Étude d'un modèle d'inférence de connaissances à partir de textes

Pierre-Antoine Jean* — **Sébastien Harispe** * — **Sylvie Ranwez*** — **Patrice Bellot**** — **Jacky Montmain***

* *LGI2P, École des mines d'Alès, 69 rue Georges Besse F-30035 Nîmes cedex 1, {prenom.nom}@mines-ales.fr,*
** *LSIS, Avenue Escadrille Normandie-Niemen F-13397 Marseille cedex 20, patrice.bellot@lsis.org,*

RÉSUMÉ. *Cet article propose une approche automatisée d'inférence de connaissances basée sur l'analyse de relations extraites à partir de textes. Son originalité repose sur la définition d'un cadre tenant compte (i) d'une structuration des objets étudiés (e.g. syntagmes nominaux) sous la forme d'un ordre partiel et (ii) de l'exploitation possible d'une connaissance a priori formalisée dans un modèle de connaissances de type ontologie (taxonomie). Ce cadre permet notamment de définir des règles de propagation de l'information basées sur la théorie des croyances afin d'inférer de nouvelles connaissances à partir des relations extraites. Bien qu'a portée plus large, notre approche est ici illustrée et évaluée au travers de la définition d'un système automatique exploitant des textes issus du Web afin de répondre à des questionnaires générés. Nous montrons notamment l'intérêt de structurer les extractions et le gain apporté par la prise en compte d'une connaissance a priori au sein d'une telle chaîne de traitement.*

ABSTRACT. *This article introduces an automated knowledge inference approach taking advantage of relationships extracted from texts. It is based on a novel framework making possible to exploit (i) a generated partial ordering of studied objects (e.g. noun phrases), and (ii) prior knowledge defined into ontologies. This framework is particularly suited for defining information propagation rules based on evidence theory in order to infer new knowledge. The proposed approach is illustrated and evaluated through the definition of a system performing question answering by analyzing texts available on the Web. This case study shows the benefits of structuring processed information (e.g. using prior knowledge) for inferring new knowledge.*

MOTS-CLÉS : *Inférence de connaissances, extraction d'information, modèle de propagation, ontologies, traitement automatique de la langue.*

KEYWORDS: *Knowledge inference, information extraction, evidence theory, ontology, natural language processing.*

1. Contexte de l'étude et positionnement

Disposer de vastes corpus de textes de nature diverses constitue une véritable opportunité dans le domaine de l'Intelligence Artificielle. Cela laisse notamment entrevoir la possibilité d'exploiter de façon systématisée et automatique les différentes connaissances et éléments d'information explicités dans ces textes par le couplage d'approches issues du Traitement Automatique des Langues et des techniques de raisonnement adaptées. Une variété de traitements et processus pourraient alors bénéficier de la richesse exprimée dans ces corpus (*e.g.* enrichissement de bases de connaissances, prise de décision, système de question-réponse).

L'approche décrite dans ce papier s'inscrit dans la lignée des travaux concernant l'enrichissement et l'exploitation de bases de connaissances. Ces derniers sont classifiés dans (Murphy *et al.*, 2014) en quatre principaux groupes. Le premier regroupe les méthodes exploitant les sources de données structurées, *e.g.* les *infobox* de Wikipedia qui servent de support à DBpedia (Auer *et al.*, 2007). Le second groupe englobe les approches utilisant les modèles d'extraction d'information en monde ouvert, *e.g.* REVERB (Etzioni *et al.*, 2011) ou OLLIE (Schmitz *et al.*, 2012). Le troisième a la particularité de combiner l'extraction d'information avec une modélisation de connaissance (ontologie), *e.g.* PROSPERA (Nakashole *et al.*, 2011) ou NELL (Carlson *et al.*, 2010). Enfin, la dernière catégorie rassemble les méthodes construisant des taxonomies – restriction des bases de connaissances générales considérant des types de prédicats multiples, *e.g.* PROBASE (Wu *et al.*, 2012).

Cette classification décrit des approches faisant intervenir des domaines et des problématiques de recherche variés pour la construction, la maintenance et l'utilisation de bases de connaissances. Notre approche s'inscrit dans la troisième catégorie. Elle repose sur l'extraction de déclarations sous la forme de triplets désambiguïsés et l'utilisation d'une ontologie et plus particulièrement de la taxonomie des concepts qu'elle modélise. À l'image des trois derniers groupes, elle exploite des données textuelles non structurées impliquant l'utilisation des méthodes d'extraction de relations et de désambiguïsation (*entity linking*). Ces méthodes permettent de collecter des relations candidates, exploitées à l'instar des travaux de (Niu *et al.*, 2012 ; Zhu *et al.*, 2009) au sein d'un processus d'inférence conduisant à l'enrichissement de la base. L'apport de notre méthode repose sur cette phase d'inférence. En effet, contrairement aux travaux existants, nous considérons un ordre taxonomique sur les extractions. Cet ordre permet d'exploiter les implications hiérarchiques entre les entités permettant ainsi une meilleure appréciation de la véracité des extractions.

