Gérer l'incertitude lors de l'extraction de relations et lors de l'inférence de nouvelles connaissances

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

Gérer l'incertitude lors de l'extraction de relations et lors de l'inférence de nouvelles connaissances

Pierre-Antoine Jean[1] Sébastien Harispe[1] Patrice Bellot[2] Sylvie Ranwez[1] Jacky Montmain[1]

[1] Laboratoire de Génie Informatique et d'Ingénierie de Production (LGI2P) - Nîmes
[2] Laboratoire des Sciences de l'Information et des Systèmes (LSIS) - Marseille

LGI2P, École des mines d'Alès, 69 rue Georges Besse
F-30035 Nîmes cedex 1
{prenom.nom}@mines-ales.fr
patrice.bellot@lsis.org

Résumé

Malgré leur volume important et leur accessibilité, de nombreuses données numériques ne peuvent être correctement exploitées car elles sont contenues dans des textes sous des formes peu ou pas structurées. L'extraction de relations est un processus qui rassemble des techniques pour extraire des entités et des relations à partir de textes, nous donnant la possibilité d'enrichir des bases de connaissances de façon automatique. Cependant le langage naturel est de façon intrinsèque porteur d'ambiguïté, ce qui constitue un premier niveau d'incertitude auquel on peut rajouter l'imprécision due aux formulations telles que "je crois que", "il semble que", etc. La base de connaissances doit donc tenir compte de cette incertitude par exemple en associant à chaque nouvelle connaissance extraite un score de confiance dépendant du degré de certitude. Cet article est une ***communication de synthèse*** *qui détaille les différentes problématiques liées à l'incertitude et à l'imprécision au cours de la chaîne de traitement allant de l'extraction d'information dans les textes à l'inférence de connaissances. Il y sera notamment question de stratégie d'agrégation des différentes sources d'incertitude et d'imprécision et de leur prise en compte dans les traitements ultérieurs (par exemple la recherche d'information ou l'aide à la décision).*

Mots Clef

TALN, extraction d'information, incertitude, inférence de règles

Abstract

Among the increasing volume of electronic resources available, non-structured texts expressed through natural language are difficult to process automatically. In this context, relation extraction techniques propose to combine various approaches to extract entities and their relations from texts, e.g. to automatically enrich a knowledge base. Nevertheless, the natural language is per se ambiguous, which makes extraction results uncertain. It can also be used to express imprecise or uncertain statements, "It seem to me", "I believe", etc. Therefore, any knowledge base enriched through text analyses must consider these uncertainties, for instance by combining a confidence score to each knowledge extraction according to its associated level of uncertainty. This information will be of major importance to infer additional knowledge from these extractions. However, how to characterize, capture and integrate the uncertainty and imprecision of natural language ? In addition, how to take into account this uncertainty to infer new knowledge ? This paper is a synthesis communication related to the consideration of uncertainty and imprecision in the context of Information Extraction from texts and knowledge inference from these extractions. We propose in particular to define the terminology, to characterize the several sources of uncertainty and to discuss strategies that can be used to capture and consider the uncertainty in knowledge extraction and knowledge inference treatments.

Keywords

NLP, information extraction, uncertainty, mining rules

1 Introduction

De nos jours, les bases de connaissances telles que DBPedia [1], YAGO [2] ou Freebase [3] contiennent des milliards de faits concernant des millions d'entités. Malgré tout, ces bases de connaissances restent incomplètes. De nombreuses informations complémentaires sont contenues dans des textes non structurés qui peuvent être utilisés pour enrichir ces bases et ainsi améliorer les traitements qui en découlent : recherche d'information, *question answering* ou aide à la décision. Trois principales approches s'offrent à

1. www.dbpedia.org
2. www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago
3. www.freebase.com

nous pour compléter ces bases de connaissances : *(i)* de manière manuelle, coûteuse et chronophage, *(ii)* de manière automatique grâce à **l'extraction d'informations** dans les textes, qui est plus incertaine mais bien plus rapide, ou *(iii)* par des méthodes **d'inférence de connaissances** à partir de faits dans la base de connaissances (*cf.* Figure 1). Dans notre étude, nous nous intéressons à ces deux dernières approches et plus particulièrement à la prise en compte de l'incertitude qui leur est inhérente.
Dans ce qui suit, nous positionnons notre étude par rapport aux approches proposées dans la littérature. Ensuite, nous discuterons plus particulièrement des méthodes d'extraction à partir de textes, puis de l'inférence de nouvelles connaissances.

