sur des dimensions de l'incertitude non exploitées lors de l'annotation originelle des corpus et l'usage de la terminologie anglaise.

	BioScope abstracts		WikiWeasel	
Epistemic	suggest	616	may	721
	may	516	probable	112
	indicate	301	suggest	108
	appear	143	possible	93
	possible	101	might	78
Doxastic	putative	43	consider	250
	think	43	believe	173
	hypothesis	43	allege	81
	believe	14	think	61
	consider	10	regard	58
Condition	if	14	if	254
	would	6	would	136

compte du contexte des mots représente une difficulté dans la tâche de détection automatique de l'incertitude.

3.2.2 Les travaux reliés à la détection de l'incertitude

La tâche de classification

Il est intéressant d'introduire les méthodes de détection de l'incertitude en présentant la problématique générale de classification. Cette dernière est définie par AGGARWAL et ZHAI, 2012 de la manière suivante.

Classification. Nous avons un ensemble de données d'entraînement (phrases, signaux, etc.), tel que chaque élément est labellisé avec la valeur d'une classe tirée des k différentes valeurs discrètes[4] indexées par $\{1...k\}$. Ces données d'entraînement sont utilisées afin de construire un *modèle de classification* qui fait le lien entre les caractéristiques d'une donnée et un label d'une classe. Une fois le modèle établi, il est exploité dans un processus de prédiction afin d'attribuer un label à des observations inconnues.

4. Nous utilisons de manière équivalente les termes "étiquette" et "label" pour définir les différentes valeurs des classes.

De multiples domaines exploitent les méthodologies de classification tels que la finance, le marketing ou l'IHM [5] (DALHOUMI et al., 2015). Dans notre cas, nous nous intéressons aux problématiques associées à la classification de textes auxquelles la détection de l'incertitude appartient. Au sein de ce cadre applicatif, les méthodes de classification ont été exploitées pour de nombreux usages. En voici quelques exemples :

— Détection de la négation : la négation est présente dans tous les langages humains et elle est utilisée pour renverser la polarité d'une partie de la proposition qui est autrement affirmative par défaut (BLANCO et MOLDOVAN, 2011). Sa détection est généralement abordée de la même manière que la détection de l'incertitude soit au travers d'une classification binaire des phrases (présence ou non d'un marqueur de négation) soit au travers de la détection de la portée des marqueurs négatifs au sein de la phrase (TANUSHI et al., 2013).
— Analyse des sentiments : ce secteur a notamment émergé lors de l'apparition des réseaux sociaux et des sites d'e-commerce. Il permet de déterminer la polarité (positive, négative, neutre) d'un texte donné. Ce type d'approche peut être exploité à des fins marketing *e.g.* pour la recommandation de produits et de publicités (ESULI et SEBASTIANI, 2006 ; PAK et PAROUBEK, 2010 ; PROIOS, EIRINAKI et VARLAMIS, 2015).
— Classification d'e-mails/spam : l'intérêt est de discriminer et filtrer les e-mails indésirables (SAHAMI et al., 1998). Ce type de classification s'appuie avant tout sur une forte précision des approches afin d'éviter la perte d'e-mails importants.
— Classification de textes en domaine spécialisé : une récente compétition présentée sur *Kaggle* [6] propose d'exploiter des comptes-rendus cliniques pour prédire le type de mutation génétique responsable du développement d'une tumeur cancéreuse chez un patient.

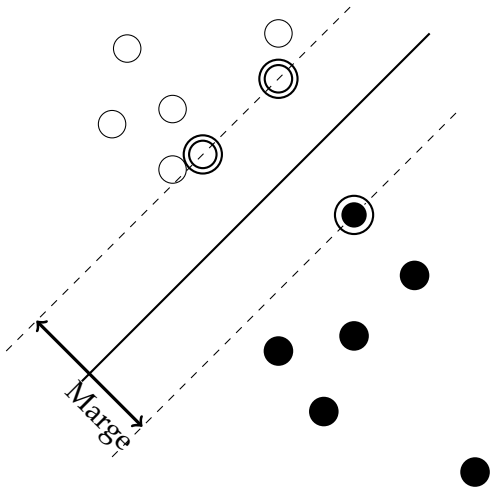
Une grande variété de techniques ont été conçues pour la classification telles que les arbres de décisions, les SVM, les classifieurs bayésiens, les CRF (*Conditional Random Fields*) ou bien les réseaux neuronaux. Toutefois dans le domaine textuel et au-delà d'une méthode de classification en particulier, la sélection des caractéristiques représente l'étape essentielle pour obtenir une phase d'apprentissage effective et plus précise (BANKO et BRILL, 2001 ; FORMAN, 2003). C'est pourquoi, dans notre processus de conception du module de détection de l'incertitude, le choix et la sélection des caractéristiques ont représenté la partie centrale du développement. La méthode de classification en elle-même a été sélectionnée au terme d'une étude empirique (non détaillée dans ce manuscrit) pour laquelle le SVM a obtenu les meilleurs résultats. Cette étude a été complétée d'une comparaison avec un algorithme récent destiné à la classification de textes, FASTTEXT (cf. sous-section 3.4.2). Les deux paragraphes

5. Interface Homme Machine.
6. Une plateforme hébergeant différentes compétitions en informatique (www.kaggle.com).

suivants présentent brièvement la méthode SVM employée dans notre approche présentée dans la section 3.3.

Le SVM (CORTES et VAPNIK, 1995) pour Séparateur à Vaste Marge, ou *Support Vecteur Machine* en anglais, est à la fois utilisé pour des problématiques de classification et de régression (valeurs continues à prédire). Cette méthode repose sur deux notions fondamentales. La première fait référence au concept de *marge maximale* qui représente la distance maximale entre la frontière de séparation, appelée hyperplan optimal, et les données les plus proches appartenant aux deux classes, appelées vecteurs supports (cf. figure 3.2).

FIGURE 3.2 – Visualisation d'un exemple de classification linéaire à deux classes avec un SVM. La droite en gras représente l'hyperplan optimal qui maximise la marge (l'espace) entre les données des deux classes. Les données entourées correspondent aux vecteurs supports.

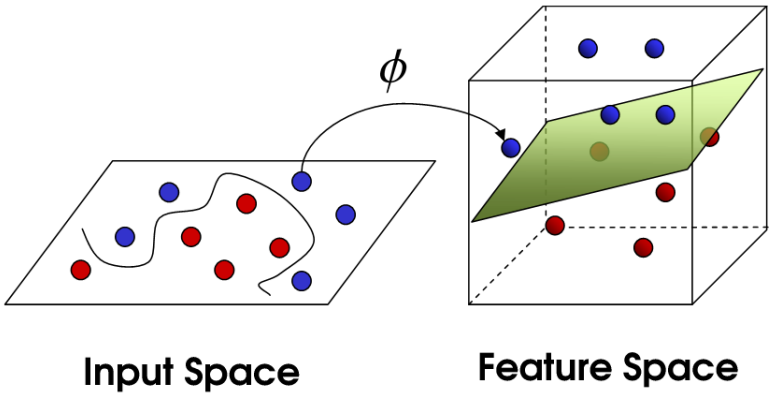


La seconde notion s'applique, quant à elle, dans le cadre où les données ne sont pas linéairement séparables. Dans cette éventualité, le SVM réalise une classification non-linéaire en utilisant ce que l'on appelle un *kernel trick*. Ce dernier permet de reconsidérer la classification dans un espace de dimension supérieure dans lequel il est possible de trouver un hyperplan optimal (cf. figure 3.3).

Les méthodes présentées à CoNLL

Différentes approches ont été suggérées dans le domaine de la détection automatique de l'incertitude. Ces approches se focalisent soit sur une détection binaire de la certitude d'une phrase (cette phrase peut-elle être qualifiée d'incertaine), soit sur la détection de la portée des marqueurs d'incertitude au sein de la phrase. Un défi proposé lors de la conférence CoNLL 2010 (FARKAS et al., 2010) a notamment permis de confronter différentes approches pour ces deux tâches respectives. L'évaluation des méthodes était réalisée au travers des deux corpus : BioScope et WikiWeasel. Dans la

FIGURE 3.3 – Transformation de l'espace de représentation des données d'entrée en un espace de plus grande dimension. Dans cet exemple, l'espace de représentation passe de deux à trois dimensions (ZARARSIZ, ELMALI et OZTURK, 2012).



suite, nous nous focaliserons sur les objectifs fixés par la première tâche de CoNLL (détection binaire).

L'approche ayant obtenu les meilleurs résultats pour la tâche de détection binaire sur le corpus BioScope (F-mesure de 86,4) a été proposée par TANG et al., 2010. Leur méthode se base sur trois classifieurs disposés en deux couches. La première couche comprend un CRF et un SVM se basant tous les deux sur un même ensemble de caractéristiques (mot, lemme, préfixe, suffixe, morphosyntaxe, syntagme) et un système d'étiquettes identiques (BIO – *Begin, In, Other* –). La seconde couche, quant à elle, est constituée d'un autre CRF et utilise des caractéristiques provenant des résultats de la première couche. Cette dernière couche réalise la détection finale des marqueurs et chaque phrase contenant un marqueur est ensuite annotée comme incertaine.

Pour le corpus WikiWeasel, GEORGESCUL, 2010 a proposé la meilleure approche avec une F-mesure de 60,2 en utilisant une classification par SVM basée sur une fonction *kernel RBF* (*Radial Basis Function*) et exploitant des caractéristiques tirées des marqueurs d'incertitude. Une méthode similaire a été mise en place dans CRUZ, TABOADA et MITKOV, 2015 et a obtenu une F-mesure de 92,3 sur le corpus SFU dont les annotations suivent celles proposées dans BioScope. On peut d'ores et déjà remarquer la différence de performance entre les meilleures méthodes en fonction de la nature du corpus (très spécialisé ou généraliste) et des dimensions de l'incertitude considérées. D'autres méthodes intéressantes de détection binaire ont été proposées et appliquées sur WikiWeasel. Par exemple, CHEN et DI EUGENIO, 2010 ont présenté une méthode hybride en deux phases. La première réalise une recherche par motif de mots consécutifs, dont certains sont généralisés par leur *morphosyntaxe* à l'aide de Lucene[7], pour récupérer des phrases candidates (potentiellement incertaines). La

7. https ://lucene.apache.org

deuxième phase utilise ces phrases candidates comme entrées pour une classification par maximum d'entropie. Cette méthode a obtenu le troisième meilleur résultat en 2010 sur 17 participants avec une F-mesure de 57,4.

Les résultats obtenus à CoNLL entre les différents corpus nous dévoilent les limites des méthodes à être performantes sur toutes les facettes de l'incertitude *e.g.* TANG et al., 2010 obtiennent 86,4 en F-mesure sur BioScope et 55 sur WikiWeasel correspondant à la meilleure moyenne (70,7) de la conférence. Dans le cadre de cette thèse, nous nous sommes intéressés à la conception d'une méthode générique de détection de l'incertitude.

Dans la section suivante, nous présentons cette méthode basée sur une représentation vectorielle concise de la phrase – ce choix de représentation a été adopté afin d'éviter les biais de sur-apprentissage identifiés dans l'utilisation de représentations vectorielles de grandes tailles (JOACHIMS, 2002a) ; la taille réduite des vecteurs assure aussi une faible complexité de la tâche de classification, caractéristique souhaitée pour le traitement de gros volumes de données. La représentation vectorielle d'une phrase synthétise différentes statistiques propres à chaque caractéristique étudiée pour la détection d'incertitude (*e.g.* unigramme, bigramme). Plusieurs mesures fréquentistes sont proposées et évaluées pour le calcul de ces caractéristiques. Elles sont par la suite comparées aux mesures classiquement retrouvées dans la littérature associée à la classification de textes. Nous déléguons ensuite la tâche de classification basée sur l'analyse des représentations vectorielles à un SVM (SEBASTIANI, 2002).

3.3 Un nouveau modèle probabiliste pour la détection de l'incertitude

3.3.1 Vue d'ensemble du modèle

L'objectif est de distinguer si une phrase exprime de l'incertitude ou non (tâche 1 de CoNLL 2010). Pour cela, nous disposons d'un ensemble de phrases annotées S provenant des corpus BioScope, WikiWeasel ou SFU. De cet ensemble, il est possible d'extraire des informations sur les particularités lexicales et syntaxiques des phrases certaines et incertaines (*e.g.* la présence de marqueurs d'incertitude, les motifs morphosyntaxiques récurrents) en tenant compte d'une liste restreinte de *stop words*. Notre méthode définit une représentation vectorielle sur un ensemble de caractéristiques d'une phrase. Chaque composante du vecteur correspond à une agrégation des poids affectés à une caractéristique locale dans la phrase (*e.g.* l'ensemble des unigrammes), en fonction d'une classe c à analyser (*e.g. est marqueur d'incertitude*). Formellement, nous évaluons un ensemble de fonctions caractéristiques $\mathcal{F} = \{f_1, f_2, ..., f_n\}$ définissant une projection $\rho_F : S \rightarrow \mathbb{R}^{|\mathcal{F}|}$ pour représenter une phrase

dans un espace vectoriel réduit. Cette représentation découle des méthodes de classification binaire de textes. En effet, dans les méthodes d'apprentissage automatique, une des principales problématiques de la classification de textes est la manière de représenter un document. Celui-ci est généralement matérialisé comme un vecteur de poids associés à ses différentes caractéristiques pouvant être les mots d'un vocabulaire dans les approches les plus simples (SEBASTIANI, 2002). Ces poids ont pour objectif de sélectionner les caractéristiques les plus pertinentes d'une classe c afin de réduire l'espace des dimensions associé à l'ensemble des caractéristiques d'un corpus (YANG et PEDERSEN, 1997).

La sous-section suivante détaille les différentes caractéristiques utilisées par notre méthode, et leur utilisation dans un modèle d'apprentissage supervisé. Les modalités d'évaluation et les résultats obtenus seront discutés dans la section 3.4.

3.3.2 Définition des caractéristiques locales et globales

Les fonctions caractéristiques sélectionnées et étudiées, bien que générales et se voulant indépendantes d'un domaine particulier, traduisent l'intuition que certains marqueurs lexicaux et morphosyntaxiques semblent importants pour la classification d'une phrase comme étant certaine ou incertaine. Deux niveaux de granularité ont été considérés pour la définition de ces fonctions.

Le premier niveau s'intéresse aux spécificités globales d'une phrase traduisant potentiellement une expression d'incertitude. Dans notre modèle, nous considérons uniquement la taille d'une phrase car nous supposons que la longueur est un indice discriminant.

Le second niveau porte sur les motifs *n-grammes* qui composent une phrase. Nous entendons par motifs *n-grammes* les séquences de n éléments de même nature, par exemple, la forme lemmatisée des mots ou leur catégorie morphosyntaxique (*Part of Speech – PoS*). A chaque *n-gramme* est associé un poids (détaillé par la suite) exprimant le fait qu'il puisse traduire une expression d'incertitude. La composante de la projection de la phrase selon la caractéristique analysée tiendra compte de l'agrégation des poids associés aux *n-grammes* qui la composent. Chacune des caractéristiques est ainsi définie par un quadruplet (type, taille, contexte, agrégation) précisant le type de *n-gramme* analysé (lemme ou motif morphosyntaxique), la taille des *n-grammes* (n), le contexte, *i.e.* si le score des *n-grammes* se base sur la fréquence d'observations des *n-grammes* dans les phrases étiquetées incertaines ou comme marqueur explicite d'incertitude, et l'agrégation utilisée pour résumer les scores des différents *n-grammes* de la phrase pour cette caractéristique. À noter que les marqueurs d'incertitude sont annotés dans le jeu d'entraînement et que les corpus sont lemmatisés en utilisant CoreNLP de l'université de Stanford [8].

8. http ://stanfordnlp.github.io/CoreNLP/

TABLEAU 3.6 – Description des caractéristiques locales utilisées. Le mode d'agrégation permet d'obtenir la valeur finale de la caractéristique donnée.

	Type	Taille	Contexte	agrégation
F_1	Lemme	1	Marqueur d'incertitude	somme
F_2	Lemme	2	Marqueur d'incertitude	somme
F_3	Lemme	1	$\in$ à une phrase incertaine	somme
F_4	*PoS*	5	$\in$ à une phrase incertaine	somme
F_6	Lemme	1	$\in$ à une phrase incertaine	max

Le tableau 3.6 résume les différentes caractéristiques basées sur l'analyse de *n-grammes*. Une phrase est donc caractérisée par un vecteur dans $\mathbb{R}^6$ avant l'étape de classification – les cinq caractéristiques présentées dans le tableau 3.6 auxquelles on rajoute la taille de la phrase (F_5).

La sous-section suivante présente les différentes mesures étudiées afin de calculer le score (poids) associé à chaque motif de *n-gramme*. Par commodité, nous illustrons désormais nos propos en considérant la caractéristique F_1 représentant les observations en tant que marqueurs d'incertitude pour un lemme. Par conséquent, les exemples ne vont pas tenir compte du type *PoS*, du type lemme avec une taille supérieure à 1 ni du contexte *appartient à une phrase incertaine* (cf. tableau 3.6). Nous partons également du postulat que la présence d'un marqueur d'incertitude traduit une phrase incertaine.

3.3.3 Définition d'une mesure probabiliste

Les données d'entraînement définissent un ensemble de phrases incertaines $S_u \subset S$ avec S l'ensemble des phrases (cf. tableau 3.7). Une phrase s_i appartenant à S est définie par une séquence de termes w (cf. formule 3.1). Ces données nous permettent d'obtenir pour chaque w son nombre d'occurrences dans le corpus, noté $\#_S(w)$, son nombre d'occurrences dans les phrases incertaines $\#_{S_u}(w)$ avec $\#_{S_u}(w) \leq \#_S(w)$, ainsi que son nombre d'occurrences en tant que marqueur d'incertitude $\#_{I_{S_u}}(w)$, avec I_{S_u} l'ensemble des marqueurs d'incertitude du corpus et $\#_{I_{S_u}}(w) \leq \#_{S_u}(w)$.

$$s_i = w_1 w_2 w_3 ... w_n \tag{3.1}$$

Connaissant le lemme w, nous pouvons définir la probabilité conditionnelle qu'il soit marqueur d'incertitude *i.e.* qu'il appartienne à la classe c (cf. équation 3.2).

$$p_I(c|w) = \#_{I_{S_u}}(w)/\#_S(w) \tag{3.2}$$

TABLEAU 3.7 – Notations utilisées dans l'étude. Nous considérons w comme un lemme.

S	L'ensemble des phrases du corpus
S_u	L'ensemble des phrases incertaines du corpus
s_i	La i^{eme} phrase du corpus
W	Tous les mots du corpus
I_{S_u}	Tous les marqueurs d'incertitude du corpus
$\#_S(w)$	Nombre d'occurrences de w dans le corpus
$\#_{S_u}(w)$	Nombre d'occurrences de w dans les phrases incertaines
$\#_{I_{S_u}}(w)$	Nombre d'occurrences de w comme marqueur d'incertitude
$p_I(c\|w)$	La probabilité que w marque une incertitude
$p_{S_u}(c\|w)$	La probabilité que w soit dans une phrase incertaine

La signification de la classe c et la définition de cette probabilité dépendent du contexte de la caractéristique considérée (marqueur d'incertitude ou *appartient à une phrase incertaine* – cf. tableau 3.6).

Cependant, l'analyse de cette probabilité dans le but de distinguer les marqueurs d'incertitude n'est pas suffisante. Du fait de la taille limitée des corpus d'entraînement, il est en effet fréquent d'obtenir des probabilités très élevées pour certains termes, malgré leur présence limitée dans le corpus d'entraînement. Prenons le cas extrême d'un lemme w qui n'apparaît qu'une seule fois dans le corpus et ce, de façon fortuite, dans un contexte incertain, sa probabilité d'appartenir à une phrase incertaine serait alors : $p_{S_u}(c|w) = 1$.

Afin de pallier cette limite, nous définissons un score de *confiance* associé à cette probabilité qui évalue la pertinence de considérer le motif analysé (ici le lemme w) comme marqueur d'incertitude.

Dans la modélisation de ce score de confiance, nous cherchons à considérer à la fois le nombre d'occurrences $\#_S(w)$ et la probabilité $p(c)$ qu'un lemme, observé dans l'ensemble des mots du corpus W et tiré aléatoirement, soit marqueur d'incertitude (*cf.* équation 3.3). Le raisonnement est que :

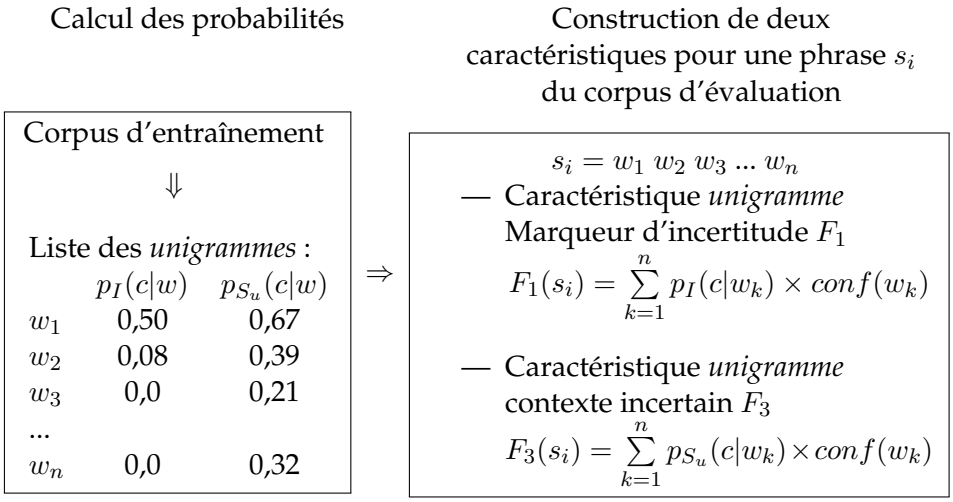
1. Nous voulons que la confiance soit fonction du nombre d'observations car nous considérons que plus le nombre d'observations est grand, plus les statistiques sont fiables.
2. Nous voulons prendre en compte $p(c)$ puisque nous considérons que plus les observations positives sont rares, moins nous avons besoin d'observations positives pour évaluer la fiabilité d'un mot. Par exemple, si nous avons une probabilité $p_I(c|w) = 0,2$ qu'un mot appartienne à une classe c spécifique (*e.g* est marqueur d'incertitude) qui est basée sur 10 observations alors la sémantique ne sera pas la même si nous considérons $p(c) = 0,5$ ou $p(c) = 0,02$.

Par conséquent, si un lemme obtient une forte probabilité d'être marqueur d'incertitude, la confiance dans ce score sera d'autant plus élevée que ce lemme est représentatif du corpus et que la probabilité $p(c)$ est faible. Enfin, dans le cas où un lemme appartenant au corpus d'évaluation n'a jamais été rencontré dans le corpus d'entraînement, nous utilisons par défaut $p(c)$ comme poids.

$$p(c) = \frac{\sum_{w \in W} \#_{I_{S_u}}(w)}{\sum_{w \in W} \#_S(w)} \tag{3.3}$$

La figure 3.4 détaille le calcul de deux caractéristiques F_1 et F_3. Elles considèrent pour des *unigrammes* de lemmes des contextes différents à savoir respectivement : est marqueur d'incertitude et appartient à une phrase incertaine.

FIGURE 3.4 – Calcul de deux caractéristiques d'une phrase s_i appartenant au corpus d'évaluation. La première phase (à gauche) permet de calculer les probabilités sur le corpus d'entraînement avec $p_I(c|w)$ la probabilité conditionnelle qu'un lemme w soit marqueur d'incertitude et $p_{S_u}(c|w)$ la probabilité conditionnelle qu'un lemme w soit présent dans une phrase incertaine. Ainsi, la signification de la classe c dépend de la probabilité conditionnelle considérée. La deuxième phase correspond à la construction des deux caractéristiques F_1 et F_3 définies dans le tableau 3.6.



Pour la modélisation du score de confiance, nous avons étudié deux mesures, possédant une sémantique propre, utilisant comme paramètre $\#_S(w)$ et $p(c)$ ainsi qu'une mesure témoin utilisant uniquement $\#_S(w)$.

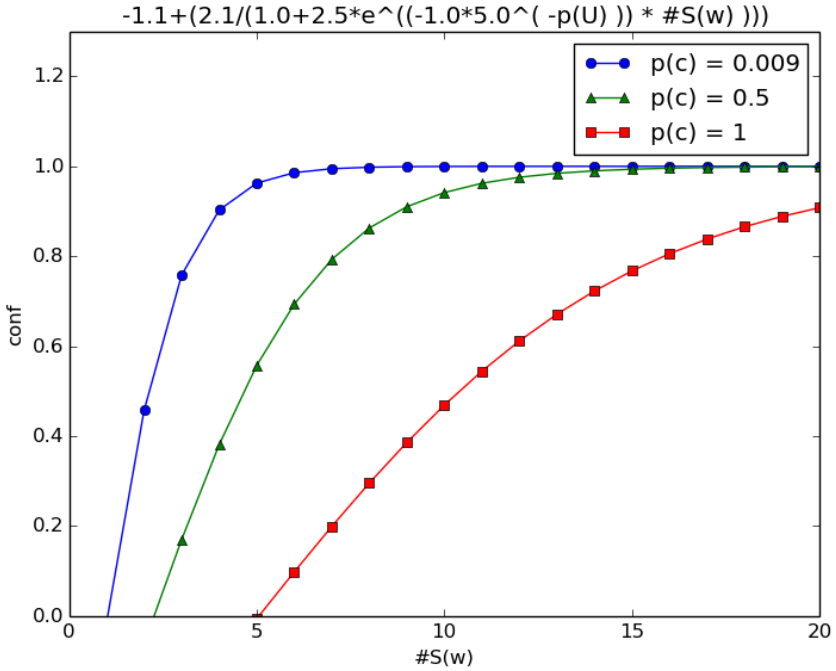
Le premier score de confiance étudié repose sur une loi de distribution binomiale cumulée utilisant : $p(c)$, la probabilité de tirer un marqueur d'incertitude dans W, le nombre d'occurrences $\#_{I_{S_u}}(w)$ du mot w observé en tant que marqueur d'incertitude et le nombre d'occurrences $\#_S(w)$ du mot w dans le corpus complet. Cette loi est définie par la probabilité de fonction de masse suivante, avec $n = \#_S(w)$, $k = \#_{I_{S_u}}(w)$ et $p = p(c)$:

$$p_b(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tag{3.4}$$

Ainsi, la confiance associée à la probabilité conditionnelle $p_I(c|w)$ est fonction de la probabilité d'effectuer un nombre d'observations identifiées comme marqueurs d'incertitude supérieur ou égal à $\#_{I_{S_u}}(w)$ (loi cumulative) en effectuant $\#_S(w)$ tirages aléatoires. Par conséquent, plus la valeur associée à la loi binomiale cumulative est élevée et moins $p_I(c|w)$ traduit une incertitude. Le score de confiance est alors modélisé par $1 - p_b(X \geq k)$. A noter que pour une valeur de $\#_S(w)$ élevée, nous approximons la loi binomiale par une loi normale selon le théorème de Moivre-Laplace (SMITH, 2012).

La seconde modélisation de la confiance que nous avons étudiée suppose intuitivement que plus la probabilité $p(c)$ est grande, plus le nombre d'occurrences d'un motif doit être conséquent pour associer un score de confiance élevé à la probabilité qu'il soit marqueur d'incertitude. Cette représentation de la confiance peut être modélisée par une fonction sigmoïde de $\#_S(w)$ dont le paramètre $p(c)$ caractérise la courbure. Plus $p(c)$ est grand et plus la pente de la courbe est lissée. (cf. figure 3.5).

FIGURE 3.5 – Modélisation de la confiance en fonction du paramètre $p(c)$. Dans le cas de la courbe bleue avec les ronds $p(c) = 0,009$, ce qui correspond à la probabilité pour un mot d'être marqueur d'incertitude dans le corpus BioScope. La courbe verte avec les triangles a un $p(c) = 0,5$ et la courbe rouge avec les carrés un $p(c) = 1$



Finalement, la dernière modélisation de la confiance, servant de mesure témoin, tient uniquement compte de $\#_S(w)$. Cette mesure nous permet d'observer l'impact de la

probabilité $p(c)$ dans les modélisations précédentes. Elle signifie que plus un lemme est fréquent, plus la confiance qui lui sera accordée sera importante (cf. équation 3.5).

$$conf(w) = 1 - \frac{1}{\#_S(w)} \tag{3.5}$$

La fonction F_1 utilisée pour calculer une des dimensions de la représentation vectorielle d'une phrase s_i, qui caractérise les unigrammes marqueurs d'incertitude est fonction de : la probabilité que l'unigramme traduise une forme d'incertitude, modulée par $conf(w)$ la confiance associée à la probabilité de l'unigramme w (cf. équation 3.6). Cette formulation est généralisable à l'ensemble des caractéristiques locales.

$$F_1(s_i) = \sum_{k=1}^{n} p_I(c|w_k) \times conf(w_k) \tag{3.6}$$

3.3.4 Sélection automatique des caractéristiques optimales

Les vecteurs représentant les phrases sont par la suite utilisés comme entrée d'un modèle d'apprentissage automatique SVM [9]. Les natures très différentes des corpus BioScope, WikiWeasel et SFU font qu'ils n'ont pas le même ensemble de caractéristiques optimales dans le SVM (cf. tableau 3.8). Une stratégie de sélection automatique des caractéristiques optimales à partir d'un corpus d'entraînement a été appliquée en suivant les travaux de CHEN et LIN, 2006. Ces travaux mettent en avant l'utilisation d'une *forêt aléatoire* (*random forest*). Ainsi, nous avons mis en place une procédure de sélection récursive où les caractéristiques en dessous d'un certain pourcentage d'importance (ici 8 %) sont supprimées jusqu'à obtenir les caractéristiques les plus pertinentes. Cette procédure est réalisée par l'intermédiaire de la bibliothèque Python *scikit-learn* [10].

Le tableau 3.8 présente uniquement les caractéristiques utilisées dans au moins un corpus. Cependant, d'autres caractéristiques ont été expérimentées. Par exemple, les trigrammes lexicaux ou les motifs *PoS* avec une taille inférieure à cinq. Ainsi, cette sélection montre que la détection de l'incertitude est plus sensible aux n-grammes courts et lexicaux. Ces résultats soutiennent les conclusions des travaux de ØVRELID, VELLDAL et OEPEN, 2010. Cependant, ils suggèrent que les caractéristiques syntaxiques ne sont pas nécessaires dans la tâche de détection de l'incertitude. Néanmoins, notre sélection démontre que pour WikiWeasel, la caractéristique des motifs *PoS* contenus dans les phrases incertaines est discriminante. Enfin, ces n-grammes

9. Ce modèle se base sur une fonction *kernel* RBF (GASPAR, CARBONELL et OLIVEIRA, 2012) et est optimisé au niveau des paramètres C et γ selon l'étude réalisée par GEORGESCUL, 2010
10. http ://scikit-learn.org/

sont principalement basés sur les marqueurs d'incertitude soit une fenêtre particulière de la phrase. Ce fait justifie la pertinence de la caractéristique F_6 qui considère le poids maximum des unigrammes marqueurs d'incertitude dans la phrase.

TABLEAU 3.8 – Les caractéristiques optimales pour BioScope, WikiWeasel et SFU obtenues par l'application d'une forêt aléatoire. La caractéristique des *trigrammes* n'est jamais pertinente quelle que soit la combinaison des caractéristiques utilisées.

Caractéristiques	BioScope	WikiWeasel	SFU
F_1 - Unigrammes, marqueurs d'incertitude	x	x	x
F_2 - Bigrammes, marqueurs d'incertitude		x	
F_3 - Unigrammes, dans les phrases incertaines	x	x	x
F_4 - Motifs *PoS* taille 5, dans les phrases incertaines		x	
F_5 - $\|s\|$ la taille de la phrase s		x	
F_6 - max(Unigrammes marqueurs d'incertitude)	x	x	x

3.4 Résultats et discussion

Cette section présente les résultats obtenus en utilisant la probabilité conditionnelle $p_I(c|w)$ couplée avec les différentes définitions de la confiance présentées dans la section précédente : la loi binomiale cumulée (cf. équation 3.4), la fonction sigmoïde (cf. figure 3.5) et la confiance témoin (cf. équation 3.5). Les résultats sont ensuite comparés en modifiant la probabilité $p_I(c|w)$ par des mesures éprouvées en théorie de l'information.