La section suivante présente une vue détaillée de l'approche. La validation proposée dans la section 3 concerne la réponse à un ensemble de questionnaires générés de façon automatique. Le protocole d'évaluation est détaillé puis les résultats obtenus sont présentés et discutés. Enfin la section 4 synthétise nos travaux et ouvre de nouvelles perspectives.

2. Approche proposée pour l'inférence de connaissances

2.1. *Vue générale de l'approche*

Dans cet article, nous proposons une approche pour l'inférence de connaissances à partir de textes. Ces connaissances sont constituées de *déclarations*[1] extraites sous la forme de triplets <*sujet*, *prédicat*, *objet*>.

L'originalité de l'approche repose sur deux particularités. La première concerne la structuration des sujets et objets étudiés au sein d'un ordre partiel. Celui-ci permet d'exprimer les liens d'implication qui peuvent être considérés entre les syntagmes nominaux, *e.g.* l'observation du syntagme nominal "maladies respiratoires chroniques" implique l'évocation de concepts plus généraux : "maladies respiratoires" et "maladies". La construction automatique de cet ordre partiel peut aussi tirer parti d'une connaissance *a priori* exprimée dans une taxonomie (TBOX d'une ontologie), afin que l'ordre partiel intègre par exemple que toute observation des termes "mucoviscidose" et "apnée du sommeil" correspond à une évocation de "maladies respiratoires chroniques". Cette structuration, centrale dans notre approche, permet de mettre en correspondance des déclarations faisant référence de façon implicite et non triviale à une même évocation. Cela est rendu possible par la construction d'un graphe d'implication, cette fois entre déclarations, défini à l'aide de la structuration des sujets et objets associés, *e.g.* les déclarations <fibrose kystique, est liée à, mutation du gène CFTR> et <maladie de Huntington, est liée à, mutation du gène HD> font implicitement référence à la déclaration éventuellement non observée <maladie chronique, est liée à, mutation gène>, qui précise que le corpus évoque qu'il existe des maladies chroniques liées à des mutations génétiques. La seconde particularité de l'approche porte sur la définition de modalités d'inférence exploitant le cadre théorique des fonctions de croyances afin de croiser les fréquences d'observations des déclarations en tenant compte des liens d'implication exprimés dans le graphe d'implication des déclarations évoqué ci-avant. La figure 1 présente les principaux blocs de la chaîne de traitement.

L'intérêt d'un tel système est double. Le premier est l'estimation de la croyance associée aux déclarations en tenant compte des observations plus spécifiques. Le second intérêt repose sur la souplesse du système. En effet, le processus d'inférence se base très largement sur des modalités de propagation de l'information amenée par les observations des différentes déclarations extraites. Ces modalités de propagation peuvent être adaptées afin de tenir compte de la qualité des extractions (incertitude des méthodes) et de propriétés intrinsèques à l'information explicitée par la déclaration – incertitude linguistique, véracité de l'information, fiabilité de la source (Dong

1. 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. 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 *relation*.

et al., 2009). Cette adaptation peut notamment être réalisée par modification de l'importance donnée à l'observation de déclarations en fonction de différents critères permettant de critiquer la pertinence d'intégrer les informations qui leurs sont associées. Ce deuxième point est détaillé dans les perspectives de l'article.

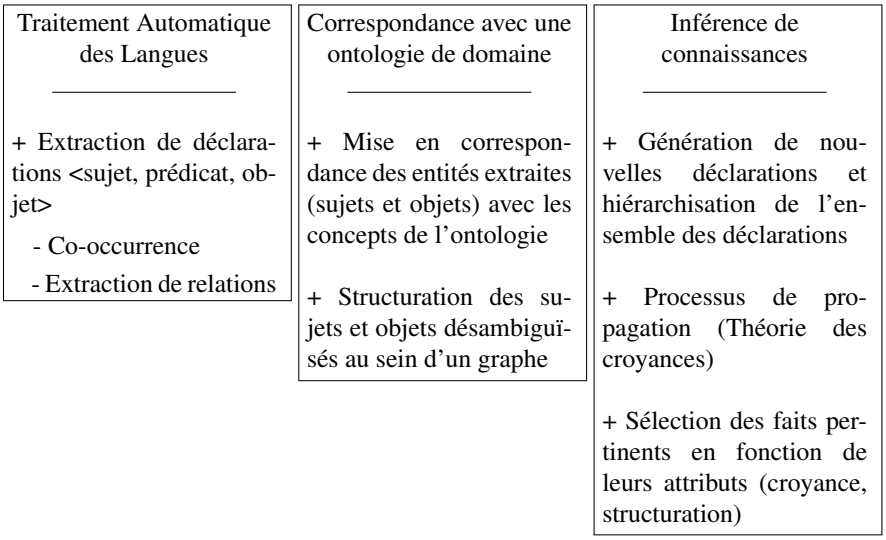

Figure 1. *Les trois principaux blocs de la chaîne de traitement : traitement automatique du langage, mise en correspondance avec la connaissance modélisée dans une ontologie et inférence de connaissances.*