2 Positionnement

2.1 L'extraction d'information

L'extraction d'information est un vaste domaine de recherche initié en 1992 par la première conférence MUC (*Message Understanding Conference*). Elle se définit comme l'extraction d'informations structurées à partir de textes écrits en langage naturel. Contrairement à la recherche d'information, elle a pour objectif de retourner directement au lecteur l'information souhaitée et non des pointeurs vers des documents. Dans le cadre de notre étude, nous nous focaliserons tout d'abord sur l'extraction de triplets (Sujet - Prédicat - Objet) ; on parle alors **d'extraction de relations binaires**, et aux marqueurs d'incertitude qui peuvent y être associés (contexte de la relation). On distingue traditionnellement trois principales tâches dans l'extraction d'informations : la **Reconnaissance d'Entités Nommées** (REN) dont le but est d'identifier et de désambiguïser des entités dans les textes, **l'extraction de relations** permettant d'extraire des connections sémantiques entre les entités identifiées et **l'extraction d'événements** qui consiste à remplir de manière automatique une structure informationnelle (un formulaire) associant différentes informations à un événement donné (par exemple pour un événement sportif, la structure informationnelle contiendrait : le nom des deux équipes, le score, le lieu, etc.). Ces trois tâches sont fortement dépendantes les unes des autres, l'extraction de relations nécessite une REN, notamment si elle est destinée à l'enrichissement d'une base de connaissances et l'extraction d'événements peut être perçue comme une extraction de relations n-aire. De nombreux travaux de la littérature proposent des approches différentes pour l'extraction d'informations. Certaines diffèrent en fonction de la disponibilité d'exemples annotés induisant l'orientation soit vers des méthodes supervisées soit vers des méthodes semi ou non supervisées et d'autres en fonction des textes analysés (c'est le cas des textes biomédicaux, par exemple, où le vocabulaire employé peut entraîner certaines difficultés, notamment pour la REN [12] - conventions de nommage non respectées, flux constant de nouvelles entités). Ces approches se distinguent aussi par les techniques employées qui peuvent être linguistiques, basées sur des patrons syntaxiques et lexicaux [5, 6], mais aussi statistiques, comme les modèles graphiques dirigés (chaînes de Markov cachées) ou non dirigés (champs markoviens aléatoires (CRFs)) [6, 11] ou encore basées sur des systèmes hybrides couplant les avantages de plusieurs techniques [15].