3.4.1 Résultats de l'approche probabiliste

Les métriques d'évaluation exploitées sont les mêmes que celles qui ont été utilisées lors de la conférence CoNLL. Elles correspondent aux mesures classiques de précision, rappel et F-mesure (cf. tableau 3.9). Dans l'expérience, un vrai positif (resp. un vrai-négatif) est défini lorsque l'approche réussit à bien authentifier une phrase comme incertaine (resp. certaine). Tandis qu'un faux-positif (resp. un faux-négatif) est défini lorsque l'approche authentifie une phrase comme incertaine (resp. certaine) alors qu'elle est certaine (resp. incertaine).

Le tableau 3.10 précise les résultats obtenus en utilisant la probabilité d'être marqueur d'incertitude couplée aux différents scores de confiance définis précédemment. Chaque entrée donne la précision, le rappel et la F-mesure obtenus pour chaque expérience.

TABLEAU 3.9 – Métriques d'évaluation utilisées durant CoNLL 2010. TP dénote les vrais positifs, FP les faux positifs et FN les faux négatifs.

Precision	$\frac{TP}{TP+FP}$
Rappel	$\frac{TP}{TP+FN}$
F-mesure	$\frac{2.Precision.Rappel}{Precision + Rappel}$

TABLEAU 3.10 – Résultats de la méthode en utilisant différentes confiances associées à la probabilité $p_I(c|w)$ sur les corpus BioScope, WikiWeasel et SFU. Les caractéristiques utilisées entre les différentes confiances sont fixées en fonction de la phase de sélection des caractéristiques optimales pour le jeu de données. Les résultats présentent la précision / le rappel / la F-mesure.

	BioScope	WikiWeasel	SFU corpus
Loi binomiale	77,9 / 82,9 / **80,3**	66,7 / 25 / 36,3	87,8 / 95,8 / 91,6
Sigmoïde	75,8 / 82,1 / 78,8	73,8 / 43,6 / 54,8	88,2 / 96,6 / **92,2**
$1 - 1/\#_S(w)$	75,8 / 81,6 / 78,6	64,9 / 48,8 / **55,7**	88,2 / 96,4 / 92,1
Sans confiance	76,3 / 81,2 / 78,7	72,1 / 11,8 / 20,2	88,1 / 95,9 / 91,9

Ces résultats amènent plusieurs remarques. Dans un premier temps, on s'aperçoit que les résultats pour la loi binomiale sur WikiWeasel sont moins bons qu'avec les autres scores de confiance. Ces résultats s'expliquent par la faible valeur de la probabilité de succès $p(c)$ (*e.g* 0,03 pour WikiWeasel) qui a pour effet d'augmenter le score de confiance de la loi binomiale. Dans un second temps, on remarque que les différents scores de confiance n'impactent pas ou peu les résultats sur les jeux de données BioScope et SFU, considérant tous les deux uniquement l'incertitude sémantique. Cependant, la prise en compte de la confiance améliore les scores sur WikiWeasel. Une analyse fine de ces corpus nous permet d'observer que la disparité des marqueurs d'incertitude pour WikiWeasel est bien plus grande que pour BioScope, elle-même plus grande que celle de SFU. Cette forte disparité, représentée par le nombre, la nature et la distribution des marqueurs d'incertitude à l'échelle du corpus ajoute un bruit important dans ces données. Pour diminuer ce bruit, nous avons appliqué un filtre de pré-classification sur le nombre d'occurrences $\#_{S_u}(w)$ des motifs *n-grammes* dans les phrases incertaines (cf. équation 3.7).

$$\#_{S_u}(w) = \begin{cases} \#_{S_u}(w) & \text{si } \#_{I_{S_u}}(w) \geq 1 \\ 0 & \text{sinon.} \end{cases} \tag{3.7}$$

Ce filtre impacte les caractéristiques considérant les lemmes et les motifs morpho-syntaxiques dans le contexte des phrases incertaines. Les résultats en appliquant ce filtre sont indiqués dans le tableau 3.11.

Enfin, les résultats sur les différents corpus d'évaluation soulignent l'importance de la définition de l'incertitude que l'on souhaite détecter et de la nature des textes.

TABLEAU 3.11 – Résultats de la méthode en utilisant un filtre sur le nombre d'occurrences des lemmes présents dans le contexte des phrases incertaines. Les caractéristiques utilisées entre les différentes confiances sont fixées en fonction de la phase de sélection des caractéristiques optimales pour le jeu de données. Les résultats présentent la précision / le rappel / la F-mesure.

	BioScope	WikiWeasel	SFU corpus
Loi binomiale	76,2 / 82,9 / 79,4	61,3 / 61,5 / 61,4	88,1 / 96,7 / 92,2
Sigmoïde	76,3 / 82,9 / **79,5**	69,7 / 55,3 / 61,7	88,3 / 96,7 / **92,3**
$1 - 1/\#_S(w)$	76,3 / 82,9 / **79,5**	68,9 / 57,7 / **62,8**	88,3 / 96,7 / **92,3**
Sans confiance	76,2 / 82,9 / 79,4	64,7 / 54,5 / 59,2	88,3 / 96,7 / **92,3**

Si l'on ne constate pas de différences significatives sur les performances obtenues sur les corpus BioScope et SFU, les résultats du tableau 3.11 démontrent l'efficacité du filtre sur les textes de WikiWeasel. En effet, en éliminant la majeure partie du bruit issu des motifs présents dans les phrases incertaines ce filtre améliore les résultats. Finalement, ce filtre permet de compenser la faiblesse de la méthode lorsqu'elle est appliquée sur des jeux de données dont la disparité des marqueurs est forte.

TABLEAU 3.12 – Comparaison des moyennes des F-mesures obtenues lors de CoNLL 2010 et de la moyenne obtenue en utilisant notre approche avec le score de confiance $1 - 1/\#_S(w)$.

	BioScope	WikiWeasel	Moyenne
TANG et al., 2010	86,4	55	70,7
GEORGESCUL, 2010	78,5	60,2	69,4
$1 - 1/\#_S(w)$	79,5	62,8	71,2

Ces résultats améliorent ceux de l'approche de GEORGESCUL, 2010 sur le corpus WikiWeasel, qui avait obtenu la première place de la tâche 1 lors de CoNLL 2010 sur ce même corpus avec une F-mesure de 60,2. De plus, nous obtenons la meilleure moyenne en terme de F-mesure, 71,2, sur les jeux BioScope et WikiWeasel par rapport à la meilleure moyenne de la conférence, 70,7 par TANG et al., 2010 (cf. tableau 3.12). Au niveau du corpus de SFU, non utilisé dans CoNLL 2010, nous avons des résultats similaires à ceux de CRUZ, TABOADA et MITKOV, 2015.

3.4.2 Comparaison avec d'autres mesures et approches

L'utilisation de la probabilité conditionnelle $p_I(c|w)$ a été confrontée à des mesures couramment utilisées dans le domaine de la classification de textes. Ces métriques considèrent un lemme w et sa relation avec une classe c. L'ensemble des valeurs obtenues à ces différents tests est donné dans le tableau 3.13.

Pointwise mutual information, PMI, mesure l'association d'un lemme w avec la classe c (cf. équation 3.8). Cette mesure est proche de la définition de notre probabilité $p_I(c|w)$. Elle pondère simplement cette probabilité par $p(c)$. Cependant, cette probabilité $p(c)$ est très faible lorsqu'on considère la classe *est marqueur d'incertitude* et aura pour conséquence de bruiter la valeur de la probabilité.

$$pmi(w,c) = log\left(\frac{p(c,w)}{p(c).p(w)}\right) = log\left(\frac{p(c|w)}{p(c)}\right) \tag{3.8}$$

Odds Ratio mesure le degré de dépendance entre un lemme w et la classe c (cf. équation 3.9). Appliqué à nos données, le *Odds Ratio* favorise les motifs avec un faible écart entre $\#_{I_{S_u}}(w)$ et $\#_S(w)$.

$$orr(w,c) = log\left(\frac{p(w|c).(1-p(w|\bar{c}))}{p(w|\bar{c}).(1-p(w|c))}\right) \tag{3.9}$$

Categorical Proportional Difference, CPD, est un ratio qui considère pour un lemme w le nombre de documents appartenant aux classes c et $\bar{c}$ qui le contiennent. L'équation 3.10 définit CPD avec dw le nombre de documents contenant w, dw_c le nombre de documents de la classe c contenant w, $dw_{\bar{c}}$ le nombre de documents de la classe $\bar{c}$ contenant w. Dans notre problématique de détection binaire de l'incertitude au niveau de la phrase, cette mesure a été adaptée, dw_c représente le nombre d'occurrences du lemme w en tant que marqueur d'incertitude ($\#_{I_{S_u}}(w)$).

$$cpd(w,c) = \frac{dw_c - dw_{\bar{c}}}{dw} \tag{3.10}$$

Weighted Log Likelihood Ratio mesure la dissimilarité de la distribution du lemme w en fonction des classes c et $\bar{c}$ (cf. équation 3.11).

$$wllr(w,c) = p(w|c).log\left(\frac{p(w|c)}{p(w|\bar{c})}\right) \tag{3.11}$$

Nous avons couplé ces différentes métriques avec les mesures de confiance définies dans la sous-section 3.3.3. Ce couplage s'apparente à l'adaptation de modèles classiquement retrouvés pour la classification de textes. Par exemple, HAMDAN, 2015

TABLEAU 3.13 – F-mesure en fonction des confiances associées à différentes métriques globales étudiées sur les corpus BioScope, WikiWeasel et SFU. Les caractéristiques utilisées entre les différentes confiances sont fixées en fonction de la phase de sélection des caractéristiques optimales pour le jeu de données. Le filtre sur le nombre d'occurrences des lemmes présents dans les phrases incertaines est appliqué.

Métrique	Confiance	BioScope	WikiWeasel	SFU
PMI	$log(\#_S(w))$	75,6	33,8	88,3
	Loi binomiale	76,6	52,3	91,5
	Sigmoïde	77,1	37,7	91
	$1 - 1/\#_S(w)$	77,3	40,6	91,1
	Sans confiance	76,4	35,1	90,6
Odds Ratio	$log(\#_S(w))$	78,1	45,5	91,1
	Loi binomiale	**79,3**	**55**	**92,2**
	Sigmoïde	**79,3**	51,5	92,1
	$1 - 1/\#_S(w)$	**79,3**	52	92,1
	Sans confiance	79,2	51,3	92,1
CPM	$log(\#_S(w))$	70,8	45,2	78,6
	Loi binomiale	69,7	48,1	80,1
	Sigmoïde	70,5	48,6	78,1
	$1 - 1/\#_S(w)$	70,4	49,9	78
	Sans confiance	69,6	48	73,3
Wllr	$log(\#_S(w))$	53,7	16,5	69,8
	Loi binomiale	55,5	45	67,1
	Sigmoïde	55,1	11,6	65,8
	$1 - 1/\#_S(w)$	55,1	11	66,3
	Sans confiance	55,1	18,9	65,7

définit le poids final d'un terme par la formule suivante :

$$w_i = localWeight \times globalWeight \times normalization \tag{3.12}$$

dans laquelle $localWeight$ est une mesure fréquentiste du terme dans le document (*e.g.* $log(termFrequency + 1)$), $globalWeight$ une métrique appliquée aux termes à l'échelle du corpus (présentée en début de section) et $normalization$ permet d'ajuster les poids en fonction de la taille du document. Une analyse a également été menée en amont sur ces mesures ainsi que sur les mesures suivantes : *Chi Square*, *Natural Entropy* et *Kullback-Leibler Divergence*. Cette étude a porté sur l'analyse du comportement de chaque mesure par rapport à la contrainte principale fixée pour notre modèle. Nous l'avons vu, cette contrainte repose sur la prise en compte, lorsque la probabilité conditionnelle est fixe, du nombre d'observations $\#_S(w)$ pour le calcul du score, tel que pour deux mots w_1 et w_2 avec $\#_S(w_1) > \#_S(w_2)$ et $p_I(c|w_1) = p_I(c|w_2)$ le score de la mesure soit supérieur pour w_1.

Le tableau 3.13 montre les différents résultats selon les métriques considérées en

fonction de l'étude menée. Le *Odd-Ratio* est la métrique la plus performante sur les trois corpus. Ainsi, une faible différence entre $\#_{I_{Su}}(w)$ et $\#_S(w)$ est une caractéristique pertinente dans la tâche de détection de l'incertitude. Cependant, la probabilité conditionnelle $p_I(c|w)$ obtient de meilleurs résultats comparée à ces autres métriques globales. De plus, cette probabilité couplée avec la confiance $1 - 1/\#_S(w)$ est la plus performante en moyenne sur les jeux de données.

Afin d'approfondir les comparaisons avec notre approche, nous avons expérimenté une méthode récente de classification de textes, intitulée FASTTEXT (JOULIN et al., 2016). Cette approche propose une extension du modèle continu *skip-gram* de MIKOLOV et al., 2013. On appelle un *k-skip-n-gram* une sous-séquence de taille n où les composants (mots, caractères) sont à une distance d'au plus k les uns des autres. Par exemple, un *1-skip-2-grams* pour la phrase *John likes my shoes* entraîne les sous-séquences : *John likes*, *likes my*, *my shoes*, *John my* et *likes shoes*. Nous avons expérimenté cette méthode en utilisant les paramètres par défaut (*epoch* = 5, représentant le nombre de fois où chaque exemple d'entraînement est vu), sur les trois différents corpus d'évaluation. Le tableau 3.14 présente les résultats obtenus avec cette méthode. On peut constater que notre approche obtient, là aussi, de meilleurs résultats.

TABLEAU 3.14 – Résultats obtenus avec l'approche FASTTEXT comparés aux résultats de notre approche. Les résultats sont présentés sous la forme précision/ rappel / F-mesure.

	BioScope	WikiWeasel	SFU corpus
FASTTEXT	81,6 / 65,6 / 72,3	80,8 / 24,8 / 37,9	93,9 / 85,6 / 89,5
$1 - 1/\#_S(w)$	76,3 / 82,9 / **79,5**	68,9 / 57,7 / **62,8**	88,3 / 96,7 / **92,3**

3.4.3 Expérimentations complémentaires

Dans le but de compléter notre vision de l'incertitude linguistique, nous avons expérimenté l'approche sur des textes appartenant à différents domaines. Pour cela, 5 styles de textes ont été utilisés : roman, économie, politique, religion et bio-médical. Les 4 premiers styles proviennent des corpus associés à la bibliothèque nltk sur Python avec respectivement, le livre Moby Dick, le journal *Wall Street*, un ensemble de discours d'inauguration des Présidents Américains et le livre de la Genèse. Concernant les textes bio-médicaux, ils proviennent de 10000 résumés d'articles bio-médicaux extraits pour le challenge BioAsq de 2015. Les résultats sont résumés dans le tableau 3.15.

À partir du tableau 3.15, une première hypothèse peut être formulée concernant l'impact de l'incertitude dans le processus d'inférence de connaissances. En effet, on observe que le pourcentage d'incertitude est très fluctuant selon le domaine textuel

TABLEAU 3.15 – Résultats obtenus avec notre approche sur différents domaines textuels.

	# de phrases	% d'incertitude
Moby Dick (roman)	5561	28
Jounal *Wall Street* (économie)	3093	22
Discours d'inauguration (politique)	4600	34
La Genèse (religion)	1276	10
Abstracts scientifique (bio-médical)	78377	20

dans lequel se situe les textes exploités. Ainsi, on peut facilement imaginer que l'influence de l'incertitude dans un contexte d'inférence sera plus importante pour les textes politiques que religieux.

Une autre étude intéressante qui pourrait être menée serait d'observer à quel point ce pourcentage d'incertitude constituerait une caractéristique discriminante d'un style de texte en particulier, ou d'un auteur à l'instar des travaux de SAVOY, 2017 sur la classification des textes par l'exploitation des *stop words*. Toutefois pour la mener, nous aurions besoin de plus de textes pour affiner ces pourcentages.

3.5 Synthèse et perspectives

Ce chapitre présente une méthode d'apprentissage automatique pour la détection binaire de l'incertitude dans le langage naturel.

Cette méthode générique est basée sur la sélection des caractéristiques optimales à partir d'un ensemble initial de caractéristiques afin d'obtenir une représentation vectorielle concise d'une phrase. Cette représentation s'appuie sur l'analyse de caractéristiques locales et globales au niveau de la phrase. Les caractéristiques locales sont construites à partir d'une agrégation spécifique des différentes probabilités conditionnelles qu'un n-gramme appartienne à un contexte donné, pondérées par un score de confiance. La longueur de la phrase a été utilisée comme caractéristique globale. Les expérimentations montrent que cette approche obtient de bons résultats sur toutes les dimensions de l'incertitude et améliore les meilleurs résultats connus sur WikiWeasel. Un composant important de la modélisation que nous proposons est la notion de confiance qui peut être associée aux observations contextuelles. En effet, dans cette étude, nous avons proposé et évalué plusieurs critères de confiance. Enfin, le modèle proposé a été comparé à plusieurs métriques classiquement utilisées en TAL et en Théorie de l'Information. L'approche est disponible au téléchargement à cette adresse *https ://github.com/pajean/uncertaintyDetection*. Celle-ci permet de reconstituer les différentes expérimentations sur les jeux de données ou de réaliser

une détection de l'incertitude sur de nouvelles phrases non annotées [11].

Plusieurs pistes d'amélioration de la méthode pourraient être envisagées pour des travaux futurs. Ces pistes concernent notamment le calcul des poids des motifs *n-grammes*. En effet, un mécanisme de propagation basé sur l'analyse des collocations permettrait une pondération contextuelle plus précise (LAVALLEY, CLAVEL et BELLOT, 2010) ; ceci dans le but d'éviter des erreurs de classification dues au poids d'un lemme trop discriminant. Une autre piste d'amélioration serait d'ajouter une caractéristique contextuelle au niveau de la phrase *i.e* indiquer par une valeur booléenne si la phrase précédente est détectée comme incertaine. On considérerait dans ce cas l'hypothèse qu'une phrase aurait plus de chance d'être incertaine si les phrases précédentes étaient incertaines. Une autre amélioration serait d'étendre la nature des motifs utilisés dans les caractéristiques. Actuellement, seulement deux types sont utilisés, les lemmes et les motifs morphosyntaxiques. Nous pourrions par exemple expérimenter les étiquettes d'un arbre des dépendances en tant qu'unité de base d'un motif ou élaborer un motif hybride de plusieurs types (CHEN et DI EUGENIO, 2010). Enfin, nous pourrions également étudier comment les similarités sémantiques entre les mots (HARISPE et al., 2015) pourraient être intégrées afin de généraliser autant que possible les connaissances extraites utilisées par le modèle *i.e* augmenter la quantité d'information portée par les observations.

Dans une volonté d'envisager d'autres aspects de la chaîne de traitement, ces perspectives sont uniquement des ouvertures pour perfectionner l'approche. Elles ne sont pas abordées dans la suite de la thèse. Le prochain chapitre porte sur la mise en place du module de raisonnement à partir des relations extraites. Ce dernier repose sur la hiérarchisation et l'enrichissement des extractions dans l'objectif d'augmenter la connaissance à notre disposition et de l'évaluer.

11. La construction des caractéristiques à partir des phrases non annotées utilise le module Python nltk (www.nltk.org) et la prédiction le module *scikit-learn* (www.scikit-learn.org).

Chapitre 4

Inférence et évaluation de la connaissance

Sommaire

Ce chapitre détaille les processus d'inférence et d'évaluation de la connaissance exploités au terme des phases d'extraction de relations et de détection de l'incertitude. L'objectif est d'accroître la portée opérationnelle des méthodes d'extraction de relations telles que REVERB et OLLIE dans le cadre d'une hypothèse en monde ouvert, en tenant compte de l'incertitude linguistique. Actuellement, ces méthodes se cantonnent à l'extraction, mais n'exploitent pas de processus particulier pour évaluer et acquérir de nouvelles connaissances. Par conséquent, nous proposons de combler ce manque au travers d'un module de raisonnement tirant profit d'un ensemble d'informations sur les relations dans le cadre d'une analyse multi-critère. Ces informations reflètent les caractéristiques choisies pour juger la pertinence d'une relation. Dans le cas de notre étude, nous exploitons la croyance de l'information transmise et son contenu informatif *i.e.* la spécificité de l'information véhiculée. Par conséquent, nous définissons un cadre formel pour l'inférence et l'évaluation des déclarations. Au sein de ce cadre la notion d'imprécision est abordée sous l'angle de l'imprécision

taxonomique (cf. sous-section 2.2.1). Ce type d'imprécision provient du degré de spécificité des concepts employés dans les phrases et de la manière de les spécifier par l'emploi d'une structure syntagmatique particulière.

La première section de ce chapitre présente le module d'inférence permettant de générer de nouvelles informations en tenant compte de la complexité des entités partiellement désambiguïsées récoltées lors de la phase d'extraction. Cette complexité attachée aux entités provient de la richesse du langage naturel et des multiples possibilités d'énoncer une information en langage naturel et de l'efficience de la phase de désambiguïsation. Ainsi, nous proposons une méthodologie pour la considérer au travers d'un graphe, que nous appelons graphe des syntagmes, permettant de structurer le sujet et l'objet des relations. C'est ce graphe qui permet, par la suite, la génération de nouvelles relations. À noter que ce processus suit une hypothèse en domaine ouvert héritée de la méthode d'extraction de relations, induisant ainsi une sémantique particulière à la relation générée qui sera discutée dans ce chapitre.

La seconde section, quant à elle, expose la méthodologie mise en place pour l'évaluation de la pertinence des déclarations. Cette dernière s'appuie sur une procédure de hiérarchisation des déclarations extraites et générées afin de calculer différents critères permettant de les évaluer. Ces critères tiennent compte de la croyance associée à la relation et son contenu informationnel. La croyance est abordée au travers d'un mécanisme de propagation *bottom-up* des observations (favorisant le principe de généralisation au sens de l'induction). Ce dernier permet d'estimer une valeur de croyance à chaque déclaration en tenant compte des observations des déclarations plus spécifiques dans le graphe des déclarations. Concernant le contenu informationnel, il est principalement retranscrit par la profondeur de la déclaration dans ce même graphe. C'est lors de cette phase que l'incertitude linguistique est prise en compte (cf. sous-section 4.2.2).

La troisième section présente les différents modèles de sélection établis pour évaluer la pertinence des déclarations. Ces modèles se basent sur les différents critères calculés pour chaque déclaration. De plus, cette section met en avant une alternative à ces modèles de sélection tenant compte des profils des utilisateurs, *i.e.* de leur niveau de connaissance et de leur façon de rechercher de l'information.

Enfin, la quatrième section propose une implémentation « jouet » de la chaîne de traitement au travers d'une interface permettant de réaliser des requêtes à partir d'un ensemble de relations issues de ReVerb extraites du corpus *ClueWeb09*. Différentes visualisations sont proposées au sein de cette section et les choix techniques de développement sont discutés.

4.1 Modalités d'inférence de connaissances

Cette section discute de l'étape d'inférence de connaissances réalisée sur les données extraites par l'intermédiaire d'un processus d'induction. Tel qu'il a été précisé dans le chapitre 1, ce processus permet de découvrir de la connaissance par la généralisation des observations. Dans ce cas, toute déclaration induite lors de ce processus et qui n'est pas explicitée dans les textes est dite découverte[1]. Toutefois avant de détailler notre étape d'inférence, il est important de présenter les différentes sémantiques qu'arbore la notion d'inférence au sein de la littérature selon le contexte d'utilisation. La première sous-section présente un état de l'art de cette notion et la seconde détaille notre processus d'inférence utilisé au sein de la chaîne de traitement.

4.1.1 Introduction et contexte

Il est important de noter la différence sémantique entre notre phase d'inférence et les méthodes classiques d'inférence appliquées aux bases de connaissances et aux textes. Les trois parties suivantes présentent diverses méthodes d'inférence de connaissances étant donné un contexte d'utilisation particulier.

L'inférence dans les bases de connaissances

Depuis quelques années, la notion de bases de connaissances est remplacée dans la littérature par la notion de graphe de connaissances. Un graphe de connaissances représente la combinaison de toutes les relations de type <s,p,o> au sein d'un graphe dans lequel chaque nœud correspond à une entité (un sujet ou un objet) et chaque arc, qui sont dirigés, à la relation entre deux entités. La direction de ce segment correspond au sens de la relation, du sujet vers l'objet. L'inférence appliquée sur ces graphes de connaissances répond à une problématique de complétion de ces derniers et des bases de connaissances par extension. NICKEL et al., 2016 exposent un état de l'art complet de ce domaine. Ils le définissent comme l'ensemble des méthodologies permettant de prédire de nouveaux faits à propos du monde, en d'autres termes, prédire de nouveaux arcs au sein du graphe considéré. Les auteurs distinguent deux principaux modèles : les modèles basés sur l'inférence de caractéristiques latentes au niveau des entités (*Latent Feature Models*) et les modèles basés sur les caractéristiques des graphes (*Graph Feature Models*). Ces deux modèles se rapportent à la notion de SRL (*Statistical Relational Learning*).

Les modèles basés sur l'inférence de caractéristiques latentes au niveau des entités permettent de capturer certaines corrélations entre les nœuds et les arcs d'un graphe

1. Le terme de découverte est généralement remplacé par "généré" pour le distinguer du domaine scientifique de la découverte de connaissances.

de connaissances dans le but de prédire de nouveaux arcs. Une caractéristique est dite latente lorsqu'elle n'est pas directement observable à partir des données. Par exemple, une explication possible pour la relation <*Alec Guinness, received, the Academy Award*> est, qu'Alec Guinness est un bon acteur (NICKEL et al., 2016). Cette explication s'appuie sur une caractéristique latente d'une entité (être un bon acteur) pour expliquer un fait observable. NICKEL, TRESP et KRIEGEL, 2011 ont proposé une approche basée sur ce principe, intitulée RESCAL. Cette dernière présente une méthode de factorisation tensorielle permettant de prédire l'existence de relations au sein du graphe de connaissances. Elle se base sur un tenseur d'ordre 3, $X \in \mathbb{R}^{n \times n \times m}$, où n correspond au nombre d'entités et m au nombre de relations. Les entrées de ce tenseur sont :

$$x_{ijk} = \begin{cases} 1 & \text{si la relation } k(\textit{entité}_i, \textit{entité}_j) \text{ existe} \\ 0 & \text{sinon.} \end{cases} \tag{4.1}$$

Leur approche réalise une factorisation spécifique où chaque dimension X_k appartenant au tenseur X est factorisée comme :

$$X_k \approx A R_k A^T \tag{4.2}$$

où A est une matrice $n \times r$ qui contient la représentation des composantes latentes des entités, A^T est la transposition de la matrice A et R_k est une matrice antisymétrique $r \times r$ qui modélise les interactions des composantes latentes pour le k-ème prédicat avec r un entier positif paramétrable. Les matrices A et R_k s'obtiennent par la minimisation du risque pénalisé décrit par l'équation 4.5 (NICKEL, TRESP et KRIEGEL, 2011).

$$\min_{A,R_k} f_{loss}(A, R_k) + f_{reg}(A, R_k) \tag{4.3}$$

où

$$f_{loss}(A, R_k) = \frac{1}{2}(\sum_k \parallel X_k - A R_k A^T \parallel_F^2) \tag{4.4}$$

et f_{reg} un terme de régularisation (une pénalité) permettant d'éviter le sur-apprentissage avec λ un paramètre d'équilibrage dont le réglage établit le niveau d'importance relatif entre les deux critères (DELPORTE, 2013)

$$f_{reg}(A, R_k) = \frac{1}{2}\lambda(\parallel A \parallel_F^2 + \sum_k \parallel R_k \parallel_F^2) \tag{4.5}$$

Par la suite, le produit vecteur-matrice $x_{ijk} = a_i^T R_k a_j$, où a_i^T est la transposition de la i-ème ligne de la matrice A correspondant à la représentation latente de l'entité i et a_j le vecteur ligne de l'entité j, peut être interprété comme le score assigné par le modèle pour représenter l'existence d'une relation $< \textit{entité}_i, k\text{-ème } \textit{prédicat}, \textit{entité}_j >$. La valeur x_{ijk} peut alors être comparée avec un seuil θ afin de déterminer si la relation donnée existe (NICKEL, TRESP et KRIEGEL, 2012).

Les modèles basés sur les caractéristiques des graphes supposent qu'il existe des arcs pouvant être prédits à partir des arcs existants du graphe. Au sein d'un tel contexte, ces modèles font généralement intervenir les concepts de la logique formelle et de la programmation logique pour évaluer et découvrir de nouvelles connaissances par le biais des relations issues de la base (LAO, MITCHELL et COHEN, 2011). Traditionnellement, ces méthodes représentent les connaissances inférées sous la forme de clause. Sachant qu'une clause est une conjonction ou une disjonction de littéraux *e.g.* $p \wedge q$ et $p \vee q$ et qu'un littéral est soit une formule atomique (appelée littéral positif) soit une formule atomique négative (appelée littéral négatif) *e.g.* p et $\neg q$. Les clauses les plus couramment utilisées sont les clauses de Horn (HORN, 1951 ; GALÁRRAGA et al., 2013). Une clause de Horn est une clause disjonctive avec au plus un littéral positif *e.g.* l'équation 4.6.

$$\neg X1 \vee \neg X2 \vee ... \vee \neg Xn \vee Y \tag{4.6}$$

Ces clauses sont exploitées afin de représenter des implications de la forme $X \rightarrow Y$ au sein des bases de connaissances. Pour réaliser le lien entre cette formule et une clause de Horn l'implication nous permet d'exprimer cette dernière formule sous la forme $\neg X \vee Y$. Ainsi, si $X = \neg X_1 \vee \neg X_2 \vee ... \vee \neg X_n$ alors d'après la loi de Morgan [2], $\neg X$ est équivalent à $X_1 \wedge X_2 \wedge ... \wedge X_n$. Ainsi, la clause 4.6 est équivalente à la clause 4.7.

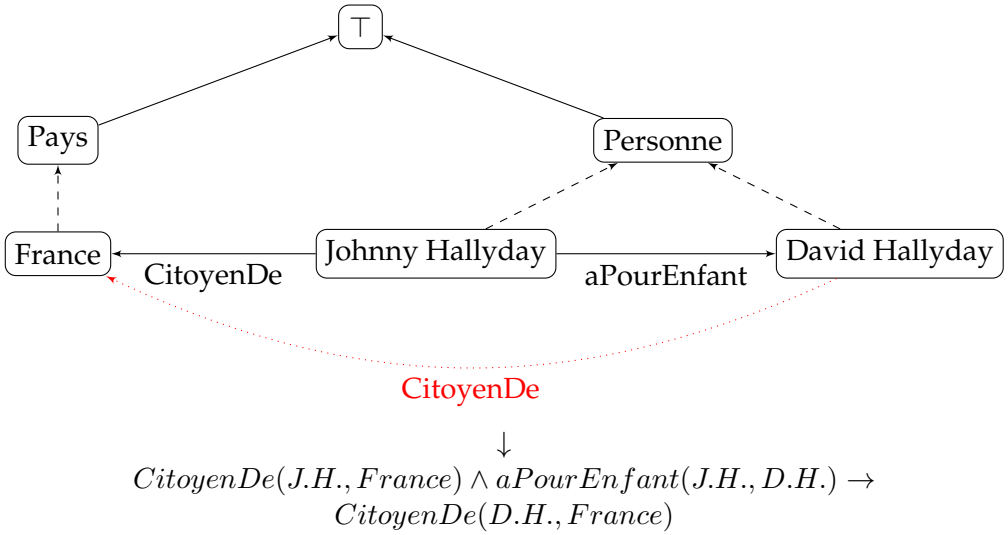
$$(X_1 \wedge X_2 \wedge ... \wedge X_n) \rightarrow Y \tag{4.7}$$

La figure 4.1 représente un exemple d'utilisation de ces clauses au sein d'une ontologie. Ces clauses peuvent être perçues comme la définition informelle de règles axiomatiques.

Plusieurs types d'approche exploitent les principes précédemment définis. TARI et al., 2010 définissent une méthode pour rechercher des interactions entre médicaments à partir de textes en exploitant les principes de la logique formelle pour accroître les connaissances extraites. Pour cela deux phases sont définies, l'une portant sur le traitement des textes et l'autre sur l'inférence, évoquant ainsi une architecture semblable à notre chaîne de traitement. La première des phases combine les résultats

2. La négation de la conjonction de deux propositions est équivalente à la disjonction des négations des deux propositions et vice versa. Respectivement $\neg(A \wedge B) \equiv \neg A \vee \neg B$ et $\neg(A \vee B) \equiv \neg A \wedge \neg B$.

FIGURE 4.1 – Illustration de l'inférence de connaissances à partir des clauses de Horn. La flèche en pointillés rouges représente la connaissance inférée. Cette étape d'inférence est généralement dépendante du support des différentes relations au sein de la base.



$$CitoyenDe(J.H., France) \wedge aPourEnfant(J.H., D.H.) \rightarrow CitoyenDe(D.H., France)$$

issus des dépendances syntaxiques d'une phrase avec un ensemble de mots clés *e.g. induce, inhibit, metabolize* pour extraire des relations de type <s,p,o>. Une fois cette extraction réalisée, la seconde phase correspond à l'inférence d'interactions médicamenteuses indirectement citées dans les textes. Cette étape de raisonnement se base sur un ensemble de clauses définies manuellement par les auteurs et décrivant les propriétés métaboliques des médicaments (cf. figure 4.2).