2.2. *Extraction de déclarations*

Les contributions mises en avant dans cet article ne concernent pas cette sous-section. La plupart des technologies utilisées ici ont déjà été présentées dans la littérature. Toutefois, pour une meilleure compréhension de l'approche globale, il nous semble indispensable de décrire les méthodes implémentées et de discuter leurs performances dans le processus d'inférence de connaissances. Les exemples considérés dans cette section proviennent d'un jeu de données en anglais collectées à partir du Web.

L'extraction de déclarations est une étape indispensable du processus. Elle détermine les informations qui seront exploitées dans la suite du traitement. Les méthodes utilisées ont pour but d'extraire des triplets des textes sous la forme <*sujet, prédicat, objet*>. Plusieurs approches d'extraction de relations ont été expérimentées. Chacune induit un traitement des données différent dans un contexte d'utilisation particulier.

La première approche expérimentée se base sur l'analyse de la co-occurrence des termes. Cette méthode est utilisée dans (Zhou *et al.*, 2014) afin d'extraire les relations entre des maladies et les symptômes qui leurs sont associés. Leur approche réalise une correspondance lexicale entre les labels associés aux concepts du MeSH (Nelson *et al.*, 2001) et les mots clés fournis dans les articles scientifiques extraits de PubMed[2]. Grâce à sa facilité de mise en oeuvre, nous avons implémenté leur solution et l'avons testée dans plusieurs contextes, en la confrontant à différents types de textes et pour diverses relations. Si de très bons résultats ont été observés sur certaines relations spécifiques pour des domaines fortement contraints *e.g.* la relation *cause* entre une maladie et un symptôme dans le domaine médical, nous déplorons un bruit parfois important dans le cas d'un prédicat factuel tel que *bornIn* (Surdeanu *et al.*, 2012).

Les autres approches que nous avons testées permettent de diminuer le bruit d'un modèle de co-occurrence. Elles appartiennent au domaine de l'extraction de relations en domaine ouvert et s'appuient sur la formulation d'un prédicat entre le sujet et l'objet. L'objectif visé ici est de développer des extracteurs indépendants d'un domaine en particulier, et capables d'être exécutés sur de larges collections de textes provenant du Web. Ces méthodes se concentrent sur différentes propriétés qui sont associées à l'expression de relations dans les textes. La première s'intitule REVERB et exploite des patrons syntaxiques basés sur des contraintes lexicales afin d'extraire des motifs verbaux (Part Of Speech – POS) spécifiques. C'est une méthode robuste et rapide mais incapable d'extraire certains types de relations telles que les relations nominales *e.g. "Microsoft co-founder Bill Gates spoke at ..."* doit entraîner l'extraction de <*Bill Gates, be co-founder of, Microsoft*>. La seconde méthode étudiée s'intitule OLLIE et permet de répondre à cette problématique. En effet, elle a la particularité par rapport à REVERB d'exploiter les dépendances syntaxiques des phrases permettant une meilleure précision des relations extraites et de conserver les relations nominales. Cependant, ces approches d'extraction de relations génèrent des sujets et des objets complexes à étudier. En effet, le sujet ou l'objet sont généralement représentés par un syntagme nominal et non par une entité distincte. Pour réduire le bruit de ces éléments, une phase de désambiguïsation et d'analyse linguistique a été intégrée. Ces procédures sont détaillées dans la sous-section 2.3.

2.3. *Exploitation de la connaissance a priori pour l'inférence de déclarations*

L'idée développée ici consiste à tirer profit d'une ontologie afin d'enrichir les informations acquises par le système. Cette phase fait le lien entre le texte et la taxonomie par le biais des méthodes de désambiguïsation. Sa finalité est de construire un ordre partiel hiérarchisant les entités des déclarations extraites (*i.e. sujets* et *objets*). Cet ordre est ensuite utilisé comme support à la génération de nouvelles déclarations. Celles-ci sont évaluées ultérieurement par nos modèles d'inférence appliqués sur un

2. https ://www.ncbi.nlm.nih.gov/pubmed

deuxième graphe hiérarchisant l'ensemble des déclarations. C'est ce processus qui est détaillé dans cette section.