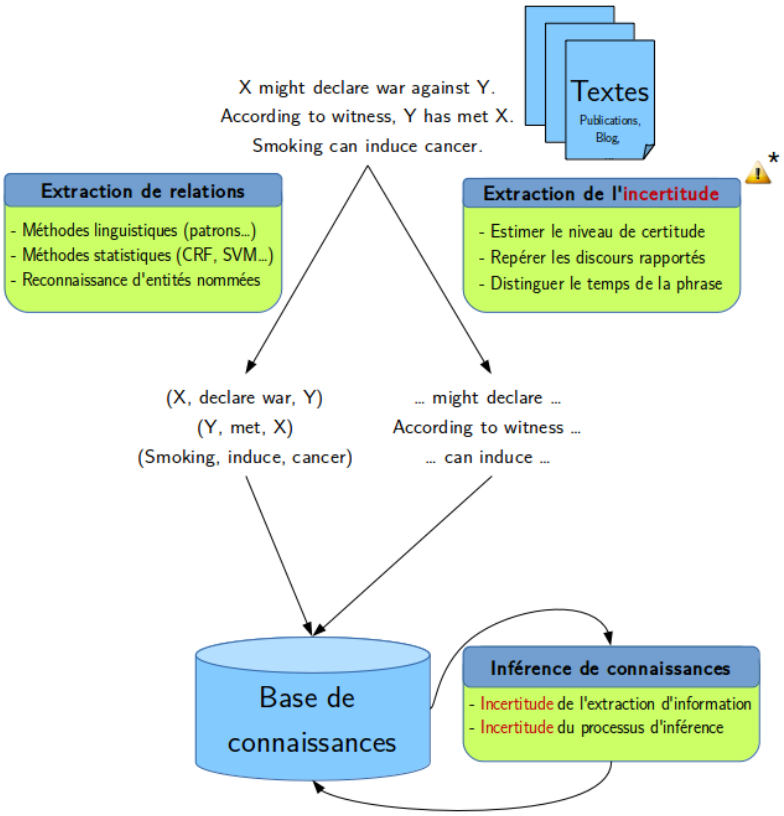

FIGURE 1 – Structure globale de l'approche envisagée. L'extraction d'informations permet de coupler les informations contenues dans les textes (relations, incertitude) avec une base de connaissances qui, elle-même, permet d'inférer de nouvelles connaissances. Le symbole (*) précise que nous nous intéressons à l'incertitude inhérente au contenu des textes. La source et le croisement contradictoire de différentes relations ne sont pas considérés.

2.2 L'inférence de connaissances

Les méthodes d'extraction d'informations permettent d'enrichir de manière automatique une base de connaissances. Cette dernière peut servir ensuite de support à différents traitements (*e.g* aide à la décision) qui seront d'autant plus performants que la base sera complète. Pour la compléter il est possible, à partir des connaissances présentes dans la base et de certaines règles d'inférence, d'inférer de nouvelles connaissances. Ce processus éprouvé depuis longtemps en *Data Mining* peut avoir plusieurs objectifs [9] : trouver de nouvelles connaissances et de nouveaux faits dans la base de connaissances, identifier d'éventuelles erreurs (d'insertion ou d'inférence) et enfin permettre d'avoir

une meilleure compréhension des données par l'intermédiaire de règles décrivant des phénomènes généraux. De plus, [9] précise que les techniques couramment utilisées en *Data Mining* (par exemple l'apprentissage par Programmation Logique Inductive -PLI- qui utilise des exemples positifs et négatifs pour trouver des hypothèses couvrant tous les exemples positifs et aucun exemple négatif) ne sont pas adaptées aux exigences des bases de connaissances de type ontologie dû aux contraintes qui y sont liées : important volume de données (passage à l'échelle) et hypothèse à domaine ouvert. Ce type d'hypothèse signifie qu'une donnée absente ne peut être utilisée comme un contre-exemple. Pour répondre à ce problème, Muggleton [13] a développé un score d'évaluation d'apprentissage basé uniquement sur les exemples positifs pour la PLI, l'approche génère des contres exemples de manière aléatoire. L'article [9] présente une approche (AMIE) qui utilise une autre stratégie pour générer des contres exemples [8] : *the Partial Completeness Assumption* (PCA) qui suppose que si une base de connaissances connaît une certaine relation pour une entité elle connaît alors toutes ses relations.