FIGURE 4.2 – Traduction en clause de Horn d'une règle définie manuellement dans l'article de TARI et al., 2010. Celle-ci se traduit de la manière suivante, si le médicament 1 ($Drug1$) induit une augmentation de l'activité de la protéine P et que P métabolise le médicament 2 ($Drug2$) alors $Drug1$ diminue les effets de $Drug2$.

$$induces(Drug1, P) \wedge enzyme(P) \wedge metabolized(Drug2, P) \wedge drug(Drug1) \wedge drug(Drug2) \rightarrow decreases(Drug1, Drug2)$$

Outre l'extraction (ou l'application) de clauses de Horn, il existe d'autres méthodes exploitant les caractéristiques des graphes de connaissances. Par exemple, LAO et COHEN, 2010 introduisent l'algorithme PRA (*Path Ranking Algorithm*) basé sur le principe de marche aléatoire (de taille limitée) au sein des graphes de connaissances. L'idée sous-jacente de PRA est d'utiliser explicitement les chemins entre deux entités comme caractéristiques afin de prédire des relations potentielles entre eux. Cet algorithme se décompose en trois principales étapes : l'extraction des chemins caractéristiques, le calcul des poids associés à ces chemins et une étape de classification spécifique pour juger si deux entités sont reliées par une relation donnée (WANG et al., 2016).

Ainsi, les méthodes employées dans le domaine de la complétion des bases de connaissances sont diverses. Elles vont de l'exploitation des caractéristiques associées aux

graphes de connaissances à celle des entités la peuplant. La partie suivante traite de la notion d'inférence appliquée, cette fois-ci, au domaine textuel.

L'inférence textuelle

Depuis quelques années les termes de *textual inference* sont apparus dans la littérature (MACCARTNEY et MANNING, 2007). Ces derniers font référence principalement aux méthodes appliquées à la tâche de reconnaissance des implications textuelles (*Textual Entailment* – TE) dont la compétition RTE (*Recognising Textual Entailment*) a favorisé les avancées. Nous avons déjà évoqué cette tâche au travers d'un exemple dans la sous-section 2.1.3 présentant l'importance de l'analyse sémantique des phrases. DAGAN, GLICKMAN et MAGNINI, 2006 définissent cette tâche de reconnaissance de la manière suivante.

> ***Textual entailment*** Étant donné deux portions de textes, la tâche nécessite de retrouver si la signification d'une des portions peut être inférée à partir de l'autre. De manière formelle, si T (*Text*) est le texte impliquant et H (*Hypothesis*) le texte impliqué, alors T *implique* H si un humain en lisant T inférerait que H est vraisemblable.

Dans la manière d'interpréter les deux portions textuelles, cette tâche recouvre plusieurs applications en TALN *e.g.* questions-réponses, extraction d'information, recherche d'information, etc. Pour illustrer cela, la figure 4.3 propose trois cas d'application touchant des domaines différents. De plus, la bibliothèque python *nltk* regroupe les données issues des trois premières années de la compétition RTE.

FIGURE 4.3 – Exemples issus du jeu de données RTE-1. T représente le texte impliquant et H le texte impliqué. Chaque exemple est issu d'une application spécifique en TAL et chaque implication est vraie.

Recherche d'information

T *The wait time for a green card has risen from 21 months to 33 months in those same regions.*

H *It takes longer to get green card.*

Questions-réponses

T *That prompted Clinton wife Hillary Rodham Clinton and daughter Chelsea to spend a cold and dreary Saturday in Minsk.*

H *Clinton's wife is called Hillary Rodham Clinton.*

Extraction d'information

T *There can be no doubt that the Administration already is weary of Aristide, a populist Roman Catholic priest who in December, 1990, won an overwhelming victory in Haiti's only democratic presidential election.*

H *Aristide became president of Haiti in 1990.*

Concernant les méthodes employées, elles sont de natures diverses. Par exemple, JIJKOUN et RIJKE, 2005 exploitent des mesures lexicales et sémantiques pour détecter l'implication entre les textes. Leur méthode, soumise lors de la première compétition RTE, avait obtenu une précision de 0,55. Un autre exemple d'approche, exploitant le même jeu de données, est celle proposée par BOS et MARKERT, 2005. Cette méthode s'appuie sur la traduction des phrases en logique de première ordre par l'intermédiaire d'un cadre théorique linguistique appelé *Discourse representation theory* (KAMP et REYLE, 1993). Elle avait obtenu une précision de 0,65.

Un modèle d'inférence hybride : *Knowledge Vault*

Il est intéressant de présenter cette approche réunissant à la fois l'extraction d'information et l'exploitation d'une connaissance *a priori*. En effet, *Knowledge Vault* (DONG et al., 2014) est une base de connaissances construite automatiquement à partir d'une chaîne de traitement composée de trois principales étapes :

1. Extraction de relations à partir de textes de diverses natures issus du Web (structurés, semi- et non-structurés).
2. Modèles SRL (*Statistical Relational Learning*) exploités sur les relations contenues dans un graphe de connaissances.
3. Évaluation de la confiance associée aux relations extraites à partir des scores obtenus des extractions et des modèles SRL.

Les modèles SRL employés au sein de cette approche utilisent une combinaison des modèles précédemment définis pour prédire les arcs : caractéristiques latentes des entités et caractéristiques des graphes de connaissances. Pour prédire le score final d'un fait, les scores des modèles SRL sont exploités avec diverses caractéristiques dérivées des faits extraits telles que les scores des extracteurs et le nombre de pages Web impliquées, au sein d'un modèle de régression logistique pour calculer la probabilité que le fait existe.

En conclusion, les trois parties précédentes montrent l'importance de la notion d'inférence au sein de la littérature. Ce processus de raisonnement couvre de nombreux domaines et est employé sur divers supports : bases de connaissances et textes. Il est sollicité dans le cadre de l'acquisition de nouvelles connaissances pouvant être implicites dans les données initiales. Les méthodes employées pour réaliser ce processus sont de natures diverses allant de la logique formelle à l'analyse sémantique des textes. Au regard de ces différentes utilisations de la notion d'inférence et dans le cadre de notre chaîne de traitement, nous justifions son emploi comme un moyen de générer de nouvelles connaissances par la généralisation des observations. Dans notre cas, la génération est effectuée en tenant compte des propriétés syntaxiques et taxonomiques des syntagmes obtenus lors de la phase d'extraction de relations qui est ensuite combinée à une phase de sélection basée sur différents critères. La

considération de la décomposition syntaxique, d'un ordre taxonomique *i.e.* l'abstraction des entités, et de l'incertitude linguistique sont les aspects différenciants de notre chaîne de traitement au regard des méthodologies précédemment définies. La sous-section suivante présente l'approche d'inférence de connaissances, tandis que la phase de sélection est décrite en section 4.3.

4.1.2 Module d'inférence de connaissances

L'idée développée dans cette section consiste à générer de nouvelles déclarations en tenant compte des entités complexes qui composent le sujet et l'objet d'une relation. Cette complexité provient, dans un premier temps, de la conservation des éléments modificateurs associés aux entités tels que les adjectifs ou les syntagmes prépositionnels. Par exemple, si nous avons la phrase suivante : "Les hommes américains blancs ont plus de chance de développer du diabète", on peut observer que l'entité d'intérêt du sujet de la relation, généralement désignée dans l'arbre syntaxique comme le mot "tête" de la partie en question (COLLINS, 1996), est le mot "hommes" et que les adjectifs "américains" et "blancs" sont des modificateurs de cette entité. Ces modificateurs sont conservés dans le sujet et l'objet des relations lors de la phase d'extraction. Dans un second temps, la complexité des entités peut croître lors de la phase de désambiguïsation. En effet, cette dernière peut désambiguïser le syntagme entier du sujet ou de l'objet comme elle peut n'en désambiguïser qu'une partie. La conséquence d'une désambiguïsation partielle serait l'obtention de relations mixant des mots et des identifiants, réalisant ainsi un lien partiel avec une taxonomie donnée (cf. tableau 4.1).

TABLEAU 4.1 – Exemple de sujets à traiter à partir du jeu de données ClueWeb09. La désambiguïsation a été réalisée à partir d'une méthode de maximum de similarité avec la mesure de Wu & Palmer sur WordNet. Le chiffre représente l'identifiant du concept correspondant dans WordNet.

Syntagme initial	Syntagme désambiguïsé avec WordNet
sexual relation	13931765
international relation	*international* 10235549

Nous formalisons la considération de ces deux principaux attributs (décomposition syntaxique et taxonomique), associés aux syntagmes des sujets et objets d'une relation extraite, comme la prise en compte des implications directes et indirectes entre les mots des syntagmes. L'implication directe découle de la propriété d'inclusion entre ces mots. Elle est retranscrite comme la considération des modificateurs syntaxiques associés à l'entité tête du sujet ou de l'objet d'une relation. Cette modification est perçue comme un moyen de spécialiser l'entité. Nous qualifions ce type d'implication de syntaxique. Concernant l'implication indirecte, elle exploite une structure taxonomique afin d'enrichir les informations acquises par le système.

Nous qualifions ce type d'implication de taxonomique. Par conséquent, cette phase est fortement dépendante du processus de désambiguïsation utilisé permettant de faire le lien entre les entités textuelles et la taxonomie. À noter que la considération de ce type d'implication n'est pas obligatoire pour la construction du graphe des syntagmes, ce qui permet à l'approche une plus grande souplesse. En effet, certains domaines spécialisés n'ont pas forcément une taxonomie à disposition. Toutefois, l'absence d'une phase de correspondance avec une taxonomie provoque une diminution de la richesse des inférences de notre approche.

L'exploitation de ces deux types d'implications a pour but d'enrichir le contenu informationnel constitué à l'issue de la phase d'extraction, en y adjoignant la génération de nouvelles relations réalisée à partir de l'ordre partiel hiérarchisant les entités de nos déclarations extraites. Les parties suivantes détaillent la construction de cet ordre partiel à partir des entités et le processus de génération de nouvelles relations.

Construction d'un ordre partiel sur les syntagmes

L'objectif de cette phase est d'ordonner les syntagmes nominaux correspondant aux sujets et objets appartenant aux relations extraites, au sein d'un graphe tenant compte des spécialisations des entités et permettant l'enrichissement des extractions par une connaissance externe. Un tel graphe peut est comparable à celui présenté dans JACQUEMIN, 1999. En considérant T un ensemble de termes (vocabulaire), on définit un syntagme σ comme une séquence de termes $\sigma = [t_1, t_2, \ldots, t_i]$ avec i la taille du syntagme et $t_{[1,\ldots,i]} \in T$. On nomme alors Σ l'ensemble des syntagmes possibles. La procédure de construction du graphe est réalisée à partir des particularités syntaxiques de la langue anglaise. Toutefois au prix de quelques ajustements, l'approche pourrait être adaptée à d'autres langues.

La phase de construction du graphe est précédée par la normalisation des syntagmes réalisée à l'aide de règles lexicales et syntaxiques permettant de simplifier la construction du graphe en éliminant des nœuds non informatifs souvent liés à des mots vides (*stop words*) *e.g. the, and*, etc. La première est une règle d'inversion centrée sur le terme *of* où une séquence ordonnée composée des trois symboles [σ_1 *of* σ_2] amènera la formation du syntagme [σ_2 σ_1]. Par exemple, le sujet *Down-regulation of AIMP2-DX2* correspond à la forme normalisée *AIMP2-DX2 down-regulation*. La seconde règle s'appuie sur la correspondance avec des patrons syntaxiques prédéfinis. Cette correspondance permet de capter l'essentiel de l'information tout en évitant de surcharger le graphe de syntagmes. Par exemple, le syntagme *younger freshwater terrapins* est conservé car il répond au patron grammatical : Adjectif – Adjectif – Nom(s). Dans l'approche développée, cette phase parcourt les étiquettes morpho-syntaxiques du sujet et de l'objet et s'arrête lorsqu'elle rencontre un label indésirable soit dans la plupart des cas un déterminant ou une préposition. Par exemple, avec le

syntagme *The walking dead* l'approche considère uniquement *walking dead*. Cette subtilité évite la création d'une feuille supplémentaire et non informative dans le graphe des syntagmes. À ces deux règles, nous pouvons ajouter la normalisation concernant la présence de la conjonction de coordination *and* et/ou la présence d'une énumération. Ces deux structures linguistiques nécessitent la décomposition de la relation en n sous-relations. Cette normalisation nécessite toutefois un traitement particulier notamment sur la délimitation des nouvelles relations. En effet, elle peut différer selon l'entité que l'on veut caractériser (cf. figure 4.4).

FIGURE 4.4 – Gestion de la conjonction de coordination *and* et de l'énumération dans la normalisation des relations. On observe que selon l'entité que l'on caractérise, il est parfois nécessaire de reprendre une même sous-partie de l'entité dans les nouvelles relations *e.g.* Phrase 1.

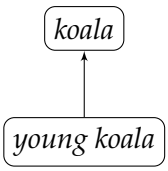
Phrase 1	*Crosstalk generates a complicated and robust signaling network.*
	↓
REVERB	<*crosstalk, generate, complicated and robust signaling network*>
Solution idéale	<*crosstalk, generate, complicated signaling network*> <*crosstalk, generate, robust signaling network*>
	———
Phrase 2	*EESB treatment could alter Bcl-2, CDK4 and p21.*
	↓
REVERB	<*EESB treatment, alter, Bcl-2, CDK4 and p21*>
Solution idéale	<*EESB treatment, alter, Bcl-2*> <*EESB treatment, alter, CDK4*> <*EESB treatment, alter, p21*>

Une fois l'obtention des relations normalisées, le graphe de syntagmes est élaboré. Ce dernier reprend les principes d'implications syntaxiques et taxonomiques énoncés dans SALVO BRAZ et al., 2006. Les auteurs de cette publication les ont exploitées pour représenter une phrase sous la forme d'un graphe de concepts. Dans l'absolu, l'ordonnancement des syntagmes a pour objectif de définir un ordre partiel sur Σ : $O_\Sigma = (\Sigma, \preceq)$. La construction de cet ordre partiel repose sur deux règles principales.

1. **Règle d'inclusion** (implications syntaxiques) : cette règle exploite une propriété d'inclusion entre les termes. Un syntagme $\sigma = [t_1, \ldots, t_i]$ spécialise tout syntagme σ' formé d'une séquence de termes contigus de σ incluant t_i, et se note $\sigma \prec \sigma'$. Par exemple *young koala* spécialise *koala* (cf. figure 4.5). La hiérarchisation des syntagmes a fait l'objet d'une tâche lors de SemEval 2015 (BORDEA et al., 2015) et le modèle le plus performant s'appuie également sur

cette règle d'inclusion des termes (GREFENSTETTE, 2015) [3].

FIGURE 4.5 – Implication directe sur le syntagme *young koala*. La feuille *young koala* spécialise l'entité *koala*.



2. **Règle d'abstraction** (implications taxonomiques) : nous considérons ici un ordre partiel sur les concepts d'une taxonomie $O_{\mathcal{C}} = (\mathcal{C}, \preceq)$ et des labels associés à ces concepts, *e.g. Phascolarctos cinereus* est un label faisant référence au concept de *koala*. Il est considéré qu'un syntagme de taille i est généralisé par un concept s'il est composé d'une séquence de termes $[t_{j\geq 0}, \ldots, t_i]$ identifiée comme expression de ce concept par un système de désambiguisation [4]. Les syntagmes sont aussi abstraits en tenant compte de l'ensemble des abstractions des concepts précisées dans la taxonomie, *e.g.* si l'on observe *young koala* et que la taxonomie définit *koala* $\prec$ *opossum*, alors le graphe de syntagmes contiendra la relation *young koala* $\prec$ *young opossum*.

Ces deux règles permettent la génération d'un graphe traduisant un ordre partiel sur les syntagmes. Par exemple si on considère $\mathcal{C} = \{'marsupial', 'opossum', 'koala'\}$ et l'ordre sur les concepts associés dans $O_{\mathcal{C}}$ (*i.e.* $koala \preceq opossum$ *et* $opossum \preceq marsupial\}$), la déclaration <*young koala, eat, eucalyptus leaves*> permet de générer l'ordre partiel de syntagmes défini dans la figure 4.6.

Cet ordre partiel sur les syntagmes enrichi par les ancêtres des concepts désambiguïsés par rapport à une taxonomie externe a pour but d'accroître le contenu informationnel du modèle de connaissance et de permettre ainsi la génération de nouvelles déclarations. Ce processus de génération fait l'objet de la partie suivante.

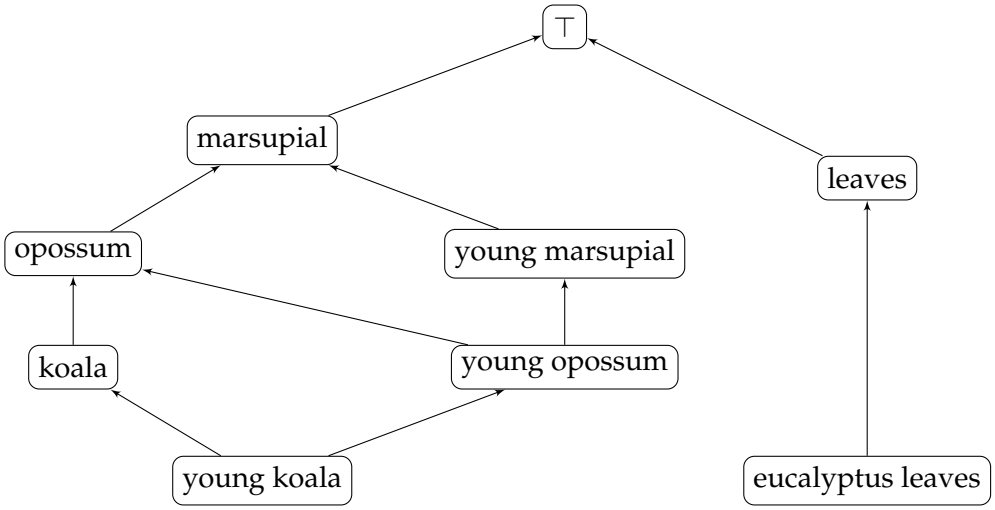
Génération de nouvelles déclarations

A partir de l'ordre partiel défini sur les syntagmes à l'étape précédente et disposant d'un ensemble de déclarations il est possible de générer de nouvelles déclarations. Ce procédé peut être réalisé par le produit cartésien entre les ensembles formés par les ascendants du sujet et ceux de l'objet d'une déclaration donnée. Par exemple, à partir de l'ordre partiel défini en figure 4.7 et de la relation initiale <*koala*, *eat*,

3. Notons que ce qui est vrai pour le traitement de textes en langue anglaise dans laquelle les adjectifs sont toujours positionnés avant le nom qu'ils qualifient, ne se vérifie pas forcément pour d'autres langues. Là encore des adaptations devront être imaginées pour une application à une autre langue.

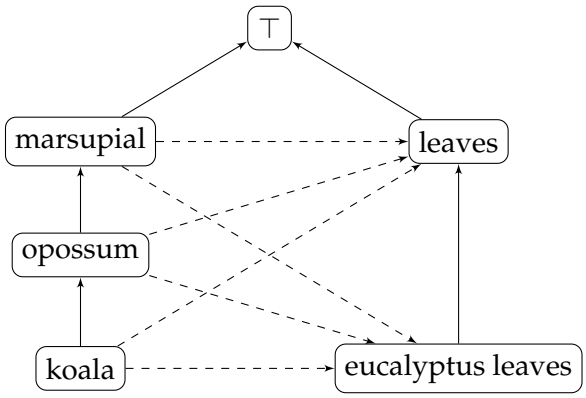
4. Par exemple, l'observation du syntagme *young Phascolarctos cinereus* induira la relation d'équivalence sur les syntagmes *young Phascolarctos cinereus* $\equiv$ *young koala* car le label *young Phascolarctos cinereus* est associé au concept *koala*.

FIGURE 4.6 – Ordre partiel défini à partir du syntagme *'young koala'* provenant de la déclaration <*young koala, eat, eucalyptus leaves*> et d'une taxonomie existante dans laquelle $koala \preceq opossum$ et $opossum \preceq marsupial$.



eucalyptus leaves>, le système est capable de générer les déclarations : <*koala, eat, leaves*>, <*opossum, eat, eucalyptus leaves*>, <*opossum, eat, leaves*>, etc.

FIGURE 4.7 – Procédé de génération de l'ensemble des relations possibles à partir de la relation initiale <*koala, eat, eucalyptus leaves*>. Les flèches en pointillés indiquent le prédicat *eat*.



Dans le cas d'une large taxonomie le nombre de relations à générer peut rendre la phase de génération fastidieuse. Une façon possible de contrôler cette étape est de réaliser au préalable une phase de propagation ascendante des nombres d'observations des déclarations au sein du graphe. L'idée est d'associer une valeur de croyance [5] à l'ensemble des concepts de l'ordre partiel sur les syntagmes afin de filtrer les concepts dont la croyance est différente de celles de leurs fils. Ainsi, ce mécanisme permet d'éviter de générer des déclarations non pertinentes. En effet, on conserve en priorité les déclarations spécifiques et celles apportant de la connaissance appuyée par des observations de plusieurs natures.

5. La notion de croyance est définie en section 4.2. Étant donnée une déclaration, elle correspond à son nombre d'observations additionné aux nombres d'observations des déclarations plus spécifiques.

4.1.3 Sémantique des déclarations

Il est important de souligner la sémantique associée aux déclarations extraites et générées. Tout d'abord, rappelons que cette sémantique diffère selon l'hypothèse de départ choisie par les auteurs d'un système. Cette hypothèse inscrit un système dans un monde soit fermé, soit ouvert. La signification derrière ces deux hypothèses diffère selon le domaine d'expertise dans lequel se situe l'approche. Par exemple, dans le domaine de la représentation des connaissances un système en monde fermé implique un espace conceptuel hermétique et strictement défini alors qu'un monde ouvert implique un espace modulable selon les contraintes établies initialement. Pour illustrer cet exemple, l'outil Protégé (GENNARI et al., 2003) conçu pour la construction de bases de connaissances supporte ces deux types de paradigme au travers des langages : OWL pour *Web Ontology Language* (MCGUINNESS et VAN HARMELEN, 2004) et le langage de Frames (MINSKY, 1974) aussi noté FRL (*Frame Representation Language*). OWL considère une hypothèse d'enrichissement en monde ouvert et FRL en monde fermé. De ce fait, FRL n'autorise aucune nouvelle entrée dans la base de connaissances tant que celle-ci n'a pas un endroit dédié dans le modèle initial et OWL autorise toute nouvelle information à moins que cette dernière viole une des contraintes établies (WANG et al., 2006). Par conséquent, dans FRL toutes les classes et individus doivent être énumérés manuellement et par avance alors qu'OWL permet la définition de nouvelles classes à partir de celles déjà établies (HORRIDGE et al., 2004).

Dans le cas de notre chaîne de traitement, le modèle d'extraction de relations (REVERB) se positionne en monde ouvert. Par conséquent, cette hypothèse est héritée pour l'ensemble des modules constituant la chaîne de traitement. Toutefois, la signification de cette hypothèse est légèrement différente par rapport à l'exemple précédent. Dans notre cas, on discute de la sémantique à attribuer aux relations extraites et générées. Cette sémantique peut être caractérisée par deux principales interprétations : l'existentielle et l'universelle. La première s'adapte plus facilement à un monde ouvert et la seconde à un monde fermé. En effet dans le cadre d'un monde ouvert, il est difficile de prétendre à une couverture universelle pour une déclaration donnée. Dans le meilleur des cas, on ne peut qu'obtenir une appréciation probabiliste d'une interprétation universelle *i.e.* l'estimation d'une probabilité définissant la fiabilité d'une déclaration générale. Ainsi au sein de notre chaîne de traitement, nous considérons une interprétation existentielle *e.g.* en sachant qu'un koala est un marsupial et que l'eucalyptus est une myrtacée alors la déclaration <*koala*, *mange*, *eucalyptus*> n'implique pas que tous les marsupiaux mangent des myrtacées mais qu'il existe au moins un marsupial mangeant au moins une myrtacée. Par conséquent, si M est l'ensemble des marsupiaux et E l'ensemble des myrtacées alors l'interprétation de la déclaration <*marsupial, mange, myrtacée*> issue du graphe des déclarations

est : $\exists m \in M$, $\exists e \in E$ pour lesquelles la déclaration mange(m, e) est vraie. Cette distinction a une importance cruciale pour la signification des relations et leur intégration par la suite dans un raisonnement.

Une fois le processus de génération achevé, l'ensemble des déclarations (extraites et générées) sont structurées au sein d'un graphe afin de calculer les critères permettant de discriminer la pertinence des déclarations lors de la phase de sélection. Les étapes de construction de ce graphe et les calculs des critères sont détaillés dans la section suivante.

4.2 Modalités d'acquisition des critères pour l'évaluation

Cette section est consacrée à la méthodologie servant de support aux calculs des critères associés aux déclarations, exploités lors de la phase de sélection. Les critères définis pour cette étude tiennent compte des propriétés taxonomiques des déclarations. Ainsi, nous proposons une procédure de hiérarchisation de l'ensemble des déclarations afin de calculer ces différents critères.

4.2.1 Construction d'une hiérarchie entre les déclarations

Cette structuration a pour objectif de hiérarchiser les relations en fonction d'une logique taxonomique entre les concepts explicités *i.e.* le sujet opossum subsume toujours le sujet koala pour un prédicat et un objet donnés. Pour cela, nous avons défini un ensemble de règles afin de diriger la manière de concevoir un tel graphe. Le tableau 4.2 expose l'ensemble des règles que nous avons pré-défini. Ces règles sont basées sur la relation taxonomique des sujets et des objets de chaque relation par rapport à l'ordre partiel établi précédemment sur les syntagmes. NOTE – on appelle s les déclarations en référence à la notation anglaise *statement*, $\mathcal{S}$ l'ensemble des déclarations observées et générées et $\mathcal{O}_{\mathcal{S}} = (\mathcal{S}, \preceq)$ l'ordre partiel sur les déclarations.

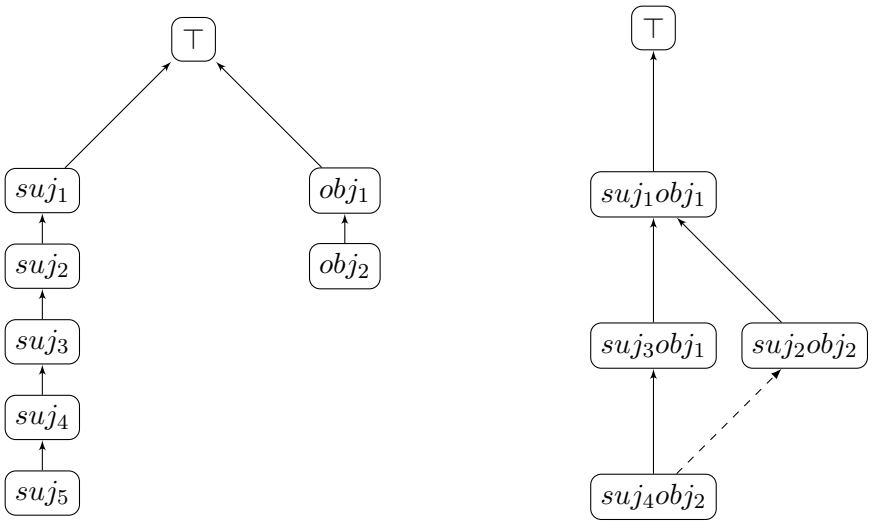
TABLEAU 4.2 – Règles de construction pour deux déclarations $s_1 =< a, p, b >$ et $s_2 =< c, p, d >$ avec a, b, c et d des syntagmes et p un même prédicat. Avec $s_2 \prec s_1$ la déclaration s_1 qui subsume la déclaration s_2, $s_1 \equiv s_2$ deux déclarations équivalentes et $s_1 ! s_2$ la déclaration s_1 non ordonnable avec s_2.

	$b \equiv d$	$b \prec d$	$d \prec b$	$\neg(b \preceq d) \wedge \neg(d \preceq b)$
$a \prec c$	$s_1 \prec s_2$	$s_1 \prec s_2$	$s_1 ! s_2$	$s_1 ! s_2$
$c \prec a$	$s_2 \prec s_1$	$s_1 ! s_2$	$s_2 \prec s_1$	$s_1 ! s_2$
$a \equiv c$	$s_1 \equiv s_2$	$s_1 \prec s_2$	$s_2 \prec s_1$	$s_1 ! s_2$
$\neg(a \preceq c) \wedge \neg(c \preceq a)$	$s_1 ! s_2$	$s_1 ! s_2$	$s_1 ! s_2$	$s_1 ! s_2$

Le graphe des déclarations est dirigé et acyclique, tel l'ordre partiel défini sur les syntagmes. Sa construction doit tenir compte de l'ensemble des relations possibles

entre les différentes déclarations. Ces possibilités incluent l'héritage multiple pour un même nœud *i.e.* au moins deux nœuds pères attribués pour un même nœud fils. Pour illustrer ce cas, prenons l'exemple de la figure 4.8. En considérant un prédicat donné, la déclaration suj_2obj_2 doit être positionnée au sein d'un graphe des déclarations dont la construction est réalisée à partir de l'ensemble des déclarations tirées de manière aléatoire. Ainsi, un algorithme de construction naïf est forcé de parcourir l'ensemble d'une branche (de la racine à une feuille donnée) pour capter l'ensemble des relations possibles par rapport à une déclaration. Par conséquent, l'algorithme ne peut pas s'arrêter au moment où il trouve, dans une branche donnée, un nœud non ordonnable avec la déclaration à placer *e.g.* suj_3obj_1 par rapport à suj_1obj_1.

FIGURE 4.8 – Les spécificités de la construction du graphe des déclarations. À gauche la structuration de l'ordre partiel construit à partir des syntagmes. À droite le graphe des déclarations. En considération d'une approche d'ajout aléatoire des nœuds la déclaration suj_2obj_2 sera embranchée uniquement à suj_1obj_1 alors qu'elle a également un lien de parenté avec la déclaration suj_4obj_2.



O_Σ : Ordre partiel sur les syntagmes

$O_\mathcal{S}$: Graphe des déclarations

Pour pallier cette contrainte de construction et pour éviter de parcourir l'ensemble d'une branche à la recherche des nœuds ordonnables, il est possible d'effectuer un tri au préalable sur les déclarations. Ce tri considère le nombre de descendants du sujet et de l'objet pour chaque déclaration au regard de l'ordre partiel sur les syntagmes et range ces valeurs de manière décroissante. Par exemple, sur la figure 4.8, la déclaration suj_1obj_1 a 5 descendants (suj_2, suj_3, suj_4, suj_5 et obj_2). Cette étape permet d'avoir l'assurance de placer les déclarations dans le graphe de la plus générale à la plus spécifique et donc d'obtenir une procédure *top-down*. Cela permet d'avoir l'assurance de générer toutes les relations possibles pour une déclaration donnée. Sur la figure 4.8 la considération d'un tri décroissant sur les déclarations en fonction du nombre de descendants permet de positionner suj_2obj_2 avant suj_4obj_2

qui ne possède qu'un seul descendant (suj_5) par rapport aux trois descendants de la déclaration suj_2obj_2 (suj_3, suj_4 et suj_5).

Cette phase de tri s'exécute sur les déclarations observées et générées à partir de l'ordre partiel sur les syntagmes, O_Σ. Une fois réalisé, l'ordre est utilisé en entrée de l'algorithme 1. Cet algorithme s'appuie sur l'ensemble de règles pré-définies exposé dans le tableau 4.2.

Algorithme 1 : Algorithme pour la construction du graphe des déclarations. La fonction $comparaison\_couple()$ s'appuie sur les règles exposées dans le tableau 4.2 et renvoie la hiérarchisation entre deux déclarations. $\mathcal{S}$ est l'ensemble des déclarations triées dans l'ordre décroissant par rapport au nombre de leurs descendants, O_Σ est la hiérarchisation entre les syntagmes et $O_\mathcal{S}$ est le graphe des déclarations.