2.3.1. *Correspondance entre les déclarations et les concepts de l'ontologie*

La désambiguïsation des termes à partir d'une taxonomie de concepts est une étape importante dans le processus d'extraction à partir des textes. Son rôle est de détecter la polysémie des termes présents dans le texte et de réaliser le lien correct entre un terme et le concept correspondant dans la taxonomie. Par exemple, dans les phrases suivantes : *"Le colin mange des graines"* et *"Le colin mange des amphipodes"*, *le colin* peut être le nom vernaculaire d'un poisson ou celui d'un oiseau. Pour lever cette ambiguïté, des outils ont été proposés, comme MetaMap dans le domaine biomédical (Aronson, 2001). Cet outil résout les ambiguïtés en choisissant un concept dans le metathesaurus UMLS[3] ayant le type sémantique le plus probable pour un contexte lexical donné (Aronson et Lang, 2010). La littérature propose également d'autres approches pour la désambiguïsation telles que des méthodes reposant sur des mesures de similarités sémantiques entre termes (Tchechmedjiev, 2012).

Outre l'ambiguïté des données, un second problème auquel la méthode se confronte est la correspondance partielle des mots avec un sujet ou un objet. En effet, une méthode d'extraction de relations telle que REVERB génère des syntagmes et non des entités. Par exemple, pour le sujet *young koala* la méthode de désambiguïsation exploitée recherche la plus longue entité d'intérêt à désambiguïser. Ici, seule la correspondance avec *koala* est réalisée et non avec *young*. Ainsi, l'information véhiculée par cette correspondance est différente de l'information retranscrite par la relation extraite, où *young koala* est plus spécifique que *koala*. Ceci peut altérer la qualité du processus d'inférence de connaissances.

Pour pallier ce problème de correspondance partielle et dans l'optique de retranscrire l'ensemble de l'information véhiculée par une relation, la méthode construit une structure intermédiaire dans laquelle un ordre partiel sur les syntagmes est enrichi grâce à l'ontologie. Cette structure permet en quelque sorte de caractériser la sémantique des déclarations et sert de support à la génération de nouvelles déclarations. Ce processus est décrit dans la section suivante.

2.3.2. *Construction d'un ordre partiel sur les syntagmes à partir d'une taxonomie*

L'objectif de cette phase est d'ordonner les syntagmes nominaux des sujets et objets, normalisés par un ensemble de règles lexicales et syntaxiques, afin d'y injecter de la connaissance issue d'une taxonomie. Les règles linguistiques utilisées reposent sur une règle d'inversion centrée sur le terme[4] *of*, où une séquence ordonnée composée des trois symboles $[\sigma_1\ of\ \sigma_2]$ amènera la formation du syntagme $[\sigma_2\ \sigma_1]$ et sur la correspondance avec des patrons syntaxiques prédéfinis. Par exemple, le syntagme '*younger freshwater terrapins*' est conservé car il répond au patron grammatical : Ad-

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

4. Nous considérons dans cette étude qu'un terme est équivalent à un simple mot.

jectif – Nom – Nom(s). Tandis que les règles syntaxiques sont basées sur des informations de partie du discours, dans lesquelles les fonctions grammaticales des syntagmes sélectionnés doivent correspondre à des motifs définis manuellement.

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 Σ l'ensemble des syntagmes possibles. Dans l'absolu, l'ordonnancement des syntagmes a pour objectif de définir un ordre partiel sur Σ : $O_\Sigma = (\preceq, \Sigma)$. La construction de cet ordre partiel repose sur deux règles principales. En pratique, ces règles s'appliqueront uniquement sur les syntagmes extraits à partir des règles linguistiques établies.

1) Règle d'inclusion : un syntagme $\sigma = [t_1, \ldots, t_i]$ spécialise tout syntagme σ' formé d'une séquence de termes contigus de σ incluant t_i, ce qui implique $\sigma \prec \sigma'$, e.g. '*young koala*' spécialise '*koala*'. 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)[5].

2) Règle d'abstraction : nous considérons ici un ordre partiel sur les concepts d'une taxonomie $O_\mathcal{C} = (\preceq, \mathcal{C})$ 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[6]. 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 en figure 2.

Cet ordre partiel sur les syntagmes enrichi par les ancêtres des concepts désambiguïsés a pour but d'accroître le contenu informationnel du modèle de connaissance et de permettre ainsi la génération de nouvelles déclarations. C'est ce qui fait l'objet de la section suivante.

5. 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.