2.3 Gestion de l'incertitude

Un des objectifs de notre étude réside dans la prise en compte et la gestion de l'incertitude inhérente au langage naturel ou liée à l'inférence de règles et de nouvelles connaissances. L'article [5] décompose l'incertitude associée aux textes en deux sous-parties. Elle distingue : l'ambiguïté et l'imprécision. L'**ambiguïté linguistique** correspond à plusieurs associations possibles entre des symboles (des mots ou des suites de mots) et leurs significations. La polysémie [4] et l'homonymie [5] sont les principales sources d'ambiguïté. La différence entre ces deux notions est que deux mots homonymes partagent une même forme orale et/ou écrite mais n'ont pas la même étymologie. Par exemple, la phrase : « J'aime le rouge. » est ambiguë car « rouge » est un polysème pouvant signifier ici, le vin ou la couleur. L'ambiguïté est principalement traitée dans les méthodes de REN. L'article [3] présente une méthode se basant sur l'environnement lexical d'un terme pour le désambiguïser. La méthodologie qui y est décrite s'appuie sur un graphe de coocurrences des termes apparaissant avec le mot ambigu. **L'imprécision** (le flou) représente un savoir incomplet qui est exprimé sur des faits ou des événements. Par exemple, la phrase « Plusieurs véhicules se dirigent vers l'est » est imprécise de par l'utilisation de l'adjectif indéfini « Plusieurs ». L'article [14] propose une catégorisation des marqueurs de certitude présents dans les phrases. Ces catégories tiennent compte du degré de certitude exprimé (fort ou faible), du temps employé, de la perspective (discours rapporté ou point de vue de l'écrivain) ou de la nature de l'information délivrée. Ces marqueurs peuvent inclure par exemple les adverbes épistémiques « selon moi, à mes yeux, à mon avis » permettant au locuteur de mentionner sa subjectivité et ainsi de relativiser son propos ou bien l'emploi du conditionnel donnant le point de vue du locuteur sur l'énoncé [4].
L'incertitude peut être appréhendée sur d'autres aspects que le texte, notamment au niveau de la source de l'information. Par exemple, si un pneumologue dit « Le tabac peut induire le cancer » et une personne *lambda* dit « Le tabac induit le cancer », malgré le marqueur d'incertitude présent dans la phrase du pneumologue, la source est plus fiable et la phrase plus exacte. L'incertitude peut être également observée lors du croisement de plusieurs documents affirmant une information contradictoire. Cependant pour cette étude, seule la première source d'incertitude liée aux textes est traitée (en ce qui concerne l'incertitude inhérente aux textes).
Cette étude se concentrera sur deux éléments centraux que sont l'**information**, qui est extraite à partir de textes sous la forme de triplets (Sujet/EN1, Prédicat/R, Objet/EN2) et les **connaissances** qui peuvent être inférées à partir de ces informations et structurées dans une base de connaissances. De plus, la gestion de l'incertitude, inhérente aux textes et au processus d'extraction d'information et liée aux approches d'inférence de connaissances et de règles, sera un élément central de notre étude.

3 Exploitation des textes

3.1 L'extraction de relations

La première partie de l'étude se focalisera sur l'extraction de relations entre deux entités nommées dans une phrase. L'objectif est de récupérer un triplet désambiguïsé et de relever les marqueurs d'incertitude associés. L'article [16] caractérise la structure d'une relation en trois principales parties construites autour des entités : la partie *Cmid* qui porte généralement la relation et les parties *Cpre* et *Cpost* qui apportent généralement des précisions sur le contexte (*cf.* Figure 2). L'article [6] précise que sur un corpus de 300 phrases aléatoires du Web, 96% des relations binaires possèdent ce format et que 4% des relations ne le respectent pas (par exemple « ***Discovered by*** *Y,X ...* » ou bien « *... the Y that X* ***discovered*** » avec X et Y représentant les deux entités). L'extraction de relations peut être à domaine ouvert [7] ou se focaliser sur des relations précises, par exemple la catégorisation de relations de régulation entre deux composants cellulaires dans des textes biomédicaux [2].

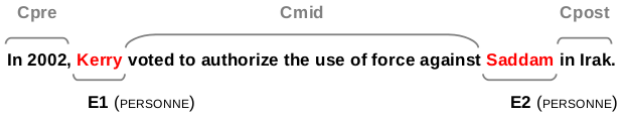

FIGURE 2 – Exemple de la structure type d'une relation extraite à partir de texte [16].

4. La polysémie est la caractéristique d'un mot ou d'une expression qui a plusieurs sens (au moins deux) ou significations différentes. Par exemple « Opéra » peut signifier la pâtisserie, le lieu ou l'art.