Data : $\mathcal{S}$, O_Σ
Result : $\mathcal{O}_\mathcal{S}$
Initialisation du graphe $\mathcal{O}_\mathcal{S}$;
Ajout de l'élément Racine à $\mathcal{O}_\mathcal{S}$;
for *chaque déclaration* $f \in \mathcal{S}$ **do**
    Initialisation de la pile $queue$ avec l'élément Racine;
    coloration $\leftarrow \emptyset$;
    while *queue est non vide* **do**
        $match \leftarrow false$;
        $f' \leftarrow queue.pop()$;
        Ajout de f' à *coloration*;
        if f' *n'est pas une feuille dans* $\mathcal{O}_\mathcal{S}$ **then**
            for *chaque* $c_{f'} \in fils(f')$ **do**
                comparaison_couple(f, $c_{f'}$, O_Σ);
                if f *est plus spécifique à* $c_{f'}$ **then**
                    $match \leftarrow true$;
                    if $c_{f'} \notin$ *coloration* **then**
                        Ajout de $c_{f'}$ à *queue*;
        else
            comparaison_couple(f, f', O_Σ);
            if f *est plus spécifique à* f' **then**
                $match \leftarrow true$;
                Ajout de f aux fils de f' dans $\mathcal{O}_\mathcal{S}$;
        if *match = false* **then**
            Ajout de f aux fils de f' dans $\mathcal{O}_\mathcal{S}$;

La sous-section suivante présente les critères exploités lors de la phase de sélection des déclarations pertinentes.

4.2.2 Les critères d'évaluation des déclarations

Nous pouvons introduire cette section au travers de l'approche PRISMATIC développée par FAN et al., 2012. Cette dernière emploie des critères d'évaluation basés sur la redondance des données textuelles pour évaluer et inférer de la connaissance.

Elle a été implantée dans la méthode *Watson Jeopardy !* (FERRUCCI et al., 2010) afin de rechercher les types d'entités les plus probables pour constituer une réponse à une question donnée. Par exemple au regard de la redondance de données textuelles, l'approche induit que les *choses* qui sont généralement *annexées* sont typiquement des régions. Par conséquent, à partir de la phrase suivante : *In 1859, it were annexed by Napoleon*, l'approche peut conforter une réponse telle que *Piémont* puisque c'est une région. Pour réaliser un tel processus, l'approche se décompose en deux principales phases, l'une portant sur l'extraction de relations textuelles et l'autre sur l'évaluation et la généralisation de ces dernières. La première phase exploite une batterie de méthodes du TAL : dépendances syntaxiques, module de reconnaissance des entités nommées (NER) et détection des co-réferences ainsi qu'un ensemble de motifs syntaxiques d'intérêt pour filtrer les relations. La seconde phase, quant à elle, permet d'estimer la valeur de pertinence de ces relations et induit des règles plus générales. Pour cela, les auteurs emploient différentes stratégies reposant sur les fréquences d'apparition. Par exemple, elle évalue la pertinence d'une relation au travers de son pmi normalisé (*Pointwise Mutual Information*) permettant de mesurer l'association entre deux éléments. Par exemple, les auteurs peuvent mesurer le degré d'association entre l'objet et le sujet/prédicat d'une relation. Le fait que le pmi soit normalisé encadre les valeurs possibles dans $[-1, 1]$ avec -1 signifiant aucune association et 1 une association complète. Par exemple, si l'objet correspond à *award*, le sujet/prédicat à *Einstein win* et que $npmi = 0,7$ alors il existe une forte co-occurrence entre ces deux parties. Cette méthode d'évaluation de la pertinence d'une relation en exploitant la redondance des données est intéressante dans le cadre d'un ensemble restreint de types d'entités. Toutefois, il serait intéressant d'observer l'impact d'une plus grande diversité de types dans les performances de *Watson Jeopardy ! e.g. Region* serait alors segmentée en *City* et *Country*. Les conséquences attendues seraient d'améliorer la précision en diminuant l'espace de recherche de la méthode. Cependant, le rappel pourrait être négativement affecté par cette plus grande diversité. Dans ce cas pour parer cette diminution de rappel, le procédé d'inférence pourrait exploiter un ordre partiel spécifique sur les types d'entités. Par exemple, si le NER distingue un type d'entité bien précis tel que *City* mais que la réponse ne coïncide pas, l'ordre partiel permettrait de remonter vers une entité plus générale associant d'autres extractions *e.g.* en utilisant la taxonomie de WordNet : *city* $\rightarrow$ *municipality* $\rightarrow$ *region*. Le degré de pertinence d'une réponse pourrait diminuer lorsque l'on remonte une branche.

Ce mécanisme de propagation *bottom-up* permettant de calculer la pertinence des déclarations est l'idée principale qui a guidé la conception du graphe des déclarations. Les deux prochaines parties détaillent les critères employés pour distinguer les déclarations et la façon de les obtenir à partir de ce graphe.

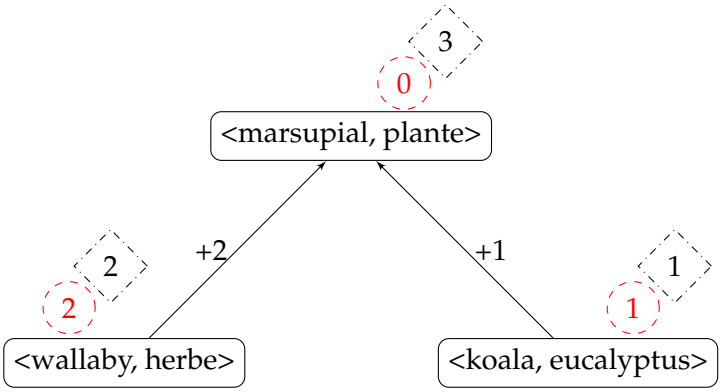
La croyance des déclarations

Le premier critère défini pour caractériser les déclarations au sein des modèles de sélection est la croyance accordée aux déclarations. Ce critère se rapproche du concept de fiabilité présenté dans VIVIANI et PASI, 2017. Dans le cas de notre étude, la croyance se base sur le nombre d'observations d'une déclaration que l'on augmente de la somme des observations des déclarations qui lui sont plus spécifiques . À partir du graphe des déclarations, cette valeur est, par conséquent, calculée au travers d'un processus de propagation *bottom-up* des observations. Ainsi, la croyance (bel) d'une déclaration s est définie comme la somme des observations de l'ensemble des descendants $\mathcal{D}$ de s avec $\mathcal{D}(s) = \{s' \in \mathcal{S} \mid s' \preceq s\}$ où $\mathcal{S}$ représente l'ensemble des déclarations. La formule 4.8 illustre la croyance d'une déclaration, avec $f(s)$ le nombre d'observations de la déclaration s.

$$bel(s) = \sum_{s' \in \mathcal{D}(s)} f(s') \tag{4.8}$$

La figure 4.9 présente un exemple dans lequel la relation <marsupial, mange, plante> n'a jamais été observée dans les textes mais pour laquelle ses relations filles lui permettent d'obtenir une croyance non nulle.

FIGURE 4.9 – Propagation *bottom-up* des observations associées à chaque relation pour le prédicat "manger". Les ronds en pointillés représentent le nombre d'observations de chaque déclaration s. Les losanges représentent la croyance bel.



Concernant l'incertitude linguistique, elle intervient lors de l'étape de propagation. Le but est de distinguer le nombre d'observations d'une déclaration s certaine $f(s)$ et incertaine $f_{inc}(s)$. Pour cela, le poids d'une observation associée à une relation incertaine peut être modifié. En effet, le poids d'une observation modélise le crédit que l'on accorde à cette déclaration. Ainsi, en diminuant cette valeur nous pouvons influencer à la fois la croyance de la déclaration et son impact dans la phase de propagation (cf. équation 4.9).

$$bel(s) = \sum_{s' \in \mathcal{D}(s)} f(s') + f_{inc}(s') \times \delta \tag{4.9}$$

où δ est un paramètre à fixer dans $[0, 1]$.

Un parallèle peut être réalisé entre notre valeur de croyance et le cadre théorique des fonctions de croyance (SHAFER, 1976). Cette analogie est décrite dans l'annexe B.

La spécificité

Le critère de spécificité correspond à la profondeur de la déclaration dans le graphe des déclarations [6] qui est d'autant plus importante si la déclaration a été désambiguïsée. En effet dans le cadre de la construction du graphe des syntagmes (guidant la construction du graphe des déclarations), le sujet et l'objet de la déclaration héritent des ascendants des concepts désambiguïsés issus de la taxonomie de référence. Par conséquent lors de l'étape de génération, des relations plus génériques vont être inférées et ces dernières constitueront des ascendants de la relation extraite dans le graphe des déclarations. Cette caractéristique est appréhendée comme une valeur reflétant le contenu informatif (IC) de la déclaration (RESNIK, 1999 ; BATET et al., 2014). Ainsi, nous pouvons envisager pour de futurs travaux différentes approches topologiques plus complexes que la simple profondeur.

Maintenant que nous avons défini les modalités d'acquisition des critères pour évaluer les déclarations, il s'agit de les exploiter au sein de modèles afin d'évaluer la pertinence des déclarations et les filtrer. La section suivante présente les modèles établis pour réaliser cette étape d'évaluation.

4.3 Évaluation de la pertinence des connaissances

Cette section aborde les modalités de sélection des déclarations au sein de la chaîne de traitement au regard de leur pertinence. Nous verrons au travers de deux principales approches que la définition de la pertinence d'une relation peut être abordée sous différents aspects. La première est basée sur des modèles de sélection que l'on a défini et évalué (cf. chapitre 5) et la seconde sur des profils utilisateurs permettant de s'adapter aux attentes et niveaux de connaissances des utilisateurs.

L'évaluation de la pertinence des connaissances est à nos yeux une phase indispensable de la chaîne de traitement qui est notamment destinée à être exécutée sur des données issues du Web. Ces données sont à l'origine de nombreuses décisions mais

6. Une fermeture transitive est appliquée aux graphes des déclarations.

peuvent être dans certains cas non vérifiées et subjectives. Nous commençons à observer dans la littérature des travaux s'intéressant à mesurer la qualité de ces contenus Web incluant les réseaux sociaux à l'instar des travaux de LEX et al., 2012. Ces derniers proposent une méthode pour évaluer un document Web basée sur la fréquence des faits qu'il contient. Ils démontrent ainsi que la qualité informative d'un document est corrélée à sa densité de faits (méthode validée sur Wikipedia). Toutefois, ces travaux ne s'intéressent pas à la croyance associée aux déclarations s'apparentant à la notion de fiabilité des faits exploitée dans les communautés étudiant l'incertitude des sources (DONG et al., 2015).

Dans notre cas, l'évaluation de la pertinence a pour objectif de filtrer les déclarations à retourner à l'utilisateur au travers de modèles de sélection. Ces modèles s'appuient sur deux principaux critères : la croyance accordée à une déclaration et sa spécificité. La notion de croyance associée à une déclaration prend en compte l'incertitude linguistique (cf. sous-section 4.2.2). La spécificité est vue comme un critère plus subjectif puisque c'est un critère dont la satisfaction différera d'un utilisateur à un autre (niveau de connaissances variables).

4.3.1 Les modèles de sélection

Dans le but d'évaluer la pertinence des relations extraites et générées, quatre modèles de sélection ont été élaborés et évalués. Ces modèles considèrent les critères cités précédemment : croyance et spécificité obtenues à partir du graphe des déclarations. Ainsi, nous définissons $\mathcal{S}_{\mathcal{M}}$ l'ensemble des déclarations pertinentes par rapport à un modèle $\mathcal{M}$ donné.

Le premier modèle d'inférence $\mathcal{M}_1$ est basé sur un seuil d'acceptation minimal α centré sur les croyances des déclarations (cf. équation 4.10). Ce modèle sert d'étalon afin d'observer l'impact de la croyance sans tenir compte de la spécificité des déclarations.

$$\mathcal{S}_{\mathcal{M}_1} = \{s \in \mathcal{S} \mid bel(s) > \alpha\} \tag{4.10}$$

Toutefois, le problème de cette approche est de conserver principalement des déclarations faiblement profondes dans le graphe qui sont, par conséquent, trop abstraites pour être exploitées.

Il est possible de raffiner ce premier modèle en conditionnant le seuil α par la profondeur maximale des déclarations dans le graphe $depth(s)$. Le second modèle $\mathcal{M}_2$ considère la moyenne des croyances associées aux déclarations à chaque niveau de profondeur x : $\overline{bel}_x$ (cf. équation 4.11).

$$\overline{bel}_x = \frac{\sum_{\{s \in \mathcal{S} | depth(s)=x\}} bel(s)}{\mid \{s \in \mathcal{S} \mid depth(s) = x\} \mid} \tag{4.11}$$

Ainsi, les déclarations les plus pertinentes $\mathcal{S}_{\mathcal{M}_2}$ correspondent à l'ensemble des déclarations ayant une croyance supérieure à la moyenne des croyances à leur niveau de profondeur : $\mathcal{S}_{\mathcal{M}_2} = \{s \in \mathcal{S} \mid bel(s) \geq \overline{bel}_{depth(s)}\}$.

Le troisième modèle $\mathcal{M}_3$ est identique au second à ceci près qu'il exploite la médiane au lieu de considérer la moyenne des croyances.

Enfin, en ce qui concerne le quatrième modèle $\mathcal{M}_4$, il se base sur l'idée suivante. Une déclaration pertinente possède une forte croyance ou, dans le cas inverse, a au moins un parent avec un nombre d'observations non nul augmentant ainsi sa vraisemblance. Ainsi, ce modèle conserve les déclarations selon deux alternatives. La première sélectionne les déclarations avec une croyance supérieure ou égale au 75^e centile des croyances des déclarations à la profondeur x ($\mathcal{Q}_3(x)$). La seconde filtre les déclarations avec une croyance supérieure au 25^e centile des croyances des déclarations à la profondeur x ($\mathcal{Q}_1(x)$) qui ont au moins un parent (ascendant direct) ayant été observé (cf. équation 4.12). Cette deuxième option permet d'augmenter le rappel potentiel de l'approche.

$$\begin{aligned}\mathcal{S}_{\mathcal{M}_4} = \{s \in \mathcal{S} \mid bel(s) \geq \mathcal{Q}_3(depth(s)) \vee \\ (bel(s) \geq \mathcal{Q}_1(depth(s)) \wedge (\exists s_{prt} \in parent(s), f(s_{prt}) > 0))\}\end{aligned} \tag{4.12}$$

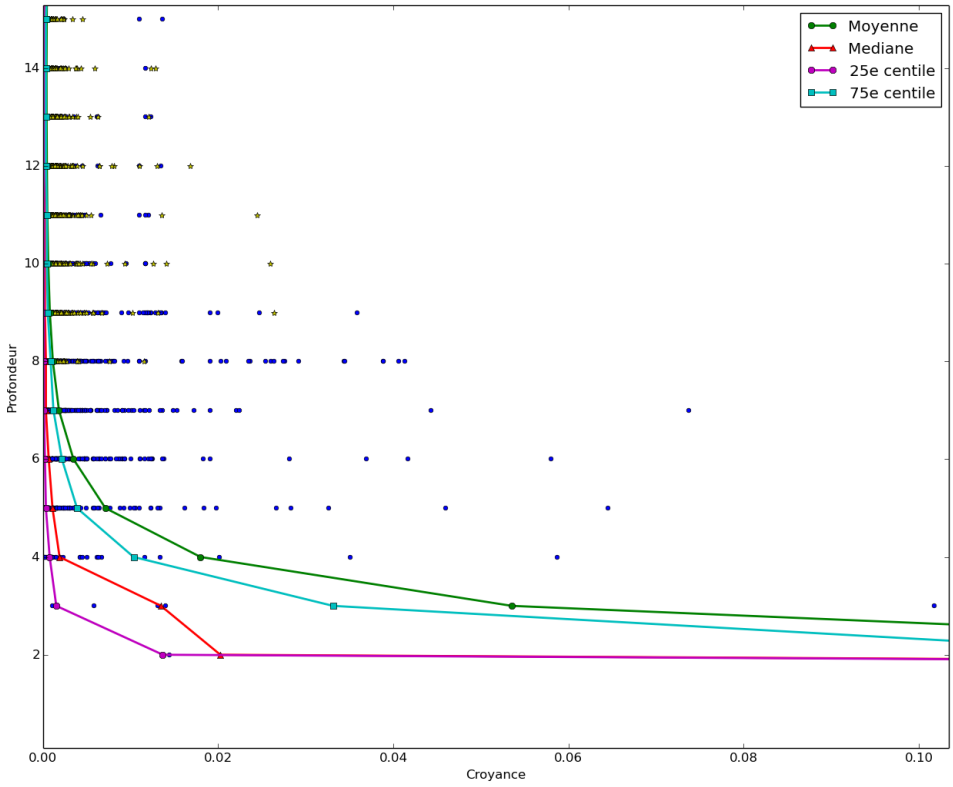
La figure 4.10 présente le comportement de la moyenne, de la médiane et des centiles en fonction des différentes profondeurs dans le graphe des déclarations.

Outre les modèles de sélection, la pertinence d'une déclaration peut également être perçue comme un paramètre ajustable en fonction des besoins et des connaissances d'un utilisateur. Privilégie-t-il les faits plus génériques ou au contraire les faits plus spécifiques. Par exemple dans le cas d'un ensemble de textes avec un vocabulaire spécialisé, un néophyte du domaine va préférer des faits plus génériques afin de faciliter sa compréhension *e.g.* au lieu de lymphome une personne n'ayant pas un vocabulaire médical préférera le concept plus générique de cancer. Par conséquent, la pertinence d'une déclaration peut être fortement dépendante d'un profil utilisateur. La sous-section suivante détaille ce cas de figure.

4.3.2 Définition de profils utilisateurs

L'évaluation de la pertinence des relations est une notion complexe à appréhender en fonction des critères à notre disposition (croyance, spécificité) et des attentes

FIGURE 4.10 – Distribution des modèles exploités : moyenne, médiane, 25^e centile et 75^e centile. Les étoiles jaunes sont les déclarations extraites et les points bleus les déclarations générées. Dans un souci de clarté, une déclaration générée sur dix est affichée et un agrandissement sur la profondeur et la croyance est réalisé.



des utilisateurs. En effet, chaque utilisateur possède un niveau de compréhension du domaine et des attentes qui lui sont propres. Pour illustrer ces propos, prenons l'exemple de la figure 4.11. La phrase utilisée, tirée de Wikipedia, permet l'extraction d'une relation et la génération de relations implicites. Nous pouvons aisément comprendre qu'un expert favorise l'information qui a été extraite puisqu'elle emploie un vocabulaire précis. Tandis qu'une personne étrangère à ce vocabulaire va sûrement préférer une des relations générées. Ce cas illustre l'intérêt de distinguer les résultats en fonction du niveau de compréhension des utilisateurs.

La compréhension est une chose mais l'attente d'un utilisateur en est une autre. En effet, nous pouvons orienter notre approche selon différentes utilisations : découverte de connaissances ou recherche d'information *e.g.* dans l'objectif de l'enrichissement automatique des bases de connaissances. Ces deux utilisations accordent à la valeur de croyance une importance particulière. En effet, au sein d'une thématique de découverte de connaissances, un utilisateur sera focalisé plus spécifiquement vers des relations issues de l'agrégation de signaux faibles tenant compte de l'incertitude linguistique, tandis qu'en considérant une thématique d'enrichissement d'une base

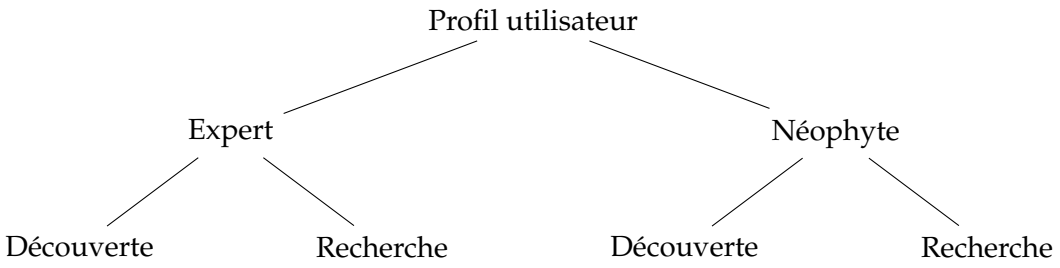
FIGURE 4.11 – Simulation d'une extraction et d'une génération de relation à partir d'une phrase tirée de Wikipedia. Un expert en médecine sera plus fortement intéressé par la relation extraite tandis qu'un néophyte sera plus à l'aise avec le vocabulaire employé par les relations générées. On rappelle ici que la signification des relations générées est de type *existentiel*.

Phrase	La kératose solaire favorise la survenue de mélanomes.
Relation extraite	<kératose solaire, favorise, mélanome>
Relations générées (Possibles)	<kératose solaire, favorise, cancer de la peau> <kératose solaire, favorise, cancer> <lésion cutanée, favorise, mélanome> <lésion cutanée, favorise, cancer de la peau> <lésion cutanée, favorise, cancer>

de connaissances, la méthode doit s'employer à exploiter des valeurs de croyance élevés pour conserver l'intégrité de la base.

Par conséquent, pour la définition de profils utilisateurs, nous distinguons à la fois le niveau d'expertise dans un domaine donné et le but de la recherche d'un utilisateur (cf. figure 4.12).

FIGURE 4.12 – Classification des profils utilisateurs en fonction d'un domaine de connaissance donné. Nous avons des profils experts et néophytes d'un domaine et sous chacune de ces catégories sont représentés leurs objectifs : découverte de connaissances ou recherche d'information.



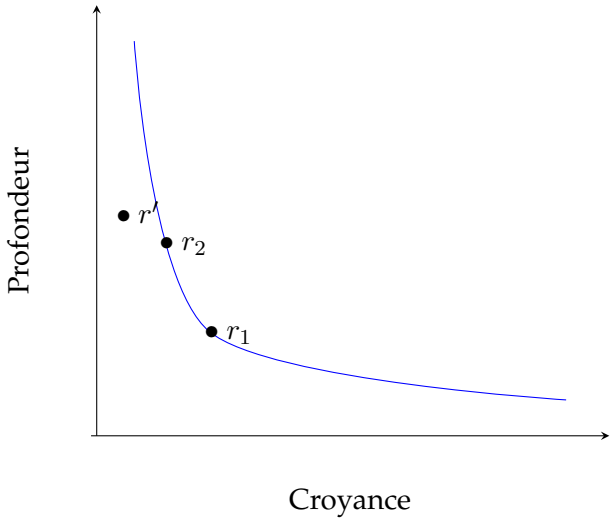
Au travers de ces profils utilisateurs, nous distinguons l'apparition d'un front de Pareto à deux dimensions sur nos données. Par conséquent, les modèles que nous pouvons définir en considération d'un profil utilisateur auraient pour objectif d'exploiter les solutions optimales au sens de Pareto. L'équation 4.13 expose la manière d'obtenir les solutions optimales au sens de Pareto r_{opt} en considérant l'ensemble des solutions possibles $\mathcal{R}$.

$$r_{opt} = r \in \mathcal{R} \text{ tel que } \neg\exists r' \in \mathcal{R} \setminus \{r\} \text{ avec } bel(r') > bel(r) \wedge depth(r') > depth(r) \tag{4.13}$$

La figure 4.13 est une représentation des alternatives selon les critères de spécificité et de croyance associés aux relations. Les alternatives r_1 et r_2 sont considérées optimales au sens de Pareto selon l'équation 4.13. Plusieurs propriétés sont définies [7] :

— Les *optima* au sens de Pareto sont incomparables entre eux.
— Il n'existe pas une alternative qui soit la meilleure.
— Une alternative non optimale (r' sur la figure 4.13) n'est pas forcément "moins bonne" qu'une alternative optimale. Elle ne peut cependant être comparée avec l'alternative optimale r_2.

FIGURE 4.13 – Représentation des états réalisables optimaux au sens de Pareto à partir de la profondeur et de la spécificité des déclarations. r_1 et r_2 sont deux états optimaux situés à la frontières des utilités.



Ces travaux concernant d'éventuels profils utilisateurs n'ont pas encore été menés ou validés. Malgré cela, cette sous-section est importante car elle permet de démontrer l'importance et la complexité associée aux notions de pertinence et de sélection.

4.4 Synthèse et implémentation de la chaîne de traitement

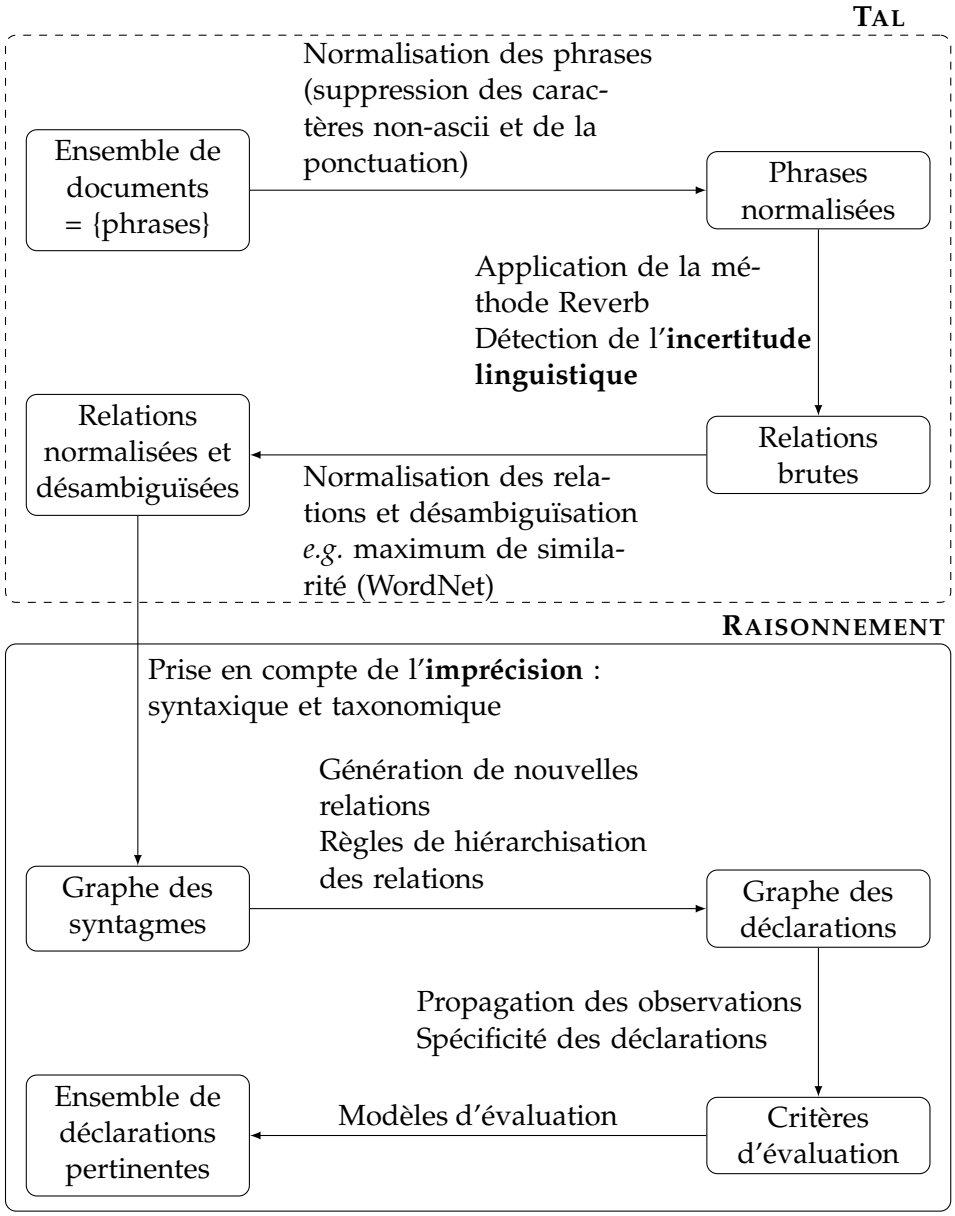
Cette section récapitule l'ensemble des étapes effectuées par la chaîne de traitement et propose différentes visualisations issues d'une implémentation réalisée au travers d'une interface graphique.

4.4.1 Récapitulatif de la chaîne de traitement

La figure 4.14 reprend et complète celle proposée au chapitre 2 (cf. figure 2.13) correspondant aux différents traitements réalisés sur les textes et les relations

7. http ://stockage.univ-brest.fr/~fdupont/deug_mass/cours2annee/files/chapitre5.pdf

FIGURE 4.14 – Récapitulation de la chaîne de traitement au regard des modules d'extraction d'information et de raisonnement.



Cette figure 4.14 récapitule l'ensemble des tâches accomplies à partir de textes non structurés pour l'extraction de relation, l'inférence et la sélection de nouvelles connaissances. L'ajout du bloc *raisonnement* résume les différentes opérations permettant la mise en place de ce processus. Ce dernier bloc considère en entrée l'ensemble des relations normalisées et désambiguïsées issues de la partie TAL et une structure taxonomique dans le cas où l'utilisateur désire augmenter les connaissances qui seront inférées. Le module de raisonnement débute avec la construction du graphe des syntagmes permettant de décomposer et structurer les relations extraites. Dans le cas où l'utilisateur apporte une structure taxonomique, ce dernier graphe est enrichi par les ascendants, issus de la taxonomie, des entités désambiguïsées au sein du graphe des syntagmes. S'ensuit la phase de génération des déclarations par le produit cartésien de l'ensemble des ascendants du sujet et de l'objet de chaque relation. Une fois réalisé, un ensemble de règles de construction s'appuyant sur le graphe des syntagmes a été proposé pour hiérarchiser l'ensemble des relations extraites et générées. Cette étape aboutit à la formation du graphe des déclarations servant de support au processus de sélection. En effet, la structuration de ce dernier graphe permet de calculer les différents critères pour évaluer la pertinence de chacune des relations. Ces critères sont la croyance, obtenue à partir d'un principe de propagation des observations et la spécificité (profondeur du fait dans le graphe).

4.4.2 Implémentation

Cette implémentation est réalisée au travers d'une interface graphique Web permettant à un utilisateur de saisir une requête constituée d'au moins un sujet ou un objet avec un prédicat. Cette dernière a été implémentée en python et exploite l'interface CGI (*Common Gateway Interface*) à l'aide des bibliothèques Python *BaseHTTPServer* et *CGIHTTPServer*. Les données exploitées par cette implémentation proviennent de l'extraction de relations réalisée avec l'outil Reverb sur le corpus *ClueWeb09* mentionné au chapitre 2 section 2.3.2. Cette implémentation exploite directement le résultat de la phase de normalisation des relations explicitée dans le chapitre précédemment cité. Ces relations sont conservées au sein de fichiers indexés par leur prédicat pour faciliter la phase de recherche (la mise en place d'une base de données serait également envisageable). Ainsi, l'interface graphique se focalise sur la gestion de la requête, la procédure d'inférence et de sélection.

Le tableau 4.3 expose les différentes visualisations présentées à l'utilisateur. La première image expose l'interface de requête. Cette dernière possède l'option *désambiguïsation* laissant le choix d'exploiter ou non une structure taxonomique impliquant la manipulation des entités désambiguïsées. L'implémentation actuelle exploite la taxonomie de WordNet 3.1 via la bibliothèque Python nltk [8]. Si l'utilisateur coche

8. http ://www.nltk.org/howto/wordnet.html

l'option, la seconde phase est la désambiguïsation du sujet et/ou de l'objet renseignés en paramètre. Cette phase est réalisée au travers de la bibliothèque *pywsd* [9]. Une fois les ambiguïtés résolues, l'application recherche les relations en adéquation avec la requête, construit le graphe des syntagmes, génère les nouvelles relations et construit le graphe des déclarations. La phase de recherche des relations, si la désambiguïsation est active, récolte l'ensemble des descendants de la ou des entités de la requête pour capter l'ensemble des supports possibles *e.g.* si la requête est *Que mangent les animaux ?* (*What animals eat ?*) correspondant au sujet *animal* et au prédicat *eat*, la phase de recherche collecte tous les animaux (entités qui ont été désambiguïsées *e.g. lion, dog,* etc.) dont l'entité *animal* elle-même, dans le fichier correspondant au prédicat *eat*. Une fois le graphe des déclarations construit, l'utilisateur peut paramétrer la phase d'inférence (cf. image 4). Les paramètres comprennent le poids à associer aux relations incertaines et le choix d'utiliser un modèle de sélection ou non. L'utilisateur a le choix entre trois modèles présentés dans la sous-section 4.3.1. Ensuite, les critères d'évaluation sont calculés pour chaque relation et si un modèle de sélection a été choisi, il est appliqué. Concernant la visualisation des résultats (cf. images 5 et 6), ils sont triés par ordre décroissant sur la somme des critères associée aux relations *i.e.* on affiche les relations fortement soutenues et spécifiques en premier. L'interface de visualisation est dotée de deux curseurs permettant à l'utilisateur de filtrer ses résultats en considération de la valeur de croyance et de spécificité des déclarations. Enfin, chaque résultat possède deux options, l'une pour détailler toutes les relations inférées à partir de cette relation (*see Also*) et l'autre pour afficher les phrases supportant cette relation en précisant s'il existe ou non un marqueur d'incertitude (*Support*).