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

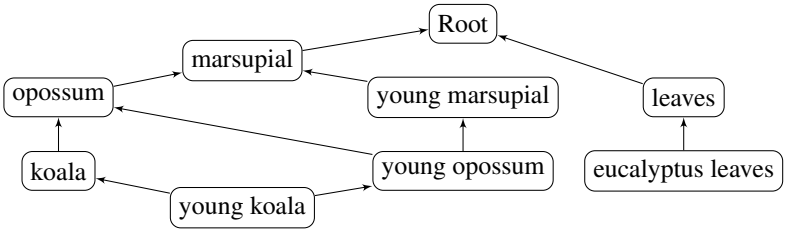

Figure 2. *Ordre partiel défini à partir des syntagmes 'young koala' et 'eucalyptus leaves' provenant de la déclaration* <*young koala, eat, eucalyptus leaves*> *et d'une taxonomie existante dans laquelle* $koala \preceq opossum$ *et* $opossum \preceq marsupial$.

2.4. *Inférer des nouvelles connaissances*

2.4.1. *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 par 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 2 et de la relation initiale <*young koala, eat, eucalyptus leaves*>, le système est capable de générer les déclarations : <*koala, eat, eucalyptus leaves*>, <*koala, eat, leaves*>, <*young opossum, eat, eucalyptus leaves*>, <*young opossum, eat, leaves*>, *etc.*

Ces déclarations sont ensuite structurées au sein d'un graphe de déclarations servant de support à la phase d'inférence. Il est important de souligner ici la sémantique associée à ces déclarations générées. En effet, si à partir de <*koala, eat, eucalyptus leaves*> et de l'ordre défini plus haut, il est possible de générer <*opossum, eat, eucalyptus leaves*>, la signification qui est associée est la suivante : il existe des opossums qui mangent des feuilles d'eucalyptus et non pas tous les opossums mangent des feuilles d'eucalyptus.

2.4.2. *Construction d'une hiérarchisation entre les déclarations*

La construction du graphe de déclarations s'appuie sur un ensemble de règles prédéfinies exposées dans le tableau 1. Ces règles sont basées sur la relation taxonomique des sujets et des objets de chaque relation. NOTE – on appelle s les déclarations en référence à la notation anglaise *statement*, $\mathcal{S}$ l'ensemble des déclarations observées et générées et $\mathcal{O}_{\mathcal{S}}(\preceq, \mathcal{S})$ l'ordre partiel sur les déclarations.

2.4.3. *Modèles utilisés pour l'inférence de nouvelles connaissances*

Afin d'identifier les connaissances pertinentes contenues dans le graphe de déclarations, l'approche exploite les fréquences des déclarations observées et la structure du graphe. Ces fréquences permettent de calculer les masses associées aux déclarations

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

Tableau 1. *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*.*

observées. Ces masses sont calculées à partir de la fréquence d'observation divisée par le nombre total d'observations (cf. équation 1).

$$m(s) = \frac{Freq(s)}{\sum_{s' \in \mathcal{S}} Freq(s')} \quad [1]$$

Par la suite, la croyance bel associée à chaque déclaration s est définie comme la somme des masses de l'ensemble des descendants $\mathcal{D}$ de s avec $\mathcal{D}(s) = \{s' \in \mathcal{S} \mid s' \preceq s\}$ (cf. équation 2). Cette étape est réalisée par une propagation *bottom-up*. Ainsi, à chaque déclaration est associée un score de croyance obtenu après cette propagation (Shafer, 1976).

$$bel(s) = \sum_{s' \in \mathcal{D}(s)} m(s') \quad [2]$$

L'exploitation du graphe de déclarations par l'attribution des valeurs de croyance a pour objectif de porter un jugement sur la véracité des déclarations. Pour cela, quatre modèles de sélection ont été élaborés et évalués. Ces modèles considèrent les différents attributs associés aux déclarations (croyance, profondeur). Ainsi, nous définissons $\mathcal{S}_\mathcal{M}$ l'ensemble des déclarations pertinentes par rapport à un modèle $\mathcal{M}$ donné. Une déclaration pertinente implique la déclaration elle-même et ses ancêtres pour l'ensemble des modèles.

Le premier modèle d'inférence $\mathcal{M}_1$, le plus simple pouvant être mis en place, est basé sur un seuil d'acceptation minimal α centré sur les croyances des déclarations : $\mathcal{S}_{\mathcal{M}_1} = \{s \in \mathcal{S} \mid bel(s) > \alpha\}$. Toutefois, le problème de cette approche est de conserver principalement des relations faiblement profondes dans le graphe en sachant que plus une déclaration est élevée dans le graphe plus elle est abstraite, donc moins informative et pertinente.