5. L'homonymie est la caractéristique de plusieurs mots partageant une même forme orale et/ou écrite mais qui ont des sens différents, par exemple « Une livre de pain qu'il livre avec un livre de recettes ».

L'article [16] propose trois traitements afin d'extraire des relations (EN1, R, EN2) entre des entités pré-définies (PERSONNE, LIEU, ORGANISATION). Le premier traitement est une analyse linguistique. Cette analyse inclut une homogénéisation du contenu textuel grâce à une lemmatisation, une désambiguïsation morpho-syntaxique et une REN. Le second traitement correspond à l'extraction des relations candidates en filtrant les phrases contenant au moins un verbe entre les deux entités. Enfin, le troisième traitement permet de filtrer les relations obtenues probablement fausses (discours rapporté, relations trop complexes) à l'aide de champs conditionnels aléatoires (CRFs) linéaires entraînés sur 1000 exemples de relations. Leur système d'extraction de relations, qui est basé sur une stratégie de récupération des entités en premier, obtient une précision de 76.2% et un rappel de 78.2% sur les relations extraites à partir d'une sous-partie du corpus AQUAINT-2 contenant 18 mois d'articles du quotidien *The New York Times*.

L'article [1] propose une méthodologie pour extraire les relations entre les types sémantiques d'entités médicales (par exemple *Diazoxide - typeOf - treatment*) en utilisant des patrons linguistiques et une base de connaissances. Leur méthodologie débute par une REN spécialisée dans le domaine médical à l'aide de MetaMap [6], leur permettant également de récupérer les types sémantiques associés à chaque entité en utilisant UMLS Metathesaurus [7]. La méthode récupère par la suite toutes les relations existantes entre deux types sémantiques dans *UMLS Semantic Network*. Puis, pour chaque type de relation, un patron linguistique est construit et utilisé pour la reconnaissance de nouvelles relations. La construction d'un patron pour une relation donnée débute par la recherche de l'ensemble des paires de termes reliées par cette relation dans UMLS. Par la suite, une collection de textes contenant ces paires est récupérée grâce au système de requêtes avancées de PubMed. Chaque requête est affinée afin d'approximer les textes qui vont potentiellement contenir la relation souhaitée à partir d'une entité donnée (par exemple *Rhinitis, Vasomotor/TH* est une requête décrivant une relation de traitement (/TH) entre un traitement non spécifié et une *Rhinitis*). Leur méthode, basée elle aussi sur l'extraction des entités, obtient un score de précision de 75,72% et un score de rappel de 60,46%, soit une F-mesure de 67,23%.

Contrairement aux deux précédents systèmes présentés, ReVerb, introduit dans [7], débute par la récupération de signature de relations. En effet, l'approche utilise un ensemble de contraintes syntaxiques, sous la forme de patrons basés sur les étiquettes morpho-syntaxiques (*part-of-speech* (POS)) (*cf.* figure 3) et de contraintes lexicales permettant d'éviter l'extraction de relations trop spécifiques. Les patrons lexicaux sont basés sur la construction d'un large dictionnaire d'entités récupérées sur 500 millions de relations à partir du Web ; ainsi le nombre d'apparitions d'une paire d'entités doit être supérieur à une valeur seuil pour qu'elle soit considérée. Cette première phase permet de récupérer 85% des relations verbales binaires. Une fois ces relations extraites, la méthode détermine les limites des entités contenues dans la phrase en utilisant trois classifieurs spécifiquement entraînés pour repérer les limites gauches et droites des différentes entités [6]. Les auteurs utilisent REPTree [8] de Weka et un CRF et obtiennent une précision d'environ 60% et un rappel de 70%.

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

FIGURE 3 – Expressions régulières basées sur le POS des relations verbales. Ces expressions extraient soit un verbe simple (*invented*), soit un verbe suivi par une préposition (*located in*) ou soit un verbe suivi par un syntagme nominal et terminant par une préposition (*has atomic weight of*) [6].