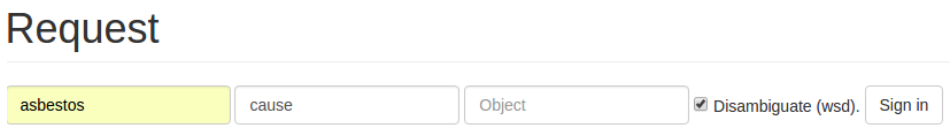
Le code est disponible à l'adresse suivante : *https ://github.com/PAJEAN/OKE* (OKE pour *Open Knowledge Extraction*). Il est fonctionnel mais pour des questions de volume, les données issues de *ClueWeb09* ont été remplacées par des données "jouets". Libre ensuite à l'utilisateur de renseigner ses propres relations et phrases ou de télécharger le corpus de relations [10].

Une autre version de la chaîne de traitement a été produite en tenant compte cette fois-ci du bloc TAL (cf. figure 4.14). Cette version prend en entrée un ensemble de phrases et à partir de Reverb extrait les relations et les normalise pour servir d'entrée à l'actuelle implémentation. Cette version est distribuée à l'adresse suivante : *https ://github.com/PAJEAN/KB_build*.

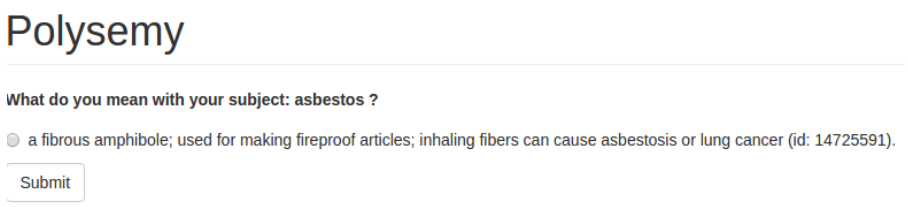
9. https ://github.com/alvations/pywsd
10. http ://reverb.cs.washington.edu/

TABLEAU 4.3 – Visualisation de l'interface graphique de la chaîne de traitement : 1. interface de requête, 2. phase de désambiguïsation, 3. recherche des relations extraites correspondantes, construction du graphe des syntagmes, génération des déclarations et construction du graphe des déclarations, 4. choix du modèle de sélection et de la valeur attribuée aux relations incertaines, 5. visualisation des résultats et ajustement possible de la valeur des critères (croyance, spécificité), 6. visualisation détaillée des résultats.

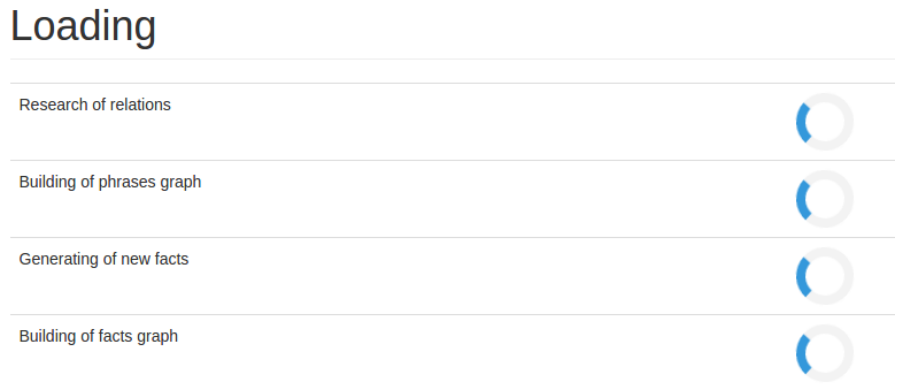
1.



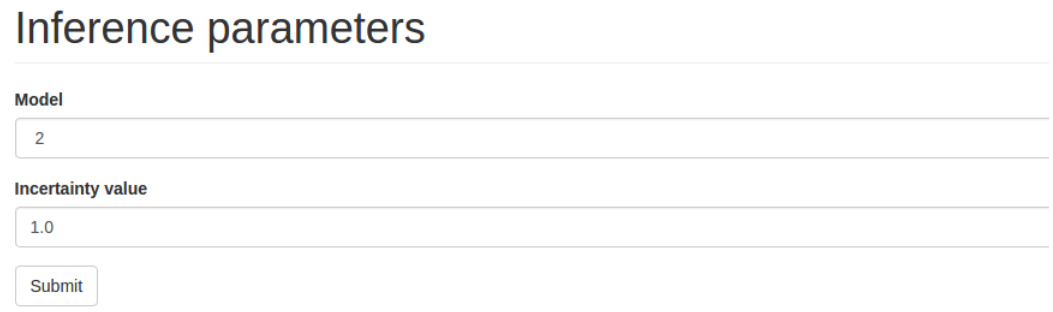
2.



3.



4.



5.

BELIEF	DEPTH	SUBJECT	OBJECT
63.0 See Also Support	22	asbestos	lung cancer
66.0 See Also Support	21	asbestos	carcinoma
69.0 See Also Support	17	asbestos	disease

6.

BELIEF	DEPTH	SUBJECT	OBJECT
63.0	22	asbestos	lung cancer
Hide (see Also)			
		amphibole	lung cancer
		mineral	lung cancer
		material	lung cancer
		substance	lung cancer
		matter	lung cancer
		physical entity	lung cancer
		entity	lung cancer

Hide (support)

ID	Seen	Certainty	Sentences
1626732	35	uncertain	Asbestos can cause lung cancer
1626439	2	certain	Asbestos also causes lung cancer
1626678	26	uncertain	Asbestos can also cause lung cancer

Ce chapitre récapitule les traitements réalisés sur les relations extraites et le module de raisonnement établi pour inférer et évaluer de nouvelles connaissances à partir de ces extractions. Le processus d'inférence est abordé comme la génération de nouvelles déclarations par la généralisation des observations. Lors de ce processus la considération de : la décomposition syntaxique des syntagmes extraits, d'un ordre taxonomique permettant l'abstraction des entités, et de l'incertitude linguistique, correspondent aux aspects différenciants de notre chaîne de traitement au regard des méthodologies abordant la notion de découverte de connaissances. Ces différents aspects sont rendus possibles par la hiérarchisation des syntagmes d'une part et des déclarations d'autre part. L'intégralité de cette chaîne de traitement a été implémentée et partagée au travers d'une interface graphique permettant à un utilisateur une meilleure lisibilité et interprétation des résultats.

Le chapitre suivant présente la phase de validation de la chaîne de traitement au travers d'une application de type questions-réponses exploitant une méthodologie de génération automatique de QCM. Une suite d'expérimentations est réalisée afin de montrer l'apport d'un module d'inférence à une approche d'extraction de relations, de comparer les différents modèles de sélection élaborés et d'observer l'impact de la prise en compte de l'incertitude linguistique.

Chapitre 5

Validation et discussion

Sommaire

La validation est une étape cruciale, particulièrement dans notre domaine de recherche où elle témoigne du passage à l'échelle et de la pertinence de l'approche. Elle permet d'analyser l'apport de l'approche proposée par rapport à un existant ou à défaut à une *baseline* de référence. La chaîne de traitement développée dans cette thèse a pour application principale la découverte de connaissances et par extension la découverte d'hypothèses de travail. Ces hypothèses sont formulées sous la forme de relations provenant du processus d'inférence et de sélection. Toutefois, il n'est pas évident d'évaluer l'approche dans un tel processus. Une façon de s'en rapprocher serait de suivre le protocole de validation de LIEKENS et al., 2011. Ces derniers proposent de retrouver de manière rétrospective certaines découvertes dans le domaine bio-médical à partir d'un ensemble de relations contenues dans des bases de données. Ainsi, ce protocole expérimental pourrait être appliqué pour la validation de notre approche. Pour cela, nous pourrions analyser des corpus de textes à un temps t donné, formuler un ensemble d'hypothèses et observer si au temps $t + n$ la chaîne de traitement arrive à confirmer ou infirmer ces hypothèses à partir d'un autre ensemble de textes. Cependant, un tel protocole de validation est difficile à mettre en place. En effet, la constitution d'un jeu de données est une étape chronophage, de même pour la réflexion que l'on doit mener sur la formulation des hypothèses, surtout en domaine spécialisé. Ces difficultés sont soulignées dans l'article de VIVIANI et PASI, 2017, dans lequel les auteurs mettent en avant divers problèmes liés à des domaines analogues à ceux abordés dans ce manuscrit. Ils évoquent notamment :

l'absence de *benchmarks* prédéfinis et de corpus *gold standard* et les difficultés à collecter et extraire d'importants volumes de données.

Par conséquent, nous avons opté pour un autre protocole d'évaluation. En effet, la découverte de connaissances peut également être exploitée dans le cadre d'un système de questions-réponses. C'est ce que nous avons choisi ici. Ce domaine a été abordé à partir d'un ensemble de questionnaires générés de manière automatique, composés de réponses correctes et incorrectes. Ces questionnaires ont pour avantage de proposer des questions dont la réponse peut être implicitement ou explicitement émise dans les textes et des questions servant de distracteur [1] permettant d'évaluer l'efficacité du modèle de raisonnement.

La première section est consacrée au protocole expérimental établi pour la validation. Le choix du type d'évaluation est argumenté et comparé par rapport à l'utilisation des *benchmarks* plus classiques dans le domaine du questions-réponses. De plus, elle détaille la procédure de génération d'un questionnaire et les données extraites pour tenter d'y répondre. La deuxième section, quant à elle, présente les résultats obtenus par la chaîne de traitement sur les questionnaires générés. Enfin, la dernière section discute de ces résultats et de la place de l'incertitude linguistique dans le processus d'inférence de notre méthode.

5.1 La recherche d'information de type questions-réponses

5.1.1 Introduction au domaine

La recherche d'information de type questions-réponses est un domaine riche et actif. Il correspond au fait de devoir trouver une réponse précise à une requête énoncée en langage naturel par un utilisateur en devant la rechercher au sein d'une large collection de documents ou sur le Web (BELLOT et al., 2014). Ces documents incluent également les bases de faits (ZELLE et MOONEY, 1996). Ce domaine implique la mise en place d'une architecture complexe permettant d'analyser les requêtes et de gérer les documents. En effet, les requêtes nécessitent généralement des méthodes appartenant au traitement automatique des langues et des approches d'enrichissement sémantique lors de leur prise en compte (FRANK et al., 2007). Tandis que les documents, lorsqu'ils sont textuels, sont généralement exploités au travers des méthodes de recherche d'information incluant l'indexation et la récupération des documents contenant potentiellement la réponse adéquate en fonction d'une requête donnée (ROBERTSON et KAREN SPÄRCK, 1976). En outre, il existe des méthodes hybrides mixant les documents textuels et les bases de faits telles que le système d'IBM, Watson (KALYANPUR et al., 2012). Cette méthode extrait une large variété d'informations

1. Question composée d'une relation incorrecte.

à partir des requêtes (entités nommées, relations, information ontologique) pour ensuite tenter de trouver des réponses candidates dans les bases de connaissances et les sources textuelles du type Wikipedia et journaux. Ces réponses candidates sont ensuite évaluées en fonction de diverses caractéristiques dont : la corrélation des candidats avec les termes de la requête, la fiabilité de la source, sa popularité et des propriétés spatiales et temporelles selon les types d'entités et prédicats utilisés dans la requête. Les relations de subsomption à partir des taxonomies sont également exploitées afin d'explorer les descendants pour un concept donné (FERRUCCI et al., 2010).

Le domaine du questions-réponses propose de nombreux *benchmarks*. Ces derniers se divisent selon trois types de questions (cf. tableau 5.1).

TABLEAU 5.1 – Ces exemples sont tirés de la tâche 1b BioAsQ de 2015 (TSATSARONIS et al., 2015). Cette tâche demandait également des réponses sous la forme de paragraphe permettant de résumer une réponse.

Type	Question	Réponse attendue
Oui/Non	Est ce que miR-21 est lié à la carcinogénèse ?	Oui
Factuel	Quelle est la maladie la plus commune attribuée à l'absence de cils cellulaires ?	Maladie rénale polykystique
Liste	Quelles sont les gènes humains les plus communs liés à la craniosynostose ?	[MSX2, RECQL4, SOX6, FGFR1, FGFR2, FGFR]

Cependant, les corpus proposés ne sont généralement pas adaptés pour démontrer l'apport de notre contribution par rapport aux méthodes d'extraction de relations en domaine ouvert. Un exemple illustrant ce propos concerne les questions factuelles provenant du corpus de questions-réponses SQuAD de Stanford (RAJPURKAR et al., 2016). Ce dernier met à disposition de la communauté des sous-ensembles de textes accompagnés d'un ensemble de questions et de réponses à propos de ces extraits. Les meilleures approches actuelles ont une F-mesure de 0,84 sachant que la F-mesure d'annotateurs humains est de 0,91 [2]. Comme nous pouvons le voir sur la figure 5.1, les réponses demandées sont contenues dans les textes et n'ont pas besoin d'un processus de raisonnement pour être élucidées. Ainsi, une méthode d'extraction de relations simple, sans processus de raisonnement est suffisante pour ce genre d'exercice. De plus, notre approche n'est pas initialement pensée pour le domaine du questions-réponses et par conséquent ne possède pas certains modules essentiels tels que l'analyse, le traitement et l'enrichissement des requêtes exprimées en langage naturel (ALMASRI, BERRUT et CHEVALLET, 2013).

2. https ://rajpurkar.github.io/SQuAD-explorer/

FIGURE 5.1 – Exemple du corpus de questions-réponses SQuAD de Stanford.

> *With advances in medicinal chemistry, most modern antibacterials are* ***semisynthetic modifications*** *of various natural compounds. These include, for example, the* ***beta-lactam antibiotics****, which include the penicillins (produced by* ***fungi*** *in the genus Penicillium), the cephalosporins, and the carbapenems. Compounds that are still isolated from living organisms are the aminoglycosides, whereas other antibacterials, for example, the sulfonamides, the quinolones, and the oxazolidinones, are produced solely by chemical synthesis. Many antibacterial compounds are relatively small molecules with a molecular weight of less than 2000 atomic mass units.*

Questions	Réponses
What are antibiotics in chemical terms ?	*semisynthetic modifications*
What type of antibiotics include penicilin ?	*beta-lactam antibiotics*
What is penicillins produced by ?	*fungi*

Un second exemple de corpus associé au domaine du questions-réponses est celui proposé dans le cadre du projet bAbI[3] de Facebook (WESTON et al., 2015). Ce dernier a pour objectif de fournir un ensemble de tâches élémentaires indépendantes les unes par rapport aux autres portant sur des problématiques du TAL et du raisonnement à partir de données textuelles, comme à l'image du corpus MCTest (RICHARDSON, BURGES et RENSHAW, 2013). La finalité du projet bAbI est de mesurer l'efficacité des approches destinées à la compréhension générale des textes. Chaque tâche est décomposée sous la forme d'un ensemble de phrases simples associées à une question. Ces tâches sont au nombre de 20 et portent chacune sur des problématiques bien identifiées *e.g.* recherche de faits, résolution de co-référence, déduction élémentaire, raisonnement lié au positionnement d'objet (cf. figure 5.2), etc.

Les meilleurs résultats sur les 20 tâches proposées dans le cadre de ce projet obtiennent une moyenne de 93% de précision (SUKHBAATAR et al., 2015), laissant une faible marge d'amélioration. Ce corpus est très intéressant dans le cadre d'un module complet destiné à la compréhension sémantique et discursif des textes. Toutefois, il est difficilement transposable à notre méthode d'extraction de connaissances. En effet, elle n'est pas initialement conçue pour gérer les co-références, la négation et les problématiques de raisonnement nécessitant une considération contextuelle et temporelle des événements. Toutes ces étapes sont des perspectives pouvant prendre la forme de modules complémentaires au sein de notre chaîne de traitement. Malgré

3. https ://github.com/facebook/bAbI-tasks

FIGURE 5.2 – Exemples de tâches élémentaires liés au projet bAbI.

Tâche 2 : Support de faits
John is in the playground.
John picks up the football.
Bob went to the kitchen.
Where is the footbal ?
Answer : *playground*

Tâche 11 : Résolution de co-référence
Daniel was in the kitchen.
Then he went to the studio.
Sandra was in the office.
Where is Daniel ?
Answer : *studio*

Tâche 16 : Induction élémentaire
Lily is a swan.
Lily is white.
Bernhard is green.
Greg is a swan.
What color is Greg ?
Answer : *white*

Tâche 17 : Positionnement
The triangle is to the right of the blue square.
The red square is on top of the blue square.
The red sphere is to the right of the blue square.
Is the red shepre to the right of the blue square ?
Answer : *yes*

tout, elle est capable de traiter certaines tâches identifiées telles que l'induction élémentaire. En effet, si nous concevons les relations *is-a* comme la taxonomie permettant d'inférer de nouvelles connaissances, la méthode est apte à fournir la réponse adéquate.

Par conséquent, la validation ne doit pas nous pénaliser sur le fait que notre approche n'est pas initialement pensée pour le domaine du questions-réponses et doit nous permettre d'évaluer à la fois : la modalité d'extraction d'information, la pertinence du modèle de propagation et l'efficacité des modèles de sélection. C'est pourquoi nous avons décidé de nous focaliser sur une validation à partir de la génération automatique de questionnaires et de la récupération d'un ensemble de textes extrait du Web. En effet, ce type de validation regroupe les critères désirés : les questions ne dépendent pas des textes extraits mais d'une base de connaissances permettant ainsi d'exploiter différents niveaux conceptuels et les questions générées sont sous la forme d'une relation ce qui n'implique pas la nécessité d'utiliser un module pour gérer la requête.

La sous-section suivante détaille le processus de génération automatique de questionnaires. Ce processus s'appuie sur un ensemble de relations d'une base de connaissances et sur sa structuration taxonomique.

5.1.2 Génération automatique d'options

Contexte

Le protocole de génération se base sur les travaux de PAPASALOUROS, KANARIS et KOTIS, 2008 portant sur l'évaluation des questionnaires à choix multiples (QCM). Toutefois, notre évaluation conçoit ce protocole uniquement dans le but de concevoir un ensemble d'affirmations vraies et fausses. Pour reprendre les termes employés

dans ce domaine, ces affirmations correspondent à des options modélisées par une déclaration de type <s,p,o>. Lorsque cette déclaration est correcte, c'est une réponse et lorsqu'elle est incorrecte, c'est un distracteur (cf. figure 5.3). Les options portent généralement sur une thématique donnée.

FIGURE 5.3 – Vocabulaire employé pour décrire un questionnaire.

Distracteur	Toulouse est la capitale de la France.
Distracteur	Marseille est la capitale de la France.
Réponse	Paris est la capitale de la France.
Distracteur	Caen est la capitale de la France.

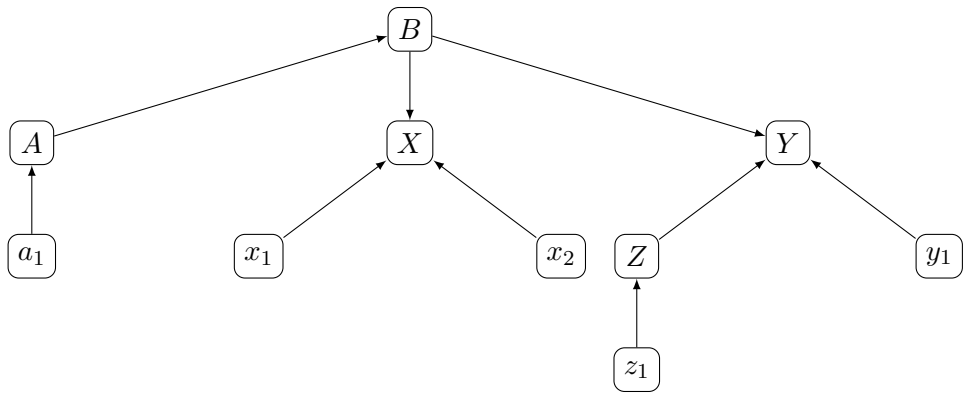
Ce protocole d'évaluation propose un nombre d'options fixé à 500, composé de 250 réponses et 250 distracteurs. Selon ABDEL-HAMEED et al., 2005, une forte densité de distracteurs permet d'améliorer la fiabilité de l'évaluation.

La stratégie de génération de PAPASALOUROS, KANARIS et KOTIS, 2008 est fondée sur la structure taxonomique des ontologies et plus particulièrement sur la relation *is-a* entre une classe (concept) et l'instanciation de cette classe en objet (individu). La stratégie stipule :

> "Si A est un concept, a un individu et $A(a)$ (a un individu de A) alors $A(a)$ est la réponse. Pour la sélection des distracteurs, si, B est un ancêtre de A, $B(b)$ avec $b \neq a$ et b n'est pas un individu de A, alors $A(b)$ est un distracteur."

Par exemple, à partir de la figure 5.4 si on considère $A(a_1)$ comme une réponse alors $A(x_1)$ est un distracteur car x_1 est un individu de X et X a un ancêtre commun avec A qui est B.

FIGURE 5.4 – Exemple de structuration de la connaissance servant de support à la génération automatique de questions. Les lettres en majuscule représentent des concepts et les lettres en minuscule des instances de ces concepts.



Cette stratégie de génération a été étendue pour permettre à l'approche de considérer les concepts en plus des instances en tant qu'options possibles. De plus, nous avons ajouté en entrée de l'algorithme un ensemble de relations initiales de type <s,p,o> permettant d'appliquer la méthode sur un prédicat autre que *is-a*. Ainsi,

nous réalisons une concordance entre un sujet (ou un objet) d'une des relations avec son nœud correspondant dans une taxonomie donnée pour générer de nouvelles relations. Nous pouvons définir cette stratégie de la manière suivante :

> "Soit so le nœud (feuille ou concept) correspondant au sujet (resp. à l'objet) d'une relation dans une taxonomie donnée et $\mathcal{A}(so)$ les ancêtres inclusifs de so. Alors, $\mathcal{A}(so)$ est constitué des réponses potentielles *i.e.* des candidats potentiels pour remplacer le sujet (resp. l'objet) de la relation. Concernant les distracteurs, ils peuvent correspondre à l'ensemble des concepts restants dans la taxonomie."

Par exemple, à partir de la figure 5.4 si on considère a_1 comme le sujet (ou l'objet) d'une relation alors A est une réponse et les concepts restants de la taxonomie des distracteurs potentiels.

La suite de cette section détaille cette stratégie de génération des options et les modalités employées pour conserver une cohérence sémantique entre les réponses et les distracteurs.

Protocole

L'ensemble des réponses correctes possibles $\mathcal{R}^+$ correspond aux concepts c appartenant à l'ensemble des concepts $\mathcal{C}$ de $O_{\mathcal{C}}$ obtenus à partir des concepts observés c_o et de leurs ascendants (cf. équation 5.1).

$$\mathcal{R}^+ = \{c \in \mathcal{C} \mid c_o \preceq c\} \tag{5.1}$$

En ce qui concerne l'ensemble des distracteurs possibles $\mathcal{R}^-$, ils correspondent à l'ensemble des concepts de $O_{\mathcal{C}}$ moins $\mathcal{R}^+$ et des concepts plausibles par rapport aux observations. L'ensemble de ces concepts plausibles équivaut à l'ensemble des descendants $\mathcal{D}$ des concepts observés (cf. équation 5.2).

$$\mathcal{R}^- = \{c \in \mathcal{C} \mid \mathcal{D}(c) \cap (\cup_{r \in R^+} \mathcal{D}(r)) = \emptyset\} \tag{5.2}$$

Cependant, une restriction sur les concepts est appliquée en exploitant les plus proches sémantiquement des réponses correctes par rapport à un seuil β (cf. équation 5.3).

$$R^-(\beta) = \{c \in \mathcal{R}^- \mid sim(c, a \in \mathcal{R}^+) \geq \beta\} \tag{5.3}$$

La recherche des concepts de $\mathcal{R}^-$ sémantiquement proches des concepts de $\mathcal{R}^+$ est réalisée en analysant les distances sémantiques entre les concepts de $\mathcal{R}^-$ et $\mathcal{R}^+$. La

méthode exploite la distance de Jaccard, où $\mathcal{A}(u)$ représente les ancêtres de u [4] (cf. équation 5.4).

$$sim_{Jaccard}(u, v) = \frac{| \mathcal{A}(u) \cap \mathcal{A}(v) |}{| \mathcal{A}(u) \cup \mathcal{A}(v) |} \tag{5.4}$$

L'utilisation de la distance de Jaccard permet de répondre aux contraintes sémantiques des options exposées par PHO, LIGOZAT et GRAU, 2015. En effet, les auteurs proposent d'évaluer les options par rapport à des critères syntaxiques et sémantiques. Ainsi, les options générées doivent avoir un profil sémantique similaire. Dans le cas contraire, ces options sont qualifiées de non-distracteur faisant référence à des distracteurs non pertinents.

FIGURE 5.5 – Les options générées doivent être similaires sémantiquement entre elles par rapport à un contexte donné.

Non-distracteur	Jacques Chirac est la capitale de la France.
Distracteur	Marseille est la capitale de la France.
Réponse	Paris est la capitale de la France.
Non-distracteur	Pôle nord est la capitale de la France.

Concernant l'homogénéité syntaxique, les options sont constituées des étiquettes des concepts de l'ontologie, soit des noms.

La prochaine sous-section détaille les données exploitées pour la génération des questionnaires et pour alimenter notre chaîne de traitement. Le processus d'extraction des données en entrée est également détaillé.

5.1.3 Données pour l'évaluation

Nous avons choisi de porter l'évaluation dans le domaine du bio-médical, il permet d'avoir accès à la fois à des bases terminologiques spécialisées (MeSH, ScnomedCT) et à des méthodes de désambiguïsation adaptées. Toutefois, la méthode est ancrée en monde ouvert ce qui inclut la possibilité de l'employer sur des corpus de nature diverse et non spécialisée.

Taxonomie et relations pour la génération automatique

La génération des options est réalisée à partir d'une liste de relations maladies/symptômes extraite de la base de données OMIM [5] (*Online Mendelian Inheritance in Man*). Ces relations ont été obtenues à partir du prédicat *has_manisfestation* (cf. tableau 5.2).

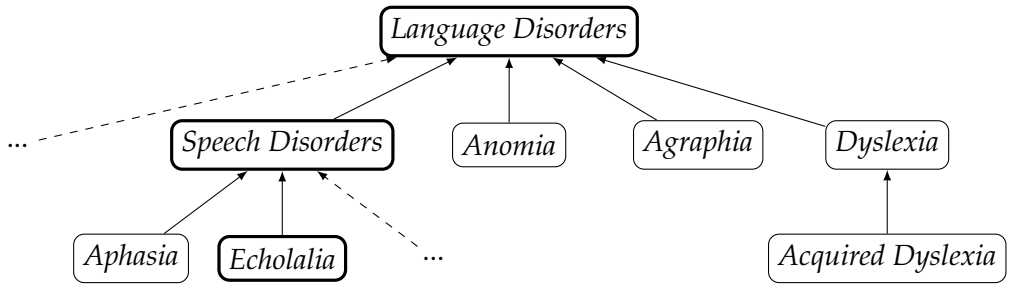
4. Le seuil β de similarité utilisé est 0,6. Ce seuil représente 60% des ancêtres en commun entre une bonne réponse et un distracteur.
5. https ://www.omim.org

TABLEAU 5.2 – Exemples de relations provenant de la base de données OMIM.

Maladie	Symptôme
Tourette Syndrome	*Echolalia*
Tourette Syndrome	*Compulsive Behavior*
Darier Disease	*Bipolar Disorder*
Darier Disease	*Papilloma*
Tyrosinemias	*Hepatomegaly*

Par la suite, ces relations sont recoupées avec l'arborescence du MeSH 2016 [6] servant de substrat au processus de génération (cf. figure 5.6). Ce dernier considère le sujet (la maladie) et l'objet (un symptôme) dans l'arborescence du MeSH puis s'emploie à générer des réponses et des distracteurs. À noter qu'en plus de la distance minimale de Jaccard, la génération est restreinte aux entités appartenant aux descripteurs C (*Diseases*) et F03 (*Mental Disorders*) du MeSH avec un filtre appliqué sur des entités trop générales telles que *Diseases* et *Signs & Symptoms*.

FIGURE 5.6 – Exemple de génération d'options à partir de l'arborescence du MeSH et de la relation OMIM <*Tourette Syndrome, has_manifestation, Echolalia*>. Les segments en pointillés représentent les concepts ou les individus du MeSH non représentés dans ce schéma pour un souci de clarté. Les candidats potentiels pour devenir une réponse sont les nœuds avec un contour en gras. Les concepts restants sont des distracteurs potentiels.



Extraction de données textuelles à partir du Web

Pour répondre aux questionnaires générés, un jeu de données a été constitué automatiquement à partir de pages Web (200 000 phrases) collectées à l'aide de requêtes adressées à Google composées seulement du nom des maladies (180 maladies utilisées) *e.g.* les noms des maladies contenus dans la première colonne du tableau 5.2. Par conséquent, la construction de ce jeu de données ne garantit pas la couverture complète des réponses des questionnaires. Les phrases récoltées sont ensuite soumises à une phase de désambiguïsation des concepts en utilisant Metamap. L'évaluation porte sur différentes approches disjointes par le mode d'extraction et l'utilisation de la connaissance *a priori*. Pour cette expérience deux modes d'extraction

6. https ://www.ncbi.nlm.nih.gov/mesh

ont été exploités et comparés : le principe de co-occurrence des concepts issus des questions et l'extraction de relations par l'intermédiaire de la méthode Reverb en exploitant une liste de prédicats prédéfinis : *include*, *due to*, et *cause*. Chacune de ces approches est évaluée avec et sans le support d'une structure taxonomique. Le tableau 5.3 présente le nombre de relations extraites selon la méthode employée.

TABLEAU 5.3 – Nombre de relations extraites en fonction de la méthode d'extraction d'information employée sur le corpus établi.

	Co-occurrence	Extraction de relations
#Relations extraites	8616	623
#Phrases	5756	237
#Phrases incertaines	986	57

La figure 5.7 présente un ensemble d'options avec deux réponses et deux distracteurs. Chacune des réponses est associée à des phrases extraites la supportant.

FIGURE 5.7 – Exemple d'options générées de manière automatique pour le prédicat *has_manifestation*. Chaque réponse est accompagnée d'exemples extraits du Web.

- **Réponse 1** : <Syndrome De Lange, Anomalies congénitales>
  - *Cornelia de Lange syndrome is a very uncommon disorder that involves delayed physical growth, as well as a variety of malformations of the face, limbs, and head.*
  - *About 20 per cent of children diagnosed with CdLs suffer from congenital cardiac abnormalities.*
- **Réponse 2** : <Syndrome De Lange, Manifestations neuro-comportementales>
  - *Behavioral disturbance is common in Cornelia de Lange syndrome and is more frequent in those with severe mental retardation.*
  - *Children with CDLS often have speech delay due to problems with the mouth, hearing impairment, and developmental delay.*
- **Distracteur** : <Syndrome De Lange, Troubles de la personnalité paranoïaque>
- **Distracteur** : <Syndrome De Lange, Paraplégie>

La section suivante présente les métriques d'évaluation exploitées ainsi que les résultats obtenus par notre approche d'extraction de connaissances sur les options générées.

5.2 Métriques et résultats

5.2.1 Métriques d'évaluation

Les métriques d'évaluation utilisées correspondent aux métriques standard : rappel, précision et F-mesure (cf. tableau 5.4). Dans ce cas, un vrai positif (resp. un faux négatif) est comptabilisé quand le système prédit vrai (resp. faux) à une question dont la réponse est vraie (resp. fausse). Un faux positif (resp. un faux négatif) est comptabilisé quand le système répond vrai (resp. faux) à une question dont la réponse est fausse (resp. vraie).

TABLEAU 5.4 – Métriques d'évaluation. TP dénote le nombre de vrais positifs, FP de faux positifs et FN de faux négatifs

Precision	$\frac{TP}{TP+FP}$
Rappel	$\frac{TP}{TP+FN}$
F-measure	$\frac{2.precision.rappel}{precision+rappel}$

Toutefois, il faut noter que nous avions initialement conçu l'évaluation sous la forme d'un QCM. Dans ce cas de figure, un moyen de pouvoir l'évaluer aurait été d'exploiter la K-mesure (cf. équation 5.5) définie lors de CLEF 2004 (HERRERA, PENAS et VERDEJO, 2004). La K-mesure $K(sys)$ d'un système sys donné implique : $R(i)$ le nombre total de réponses correctes attendues pour la question i, $answered(sys, i)$ le nombre de réponses données par le système sys pour la question i, $score(r)$ le score de confiance assigné par le système à l'option r (*e.g.* 1 pour une option jugée correcte, 0 sinon) et $eval(r)$ l'évaluation de r en fonction de la génération du QCM (*e.g.* -1 pour un distracteur et 1 pour une réponse).

$$K(sys) = \frac{1}{\#questions} \times \sum_{i \in questions} \frac{\sum_{r \in answers(sys,i)} score(r) \times eval(r)}{max(R(i), answered(sys, i))} \tag{5.5}$$

$$K(sys) \in [-1, 1]$$

Cette mesure reflète le niveau de connaissance d'un système. Quand $K(sys)$ est égal à 0, cela correspond à un système sans connaissance qui assigne un score de confiance de 0 pour toutes les réponses. Ainsi, $K(sys) = 0$ aurait pu servir de *baseline* à cette évaluation à l'instar de CLEF 2004. Des expérimentations complémentaires, pour conforter nos conclusions, ont été réalisées en exploitant cette mesure d'évaluation (cf. section 5.3).