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 3).

$$\overline{bel}_x = \frac{\sum_{\{s \in \mathcal{S} | depth(s)=x\}} bel(s)}{\mid \{s \in \mathcal{S} \mid depth(s) = x\} \mid} \quad [3]$$

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 a une forte croyance ou, dans le cas inverse, a au moins un parent avec une masse non nulle. 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 une masse non nulle (cf. équation 4). Cette deuxième option permet d'augmenter le rappel potentiel de l'approche.

$$\begin{aligned}\mathcal{S}_{\mathcal{M}_4} = \{s \in \mathcal{S} \mid bel(s) \geq \mathcal{Q}_3(depth(s)) \vee \\ (bel(s) \geq \mathcal{Q}_1(depth(s)) \wedge (\exists s_{prt} \in parent(s), m(s_{prt}) > 0))\}\end{aligned} \quad [4]$$

3. Évaluation

En terme de validation nous avons choisi de définir des expériences reproductibles dont les résultats sont quantifiables par des métriques admises. Nous souhaitons évaluer la globalité de la chaîne de traitement : la modalité d'extraction d'information, la pertinence du modèle de propagation et l'efficacité des modèles de sélection. Les principales évaluations existantes proposent un ensemble de requêtes sans fournir les textes pour y répondre. Ces évaluations reposent uniquement sur les connaissances préalablement extraites par les approches (Kalyanpur *et al.*, 2012 ; Berant *et al.*, 2013). Sans ces corpus, il est difficile d'évaluer au même niveau les approches. Par conséquent, cette évaluation est appréhendée au travers de la composition d'un corpus non contrôlé extrait du Web et d'un ensemble de questionnaires, composés de réponses correctes et incorrectes, généré de manière automatique. Les sous-sections suivantes détaillent la procédure de génération d'un questionnaire, les données utilisées et les résultats obtenus.

3.1. *Génération automatique de questionnaires*

L'évaluation se base sur les travaux de (Pho *et al.*, 2015) portant sur l'évaluation des questionnaire à choix multiples (QCM). Chaque questionnaire généré est composé d'un ensemble d'options. Une option est modélisée par une déclaration de type <*sujet, prédicat, objet*>. Contrairement aux travaux de (Seyler *et al.*, 2016), nous ne réalisons pas de verbalisation des déclarations car l'objectif est principalement d'évaluer la méthode sur l'analyse des textes et non des requêtes. Lorsqu'une déclaration est correcte, c'est une réponse et lorsqu'elle est incorrecte, c'est un distracteur. Les options portent sur une thématique donnée et les modèles doivent retrouver l'ensemble des réponses. Ce protocole d'évaluation propose un ensemble de questionnaires avec un nombre d'options fixé à 500, composées de 250 réponses/distracteurs. Le questionnaire est réalisé de manière automatique à partir d'une taxonomie $O_{\mathcal{C}}$ et d'un ensemble de relations considérées représentatives issues d'une ontologie de domaine. La stratégie de génération automatique des questionnaires provient d'une stratégie de génération automatique de QCM proposée par (Papasalouros *et al.*, 2008). Cette dernière a été étendue pour permettre à l'approche de considérer les concepts en plus des instances en tant qu'options possibles. La suite de cette section détaille ce processus avec la génération des déclarations associées aux réponses $\mathcal{R}^+$ et aux distracteurs $\mathcal{R}^-$.

L'ensemble des réponses correctes possibles $\mathcal{R}^+$ pour un questionnaire donné 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 : $\mathcal{R}^+ : \{c \in \mathcal{C} \mid c_o \preceq c\}$.

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 les concepts plausibles par rapport aux observations. L'ensemble de ces concepts plausibles équivaut à l'ensemble des descendants $\mathcal{D}$ des concepts observés *i.e* $\mathcal{R}^- : \{c \in \mathcal{C} \mid \mathcal{D}(c) \cap (\cup_{r \in \mathcal{R}^+} \mathcal{D}(r)) = \emptyset\}$. Cependant, une restriction sur les concepts est appliquée en exploitant les plus proches sémantiquement des réponses correctes par rapport à un seuil β, $R^-(\beta) : \{c \in \mathcal{R}^- \mid sim(c, a \in \mathcal{R}^+) \geq \beta\}$. 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[7].

En plus de cette distance de Jaccard minimale, 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*. L'utilisation de la distance de Jaccard et de la restriction sur certains types de concepts permet de répondre aux contraintes sémantiques des options exposées dans la thèse de (Pho *et al.*, 2015). En effet, il y est question de la validation des QCM par l'homogénéité syntaxique et sémantique des options. Concernant, l'homogénéité syntaxique, les options sont constituées des concepts de l'ontologie, soit des noms.

7. Le seuil β de similarité utilisé est 0.6. Ce seuil représente 60% des ancêtres en commun entre une bonne réponse et un distracteur.