Nous pouvons observer que les trois méthodes présentées considèrent uniquement l'extraction de relations sans tenir compte des marqueurs d'incertitudes qui peuvent être présents dans une phrase. Dans ces applications, la phrase « *According to witness, Y has met X* » est réduite à sa seconde partie [10]. La perte de l'expression de l'incertitude modifie la compréhension et la fiabilité de l'information délivrée par la phrase. La sous-section suivante présente des approches abordant le problème de l'extraction de marqueurs d'incertitude.

3.2 Évaluation et extraction de l'incertitude

Notre objectif initial est d'inférer de la connaissance en prenant en compte les différents niveaux d'incertitude : au niveau des textes et des connaissances inférées. En effet, si nous ajoutons un fait dans une base de connaissances sans tenir compte de l'incertitude, le résultat n'aura pas la même signification que l'information originelle exprimée dans le texte. La prise en compte de l'incertitude au niveau des textes passe par la reconnaissance de différents éléments phrastiques ou expressions qui nuancent l'information délivrée par une relation. Rubin et *al.* propose dans [14] un modèle de catégorisation de la certitude au travers de quatre dimensions :

— **le niveau de certitude** (*Absolute, High, Moderate, Low*). Par exemple, l'expression suivante « *... will almost certainly have to ...* » représente un niveau de certitude *Absolute* tandis que l'utilisation du modal *might* dans « *... might buy ...* » représente un niveau de certitude *Low* ;

— **la perspective** (point de vue de l'écrivain, point de vue rapporté). Par exemple, la certitude dans la phrase « *More evenhanded coverage of the presidential race would help enhance the legitimacy of*

6. www.metamap.nlm.nih.gov/
7. www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
8. www.weka.sourceforge.net/doc.dev/weka/classifiers/trees/REPTree.html

the eventual winner, which ***now appears likely to be*** *Putin.* » est attribuée à l'écrivain selon ses connaissances et son expérience au moment où il écrit la phrase tandis que « *According to witness, Y has met X* » est un discours rapporté dans lequel la fiabilité de l'information est associée à celle de la source secondaire « *witness* » ;

— **le *focus*** de la certitude. Cette dimension est divisée en deux parties en fonction de la nature de l'information délivrée, soit abstraite (jugements, opinions, croyances, émotions) qui reflète une idée qui ne représente pas une réalité mais plutôt un monde hypothétique ou soit factuelle qui rapporte des états ou des événements et des faits connus ;

— **le temps.** Cette dimension prend en compte la pertinence du temps (passé, présent, futur). Le passé inclut des événements complets, le présent des états immédiats et incomplets et le futur est une prédiction ou une action suggérée.

L'article [10] s'inspire et enrichit ce modèle de catégorisation afin d'évaluer la certitude dans les phrases. Leur approche ajoute une nouvelle dimension qui est l'identification de la source (si la phrase est un discours rapporté) et modifie la troisième dimension (*focus*) par la notion de « réalité » qui permet de différencier l'affirmatif et le négatif (le modèle précédent ne traitait pas ces notions). Par la suite, leur approche utilise des patrons linguistiques pour détecter dans les textes l'incertitude selon ces cinq dimensions. Ces patrons se basent sur une association entre des termes ou des expressions et une dimension particulière, par exemple, l'adjectif « présumé » est considéré dans leur modèle comme un niveau d'incertitude *Moderate* ou bien les structures « selon, d'après, de source(s) ... » indique un point de vue rapporté. Enfin, chacune des identifications ajoute une annotation en entête d'un fichier RDF permettant de préciser les cinq dimensions (par exemple : *<onto :Level>Moderate</onto :Level>*). Leur approche a été évaluée sur chacune des dimensions : les dimensions Source et Niveau sont les dimensions obtenant la plus faible F-mesure (64% et 69%) tandis que la dimension temps obtient une F-mesure de 100%.

Ces différentes annotations pourraient permettre une pondération associée aux triplets lorsqu'ils sont insérés dans la base de connaissances. Cette pondération serait une information importante à considérer au sein du processus d'inférence.