La sous-section suivante expose les résultats obtenus (précision, rappel et F-mesure) sur les options générées automatiquement à partir des relations d'OMIM et de la taxonomie du MeSH en considérant un ensemble de phrases extraites du Web.

5.2.2 Résultats

Les résultats obtenus portent sur les différentes configurations de paramètres de notre chaîne de traitement. Ces paramètres sont les suivants.

— La méthode d'extraction de relations.
— L'utilisation d'une structure taxonomique.
— L'application d'un modèle de sélection.
— L'attribution d'une valeur d'incertitude aux relations incertaines.

L'ensemble des résultats portent sur les modèles de sélection 2, 3 et 4 présentés au chapitre 4. Le modèle 1 basé sur un seuil de confiance est sujet à beaucoup de variation due à l'estimation empirique du seuil en fonction des questions, du prédicat et du corpus étudié. De plus, les résultats présentés correspondent à la moyenne des mesures obtenues sur l'ensemble des questionnaires générés.

Le tableau 5.5 résume les résultats obtenus avec les différentes méthodes d'extraction et l'utilisation ou non d'une structuration taxonomique mais sans la prise en compte de l'incertitude.

TABLEAU 5.5 – Moyenne des résultats et leur écart-type respectif (STD) obtenus sur 100 questionnaires. Plusieurs configurations ont été expérimentées en fonction des méthodes d'extraction, d'une étape de propagation sur O_C et d'une phase de sélection. $\mathcal{M}_2$ est le modèle basé sur la moyenne, $\mathcal{M}_3$ le modèle basé sur la médiane et $\mathcal{M}_4$ le modèle exploitant la croyance du fait et de ses parents. Lorsque *Sélection* est équivalent à *Non*, aucun modèle n'est utilisé.

	O_C	Sélection	Précision (STD)	F-mesure (STD)
Co-occurrence	Non	Non	0,96 (0,03)	0,21 (0,03)
	Oui	Non	0,95 (0,02)	0,40 (0,03)
	Oui	$\mathcal{M}_2$	0,98 (0,03)	0,15 (0,03)
	Oui	$\mathcal{M}_3$	0,97 (0,02)	0,35 (0,03)
	Oui	$\mathcal{M}_4$	0,97 (0,02)	0,26 (0,03)
Ext. de relations	Non	Non	0,97 (0,01)	0,05 (0,02)
	Oui	Non	0,99 (0,03)	0,13 (0,03)
	Oui	$\mathcal{M}_2$	0,99 (0,03)	0,05 (0,02)
	Oui	$\mathcal{M}_3$	0,98 (0,03)	0,12 (0,03)
	Oui	$\mathcal{M}_4$	0,99 (0,03)	0,05 (0,02)

Le tableau 5.6 considère les mêmes modèles avec la prise en compte de l'incertitude linguistique. Celle-ci impacte les modalités de propagation pour les déclarations incertaines. En effet, les valeurs associées a ces déclarations lors de la propagation vont être comprises entre $[0, 1[$.

TABLEAU 5.6 – Valeur d'incertitude différente avec les modèles $\mathcal{M}_2$, $\mathcal{M}_3$ et $\mathcal{M}_4$ exploitant l'extraction par co-occurrence et O_C.

Modèle	Val. incert.	Précision	F-mesure
$\mathcal{M}_2$	0,5	0,98	0,15
$\mathcal{M}_2$	0,0	0,98	0,14
$\mathcal{M}_3$	0,5	0,96	0,36
$\mathcal{M}_3$	0,0	0,96	0,36
$\mathcal{M}_4$	0,5	0,98	0,26
$\mathcal{M}_4$	0,0	0,97	0,26

5.3 Discussion

La principale difficulté de cette validation repose sur l'élaboration de la collection de phrases nécessaire pour répondre aux questions. En effet, nous n'avons pas la garantie d'avoir les informations adéquates au sein de ce jeu de données pour répondre à l'ensemble des options. De ce fait, dans la majorité des cas une réponse classée comme distracteur doit être perçue comme une option n'ayant pas de support pour émettre un jugement sur la déclaration. Cet aspect de la validation impacte négativement le rappel associé aux approches. En effet, le rappel obtenu pour la méthode de co-occurrence avec propagation et sans modèle de sélection est de 0,25. Cette valeur représente la couverture maximale pouvant être obtenue par l'approche. Par conséquent, la discussion porte principalement sur la comparaison relative entre les précisions des différentes configurations expérimentées.

Le tableau 5.5 montre que la principale influence sur la F-mesure est conditionnée par l'utilisation de la connaissance *a priori* au regard des déclarations extraites pour un système d'extraction donnée. Par ailleurs, ces systèmes d'extraction conditionnent également les résultats de manière significative, notamment par une couverture des relations plus vaste. A noter que ce phénomène est propre à cette expérimentation et que les résultats ne sont pas généralisables à d'autres prédicats. Par exemple, le prédicat *bornIn* implique des entités nommées (personne, lieu) entraînant des problématiques de prédicats multiples (SURDEANU et al., 2012). Les conditions du cadre expérimental suscitent cette propriété de spécificité des résultats. En effet, nous devons ajouter à un domaine textuel fortement contraint, une co-occurrence des types d'entités (maladie/symptôme) largement conditionnée à l'observation d'une même sémantique associée à un prédicat donné (<maladie, cause, symptôme>). En ce qui concerne l'hétérogénéité des F-mesures entre les modèles

d'extraction, elle s'explique en partie par la différence dans le nombre de phrases captées par la méthode d'extraction de relations (cf. tableau 5.3). L'écart indique que la relation entre une maladie et un symptôme peut être exprimée par un nombre de prédicats plus grand au regard de la liste utilisée pour filtrer les relations (cf. tableau 5.7).

TABLEAU 5.7 – Exemples de phrases dans lesquelles une maladie est associée à des symptômes. Les phrases sont issues de notre jeu de données collecté sur le Web. Le symbole ✓ signifie que l'on considère ce prédicat pour la phase d'extraction de relations. Le symbole ✗ illustre la diversité des prédicats pour une sémantique donnée.

Prédicats	Exemples	
include	*The complications of* ***Crohn's disease include bowel obstruction, abscesses, free perforation and hemorrhage****, which in rare cases may be fatal.*	✓
cause	***Tay Sachs disease*** *is a genetic disorder that can* ***cause*** *both the* ***mental and the physical deterioration*** *of the patient's body*	✓
due to	*Babies with* ***diGeorges syndrome*** *may experience feeding problems* ***due to high arched*** *and* ***cleft palate****.*	✓
resulting	***Angelman syndrome*** *is caused by the loss of function of a particular gene during fetal development,* ***resulting*** *in severe* ***neurological impairment*** *present at birth and lasting for a lifetime.*	✗
lead to	***Distal muscular dystrophy*** *can* ***lead to weakness and wasting of muscles of the hands, forearms and lower legs****.*	✗
associating	***Netherton's syndrome*** *is a recessive autosomal disease* ***associating ichthyosiform dermatosis, hair dysplasia and systemic involvement****.*	✗
experience	*A person with* ***schizophrenia*** *may* ***experience psychotic symptoms****.*	✗

Une analyse fine des faux-positifs démontre l'importance des modèles de sélection. Ils nous ont permis de remarquer que la génération automatique des questionnaires appliquée au domaine médical implique un risque d'erreur si les relations pour une maladie donnée ne sont pas exhaustives dans l'ontologie. En effet, on observe que certains échecs sont contestables. Par exemple, la méthode infère avec un fort support que la maladie *Alkaptonuria* induit dans certains cas des maladies articulaires. Cette inférence a été validée après lecture dans Orphanet[7]. Cependant, OMIM ne renseigne pas ce symptôme. Cette découverte de connaissances est un résultat encourageant pour notre chaîne de traitement. En ce qui concerne les autres faux positifs, ils sont généralement induits par le biais de la méthode de co-occurrence, *e.g.* le symptôme *Bacterial Infections and Mycoses* est inféré pour le *syndrome de Down* alors que ce n'est qu'une conséquence de ce dernier : *Pneumonia is one of the most common infections to affect Down syndrome patients*.

7. www.orpha.net

En ce qui concerne la prise en compte de l'incertitude (cf. tableau 5.6) on remarque que son influence n'impacte pas les résultats obtenus pour les modèles $\mathcal{M}_2$ et $\mathcal{M}_4$. Toutefois, le score obtenu avec le modèle $\mathcal{M}_3$ est significatif selon un test de Student. En effet, les écarts-types sur chacune des configurations sont de l'ordre de 10^{-2}. Malgré cela et au regard des résultats obtenus, la considération de l'incertitude linguistique dans les modalités de propagation peut être considérée comme un processus secondaire de la chaîne de traitement. La méthode d'extraction de relations, le processus d'inférence en tenant compte d'une structure taxonomique et l'utilisation d'un modèle de sélection constituent les sources apportant le plus de variabilité au niveau des résultats. De plus, le domaine bio-médical et l'utilisation de Google *Search* pour obtenir un corpus de données ont une part de responsabilité dans ces conclusions. En effet, même si le taux d'incertitude est relativement élevé dans ce domaine, les relations extraites sont généralement vraies et ce phénomène est amplifié par le fait d'exploiter l'algorithme de recherche d'information de Google favorisant des résultats populaires.

Une autre expérimentation confirme ce constat sur la prise en compte de l'incertitude. Le protocole expérimental établi exploite des QCM générés de manière automatique afin d'observer l'impact des différents paramètres sur la K-mesure, présentée en sous-section 5.2.1. Chaque QCM est représenté comme un ensemble de questions dont chacune est associée à 10 options possibles comprenant 1 à 4 réponses. La construction des QCM suit le même processus que celui des questionnaires (cf. section 5.1.2). Le tableau 5.8 résume les résultats obtenus dont la conclusion sur l'incertitude linguistique est similaire à celle précédemment évoquée.

TABLEAU 5.8 – Résultats en terme de K-mesure sur un QCM généré automatiquement à partir des relations d'OMIM et de la taxonomie du MeSH. Les données exploitées pour y répondre proviennent d'une phase d'extraction de relations par co-occurrence. Plusieurs configurations ont été expérimentées en fonction d'une étape de propagation sur $O_{\mathcal{C}}$ et de la présence ou non des différents modèles de sélection ($\mathcal{M}$). Les succès correspondent au nombre de fois où l'approche propose une option correcte tandis que les échecs correspondent au nombre de fois où l'approche propose une option incorrecte.

Modèle	Val. incert.	Succès	Échecs	K-mesure
$\neg O_{\mathcal{C}} \bigwedge \neg \mathcal{M}$	/	53	11	0,11
$O_{\mathcal{C}} \bigwedge \neg \mathcal{M}$	/	217	33	0,37
$O_{\mathcal{C}} \bigwedge \mathcal{M}_2$	1,0	94	5	0,19
$O_{\mathcal{C}} \bigwedge \mathcal{M}_2$	0,5	98	3	0,21
$O_{\mathcal{C}} \bigwedge \mathcal{M}_2$	0,0	96	3	0,20
$O_{\mathcal{C}} \bigwedge \mathcal{M}_3$	1,0	199	28	0,34
$O_{\mathcal{C}} \bigwedge \mathcal{M}_3$	0,5	202	29	0,36
$O_{\mathcal{C}} \bigwedge \mathcal{M}_3$	0,0	201	28	0,36
$O_{\mathcal{C}} \bigwedge \mathcal{M}_4$	1,0	168	15	0,31
$O_{\mathcal{C}} \bigwedge \mathcal{M}_4$	0,5	165	15	0,30
$O_{\mathcal{C}} \bigwedge \mathcal{M}_4$	0,0	154	12	0,30

Toutefois, l'incertitude linguistique peut constituer une information utile dans le cadre d'une interaction avec un utilisateur au travers d'une interface graphique (cf. chapitre 4). En effet, elle peut apporter une information supplémentaire aux supports des relations inférées pouvant guider un utilisateur dans sa prise de décision.

Les résultats exposés dans le tableau 5.8 permettent également d'apporter des éléments d'information pour comparer et appuyer l'intérêt d'exploiter les modèles de sélection. Malgré une K-mesure supérieure pour le cas de la propagation sans l'utilisation de modèle de sélection [8], l'observation des succès et des échecs est une information importante à prendre en compte selon le contexte d'utilisation. En effet, ces modèles ont été initialement modélisés pour le domaine de la découverte de connaissances (et non du questions-réponses) dans lequel ils sont confrontés au filtrage d'un important volume de données textuelles. Ainsi au sein d'un tel contexte, nous avons privilégié la précision des résultats retournés (favoriser les fortes croyances) par rapport au rappel *i.e.* au nombre de résultats retournés.

8. Ce résultat peut être expliqué soit par un nombre de distracteurs trop faible soit par le consensus des données exploités (relations maladies/symptômes).

Chapitre 6

Conclusion

Sommaire

Dans cette thèse, nous avons présenté une chaîne de traitement capable de tirer parti des informations explicites et implicites de vastes corpus textuels non-structurés en considération de l'incertitude linguistique. Elle exploite pour cela une méthode éprouvée dans le domaine de l'extraction de relations en domaine ouvert couplée à un module de raisonnement inductif permettant d'inférer de nouvelles connaissances par la généralisation des observations. Une fois l'inférence réalisée, l'approche évalue la pertinence de l'ensemble des déclarations par l'intermédiaire de modèle de sélection. Son intérêt porte à la fois sur le domaine de l'enrichissement des bases de connaissances, du questions-réponses et de la génération d'hypothèses.

6.1 Incertitude linguistique et imprécision taxonomique dans un processus d'extraction de connaissances

6.1.1 Récapitulatif de la chaîne de traitement

L'architecture de la chaîne de traitement proposée se décompose en plusieurs parties allant de l'extraction de relations à la sélection des relations pertinentes en passant par l'inférence de connaissances rendue possible par l'organisation, l'enrichissement et la caractérisation des extractions. Par conséquent, la contribution de cette thèse est basée sur ce système de structuration de la connaissance et sur la gestion de l'incertitude dans le processus d'induction et de la présentation des résultats graphiques (JEAN et al., 2015 ; JEAN et al., 2017).

La structuration de l'information extraite a deux objectifs, le premier est de générer de nouvelles relations et le second est de servir de support au processus d'évaluation de la pertinence. Cette étape de structuration se conçoit au travers de la construction de deux graphes que l'on appelle respectivement : graphe des syntagmes et graphe des déclarations. L'agencement du premier graphe débute par la décomposition syntaxique des syntagmes nominaux constituant les sujets et objets des relations extraites. Cette première dissociation conceptuelle permet déjà de réaliser de simples inductions *e.g.* le syntagme nominal "*CFTR gene mutation*" sous-entend les concepts plus génériques, et potentiellement non exprimés : "*gene mutation*" et "*mutation*". Par la suite, ce graphe peut servir d'armature à une phase d'enrichissement par le biais d'une source de connaissances structurées externe de type taxonomique. En effet, une phase de désambiguïsation lexicale, exploitant par exemple des mesures de similarité, permet de réaliser certaines correspondances entre les concepts du graphe de syntagmes et une taxonomie donnée. Par conséquent, le graphe peut tirer parti de la connaissance contenue dans cette taxonomie pour élargir sa portée conceptuelle *e.g.* si on extrait les mots "*cystic fibrosis*" (mucoviscidose), la taxonomie de WordNet nous permet d'ajouter un nouveau concept plus général dans le graphe de syntagmes qui est "*monogenic disease*". Une fois la construction terminée, la phase de génération débute. Cette phase permet d'élargir l'ensemble des relations extraites en exploitant les ascendants des sujets et des objets de ces relations dans le graphe des syntagmes. Il faut noter que ces nouvelles relations ainsi que les relations extraites ont une signification particulière. En effet, la chaîne de traitement exploite une hypothèse existentielle et non universelle, *e.g.* en sachant que le *nifuroxazide* est un composé de la famille des *nitrofuranes* alors l'observation de *Le nifuroxazide est employé dans le traitement de la colite* soutient l'idée qu'il existe au moins un composé de la famille des *nitrofuranes* qui est employé dans le traitement de la colite.

Une fois les relations générées, le processus d'évaluation de la pertinence des déclarations tient compte de différents critères. Dans le cadre de cette thèse, nous exploitons la croyance et la spécificité des relations permettant à l'approche de s'ancrer à la fois dans le domaine de l'enrichissement des bases de connaissances et de la découverte de connaissances tout en tenant compte des besoins des utilisateurs. La valeur de croyance est abordée au sein d'un mécanisme de propagation *bottom-up* des observations que l'on a adapté pour tenir compte de l'incertitude. Par conséquent, l'obtention de l'ensemble de ces critères doit passer par la structuration d'un second graphe que l'on nomme graphe des déclarations. Ce dernier ordonne, selon des règles de subsomption définies manuellement sur les sujets et objets, les relations extraites et générées. Ainsi, ce graphe offre l'opportunité d'évaluer la croyance et la spécificité (profondeur au sein du graphe) des déclarations.

Les critères associés à chacune des relations sont exploités pour définir la pertinence à conserver une relation. Toutefois, à partir d'ici nous pouvons distinguer deux principaux chemins. Le premier est la définition de modèles formels non paramétrables.

Ces derniers s'adaptent plus facilement à un contexte d'enrichissement de base de connaissances dans lequel les valeurs de croyance sont maximisées. Toutefois, dans un contexte utilisateur, il est plus judicieux d'adapter les résultats en fonction d'un profil donné. En effet dans ces cas là, la spécificité peut être déterminante selon le niveau de connaissance d'un utilisateur dans un domaine donné (expert *versus* néophyte) *e.g.* on préférera utiliser les termes "cancer de la peau" plutôt que "mélanome" qui est une forme spécifique, dans le but de répondre à un public non expert.

Maintenant que l'architecture générale de la chaîne de traitement a été résumée, exposons les raisons d'exploiter l'incertitude linguistique et les modalités mises en place pour sa détection.

6.1.2 Détection et prise en compte de l'incertitude linguistique

Le choix de considérer l'incertitude linguistique est double. Le premier est le rôle qu'elle peut jouer dans le domaine de la découverte de connaissances. Ce domaine a pour objectif de déceler des agrégations de signaux faibles entre relations afin de proposer des hypothèses de recherche pour un domaine donné. Ainsi, au sein de ce contexte précis de découverte et d'inférence de connaissances, l'exploitation de cette forme d'incertitude peut jouer un rôle essentiel. Le second intérêt provient du manque de publications dans ce domaine. Là où d'autres sources d'incertitude ont été étudiées, l'incertitude linguistique a été relativement délaissée dans la littérature. Enfin dans un contexte plus large, l'incertitude est une indication primordiale pour un utilisateur. En effet, l'indication de la provenance d'une inférence est une information importante pour un utilisateur à l'instar de l'outil de traduction Linguee [1] dans lequel l'incertitude d'une traduction est labellisée par un symbole spécifique. Cet intérêt s'étend également aux autres dimensions de l'incertitude telles que l'incertitude liée aux méthodes ou à la véracité des faits.

Comme nous l'avons vu, l'incertitude linguistique peut être identifiée à partir de méthodes spécifiques adaptées au langage naturel. Dans le cadre de cette thèse, une application de détection de l'incertitude a été élaborée (JEAN et al., 2016a ; JEAN et al., 2016b). Cette dernière exploite une méthode d'apprentissage à partir d'une représentation vectorielle des phrases basée sur une agrégation des poids attribués à chaque unité de la phrase (uni-grammes, bi-grammes, etc.). Le poids de ces unités est représenté en partie par la probabilité qu'ils ont d'appartenir à une classe donnée de l'incertitude *i.e.* d'appartenir à une phrase incertaine ou d'être marqueur d'incertitude. Toutefois, du fait de la taille limitée des corpus d'entraînement, un lemme n'apparaissant qu'une seule fois dans le corpus et ce, de façon fortuite, dans un contexte incertain, aura une probabilité d'appartenir à une phrase incertaine de 1. Pour pallier ce problème nous avons nuancé ces probabilités par la modélisation

1. http ://www.linguee.fr/

d'un score de confiance tenant compte du nombre d'apparitions. Par la suite, ce modèle a été évalué par rapport à des jeux de données issus de la conférence CoNLL en 2010. Nos résultats améliorent la moyenne des performances des méthodes soumises lors de cette compétition.

6.1.3 Validation de la chaîne de traitement

La validation de l'approche est réalisée dans le cadre d'une tâche de questions-réponses. Pour cela, nous nous sommes appuyés sur la construction automatique d'un ensemble de questionnaires contenant des questions nécessitant une réponse positive et des questions nécessitant une réponse négative. La constitution de ces questionnaires a été basée sur la modification d'un algorithme de construction automatique de QCM. Pour les générer, nous avons exploité un ensemble de relations Maladie/Symptômes provenant de la base de données OMIM et de la taxonomie du MeSH. Pour tenter d'y répondre, nous avons récolté des données à partir du Web en utilisant comme requêtes les noms des maladies à notre disposition : 200 000 phrases ont été ainsi récoltées. Les résultats obtenus démontrent l'intérêt de notre chaîne de traitement et du module d'inférence par rapport à l'utilisation du module d'extraction de relations seul.

La phase de validation a également permis de juger de l'intérêt de considérer l'incertitude linguistique dans la chaîne de traitement. Dans les conditions expérimentales définies : domaine bio-médical, recherche de données Web avec Google *Search*, nous avons remarqué que l'incertitude linguistique n'a pas réellement impacté les performances de l'approche, tandis que la modification du type d'extraction de relations et la décision d'exploiter ou non un modèle de sélection avait une importance bien plus cruciale. Par conséquent, la prise en compte de l'incertitude linguistique est difficile à mettre en valeur du fait de sa représentativité dans le langage naturel et de la nature des données.

6.2 Perspectives

Au niveau des perspectives, l'approche est généralisable à d'autres dimensions de l'incertitude potentiellement plus impactantes. La première dimension que nous envisageons est celle issue de la contradiction des données pour l'évaluation des relations et par extension des sources. La seconde est la gestion de l'incertitude propre aux méthodes d'extraction d'information employées. En effet, chaque méthode engage une fiabilité différente au niveau des extractions. Par exemple, un modèle de co-occurrence est susceptible de générer une incertitude plus grande qu'une méthode d'extraction de relations. Ainsi, la gestion de ces degrés d'incertitude liés aux

méthodes d'extraction permettrait d'envisager un modèle manipulant plusieurs extracteurs de manière simultanée. Par exemple, DONG et al., 2014 ont proposé des pistes de réflexion concernant ce type d'incertitude. En effet, ces derniers considèrent quatre types d'extracteurs, chacun adapté à différentes manières de présenter l'information au sein d'une page Web : texte brut, tableau, exploitation du DOM (*Document Object Model*) et des pages Web annotées. Les sorties de ces extracteurs sont ensuite combinées pour construire un vecteur de caractéristiques spécifique à un <s,p,o>. Ce vecteur est utilisé au sein d'un classifieur binaire afin de déterminer la probabilité qu'il existe un lien de type p entre s et o. Le classifieur apprend des poids différents pour chaque composant du vecteur permettant ainsi d'obtenir la fiabilité relative de chaque système. Par conséquent, nous pourrions imaginer une extension de la chaîne de traitement en exploitant plusieurs extracteurs de manière simultanée, l'incertitude des sources et l'incertitude linguistique.

Nous pourrions également envisager un raffinement de la chaîne de traitement par la modification des modèles de sélection. En effet, notre approche pourrait être en mesure de raffiner le moyen d'inférer de la connaissance au travers de plusieurs perspectives. La première serait de tenir compte de la dispersion des observations au niveau des descendants des nœuds. En effet, une discussion peut être ouverte sur la provenance de ces observations. Par exemple, si nous voulons inférer la relation <*marsupiaux, mangent, eucalyptus*>, serait-il préférable que la grande majorité des observations proviennent de <*koalas, mangent, eucalyptus*> ou d'une dispersion, à nombre total d'observations égal, des observations entre de nombreuses relations plus spécifiques à la relation que l'on désire inférer *e.g.* <*bandicoots, mangent, eucalyptus*>, etc. La seconde perspective serait de pouvoir distinguer les relations négatives. En effet, les modèles de sélection pourraient ainsi être capables de pondérer la pertinence d'un fait en fonction du nombre de fois où il a été trouvé négatif. Pour cela, nous pourrions exploiter notre approche de détection de l'incertitude sur des données spécifiquement labellisées pour considérer les marqueurs de négation. Enfin, la troisième perspective aborde une discussion à propos du poids des extractions dans le module d'inférence. Prenons l'exemple de deux phrases.

— *Les bandicoots mangent de l'eucalyptus*.
— *Tous les marsupiaux mangent de l'eucalyptus*.

Au sein des modèles actuels aucune distinction n'est réalisée et chaque relation extraite aura un poids identique. Toutefois, ne serait-il pas plus judicieux de distinguer une observation spécifique, d'une observation plus générique (bandicoots *vs.* marsupiaux) ? En effet, l'une est plus étendue que l'autre car elle considère tous les marsupiaux. Toutefois dans le cadre de données Web, il y a risque de considérer des poids différents en fonction des déclarations à cause de leur provenance *e.g.* personnes malicieuses ou manque d'objectivité et de connaissance de la part d'un utilisateur, etc.

Enfin, l'annexe B propose d'élargir la notion de découverte de connaissances, introduite dans cette thèse, en tenant compte d'un encadrement de probabilité (croyance et plausibilité) à associer aux relations par l'intermédiaire du cadre théorique des fonctions de croyance.

La ligne de conduite adoptée tout au long de cette thèse a été l'applicabilité et le partage des méthodes pensées et développées. Les compétitions en informatique telles que CoNLL et COLIEE (cf. Annexe A) ont été des dispositifs scientifiques importants pour la conception et le développement des approches partagées. En effet, la mise à disposition de *benchmarks* permet aux différentes communautés scientifiques de comparer les approches et de mieux comprendre et délimiter les enjeux actuels de ces domaines. Cette réflexion s'est fait ressentir lors de la validation de la chaîne de traitement dans son ensemble pour laquelle aucun *benchmark* n'est réellement significatif. Ainsi de mon point de vue, il est important d'encourager les doctorants, et plus généralement le personnel académique, à participer ou proposer de tels événements scientifiques, dans le but de partager et animer ces différentes communautés.

Annexe A

COLIEE 2017

Sommaire

COLIEE 2017[1] (*Competition on Legal Information Extraction/Entailment*), est une compétition sur l'extraction et l'implication d'information dans le domaine du droit. Lors de cette dernière, deux tâches ont été proposées : l'une appartenant à la recherche d'information et à l'implication des données et l'autre au questions-réponses de type booléen. Cette annexe se concentre sur la première tâche dont l'objectif était de proposer un système permettant d'analyser un cas pratique juridique écrit en langage naturel afin de retrouver un ou plusieurs articles du code civil japonais (traduit en anglais) permettant à un expert de prendre une décision sur la légalité de ce cas. Par extension, cette première tâche pouvait servir de support à la seconde dans laquelle les participants devaient fournir un système expert devant prendre une décision (Oui ou Non) sur la légalité du cas proposé. Ma participation à cette compétition a été motivée par un doctorant du LGI2P, Gildas Tagny-Ngompe, dont le sujet de thèse, intitulé "Analyse sémantique d'un corpus de décisions jurisprudentielles pour l'élaboration de modèles prédictifs du risque judiciaire", aborde les notions de segmentation de textes et de prise de décision dans le domaine du droit. Son objectif est d'extraire la sémantique des textes de décisions jurisprudentielles afin de modéliser une approche permettant de prédire de futures décisions. Ma collaboration est le

1. http ://webdocs.cs.ualberta.ca/∼miyoung2/COLIEE2017/

fruit de thématiques connexes et complémentaires entre ma thèse et la sienne, plus précisément sur les parties TAL et extraction d'information. De plus, Sébastien Harispe, co-encadrant de ces deux thèses, a permis de renforcer les compétences TAL et recherche d'information de l'équipe.

A.1 Tâche sur la recherche et l'implication d'information

A.1.1 Description des données

Le code civil japonais

Les données de référence proviennent d'une traduction en anglais d'un sous-ensemble du code civil japonais contenant ∼1100 articles. La décomposition de ce code est relativement complexe. Celle-ci prend la forme de textes comportant des segments facultatifs rendant l'étape de formatage automatique fastidieuse. La figure A.1 montre la décomposition des différentes portions du code civil fournies par les organisateurs de l'événement.

TABLEAU A.1 – Décomposition du code civil en entrée de la compétition COLIEE. Certaines sections sont facultatives. Le symbole √ indique que le segment est facultatif.

Segments	Exemples	Facultatif
- Partie	*Claims*	
- Chapitre	*General Provisions*	
- Section	*Subject of claim*	√
- Sous-Section	*Guarantee obligation*	√
- Division	*General provisions*	√
- Sous-Division	*Responsibility of guarantor*	
- Article	*Article 446*	
- Paragraphe	*1 to 3*	√

Ce code se décompose principalement en articles organisés en sous-divisions puis en chapitres et enfin en parties. Les articles sont eux-mêmes généralement décomposés en paragraphes contenant une ou plusieurs phrases. Au niveau du contenu des articles, il est important de noter différentes structures particulières. La première est l'intention des juristes à vouloir généraliser certaines variables appartenant aux articles du code civil afin d'éviter la redondance des articles (cf . figure A.1). Il est important d'avoir conscience de ce procédé car il peut influencer les méthodes de recherche d'information basées sur une correspondance exacte entre les termes de la requête et des documents (*tf-idf*, *BM25*).

La seconde particularité du code civil est l'utilisation d'un vocabulaire riche appliqué à des domaines bien distincts. En effet, certains articles décrivent une situation précise en employant un vocabulaire adapté (cf. figure A.2). L'utilisation de termes

FIGURE A.1 – Exemple d'article utilisant une règle de généralisation (en gras).

*Article 611 - Paragraph 1 : If any part of a leased **thing** is lost due to reasons not attributable to the negligence of the lessee, the lessee may demand a reduction of the rent in proportion to the value of the lost part.*

spécifiques au sein des articles peut être un atout pour les identifier dans le cas où une requête spécifie l'un de ces termes.

FIGURE A.2 – Utilisation d'un vocabulaire diversifié permettant de couvrir diverses situations particulières.

*Article 219 - Paragraph 1 : An owner of **land** containing a **stream** including a **channel** or **moat** may not change the course or width of the same if the land on the other side is owned by others.*

Enfin, la dernière structure particulière au sein des articles est la référence vers d'autres articles. Ces références peuvent être réalisées de deux façons principales. La première est une citation exacte du paragraphe ou de l'article en question, tandis que la seconde est une citation relative par rapport à l'article en question. Cette référence relative peut se référer à un ou plusieurs paragraphes ou articles précédents (cf. figure A.3).

FIGURE A.3 – Exemple de références exactes et relatives pouvant apparaître dans les articles du code civil.

- Références exactes :
  - [...] *Paragraph 2 of Article 196 ;* [...]
  - [...] *the provision of Article 128 and Article 129* [...]
- Références relatives :
  - [...] *the provisions to the preceding paragraph* [...]
  - [...] *the ruling under paragraph 2 of the preceding Article* [...]
  - [...] *the provisions of the preceding* [...] *two paragraphs*
  - [...] *the provision of the preceding two Articles* [...]

La référence vers d'autres articles implique une façon particulière de concevoir et analyser le code civil. D'une façon linéaire, nous passons à une façon récursive. Pour notre approche, nous nous sommes focalisés sur ce processus et sur la manière de le manipuler afin d'y appliquer des méthodologies classiques en recherche d'information.