3.2. *Données*

La génération des options est réalisée à partir d'une liste de relations maladies/symptômes extraite de la base de données OMIM (*Online Mendelian Inheritance in Man*). Ces relations sont recoupées avec l'arborescence du MeSH 2016. La figure 3 présente un ensemble d'options avec deux réponses et deux distracteurs. Chacune des réponses est associée à des phrases extraites la supportant.

– **Réponse 1** : <Syndrome De Lange, Anomalies congénitales>

- Cornelia de Lange syndrome 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>

Figure 3. *Exemple d'options générées de manière automatique pour le prédicat has_manifestation. Chaque réponse est accompagnée d'exemples extraits du Web.*

À partir des relations recueillies, 100 questionnaires sont générés en piochant de manière aléatoire dans l'ensemble des réponses et des distracteurs générés. Pour tenter d'y répondre 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 du nom des maladies (180 maladies utilisées). Par conséquent, la construction de ce jeu de données ne garantit pas la couverture complète aux 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. Deux modes d'extraction sont comparés, le principe de co-occurrence des concepts issus des questions et l'extraction de relations selon les prédicats *include*, *due to*, et *cause*. Chacune de ces approches est évaluée avec et sans le support de la connaissance *a priori*. Le tableau 2 présente le nombre de relations extraites selon la méthode employée.

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

Tableau 2. *Nombre de relations extraites en fonction de la méthode d'extraction d'information employée sur le corpus établi.*

3.3. *Résultats et discussion*

Les résultats portent sur les modèles $\mathcal{M}_2$, $\mathcal{M}_3$ et $\mathcal{M}_4$ présentés en sous-section 2.4. Le modèle $\mathcal{M}_1$ 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. Le tableau 3 résume ces résultats obtenus avec les différentes méthodes d'extraction et l'utilisation ou non de la connaissance *a priori*. Les métriques d'évaluation utilisées correspondent aux métriques standard : rappel, précision et F-mesure. Les résultats finaux correspondent à la moyenne des mesures obtenue sur l'ensemble des questionnaires. 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).

	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)

Tableau 3. *Moyenne des résultats et leur écart-type respectif (STD) obtenus sur les 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.*

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 proposée dans l'article. Par conséquent, la discussion porte principalement sur la comparaison relative entre les précisions obtenues par les différentes configurations expérimentées.

Le tableau 3 montre que la principale influence sur la F-mesure est conditionnée par l'utilisation de la connaissance *a priori* au regard des simples déclarations extraites pour un système d'extraction donné. 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 sont à l'origine de la 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 de phrases captées par la méthode d'extraction de relations (cf. tableau 2). L'écart indique que la relation entre une maladie et un symptôme peut être exprimée par un nombre de prédicats supérieur à celui utilisé dans l'expérimentation.

Une analyse fine des faux-positifs démontre l'importance des modèles de sélections. En effet, 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 des maladies des articulations. Cette inférence a été validée après lecture dans Orphanet[8]. Cependant, OMIM ne renseigne pas ce symptôme. 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*.

4. Conclusion et perspectives

Dans cet article, nous avons proposé une méthode d'inférence de connaissances à partir des textes. Elle peut être exploitée dans un contexte d'enrichissement des bases de connaissances ou de génération d'hypothèses. Cette méthode se découpe en trois principaux blocs permettant de *i*) recueillir des informations dans les textes, *ii*) de

8. www.orpha.net

réaliser un lien avec une ontologie de domaine et *iii*) d'inférer de nouvelles connaissances. Le processus d'inférence est appliqué sur un graphe intermédiaire hiérarchisant les déclarations extraites et enrichies par l'ontologie. Ce dernier est un support au processus de propagation issu de la théorie des croyances permettant d'appliquer une valeur de confiance aux déclarations. Les résultats obtenus démontrent l'influence des différentes approches d'extraction et l'apport de la connaissance *a priori* au sein de la chaîne de traitement.

L'intérêt majeur du modèle de propagation proposé dans ce papier repose sur la possibilité de l'adapter en fonction des données auxquelles il est confronté. En effet, nous pouvons envisager la modification des modalités de propagation des déclarations au regard de nouvelles caractéristiques les décrivant. Cette perspective permettrait à l'approche d'apporter une manière alternative d'évaluation de la connaissance en tenant compte notamment de l'incertitude liée aux déclarations et à la façon de les extraire. En effet, l'incertitude est un phénomène s'exprimant de plusieurs manières. La première est liée au langage naturel, soit la valeur de vérité associée à la déclaration extraite. Ce type d'incertitude peut notamment être obtenu à partir des méthodes de détection de l'incertitude dans le langage naturel (Jean *et al.*, 2016). La seconde manière est de considérer 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, (Murphy *et al.*, 2014) ont proposé des pistes de réflexion concernant ce type d'incertitude. Par conséquent, nous pourrions imaginer une extension de la chaîne de traitement, exploitant simultanément l'imprécision des méthodes et l'incertitude linguistique. Ces types d'incertitude seraient retranscrits comme une valeur spécifique associée aux relations lors de la phase de propagation.