4 Inférence de connaissances

Les bases de connaissances sont au carrefour de plusieurs disciplines : la recherche d'information, le traitement automatique des langues avec les systèmes de *question answering* et le raisonnement. La deuxième partie de nos travaux consistera à exploiter les relations extraites afin d'enrichir une base de connaissances et d'y appliquer des mécanismes de raisonnement en considérant l'incertitude liée aux phrases et à la connaissances et/ou règles qui pourront être inférées à partir de cet enrichissement. Actuellement, nous nous concentrons sur l'extraction d'informations, aussi cette seconde partie présente uniquement au travers d'un exemple les résultats que l'on attendrait.

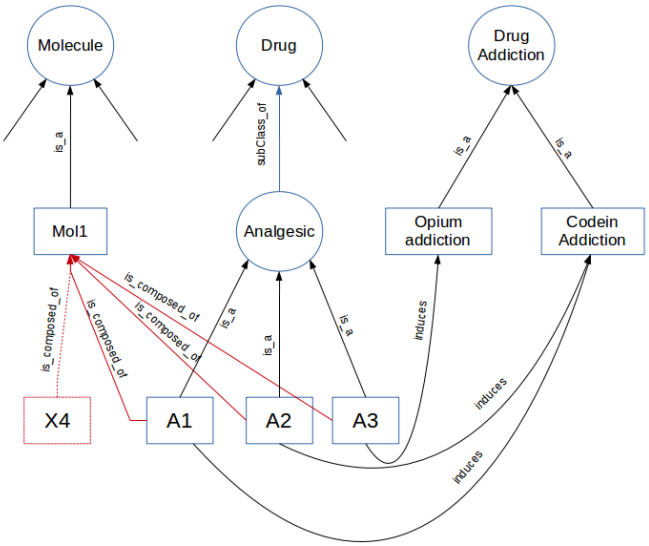

FIGURE 4 – Base de connaissances partielle d'une relation entre un analgésique et une molécule et entre un analgésique et une addiction.

La figure 4 décrit une base de connaissances partielle sur les médicaments, les molécules et les addictions aux médicaments. On peut observer que les trois instances de la classe *Analgesic* contiennent la molécule Mol1. Supposons que cette molécule soit principalement rattachée aux analgésiques dans la base de connaissances complète. Imaginons maintenant que nous avons extrait la phrase suivante : « X4 est un médicament composé de la molécule Mol1. ». Une induction relativement évidente, au vu de la base de connaissances, est d'inférer que X4 est un analgésique, donc inférer de la connaissance. Cette inférence doit être couplée avec une estimation de l'incertitude afin d'apporter à la connaissance une mesure de fiabilité. Cette incertitude doit dépendre du contexte d'incertitude des phrases, par exemple : « X4 est un médicament qui pourrait être composé de Mol1 » ne doit pas être évaluée de le même manière que la phrase précédente car elle nuance la certitude de la relation. De plus, les scores de confiance résultant des phases de désambiguïsation des entités correspond également à une information importante à prendre en compte. L'idée sous-jacente, qui constitue les prémices de nos travaux, est de réaliser une fonction pour agglomérer les différentes formes d'incertitude pouvant être extraites des phrases afin de pondérer une relation entre deux entités dans une base de connaissances.

De plus dans ce schéma de connaissances, on observe que chaque instance de la classe *Analgesic* induit une addiction aux médicaments. Dans le cas présent, inférer la règle suivante : « *Analgesic induce Drug Addiction* » est intui-

tif. L'article de Leaman et al. [9] présente une méthode d'extraction de règles qui explore l'espace de recherche des règles possibles à l'aide d'une extension itérative des règles de Horn en ajoutant des opérateurs d'extraction spécifiques dans le corps de la règle. Ce type de règle possède une tête représentée par une unique relation (*fatherOf(f ;c)*) et un corps constitué d'une conjonction de faits, par exemple :

$$motherOf(m\,;c) \wedge marriedTo(m\,;f)) \Rightarrow fatherOf(f\,;c)$$

Afin de réduire l'espace de recherche leur approche extrait des règles de Horn dites fermées correspondant au fait que chaque variable apparaisse au moins deux fois dans la règle (comme dans l'exemple précédent). De plus, les auteurs proposent une mesure de confiance associée à la règle extraite. Cette mesure est basée sur le nombre d'instanciations de la règle qui apparaissent dans la base de connaissances normalisée avec l'ensemble des faits positifs (en utilisant la *Partial Completeness Assumption*). Cette mesure considère uniquement le processus d'inférence (la base de connaissances est certaine).