Les cas juridiques

Les données d'entraînement correspondent à un ensemble de cas juridiques. Elles ont été fournies dans un format XML dans lequel chaque cas juridique (requête) est associé avec un ou plusieurs articles du code civil et la réponse booléenne attendue pour ce cas donné (cf. figure A.4). Au total 578 cas ont été fournis. Il faut noter l'importance de la négation dans les formulations des différents cas. Cette spécificité

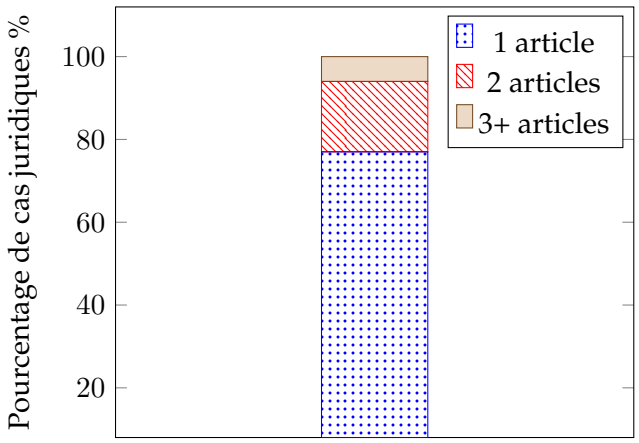
peut être déterminante dans la tâche 2 de cette compétition (non abordée dans cette annexe).

FIGURE A.4 – Exemple d'une donnée d'entraînement. Le label correspond à la réponse attendue et la balise <t1> aux articles attendus. Tandis que la balise <t2> correspond au cas juridique.

```
<?xml version="1.0" endoding="UTF-8"?>
  <pair id="H27-12-O" label="N">
    <t1>
      <article id="317"/>
      <article id="319"/>
      <article id="192"/>
    </t1>
    <t2>
        In cases where the hand luggage which a hotel guest brought to the
          hotel is not owned by him/her, the owner of the hotel may not
          exercise his/her statutory lien with respect to the hand luggage
          even if he/she believed without negligence that it was owned by
          the hotel guest.
    </t2>
  </pair>
```

Une analyse intéressante menée sur le corpus d'entraînement montre que plus de 75% des requêtes ne nécessitent qu'un seul article pour répondre à un cas juridique donné. Cette statistique peut jouer un rôle important sur la modélisation des approches. En effet, elle permet d'estimer la prise de risque de considérer plusieurs articles en sortie d'une chaîne de traitement (cf. figure A.5).

FIGURE A.5 – Pourcentages de cas juridiques nécessitant un, deux ou trois et plus articles pour pouvoir être élucidés.



A.1.2 Approches proposées dans ce domaine

Les travaux présentés lors des sessions précédentes de COLIEE proposent des chaînes de traitement principalement basées sur des modèles d'apprentissage. KIM, XU et

GOEBEL, 2015 ont soumis une approche reposant sur la méthode de *ranking SVM* issue des travaux de JOACHIMS, 2002b. Au sein de leur approche, chaque article du code civil est considéré comme un vecteur de caractéristiques. Ce dernier contient trois entrées représentées par une caractéristique lexicale, syntaxique et une caractéristique basée sur la somme des *tf-idf* entre les termes en commun de la requête et des articles. Ce type de représentation vectorielle est commune en recherche d'information. Une liste de caractéristiques pertinentes [2] pour la recherche de documents Web est décrite dans l'article de QIN et LIU, 2013. Ainsi, cette phase d'apprentissage a pour objectif de découvrir l'article le plus pertinent pour répondre à une requête donnée. Leur approche a obtenu une F-mesure de 0,55 lors de COLIEE 2015. Il est important de noter que les auteurs recherchaient un seul article par requête. Ainsi, ils ne tenaient pas compte de l'imbrication des articles.

À cela, CARVALHO et al., 2015 proposent une approche capable de considérer plusieurs articles selon différents critères spécifiques. Leur méthode est basée sur une adaptation de la méthode *tf-idf* aux *n-grammes*. L'équation A.1 expose la fonction de score dans laquelle I_q et I_{art} sont des paramètres de significativité respectivement associés aux *n-grammes* de la question et des articles, q_ng_set l'ensemble des *n-grammes* pour la question et art_ng_set l'ensemble des *n-grammes* pour l'article a.

$$score(a) = \frac{\sum_{\forall t} idf(t)}{I_q \times \mid q\_ng\_set \mid + I_{art} \times \mid art\_ng\_set \mid}, t \in \{q\_ng\_set \cap art\_ng\_set\} \tag{A.1}$$

$idf(t)$ est l'*Inverse Document Frequency* pour le *n-gramme* t sur tous les articles de la collection (cf. équation A.2). Avec N le nombre total d'articles et df_t le nombre d'articles dans lequel t apparaît.

$$idf(t) = log \frac{N}{df_t} \tag{A.2}$$

Ainsi, par le biais de la fonction de score, ils vont pouvoir attribuer une pertinence à chaque article dans le but de conserver les 10 meilleurs articles. Ensuite, ils réalisent une phase de filtrage. Cette dernière vérifie si le score du meilleur article dépasse un seuil s_1, si c'est le cas alors tous les articles référés par ce dernier (références vers d'autres articles) et dont les scores excèdent un seuil s_2 sont conservés pour la réponse finale. Leur méthode avait obtenu une f-mesure de 0,51 lors de COLIEE 2016. À noter que les seuils s_1 et s_2 ainsi que les paramètres de la fonction de score et le nombre n d'un *n-gramme* ont été déterminés de manière empirique.

2. https ://www.microsoft.com/en-us/research/project/mslr/

A.2 Description de la chaîne de traitement

L'approche que nous proposons se base sur une décomposition du code civil en vue d'une restructuration hybride des articles, par la suite soumis à la fonction de pondération BM25. Cette fonction a été choisie après une phase expérimentale réalisée avec Terrier [3] (cf. figure A.6).

FIGURE A.6 – Structuration des fichiers xml des articles du code civil et des requêtes en entrée de Terrier.

Articles

<DOC>	Balise de début d'un article
<DOCNO> </DOCNO>	Identifiant unique de l'article
<TEXT> </TEXT>	Contenu textuel de l'article
</DOC>	Balise de fin de l'article

Requêtes

<top>	Balise de début d'une requête
<num></num>	Identifiant unique de la requête
<title></title>	Contenu de la requête
</top>	Balise de fin de la requête

Cette phase a consisté en l'évaluation empirique de plusieurs mesures de pondération proposées par Terrier sur les requêtes et les articles du code civil. Au total 17 mesures ont été testées telles que DLH (*hyper-geometric DFR model*), IFB2 (*Inverse Term Frequency model for randomness*) ou la version de Lemur du *tf-idf* (ZHAI, 2001).

A.2.1 Méthode de pondération : BM25

La méthode du BM25 est une méthode *tf-idf like* proposée par ROBERTSON et KAREN SPÄRCK, 1976. Elle exprime le poids w d'un document D en fonction de la requête Q selon l'équation A.3 :

$$w(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{tf(q_i,D)(k_1+1)}{tf(q_i,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \tag{A.3}$$

dans laquelle q_i est un mot de Q, $tf(q_i, D)$ est la fréquence d'apparition de q_i dans D, $|D|$ la longueur du document D (nombre de mots), $avgdl$ est la moyenne des longueurs des documents dans la collection et k_1 et b sont des paramètres avancés du modèle.

$$IDF(q_i) = log \frac{N - n(q_i) + 0,5}{n(q_i) + 0,5} \tag{A.4}$$

3. http ://terrier.org/

Au sein de l'équation A.4, N est le nombre total de documents dans la collection et $n(q_i)$ est le nombre de documents contenant q_i (ROBERTSON et WALKER, 1997).

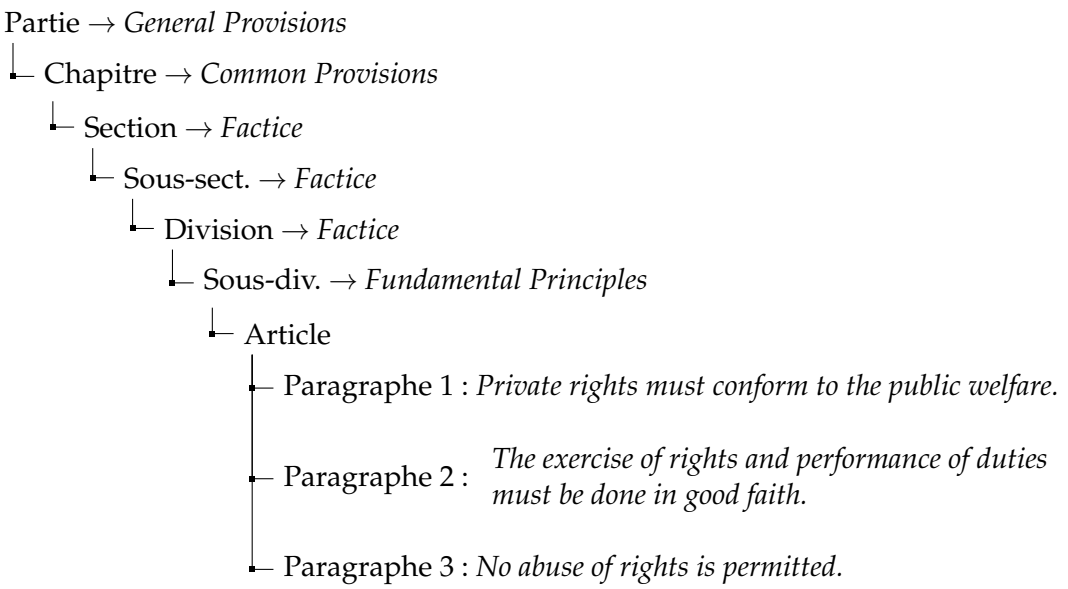
A.2.2 Restructuration du code civil et évaluation des articles candidats

La philosophie de la chaîne de traitement est de s'approprier la structure du code civil en tenant compte des liens d'implication entre les articles afin d'affiner les performances du BM25. Pour cela, nous avons développé une approche reposant sur trois principales étapes : génération de nouveaux documents, pondération de ces documents par BM25 et sélection des meilleurs résultats.

Génération de documents hybrides

Une première phase a consisté à restructurer le code civil en tenant compte des différentes références au sein des articles. Dans l'objectif de faciliter cette phase, elle est précédée par une normalisation de l'architecture du code civil. Cette procédure permet de combler les manques au niveau de la structure du code civil en générant des sections factuelles afin de standardiser l'origine de chaque article et obtenir ainsi une structure arborescente constante du code civil (cf. figure A.7). Cette normalisation permet d'améliorer la précision lorsqu'il s'agit d'interpréter des références relatives au sein des articles. En effet, la navigation au sein de l'arborescence allège le processus de correspondance entre les articles.

FIGURE A.7 – Construction d'une arborescence à partir du code civil.



Concernant la génération de nouveaux documents, la procédure considère à la fois les références directes et indirectes entre les articles. La structure de ces documents diffère selon trois modalités :

— l'article simple ;
— la décomposition de l'article en plusieurs paragraphes ;
— le mélange entre deux articles dont l'un fait référence à l'autre.

Les idées sous-jacentes derrière ces trois modalités sont à la fois de considérer les implications entre articles via la construction d'articles hybrides et de préciser les résultats d'une recherche au travers de la décomposition des articles en paragraphes. Par conséquent, le code civil généré est plus volumineux. Les documents générés ont également subi une phase d'élimination des *stop-word*, de racinisation et d'élimination d'un sous-ensemble de termes redonnant dans le code civil ne dissociant pas les articles [4].

BM25 et sélection des articles

Une fois la génération des nouveaux documents réalisée, ils sont soumis à la fonction de pondération BM25. Cette étape permet de classifier les documents du plus au moins pertinents. À partir de ce point plusieurs algorithmes ont été élaborés pour rechercher les articles les plus pertinents. Par exemple, nous avons expérimenté une approche basée sur une analyse avancée des k meilleurs résultats du BM25. L'évaluation portait sur la différence de poids entre le premier résultat et un document situé dans l'intervalle $[2, k]$. Si cette différence est supérieure à un seuil défini manuellement alors le document en question est retourné. Un autre exemple de méthode expérimentée a été la mise en place d'un apprentissage par SVM en tenant compte de différentes caractéristiques à l'instar des travaux de QIN et LIU, 2013. Les caractéristiques exploitées pour représenter un document d en fonction d'une requête r ont été les suivantes :

— Nombre de mots en commun entre r et q et sa normalisation par rapport au nombre de mots de r.
— Nombre de mots de d.
— La somme, le minimum, le maximum, la moyenne et la variance à la fois des fréquences de mots (*tf*) de d et des *tf-idf* de d.
— Poids du BM25.

Toutefois, l'approche la plus efficiente est l'utilisation du premier résultat du BM25 pouvant potentiellement contenir plusieurs articles. Les résultats obtenus sont présentés dans la section suivante.

4. *provision, shall, apply, mutatis, mutandis, case, article, articles, act, paragraph, paragraphs, no, preceding, precede, provision*

A.3 Expérimentations et résultats

Les conditions expérimentales sont les mêmes tout au long des expérimentations. Nous utilisons le même ensemble de paramètres pour le BM25 soit $b = 0,75$ et $k_1 = 1,2$ correspondant aux paramètres par défaut dans *terrier*.

A.3.1 Métriques d'évaluations

Les résultats sont évalués selon les métriques de performance proposées par la conférence. Ces dernières sont les métriques classiques en recherche d'information : Précision, Rappel et F-mesure (cf. figure A.8).

FIGURE A.8 – Métriques d'évaluation établies par l'organisation de COLIEE 2017.

$$\text{Précision} \quad \frac{\text{Nombre d'articles correctement retrouvés pour toutes les requêtes}}{\text{Nombre d'articles retrouvés pour toutes les requêtes}}$$

$$\text{Rappel} \quad \frac{\text{Nombre d'articles correctement retrouvés pour toutes les requêtes}}{\text{Nombre d'articles attendus pour toutes les requêtes}}$$

$$\text{F-mesure} \quad \frac{(2 \times \text{Précision} \times \text{Rappel})}{(\text{Précision} + \text{Rappel})}$$

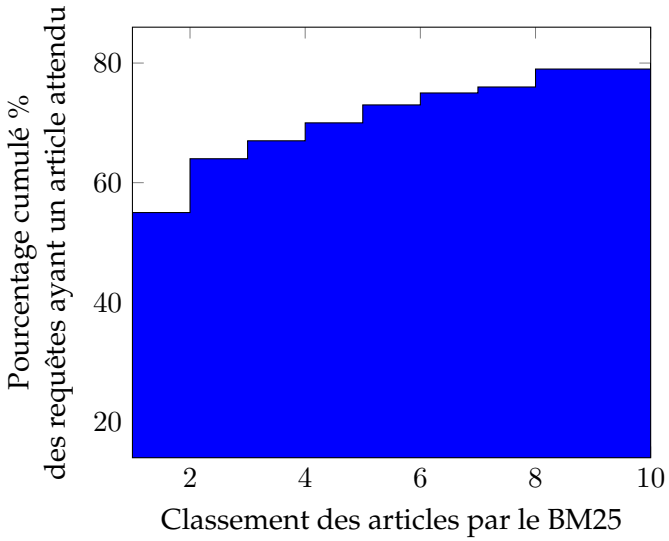
A.3.2 *Baseline*

La *baseline* adoptée est basée sur les performances du BM25 à partir des articles non modifiés du code civil. Les résultats considérés sont les articles avec le poids le plus élevé pour chaque requête. Cette *baseline* permet à la fois d'observer les capacités du BM25 sur les données non transformées et permet de comparer les performances liées à la restructuration du code civil. Après l'analyse des résultats obtenus avec la *baseline*, il est intéressant d'observer la répartition des articles attendus au sein des k meilleurs résultats proposés par l'approche (cf. figure A.9). Cette répartition montre le pourcentage cumulé des requêtes ayant un article correct dans les k meilleurs résultats du BM25. Ainsi, on observe que 55% des requêtes ont un article attendu en première position et que 73% des requêtes ont au moins un article attendu dans les cinq premiers résultats du BM25. Ces informations couplées à celles sur le nombre d'articles attendus par requête sont précieuses pour la conception de l'approche car elles permettent d'estimer les risques de rechercher et considérer plusieurs articles pour une requête dans les résultats du BM25.

A.3.3 Résultats

La méthode développée a été évaluée sur les cas juridiques d'entraînement puis testée sur ceux de la compétition COLIEE 2017.

FIGURE A.9 – Pourcentage cumulé des requêtes ayant un article attendu au sein des k meilleurs classements du BM25.



Résultats sur les données d'entraînement

Le tableau A.2 récapitule à la fois les performances du BM25 sur le code civil non modifié et les performances de notre approche sur les 578 cas juridiques constituant le jeu d'entraînement.

TABLEAU A.2 – Performance de la *baseline* et de notre approche (KID17) sur l'ensemble des données d'entraînement.

	Précision	Rappel	F-mesure
Baseline	62,8	48,4	54,7
KID17	64,9	51,4	57,4

Résultats sur les données de la compétition 2017

Le tableau A.3 récapitule les performances des participants sur les données de la compétition. Ces données étaient constituées de 80 cas juridiques inconnus.

A.4 Conclusion

Cette annexe explore les problématiques rencontrées dans les domaines de la recherche et l'implication d'information appliqués au droit. Nous avons proposé une modélisation centrée sur une réorganisation du code civil et la génération d'articles hybrides permettant de tenir compte des références directes et indirectes entre les articles. Ce code civil modifié a été ensuite soumis à la méthode de pondération BM25.

TABLEAU A.3 – Résultats pour la tâche 1 de COLIEE 2017. KID17 représente les résultats obtenus par notre approche.

Identifiant	Rappel	Précision	F-mesure	Langage
iLis7-1	**0,734940**	**0,554545**	**0,632124**	English
JNLP1-RT	0,689655	0,545455	0,609137	English
JNLP1-R	0,686047	0,536364	0,602041	English
KID17	0,703704	0,518182	0,596859	English
iLis7-2	0,654762	0,500000	0,567010	English
UA-TFIDF	0,666667	0,472727	0,553191	English
HUKB-1	0,658228	0,472727	0,550265	Japanese
HUKB-3	0,551402	0,536364	0,543779	Japanese
HUKB-2	0,586957	0,490909	0,534653	Japanese
JAISTNLP2-1a-norerank	0,628205	0,445455	0,521277	English
JAISTNLP2-1b-rerank	0,615385	0,436364	0,510638	English
UA-LM	0,602564	0,427273	0,500000	English
NOR17	0,462185	0,500000	0,480349	English
JNLP1-T	0,500000	0,354545	0,414894	English
VNPT	0,430556	0,281818	0,340659	English
KIS-IE-NM	0,346154	0,245455	0,287234	Japanese
KIS-IE-M	0,263158	0,272727	0,267857	Japanese

Nous avons observé une amélioration de performance par le biais de cette réorganisation, permettant d'accéder à la quatrième place de cette compétition sur dix-sept soumissions. L'ensemble des implémentations réalisé au cours de cette compétition est disponible au téléchargement à l'adresse suivante : *https ://github.com/PAJEAN/Coliee2017*.

Cette compétition m'a permis de mettre en pratique, sur des types de données différentes, certaines approches développées lors de la mise en place de la chaîne de traitement présentée dans ce manuscrit. En effet, certaines implémentations ont été mises à profit pour nettoyer et normaliser les données textuelles. De plus, l'expertise acquise dans le domaine de l'apprentissage automatique, grâce au développement du module de détection de l'incertitude linguistique, a été précieuse pour explorer différentes pistes de recherche lors de cette compétition.

Annexe B

La théorie des fonctions de croyance

Le cadre théorique des fonctions de croyance fournit des outils mathématiques afin de rechercher la réponse la plus probable à une question donnée en évaluant la croyance et la plausibilité des solutions possibles. Introduisons ce principe sur un exemple de dégustation de vin réalisé par un ensemble de dégustateurs. Cet exemple est inspiré de celui de BAUDRIT, 2005. Les pourcentages associés aux avis des dégustateurs sont rapportés ci-dessous :

- 20% des dégustateurs sont certains d'avoir perçu un goût de cassis (C)
- 30% des dégustateurs sont certains d'avoir perçu un goût de mûre (M)
- 10% des dégustateurs sont certains d'avoir perçu un goût de groseille (G)
- 10% des dégustateurs sont certains d'avoir perçu un goût de fruits rouges FR (comprenant les groseilles mais également les framboises F pour cet exemple)
- 10% des dégustateurs sont certains d'avoir perçu un goût de fruits noirs FN (comprenant le cassis et les mûres)
- 20% des dégustateurs ne se sont pas prononcés

À partir de cette information, nous pouvons définir une distribution de masse m telle que :

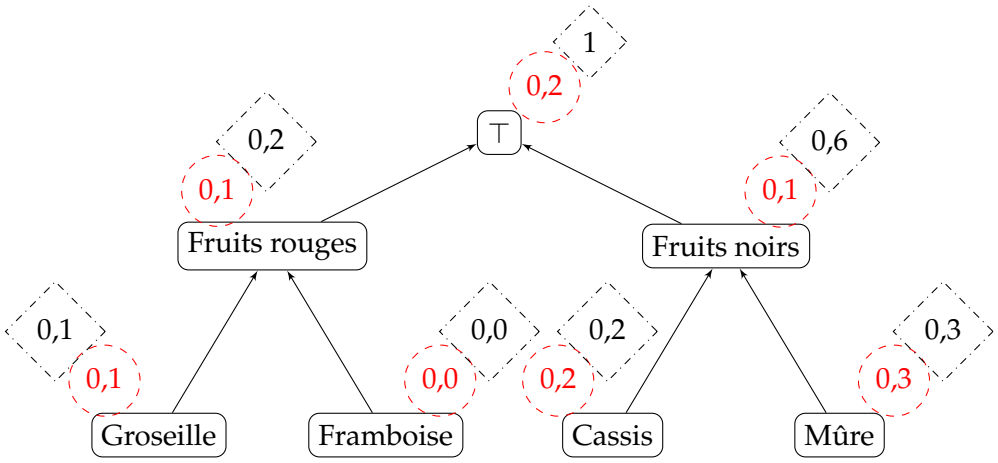
$$
\begin{aligned}
&m(C) = 0,2 \\
&m(M) = 0,3 \\
&m(C \cup M) = m(FN) = 0,1 \\
&m(G) = 0,1 \\
&m(F) = 0,0 \\
&m(G \cup F) = m(FR) = 0,1 \\
&m(C \cup M \cup G \cup F) = 0,2
\end{aligned}
$$

Maintenant, si on cherche à évaluer la probabilité qu'un dégustateur décèle un goût de fruits noirs (Cassis ou Mûre), on a alors :

— Au minimum : 20%+30%+10% des dégustateurs qui trouveront cette saveur. Ceci forme alors une borne minimale de probabilité Bel, fixée à 60% (nommée croyance ou crédibilité).
— Au maximum : 20%+30%+10% auxquels il convient d'ajouter ceux qui ne se sont pas prononcés. On obtient alors une borne maximale de probabilité Pl, fixée à 80% (nommée Plausibilité).

Par conséquent, la probabilité P_{FN} qu'un dégustateur perçoive un goût de fruits noirs est : $60\% \leq P_{FN} \leq 80\%$. La figure B.1 illustre l'exemple précédent en tenant compte uniquement de la propagation des masses au sein d'un ordre taxonomique hiérarchisant les différents concepts de l'exemple.

FIGURE B.1 – Visualisation du problème sous la forme graphique. Les ronds en pointillés représentent la masse m (c'est le nombre d'occurrences du nœud divisé par le nombre total d'observations) associée à chaque nœud du graphe. Les losanges représentent la croyance (au sens de la théorie des fonctions de croyance) pour chaque nœud.



L'objectif de la théorie des fonctions de croyance est de permettre de combiner des preuves distinctes pour calculer un encadrement de probabilité d'un événement A borné par la croyance (cf. équation B.3) et la plausibilité (cf. équation B.4) :

$$Croyance(A) \leq P(A) \leq Pl(A) \tag{B.1}$$

Pour que l'on puisse se placer dans ce cadre formel, il est nécessaire de travailler à prédicat donné (dans l'exemple, les dégustateurs perçoivent un goût de) et la question doit être factuelle, i.e., une réponse unique et la plus spécifique possible est attendue. Dans le cadre de nos travaux et au regard d'un sujet (ou objet) et d'un prédicat fixés pour une question factuelle, la masse associée à chaque déclaration s serait calculée à partir des fréquences d'observation (cf. équation B.2).

$$m(s) = \frac{Freq(s)}{\sum_{s' \in \mathcal{S}} Freq(s')} \tag{B.2}$$

La croyance, quant à elle, serait définie comme la somme des masses de l'ensemble des descendants $\mathcal{D}$ de s avec $\mathcal{D}(s) = \{s' \in \mathcal{S} \mid s' \preceq s\}$. Cette étape est le produit d'une propagation *bottom-up*.

$$Croyance(s) = \sum_{s' \in \mathcal{D}(s)} m(s') \tag{B.3}$$

Finalement, la plausibilité serait définie comme la somme des masses des éléments ne contredisant pas la déclaration s.

$$Pl(s) = \sum_{s' \cap s \neq \emptyset} m(s') \tag{B.4}$$

Ainsi à sujet et prédicat fixés, la théorie des fonctions de croyances peut être appliquée à notre problématique. Cette dernière pourrait être intéressante au travers d'une extension de la chaîne de traitement exploitant la notion de plausibilité permettant la découverte de déclarations plus spécifiques. Par conséquent, cela étendrait la définition de notre notion de découverte jusqu'à présent définie comme la génération de déclarations abstraites.

Bibliographie

ABDEL-HAMEED, Ahmad A, Eiad A AL-FARIS, Ibrahim A ALORAINY et Mohammed O AL-RUKBAN (2005). « The criteria and analysis of good multiple choice questions in a health professional setting. » In : *Saudi medical journal* 26.10, p. 1505–1510.

ACKOFF, Russell L (1989). « From data to wisdom ». In : *Journal of applied systems analysis* 16.1, p. 3–9.

AGGARWAL, Charu C et ChengXiang ZHAI (2012). « A survey of text classification algorithms ». In : *Mining text data*, p. 163–222.

AGIRRE, Eneko, Mona DIAB, Daniel CER et Aitor GONZALEZ-AGIRRE (2012). « Semeval-2012 task 6 : A pilot on semantic textual similarity ». In : *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, p. 385–393.

ALMASRI, Mohannad, Catherine BERRUT et Jean-Pierre CHEVALLET (2013). « Wikipedia-based semantic query enrichment ». In : *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*. ACM, p. 5–8.

ANDREEWSKY, Evelyne et Danièle BOURCIER (2000). « Abduction in language interpretation and law making ». In : *Kybernetes* 29, p. 836–845.

ARONSON, Alan R (2001). « Effective mapping of biomedical text to the UMLS Metathesaurus : the MetaMap program. » In : *Proceedings of the AMIA Symposium*. American Medical Informatics Association, p. 17–21.

ARONSON, Alan R et François-Michel LANG (2010). « An overview of MetaMap : historical perspective and recent advances ». In : *Journal of the American Medical Informatics Association* 17.3, p. 229–236.

AUER, Sören, Christian BIZER, Georgi KOBILAROV, Jens LEHMANN, Richard CYGANIAK et Zachary IVES (2007). « Dbpedia : A nucleus for a web of open data ». In : *The semantic web*, p. 722–735.

BAADER, Franz (2003). *The description logic handbook : Theory, implementation and applications*. Cambridge university press.

BANERJEE, Satanjeev et Ted PEDERSEN (2002). « An adapted Lesk algorithm for word sense disambiguation using WordNet ». In : *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, p. 136–145.

BANKO, Michele et Eric BRILL (2001). « Scaling to very very large corpora for natural language disambiguation ». In : *Proceedings of the 39th annual meeting on association for computational linguistics*. Association for Computational Linguistics, p. 26–33.

BATET, Montserrat, Sébastien HARISPE, Sylvie RANWEZ, David SÁNCHEZ et Vincent RANWEZ (2014). « An information theoretic approach to improve semantic similarity assessments across multiple ontologies ». In : *Information Sciences* 283, p. 197–210.

BAUDRIT, Cédric (2005). « Représentation et propagation de connaissances imprecises et incertaines : Application à l'évaluation des risques liés aux sites et aux sols pollués ». Thèse de doct. Université Toulouse III.

BAZIZ, Mustapha, Nathalie AUSSENAC-GILLES et Mohand BOUGHANEM (2003). « Exploitation des Liens Sémantiques pour l'Expansion de Requêtes dans un Système de Recherche d'Information. » In : *INFORSID*, p. 121–134.

BELLINGER, Gene, Durval CASTRO et Anthony MILLS (2004). « Data, information, knowledge, and wisdom ». In : URL : https://pdfs.semanticscholar.org/b553/609347b9b8bc5698ccaef823b3acc1128dd7.pdf.

BELLOT, Patrice, L. BONNEFOY, V. BOUVIER, F. DUVERT et YM K KIM (2014). *Large scale text mining approaches for information retrieval and extraction*. Springr.

BEN ABACHA, Asma (2012). « Recherche de réponses précises à des questions médicales : le système de questions-réponses MEANS ». Thèse de doct. Université Paris Sud-Paris XI, p. 1–162.

BEN ABACHA, Asma et Pierre ZWEIGENBAUM (2015). « MEANS : A medical question-answering system combining NLP techniques and semantic Web technologies ». In : *Information Processing & Management* 51.5, p. 570–594.

BERETTA, Valentina, Sébastien HARISPE, Sylvie RANWEZ et Isabelle MOUGENOT (2016). « How Can Ontologies Give You Clue for Truth-Discovery ? An Exploratory Study ». In : *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*. ACM, p. 15.

BERNERS-LEE, Tim, James HENDLER et Ora LASSILA (2001). « The semantic web ». In : *Scientific american* 284.5, p. 28–37.

BLANCO, Eduardo et Dan MOLDOVAN (2011). « Semantic representation of negation using focus detection ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*. Association for Computational Linguistics, p. 581–589.

BODENREIDER, Olivier (2004). « The unified medical language system (UMLS) : integrating biomedical terminology ». In : *Nucleic acids research* 32.1, p. 267–270.

BOLLACKER, Kurt, Colin EVANS, Praveen PARITOSH, Tim STURGE et Jamie TAYLOR (2008). « Freebase : a collaboratively created graph database for structuring human knowledge ». In : *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, p. 1247–1250.

BORDEA, Georgeta, Paul BUITELAAR, Stefano FARALLI et Roberto NAVIGLI (2015). « Semeval-2015 task 17 : Taxonomy extraction evaluation (texeval) ». In : *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 902–910.

BOS, Johan et Katja MARKERT (2005). « Recognising textual entailment with logical inference ». In : *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, p. 628–635.

BOUCHON-MEUNIER, Bernadette (1995). *La logique floue et ses applications*. Addison-Wesley France.

BOUCHON-MEUNIER, Bernadette et Hung T NGUYEN (1996). *Les incertitudes dans les systèmes intelligents*. Presses universitaires de France.

BRIN, Sergey (1998). « Extracting patterns and relations from the world wide web ». In : *International Workshop on The World Wide Web and Databases*. Springer, p. 172–183.

BRUNET, Etienne (2002). « Le lemme comme on l'aime ». In : *JADT*, p. 221–232.

BUDANITSKY, Alexander et Graeme HIRST (2001). « Semantic distance in WordNet : An experimental, application-oriented evaluation of five measures ». In : *Workshop on WordNet and other lexical resources*. T. 2, p. 2–2.

BUITELAAR, Paul et Philipp CIMIANO (2008). *Ontology learning and population : bridging the gap between text and knowledge*. T. 167. Ios Press.

BUITELAAR, Paul, Philipp CIMIANO et Bernardo MAGNINI (2005). « Ontology learning from text : An overview ». In : *Ontology learning from text : Methods, evaluation and applications* 123, p. 3–12.

BUNESCU, Razvan C et Raymond J MOONEY (2005). « A shortest path dependency kernel for relation extraction ». In : *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, p. 724–731.

CARLSON, Andrew, Justin BETTERIDGE, Bryan KISIEL, Burr SETTLES, Estevam R HRUSCHKA JR et Tom M MITCHELL (2010). « Toward an Architecture for Never-Ending Language Learning. » In : *AAAI*. T. 5, p. 3.

CARVALHO, Danilo S, Minh-Tien NGUYEN, Chien-Xuan TRAN et Minh-Le NGUYEN (2015). « Lexical-Morphological Modeling for Legal Text Analysis ». In : *JSAI International Symposium on Artificial Intelligence*. Springer, p. 295–311.

CATELLIN, Sylvie (2004). « L'abduction : une pratique de la découverte scientifique et littéraire ». In : *Hermès, La Revue* 2, p. 179–185.

CHAN, Yee Seng et Dan ROTH (2011). « Exploiting syntactico-semantic structures for relation extraction ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*. Association for Computational Linguistics, p. 551–560.

CHEN, Lin et Barbara DI EUGENIO (2010). « A lucene and maximum entropy model based hedge detection system ». In : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*. Association for Computational Linguistics, p. 114–119.