5. Bibliographie

Aronson A. R., « Effective mapping of biomedical text to the UMLS Metathesaurus : the MetaMap program », *AMIA Symposium*, vol. , p. 17, 2001.

Aronson A. R., Lang F. M., « An overview of MetaMap : historical perspective and recent advances », vol. 17, American Medical Informatics Association, p. 229-236, 2010.

Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., « DBpedia : A nucleus for a web of open data », *The semantic web*, p. 722-735, 2007.

Berant J., Chou A., Frostig R., Liang P., « Semantic Parsing on Freebase from Question-Answer Pairs. », *EMNLP*, vol. 2, p. 6, 2013.

Bordea G., Buitelaar P., Faralli S., Navigli R., « Semeval-2015 task 17 : Taxonomy extraction evaluation (texeval) », *EMNLP*, vol. , p. 902-910, 2015.

Carlson A., Betteridge J., Kisiel B., Settles B., Jr. E. H., Mitchell T., « Toward an architecture for never-ending language learning », *AAAI*, vol. 5, p. 3, 2010.

Dong X. L., Berti-Equille L., Srivastava D., « Integrating conflicting data : The role of source dependence », *PVLDB*, vol. 2, p. 550-561, 2009.

Etzioni O., Fader A., Christensen J., Soderland S., Mausam M., « Open Information Extraction : The Second Generation », *IJCAI*, vol. 11, p. 3-10, 2011.

Grefenstette G., « INRIASAC : Simple hypernym extraction methods », *EMNLP*,p. 911-914, 2015.

Jean P. A., Harispe S., Ranwez S., Bellot P., Montmain J., « Uncertainty detection in natural language : a probabilist model », *WIMS*, 2016.

Kalyanpur A., K.Boguraev B., Patwardhan S., Murdock J. W., Lally A., Welty C., Prager J. M., Coppola B., Fokoue-Nkoutche A., Zhang L., « Structured data and inference in DeepQA », *IBM Journal of Research and Development*, vol. 56, n° 3.4, p. 1-14, 2012.

Murphy K., Dong X., Gabrilovich E., Heitz G., Horn W., Lao N., Zhang W., « Knowledge Vault : A web-scale approach to probabilistic knowledge fusion », *20th ACM SIGKDD international conference on Knowledge discovery and data mining*, vol. , p. 601-610, 2014.

Nakashole N., Theobald M., Weikum G., « Integrating conflicting data : The role of source dependence », *ACM conference on Web search and data mining*, p. 227-236, 2011.

Nelson S. J., Johnston D., Humphreys B. L., « Relationships in Medical Subject Headings », *C.A. Bean and R. Green, editors, Relationships in the Organization of Knowledge*, p. 171–184, 2001.

Niu F., Zhang C., Ré C., Shavlik J., « Elementary : Large-scale knowledge-base construction via machine learning and statistical inference », *International Journal on Semantic Web and Information Systems*, p. 42-73, 2012.

Papasalouros A., Kanaris K., Kotis K., « Automatic Generation Of Multiple Choice Questions From Domain Ontologies », *e-Learning IADIS*, p. 427-434, 2008.

Pho V., Ligozat A. L., Grau B., « Distractor quality evaluation in multiple choice questions », *International Conference on Artificial Intelligence in Education*, p. 377-386, 2015.

Schmitz M., Bart R., Soderland S., Etzioni O., « Open language learning for information extraction », *EMNLP-CoNLL*, p. 523-534, 2012.

Seyler D., Yahya M., Berberich K., « Knowledge Questions from Knowledge Graphs », *arXiv preprint arXiv :1610.09935*, 2016.

Shafer G., « A mathematical theory of evidence », *Princeton university press*, vol. 1, 1976.

Surdeanu M., Tibshirani J., Nallapati R., Manning C. D., « Multi-instance multi-label learning for relation extraction », *EMNLP-CoNLL*, p. 455-465, 2012.

Tchechmedjiev A., « État de l'art : mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances », JEP-TALN-RECITAL, p. 295, 2012.

Wu W., Li H., Wang H., Zhu K. Q., « Probase : a probabilistic taxonomy for text understanding », *ACM SIGMOD International Conference on Management of Data*, p. 20-24, 2012.

Zhou X., Menche J., Barabási A., Sharma A., « Human symptoms–disease network », *Nature communications*, vol. 5, 2014.

Zhu J., Nie Z., Liu X., Zhang B., Wen J. R., « StatSnowball : a statistical approach to extracting entity relationships », *18th international conference on World wide web*, p. 101-110, 2009.