5 Conclusion

Nous avons présenté l'état actuel de notre positionnement par rapport aux approches de la littérature concernant l'extraction de relations à partir de textes et la prise en compte de l'incertitude. Cette incertitude peut provenir de différentes sources et impacter le processus d'extraction de connaissances, à différents niveaux. Nous avons observé que les marqueurs d'incertitude présents dans les phrases peuvent être positionnés selon plusieurs dimensions principales : le niveau, la perspective, le *focus* et le temps. Ces marqueurs peuvent être extraits par l'intermédiaire d'approches de Traitement Automatique des Langues, telles que les patrons linguistiques appris automatiquement ou non, et permettent d'ajouter une annotation sur la certitude de la phrase. Par la suite, une base de connaissances construite avec ces données incertaines représente une mine de connaissances dans laquelle nous pourrions inférer des règles pouvant prendre en compte d'une part l'incertitude inhérente aux textes et d'autre part l'incertitude de l'inférence elle-même.

Références

[1] A. B. Abacha and P. Zweigenbaum. Automatic extraction of semantic relations between medical entities : a rule based approach. In *Fourth International Symposium on Semantic Mining in Biomedicine*, 2011.

[2] S. Ananiadou, P. Thompson, R. Nawaz, J. McNaught, and D. B. Kel. Event-based text mining for biology and functional genomics. In *Briefings in Functional Genomics*, 2014.

[3] B. Andreopoulos, D. Alexopoulou, and M. Schroeder. Word sense disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering. *Int. J. Data Mining and Bioinformatics*, 2008.

[4] A. Borillo. Les « adverbes d'opinion forte » selon moi, à mes yeux, à mon avis,... : point de vue subjectif et effet d'atténuation. *Langue française*, 2004.

[5] V. Dragos. An ontological analysis of uncertainty in soft data. *Information Fusion (FUSION), 2013 16th International Conference on*, 2013.

[6] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and Mausam. Open information extraction : the second generation. *IJCAI*, 2011.

[7] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. *Conference on Empirical Methods in Natural Language Processing*, 2011.

[8] L. Galárraga, N. Preda, and F. Suchanek. Mining rules to align knowledge bases. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, 2013.

[9] L. Galárraga, C. T. F. Suchanek, and K. Hose. Amie : Association rule mining under incomplete evidence in ontological knowledge bases. 2013.

[10] B. Gounjon. Uncertainty detection for information extraction. In *International conference RANLP*, 2009.

[11] Y. Kim, P. Bellot, E. Faath, and M. Dacos. Automatic annotation of bibliographical reference in digital humanities books, articles and blogs. *Proceedings of the CIKM*, 2011.

[12] R. Leaman and G. Gonzalez. Banner : an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, 2008.

[13] S. Muggleton. Learning from positive data. In *Inductive Logic Programming Workshop*, 1996.

[14] V. Rubin, E. Liddy, and N. Kando. Certainty identification in texts : Categorization model and manual tagging results. *Computing Attitude and Affect in Text : Theory and Applications, The Information Retrieval Series*, 2005.

[15] C. Wang, A. Kalyanpur, J. Fan, B. K. Boguraev, and D. C. Gondek. Relation extraction and scoring in deepqa. In *IBM Journal of Research and Development*, 2012.

[16] W. Wang, R. Besançon, O. Ferret, and B. Grau. Regroupement sémantique de relations pour l'extraction d'information non supervisée. *TALN-Récital*, 2013.