CHEN, Yi-Wei et Chih-Jen LIN (2006). « Combining SVMs with various feature selection strategies ». In : *Feature extraction*, p. 315–324.

CHOMSKY, Noam (1964). *Aspects of the Theory of Syntax*. T. 11. MIT press.

CIMIANO, P et J VÖLKER (2005). « Text2Onto. Natural language processing and information systems ». In : *10th International Conference on Applications of Natural Language to Information Systems, NLDB*, p. 15–17.

COLLINS, Michael John (1996). « A new statistical parser based on bigram lexical dependencies ». In : *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 184–191.

CORTES, Corinna et Vladimir VAPNIK (1995). « Support-vector networks ». In : *Machine learning* 20.3, p. 273–297.

COULET, Adrien (2008). « Construction et utilisation d'une Base de Connaissances pharmacogénomique pour l'intégration de données et la découverte de connaissances ». Thèse de doct. Université Henri Poincaré, Nancy 1.

CRAVEN, Mark, Johan KUMLIEN et al. (1999). « Constructing biological knowledge bases by extracting information from text sources. » In : *ISMB*. T. 1999, p. 77–86.

CRUZ, N, Maite TABOADA et Ruslan MITKOV (2015). « A machine learning approach to negation and speculation detection ». In : *Association for Information Science and Technology*.

CULOTTA, Aron et Jeffrey SORENSEN (2004). « Dependency tree kernels for relation extraction ». In : *Proceedings of the 42nd annual meeting on association for computational linguistics*. Association for Computational Linguistics, p. 423.

DAGAN, Ido, Oren GLICKMAN et Bernardo MAGNINI (2006). « The PASCAL recognising textual entailment challenge ». In : *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*. Springer, p. 177–190.

DAIBER, Joachim, Max JAKOB, Chris HOKAMP et Pablo N MENDES (2013). « Improving efficiency and accuracy in multilingual entity extraction ». In : *Proceedings of the 9th International Conference on Semantic Systems*. ACM, p. 121–124.

DALHOUMI, Sami, Gérard DRAY, Jacky MONTMAIN, Gérard DEROSIÈRE et Stéphane PERREY (2015). « An adaptive accuracy-weighted ensemble for inter-subjects classification in brain-computer interfacing ». In : *Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on*. IEEE, p. 126–129.

DE GIACOMO, Giuseppe et Maurizio LENZERINI (1996). « TBox and ABox reasoning in expressive description logics. » In : *KR* 96, p. 316–327.

DE SAUSSURE, Ferdinand (1916). *Cours de linguistique générale : Édition critique*. T. 1. Otto Harrassowitz Verlag.

DELPORTE, Julien (2013). « Factorisation Matricielle, Application à la Recommandation Personnalisée de Préférences ». Thèse de doct. Institut National des Sciences Appliquées de Rouen, p. 1–124.

DINH, Duy et Lynda TAMINE (2010). « Recherche d'information sémantique dans les documents biomédicaux : approche basée sur le sens précis des concepts ». In : *INFormatique des Organisations et Systemes d'Information et de Decision, INFORSID 2010*, p. 261–274.

DONG, Xin, Evgeniy GABRILOVICH, Geremy HEITZ, Wilko HORN, Ni LAO, Kevin MURPHY, Thomas STROHMANN, Shaohua SUN et Wei ZHANG (2014). « Knowledge vault : A web-scale approach to probabilistic knowledge fusion ». In : *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 601–610.

DONG, Xin Luna, Laure BERTI-EQUILLE et Divesh SRIVASTAVA (2009). « Integrating conflicting data : the role of source dependence ». In : *Proceedings of the VLDB Endowment* 2.1, p. 550–561.

DONG, Xin Luna, Evgeniy GABRILOVICH, Kevin MURPHY, Van DANG, Wilko HORN, Camillo LUGARESI, Shaohua SUN et Wei ZHANG (2015). « Knowledge-based trust : Estimating the trustworthiness of web sources ». In : *Proceedings of the VLDB Endowment* 8.9, p. 938–949.

DRAGOS, Valentina (2013). « An ontological analysis of uncertainty in soft data ». In : *Information Fusion (FUSION), 2013 16th International Conference on*. IEEE, p. 1566–1573.

DRUMOND, Lucas et Rosario GIRARDI (2008). « A Survey of Ontology Learning Procedures. » In : *WONTO* 427, p. 1–13.

ESULI, Andrea et Fabrizio SEBASTIANI (2006). « SENTIWORDNET : A high-coverage lexical resource for opinion mining ». In : *Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR)*.

ETZIONI, Oren, Michael CAFARELLA, Doug DOWNEY, Stanley KOK, Ana-Maria POPESCU, Tal SHAKED, Stephen SODERLAND, Daniel S WELD et Alexander YATES (2004). « Web-scale information extraction in knowitall :(preliminary results) ». In : *Proceedings of the 13th international conference on World Wide Web*. ACM, p. 100–110.

ETZIONI, Oren, Anthony FADER, Janara CHRISTENSEN, Stephen SODERLAND et Mausam MAUSAM (2011). « Open Information Extraction : The Second Generation. » In : *IJCAI*. T. 11, p. 3–10.

FADER, Anthony, Stephen SODERLAND et Oren ETZIONI (2011). « Identifying relations for open information extraction ». In : *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, p. 1535–1545.

FAN, James, Aditya KALYANPUR, David C GONDEK et David A FERRUCCI (2012). « Automatic knowledge extraction from documents ». In : *IBM Journal of Research and Development* 56.3.4, p. 5.

FARKAS, Richárd, Veronika VINCZE, György MÓRA, János CSIRIK et György SZARVAS (2010). « The CoNLL-2010 shared task : learning to detect hedges and their scope in natural language text ». In : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*. Association for Computational Linguistics, p. 1–12.

FERRET, Olivier et Hervé LE BORGNE (2016). « Utilisation des relations d'une base de connaissances pour la désambiguïsation d'entités nommées ». In : *TALN*.

FERRUCCI, David et al. (2010). « Building Watson : An overview of the DeepQA project ». In : *AI magazine* 31.3, p. 59–79.

FERRUCCI, David A (2012). « Introduction to "this is watson" ». In : *IBM Journal of Research and Development* 56.3.4, p. 1–1.

FERSON, Scott, Jason O'RAWE, Andrei ANTONENKO, Jack SIEGRIST, James MICKLEY, Christian C LUHMANN, Kari SENTZ et Adam M FINKEL (2015). « Natural language of uncertainty : numeric hedge words ». In : *International Journal of Approximate Reasoning* 57, p. 19–39.

FORMAN, George (2003). « An extensive empirical study of feature selection metrics for text classification ». In : *Journal of machine learning research* 3.Mar, p. 1289–1305.

FRANK, Anette, Hans-Ulrich KRIEGER, Feiyu XU, Hans USZKOREIT, Berthold CRYSMANN, Brigitte JÖRG et Ulrich SCHÄFER (2007). « Question answering from structured knowledge sources ». In : *Journal of Applied Logic* 5.1, p. 20–48.

FUCHS, Catherine (2008). « L'incertitude interprétative dans l'activité de langage ». In : *Actes de savoirs* 5, p. 41–57.

GALÁRRAGA, Luis Antonio, Christina TEFLIOUDI, Katja HOSE et Fabian SUCHANEK (2013). « AMIE : association rule mining under incomplete evidence in ontological knowledge bases ». In : *Proceedings of the 22nd international conference on World Wide Web*. ACM, p. 413–422.

GANTER, Viola et Michael STRUBE (2009). « Finding hedges by chasing weasels : Hedge detection using Wikipedia tags and shallow linguistic features ». In : *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, p. 173–176.

GASPAR, Paulo, Jaime CARBONELL et José Luís OLIVEIRA (2012). « On the parameter optimization of Support Vector Machines for binary classification ». In : *Journal of Integrative Bioinformatics (JIB)* 9.3, p. 33–43.

GELBUKH, Alexander, Grigori SIDOROV et San-Yong HAN (2003). « Evolutionary approach to natural language word sense disambiguation through global coherence optimization. » In : *WSEAS Transactions on Computers* 2.1, p. 257–265.

GELBUKH, Alexander, Grigori SIDOROV et Sang-Yong HAN (2005). « On some optimization heuristics for lesk-like wsd algorithms ». In : *Lecture Notes in Computer Science* 3513, p. 402–406.

GENNARI, John H, Mark A MUSEN, Ray W FERGERSON, William E GROSSO, Monica CRUBÉZY, Henrik ERIKSSON, Natalya F NOY et Samson W TU (2003). « The evolution of Protégé : an environment for knowledge-based systems development ». In : *International Journal of Human-computer studies* 58.1, p. 89–123.

GEORGESCUL, Maria (2010). « A hedgehop over a max-margin framework using hedge cues ». In : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*. Association for Computational Linguistics, p. 26–31.

GREEN JR, Bert F, Alice K WOLF, Carol CHOMSKY et Kenneth LAUGHERY (1961). « Baseball : an automatic question-answerer ». In : *Western joint IRE-AIEE-ACM computer conference*. ACM, p. 219–224.

GREFENSTETTE, Gregory (2015). « INRIASAC : Simple hypernym extraction methods ». In : *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 911–914.

HAMDAN, Hussam (2015). « Sentiment Analysis in Social Media ». Thèse de doct. Université d'Aix-Marseille, p. 1–165.

HAN, Xianpei et Jun ZHAO (2009). « NLPR_KBP in TAC 2009 KBP Track : A Two-Stage Method to Entity Linking. » In : *TAC*.

HARISPE, Sébastien, Sylvie RANWEZ, Stefan JANAQI et Jacky MONTMAIN (2015). « Semantic similarity from natural language and ontology analysis ». In : *Synthesis Lectures on Human Language Technologies* 8.1, p. 1–254.

HARRIS, Roand Johnston (1962). « An Experimental Inquiry Into the Functions and Value of Formal Grammar in the Teaching of English : With Special Reference to the Teaching of Correct Written English to Children Aged Twelve to Fourteen ». Thèse de doct. University of London.

HEARST, Marti A (1992). « Automatic acquisition of hyponyms from large text corpora ». In : *Proceedings of the 14th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, p. 539–545.

HERRERA, Jesús, Anselmo PENAS et Felisa VERDEJO (2004). « Question answering pilot task at CLEF 2004 ». In : *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, p. 581–590.

HOEK, Remko van, Håkan ARONSSON, Gyöngyi KOVÁCS et Karen M SPENS (2005). « Abductive reasoning in logistics research ». In : *International Journal of Physical Distribution & Logistics Management* 35.2, p. 132–144.

HORN, Alfred (1951). « On sentences which are true of direct unions of algebras ». In : *The Journal of Symbolic Logic* 16.01, p. 14–21.

HORRIDGE, Matthew, Holger KNUBLAUCH, Alan RECTOR, Robert STEVENS et Chris WROE (2004). « A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools Edition 1.0 ». In : *University of Manchester*.

HYLAND, Ken (1998). *Hedging in scientific research articles*. T. 54. John Benjamins Publishing.

JACQUEMIN, Christian (1999). « Syntagmatic and paradigmatic representations of term variation ». In : *The 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p. 341–348.

JEAN, Pierre-Antoine, Sébastien HARISPE, Patrice BELLOT, Sylvie RANWEZ et Jacky MONTMAIN (2015). « Gérer l'incertitude lors de l'extraction de relations et lors de l'inférence de nouvelles connaissances ». In : *RJCIA*, p. 1–6.

JEAN, Pierre-Antoine, Sébastien HARISPE, Sylvie RANWEZ, Patrice BELLOT et Jacky MONTMAIN (2016a). « Un modèle probabiliste pour la détection de l'incertitude dans le langage naturel ». In : *Document numérique* 19.2, p. 53–73.

– (2016b). « Uncertainty detection in natural language : a probabilistic model ». In : *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*. ACM, p. 10.

– (2017). « Étude d'un modèle d'inférence de connaissances à partir de textes ». In : *CORIA*, p. 201–217.

JEAN-LOUIS, Ludovic (2011). « Approches supervisées et faiblement supervisées pour l'extraction d'événements et le peuplement de bases de connaissances ». Thèse de doct., p. 193.

JIJKOUN, Valentin et Maarten de RIJKE (2005). « Recognizing textual entailment using lexical similarity ». In : *Pascal RTE*.

JOACHIMS, Thorsten (2002a). *Learning to classify text using support vector machines : Methods, theory and algorithms*. Kluwer Academic Publishers.

– (2002b). « Optimizing search engines using clickthrough data ». In : *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 133–142.

JOULIN, Armand, Edouard GRAVE, Piotr BOJANOWSKI et Tomas MIKOLOV (2016). « Bag of Tricks for Efficient Text Classification ». In : *CoRR*.

JOUSSELME, Anne-Laure, Patrick MAUPIN et Éloi BOSSÉ (2003). « Uncertainty in a situation analysis perspective ». In : *Proceedings of the Sixth International Conference of Information Fusion*.

JURAFSKY, Dan et James H MARTIN (2014). *Speech and language processing*. T. 3. Pearson.

KALYANPUR, Aditya et al. (2012). « Structured data and inference in DeepQA ». In : *IBM Journal of Research and Development* 56.3.4, p. 10.

KAMP, Hans et Uwe REYLE (1993). *From discourse to logic ; An introduction to model-theoretic semantics of natural language, formal logic and DRT*. Kluwer, Dordrecht.

KERDJOUDJ, Fadhela et Olivier CURÉ (2015). « Gestion de l'incertitude dans le cadre d'une extraction des connaissances à partir de texte. » In : *EGC*, p. 477–478.

KETOKIVI, Mikko et Saku MANTERE (2010). « Two strategies for inductive reasoning in organizational research ». In : *Academy of Management Review* 35.2, p. 315–333.

KIM, Mi-Young, Ying XU et Randy GOEBEL (2015). « A Convolutional Neural Network in Legal Question Answering ». In : *Ninth International Workshop on Juris-informatics (JURISIN)*.

KONSTANTINOVA, Natalia, Sheila CM DE SOUSA, Noa P Cruz DÍAZ, Manuel J Mana LÓPEZ, Maite TABOADA et Ruslan MITKOV (2012). « A review corpus annotated for negation, speculation and their scope. » In : *LREC*, p. 3190–3195.

KULICK, Seth et al. (2004). « Integrated annotation for biomedical information extraction ». In : *Proc. of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, p. 61–68.

LAFOURCADE, Mathieu et Eugène SANDFORD (1999). « Analyse et désambiguïsation lexicale par vecteurs sémantiques ». In : *Proc. of TALN*. T. 99, p. 351–356.

LAKOFF, George (1975). « Hedges : a study in meaning criteria and the logic of fuzzy concepts ». In : *Contemporary Research in Philosophical Logic and Linguistic semantics*. Springer, p. 221–271.

LAO, Ni et William W COHEN (2010). « Relational retrieval using a combination of path-constrained random walks ». In : *Machine learning* 81.1, p. 53–67.

LAO, Ni, Tom MITCHELL et William W COHEN (2011). « Random walk inference and learning in a large scale knowledge base ». In : *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, p. 529–539.

LAVALLEY, Rémi, Chloé CLAVEL et Patrice BELLOT (2010). « Extraction probabiliste de chaînes de mots relatives à une opinion ». In : *Traitement Automatique des Langues* 51, p. 101–130.

LEMAIRE, Benoît (2008). « Limites de la lemmatisation pour l'extraction de significations ». In : *9e Journées internationales d'Analyse Statistique des Données Textuelles*, p. 725–732.

LESK, Michael (1986). « Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone ». In : *Proceedings of the 5th annual international conference on Systems documentation*. ACM, p. 24–26.

LEX, Elisabeth, Michael VOELSKE, Marcelo ERRECALDE, Edgardo FERRETTI, Leticia CAGNINA, Christopher HORN, Benno STEIN et Michael GRANITZER (2012). « Measuring the quality of web content using factual information ». In : *Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality*. ACM, p. 7–10.

LI, Xin et Dan ROTH (2006). « Learning question classifiers : the role of semantic information ». In : *Natural Language Engineering* 12.3, p. 229–249.

LI, Yaliang, Jing GAO, Chuishi MENG, Qi LI, Lu SU, Bo ZHAO, Wei FAN et Jiawei HAN (2016). « A survey on truth discovery ». In : *ACM Sigkdd Explorations Newsletter* 17.2, p. 1–16.

LIEKENS, Anthony ML, Jeroen DE KNIJF, Walter DAELEMANS, Bart GOETHALS, Peter DE RIJK et Jurgen DEL-FAVERO (2011). « BioGraph : unsupervised biomedical knowledge discovery via automated hypothesis generation ». In : *Genome biology* 12.6, R57.

LIGHT, Marc, Xin Ying QIU et Padmini SRINIVASAN (2004). « The language of bioscience : Facts, speculations, and statements in between ». In : *Proceedings of BioLink*

2004 workshop on linking biological literature, ontologies and databases : tools for users. Association for Computational Linguistics, p. 17–24.

MACCARTNEY, Bill et Christopher D MANNING (2007). « Natural logic for textual inference ». In : *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Association for Computational Linguistics, p. 193–200.

MAEDCHE, Alexander, Steffen STAAB, Nenad STOJANOVIC, Rudi STUDER et York SURE (2001). « Seal—a framework for developing semantic web portals ». In : *Advances in Databases*, p. 1–22.

MANNING, Christopher D, Prabhakar RAGHAVAN et Hinrich SCHÜTZE (2008). *Introduction to information retrieval*. T. 1. 1. Cambridge university press Cambridge.

MANNING, Christopher D et Hinrich SCHÜTZE (1999). *Foundations of statistical natural language processing*. T. 999. MIT Press.

MARCUS, Mitchell P, Mary Ann MARCINKIEWICZ et Beatrice SANTORINI (1993). « Building a large annotated corpus of English : The Penn Treebank ». In : *Computational linguistics* 19.2, p. 313–330.

MCCARTHY, John, Marvin L MINSKY, Nathaniel ROCHESTER et Claude E SHANNON (2006). « A proposal for the Dartmouth summer research project on artificial intelligence, august 31, 1955 ». In : *AI magazine* 27.4, p. 12.

MCGUINNESS, Deborah L et Frank VAN HARMELEN (2004). « OWL web ontology language overview ». In : *W3C recommendation* 10.10, p. 2004.

MCNAMEE, Paul et Hoa Trang DANG (2009). « Overview of the TAC 2009 knowledge base population track ». In : *Text Analysis Conference (TAC)*. T. 17, p. 111–113.

MENDES, Pablo N, Hannes MÜHLEISEN et Christian BIZER (2012). « Sieve : linked data quality assessment and fusion ». In : *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. ACM, p. 116–123.

MIKOLOV, T., I. SUTSKEVER, K. CHEN, G. S. CORRADO et J. DEAN (2013). « Distributed representations of words and phrases and their compositionality ». In : *Adv. NIPS*.

MILLER, George A (1995). « WordNet : a lexical database for English ». In : *Communications of the ACM* 38.11, p. 39–41.

MINSKY, Marvin (1974). *A Framework for Representing Knowledge*. Rapp. tech.

MINTZ, Mike, Steven BILLS, Rion SNOW et Dan JURAFSKY (2009). « Distant supervision for relation extraction without labeled data ». In : *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2*. Association for Computational Linguistics, p. 1003–1011.

MOESCHLER, Jacques et Anne REBOUL (1998). « La pragmatique aujourd'hui ». In : *Paris : Points Essais*.

MUSLEA, Ion et al. (1999). « Extraction patterns for information extraction tasks : A survey ». In : *The AAAI-99 workshop on machine learning for information extraction*. T. 2. 2. Orlando Florida.

NAKASHOLE, Ndapandula, Martin THEOBALD et Gerhard WEIKUM (2011). « Scalable knowledge harvesting with high precision and high recall ». In : *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, p. 227–236.

NICKEL, Maximilian, Volker TRESP et Hans-Peter KRIEGEL (2011). « A three-way model for collective learning on multi-relational data ». In : *Proceedings of the 28th international conference on machine learning (ICML-11)*, p. 809–816.

– (2012). « Factorizing Yago : scalable machine learning for linked data ». In : *Proceedings of the 21st international conference on World Wide Web*. ACM, p. 271–280.

NICKEL, Maximilian, Kevin MURPHY, Volker TRESP et Evgeniy GABRILOVICH (2016). « A review of relational machine learning for knowledge graphs ». In : *Proceedings of the IEEE* 104.1, p. 11–33.

NIU, Feng, Ce ZHANG, Christopher RÉ et Jude SHAVLIK (2012). « Elementary : Large-scale knowledge-base construction via machine learning and statistical inference ». In : *International Journal on Semantic Web and Information Systems (IJSWIS)* 8.3, p. 42–73.

ONTANÓN, Santiago, Gabriel SYNNAEVE, Alberto URIARTE, Florian RICHOUX, David CHURCHILL et Mike PREUSS (2013). « A survey of real-time strategy game AI research and competition in starcraft ». In : *IEEE Transactions on Computational Intelligence and AI in games* 5.4, p. 293–311.

ØVRELID, Lilja, Erik VELLDAL et Stephan OEPEN (2010). « Syntactic scope resolution in uncertainty analysis ». In : *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, p. 1379–1387.

PAK, Alexander et Patrick PAROUBEK (2010). « Twitter as a corpus for sentiment analysis and opinion mining. » In : *LREC*. T. 10. 2010.

PANG, Bo et Lillian LEE (2004). « A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts ». In : *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 271.

PAPASALOUROS, Andreas, Konstantinos KANARIS et Konstantinos KOTIS (2008). « Automatic Generation Of Multiple Choice Questions From Domain Ontologies ». In : *e-Learning IADIS*, p. 427–434.

PEREIRA, Suzanne, Aurélie NÉVÉOL, Gaétan KERDELHUÉ, Elisabeth SERROT, Michel JOUBERT et Stéfan J DARMONI (2008). « Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue ». In : *AMIA Annual Symposium Proceedings*. T. 2008. American Medical Informatics Association, p. 586.

PHO, Van-Minh, Anne-Laure LIGOZAT et Brigitte GRAU (2015). « Distractor quality evaluation in multiple choice questions ». In : *International Conference on Artificial Intelligence in Education*. Springer, p. 377–386.

PORTER, Martin F (1980). « An algorithm for suffix stripping ». In : *Program* 14.3, p. 130–137.

PRESUTTI, Valentina, Francesco DRAICCHIO et Aldo GANGEMI (2012). « Knowledge Extraction Based on Discourse Representation Theory and Linguistic Frames ». In : *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management*, p. 114–129.

PROIOS, Dimitris, Magdalini EIRINAKI et Iraklis VARLAMIS (2015). « TipMe : Personalized advertising and aspect-based opinion mining for users and businesses ». In : *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, p. 1489–1494.

QIN, Tao et Tie-Yan LIU (2013). « Introducing LETOR 4.0 Datasets ». In : *CoRR*.

RAJPURKAR, Pranav, Jian ZHANG, Konstantin LOPYREV et Percy LIANG (2016). « Squad : 100,000+ questions for machine comprehension of text ». In : *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

RESNIK, Philip (1999). « Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural language ». In : *J. Artif. Intell. Res.(JAIR)* 11, p. 95–130.

RIALLE, Vincent (1996). « L'intelligence artificielle et sa place dans les sciences de la cognition ». In : *Bulletin de l'Association Française de l'Intelligence Artificielle* 26, p. 8–12.

RICHARDSON, Matthew, Christopher JC BURGES et Erin RENSHAW (2013). « MCTest : A Challenge Dataset for the Open-Domain Machine Comprehension of Text. » In : *EMNLP*. T. 3, p. 4.

ROBERTSON, Stephen E et Jones KAREN SPÄRCK (1976). « Relevance weighting of search terms ». In : *Journal of the American Society for Information Science*. T. 27. 3, p. 129–146.

ROBERTSON, Stephen E et Steve WALKER (1997). « On relevance weights with little relevance information ». In : *ACM SIGIR Forum*. T. 31. SI. ACM, p. 16–24.

SACK, Harald, Stefan DIETZE, Anna TORDAI et Christoph LANGE (2016). *Semantic Web Challenges : Third SemWebEval Challenge at ESWC 2016, Heraklion, Crete, Greece, May 29-June 2, 2016, Revised Selected Papers*. T. 641. Springer.

SAHAMI, Mehran, Susan DUMAIS, David HECKERMAN et Eric HORVITZ (1998). « A Bayesian approach to filtering junk e-mail ». In : *Learning for Text Categorization : Papers from the 1998 workshop*. T. 62, p. 98–105.

SALVO BRAZ, Rodrigo de, Roxana GIRJU, Vasin PUNYAKANOK, Dan ROTH et Mark SAMMONS (2006). « An inference model for semantic entailment in natural language ». In : *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*. Springer, p. 261–286.

SANTORINI, Beatrice (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision)*. Rapp. tech. University of Pennsylvania Department of Computer et Information Science Technical Report No. MS-CIS-90-47.

SAVOY, Jacques (2017). « Catégorisation de textes avec style ». In : *COnférence en Recherche d'Informations et Applications - CORIA 2017, 14th French Information Retrieval Conference.*

SCHMITZ, Michael, Robert BART, Stephen SODERLAND, Oren ETZIONI et al. (2012). « Open language learning for information extraction ». In : *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, p. 523–534.

SCHWAB, Didier, Jérôme GOULIAN, Andon TCHECHMEDJIEV et Hervé BLANCHON (2012). « Ant colony algorithm for the unsupervised word sense disambiguation of texts : Comparison and evaluation ». In : *COLING 2012*, p. 2389–2404.

SEBASTIANI, Fabrizio (2002). « Machine learning in automated text categorization ». In : *ACM computing surveys (CSUR)* 34.1, p. 1–47.

SHAFER, Glenn et al. (1976). *A mathematical theory of evidence*. T. 1. Princeton university press Princeton.

SHAW, Ralph R (1948). « Royal society scientific information conference ». In : *The American Statistician* 2.4, p. 14–16.

SMETS, Philippe (1997). « Imperfect information : Imprecision-uncertainty ». In : *Uncertainty management in information systems. from needs to solutions*, p. 225–254.

SMITH, David Eugene (2012). *A source book in mathematics*. Courier Corporation.

SUCHANEK, Fabian M, Gjergji KASNECI et Gerhard WEIKUM (2007). « Yago : a core of semantic knowledge ». In : *Proceedings of the 16th international conference on World Wide Web*. ACM, p. 697–706.

SUKHBAATAR, Sainbayar, Arthur SZLAM, Jason WESTON et Rob FERGUS (2015). « Weakly supervised memory networks ». In : *CoRR*.

SUNDHEIM, Beth M (1992). « Overview of the fourth message understanding evaluation and conference ». In : *Proceedings of the 4th conference on Message understanding*. Association for Computational Linguistics, p. 3–21.

SURDEANU, Mihai, Julie TIBSHIRANI, Ramesh NALLAPATI et Christopher D MANNING (2012). « Multi-instance multi-label learning for relation extraction ». In : *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, p. 455–465.

SZARVAS, György, Veronika VINCZE, Richárd FARKAS et János CSIRIK (2008). « The BioScope corpus : annotation for negation, uncertainty and their scope in biomedical texts ». In : *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Association for Computational Linguistics, p. 38–45.

SZARVAS, György, Veronika VINCZE, Richárd FARKAS, György MÓRA et Iryna GUREVYCH (2012). « Cross-genre and cross-domain detection of semantic uncertainty ». In : *Computational Linguistics* 38.2, p. 335–367.

TAN, Liling (2014). *Pywsd : Python Implementations of Word Sense Disambiguation (WSD) Technologies [software]*. https ://github.com/alvations/pywsd.

TANG, Buzhou, Xiaolong WANG, Xuan WANG, Bo YUAN et Shixi FAN (2010). « A cascade method for detecting hedges and their scope in natural language text ». In :

Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task. Association for Computational Linguistics, p. 13–17.

TANUSHI, Hideyuki, Hercules DALIANIS, Martin DUNELD, Maria KVIST, Maria SKEPPSTEDT et Sumithra VELUPILLAI (2013). « Negation scope delimitation in clinical text using three approaches : NegEx, PyConTextNLP and SynNeg ». In : *19th Nordic Conference of Computational Linguistics (NODALIDA 2013), May 22-24, 2013, Oslo, Norway*. Linköping University Electronic Press, p. 387–474.

TARI, Luis, Saadat ANWAR, Shanshan LIANG, James CAI et Chitta BARAL (2010). « Discovering drug–drug interactions : a text-mining and reasoning approach based on properties of drug metabolism ». In : *Bioinformatics* 26.18, p. i547–i553.

TCHECHMEDJIEV, Andon (2012). « État de l'art : mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances ». In : JEP-TALN-RECITAL, p. 295.

TSATSARONIS, George et al. (2015). « An overview of the BioAsq large-scale biomedical semantic indexing and question answering competition ». In : *BMC bioinformatics* 16.1, p. 138.

VINCZE, Veronika (2013). « Weasels, hedges and peacocks : Discourse-level uncertainty in Wikipedia articles ». In : *Sixth International Joint Conference on Natural Language Processing*, p. 383–391.

– (2015). « Uncertainty detection in natural language texts ». Thèse de doct. University of Szeged, p. 1–141.

VIVIANI, Marco et Gabriella PASI (2017). « Credibility in social media : opinions, news, and health information—a survey ». In : *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*.

VÖLKER, Johanna, Pascal HITZLER et Philipp CIMIANO (2007). « Acquisition of OWL DL axioms from lexical resources ». In : *European Semantic Web Conference*. Springer, p. 670–685.

WANG, Hai H et al. (2006). « Frames and OWL side by side ». In : *Presentation Abstracts*, p. 54.

WANG, Quan, Jing LIU, Yuanfei LUO, Bin WANG et Chin-Yew LIN (2016). « Knowledge Base Completion via Coupled Path Ranking. » In : *ACL*.

WANG, Wei, Romaric BESANÇON, Olivier FERRET et Brigitte GRAU (2013). « Semantic relation clustering for unsupervised information extraction ». In : *Proceedings of TALN 2013* 1, p. 353–366.

WEISSENBACHER, David (2008). « Effects of imperfect annotations on Natural Language Processing systems, an applicative case study : the pronominal anaphora resolution ». Thèse de doct., p. 185.

WESTON, Jason, Antoine BORDES, Sumit CHOPRA, Alexander M RUSH, Bart van MERRIËNBOER, Armand JOULIN et Tomas MIKOLOV (2015). « Towards AI-complete question answering : A set of prerequisite toy tasks ». In : *CoRR*.

WINOGRAD, Terry (1971). *Procedures as a representation for data in a computer program for understanding natural language*. Rapp. tech. Massachusetts Inst of Tech Cambridge Project MAC.

WOODS, William A, Ronald M KAPLAN et Bonnie NASH-WEBBER (1972). « The Lunar Sciences : Natural Language Information System : Final Report ». In : *BBN Report 2378 Bolt Beranek and Newman Inc. Cambridge, Massachusetts.*

WOUTERS, Els (1998). *Maigret :" je ne déduis jamais" : la méthode abductive chez Simenon*. T. 8. Editions du CEFAL.

WU, Andrew S, Bao H DO, Jinsuh KIM et Daniel L RUBIN (2011). « Evaluation of negation and uncertainty detection and its impact on precision and recall in search ». In : *Journal of digital imaging* 24.2, p. 234–242.

WU, Fei et Daniel S WELD (2010). « Open information extraction using Wikipedia ». In : *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, p. 118–127.

WU, Wentao, Hongsong LI, Haixun WANG et Kenny Q ZHU (2012). « Probase : A probabilistic taxonomy for text understanding ». In : *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, p. 481–492.

WU, Zhibiao et Martha PALMER (1994). « Verbs semantics and lexical selection ». In : *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 133–138.

YANG, Yiming et Jan O PEDERSEN (1997). « A comparative study on feature selection in text categorization ». In : *Icml*. T. 97, p. 412–420.

YATES, Alexander, Michael CAFARELLA, Michele BANKO, Oren ETZIONI, Matthew BROADHEAD et Stephen SODERLAND (2007). « Textrunner : open information extraction on the web ». In : *Proceedings of Human Language Technologies : The Annual Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*. Association for Computational Linguistics, p. 25–26.

ZARARSIZ, Gokmen, Ferhan ELMALI et Ahmet OZTURK (2012). « Bagging support vector machines for leukemia classification ». In : *International Journal of Computer Science Issues* 9.1.

ZELLE, John M et Raymond J MOONEY (1996). « Learning to parse database queries using inductive logic programming ». In : *Proceedings of the national conference on artificial intelligence*, p. 1050–1055.

ZHAI, Chengxiang (2001). « Notes on the Lemur TFIDF model ». In : *Unpublished report*.

ZHOU, XueZhong, Jörg MENCHE, Albert-László BARABÁSI et Amitabh SHARMA (2014). « Human symptoms–disease network ». In : *Nature communications* 5.