État de l'art sur l'acquisition de relations sémantiques entre termes : contextualisation des relations de synonymie

Mounira Manser
LIM&BIO (EA3969)
Université Paris 13
93017 Bobigny Cedex
France
manser.mounira@gmail.com

RÉSUMÉ

L'accès au contenu des textes de spécialité est une tâche difficile à réaliser. Cela nécessite la définition de méthodes automatiques ou semi-automatiques pour identifier des relations sémantiques entre les termes que contiennent ces textes. Nous distinguons les approches de TAL permettant d'acquérir ces relations suivant deux types d'information : la structure interne des termes ou le contexte de ces termes en corpus. Afin d'améliorer la qualité des relations acquises et faciliter leur réutilisation en corpus, nous nous intéressons à la prise en compte du contexte dans une méthode d'acquisition de relations de synonymie basée sur l'utilisation de la structure interne des termes. Nous présentons les résultats d'une expérience préliminaire tenant compte de l'usage des termes dans un corpus biomédical en anglais. Nous donnons quelques pistes de travail pour définir des contraintes sémantiques sur les relations de synonymie acquises.

ABSTRACT

State of the Art on the Acquisition of Semantic Relations between Terms : Contextualisation of the Synonymy Relations

Accessing to the context of specialised texts is a crucial but difficult task. It requires automatic or semi-automatic methods dedicated to the identification of semantic relations between terms appearing in the texts. NLP approaches for acquiring semantic relations between terms can be distinguished according to the type of information : the internal structure of the terms and the term context. In order to improve the quality of the acquired synonymy relations and their reusability in other corpora, we aim at taking into account the context into an approach based on the internal structure of the terms. We present the results of a preliminary experiment taking into account the use of the terms in a English biomedical corpora. This experiment will be helpful to add semantic constraints to the already acquired synonymy relations.

MOTS-CLÉS : Acquisition de relations, Synonymie, Relations sémantiques, Terminologie, Domaine Biomédical, Corpus de spécialité.

KEYWORDS: Relation Acquisition, Synonymy, Semantic Relations, Terminology, Biomedical Domain, Specialised corpora.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 163–175,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Devant la masse considérable de donnés disponibles, la difficulté d'extraire et de rechercher l'information à partir des textes devient de plus en plus importante et demande un effort de structuration et d'ordonnancement des données. Les ressources terminologiques répondent partiellement à ce besoin en proposant les termes du domaine et différents types de relations (synonymie, hypéronymie ou des relations plus spécifiques au domaine comme *a-pour-symptôme*). La projection des termes issus d'une terminologie ne suffit pas (Cabré, 1999; Hamon, 2005; Spasic *et al.*, 2005; McIntosh et Curran, 2009). Il est nécessaire de disposer de terminologies structurées afin d'être capable de s'adapter aux usages dans les documents manipulés. Pour cela, des approches automatiques ou semi-automatiques doivent être mises en œuvre. Des approches de TAL ont ainsi été proposées pour aider à l'identification des termes et des relations permettant de structurer ces listes de termes. Cependant, la qualité des résultats peut varier en termes de rappel (les méthodes sont trop restrictives) ou de précision (les ambiguïtés ou la polysémie sont peu ou mal prises en compte).

Nous nous intéressons ici à l'acquisition de relations de synonymie entre termes. Notre travail consiste à proposer et à tester des méthodes dans le but d'améliorer l'acquisition de relations de synonymie produites par SynoTerm (Hamon et Nazarenko, 2001). Il s'agit de filtrer des relations de synonymie en exploitant les informations contextuelles et sémantiques liées aux termes ou à leur composants. Le travail est réalisé sur des données issues du domaine biomédical en langue anglaise. Nous travaillons à la fois avec les ressources terminologiques Gene Ontology et UMLS, ainsi qu'avec un ensemble des résumés Medline (voir section 3).

Dans cet article nous présentons d'abord les relations sémantiques pouvant être fournies par une ressource terminologique (section 2.1). Puis nous présentons un état de l'art des approches d'acquisition de ces relations (section 2). La section 3 est consacrée au matériel que nous avons utilisé pour nos expériences. Nous décrivons à la section 4 les pistes de travail pour l'amélioration de l'acquisition de relations de synonymie et les résultats d'une expérience préliminaire (section 5).

2 Approches pour l'identification des relations sémantiques entre les termes

Différents types d'approches de TAL permettent d'extraire des relations sémantiques entre les termes. L'acquisition de telles relations est réalisée soit en exploitant la structure interne des termes issus de corpus ou de terminologie (section 2.2), soit en s'appuyant sur le contexte de ces termes en corpus (section 2.3). Dans cette section, nous présentons tout d'abord les types de relations pouvant apparaître dans une terminologie puis un panorama des approches et des types de relations qu'elles peuvent permettre d'acquérir.

2.1 Relations sémantiques dans une terminologie

Les terminologies visent à recenser les termes d'un domaine de spécialité, c'est-à-dire les unités linguistiques désignant un concept, un objet ou un processus (Bourigault et Jacquemin, 2000),

mais aussi les relations sémantiques qu'entretiennent ces termes entre eux. Plusieurs types de relations sémantiques sont proposées par les ressources terminologiques (Sager, 1990), le choix d'inclure un type de relation étant surtout dépendant de l'usage de la ressource :

– les **relations taxinomiques** : Ce type de relation structure des termes dans une arborescence.

  Les relations d'**hypéronymie** (`is-a/est-un`) relient un terme général à un terme spécifique. Par exemple, nous avons les relations *oxidase* `is-a` *enzyme* ou *contractil fiber* `is-a` *fiber*.

  Les **relations partitives** (méronymie ou partie-tout) sont utilisées pour définir une relation entre deux termes ou l'un est une partie de l'autre. On a par exemple la relation de méronymie *nucleus* `partie-de` *cell*.

– Les **relations sémantiques lexicales** regroupent deux types de relations :

  les relations de `synonymie` ou d'équivalence qui relient les termes possédant le même sens, par exemple *red blood* et *erythrocyte* ;

  les relations d'`antonymie` ou d'opposition qui relient les termes ayant des sens contraires, par exemple *anabolosm* et *catabolism*

– Les **relations inter-hiérarchiques** (transversales) relient les termes appartenants à des branches distinctes d'une ou plusieurs hiérarchies. Ces relations sont très variables suivant les domaines. Par exemple, la relation `localisée-dans` permet de lier les termes *division cellulaire* et *cellule*.

2.2 Exploitation de la structure interne des termes

L'acquisition des relations sémantiques basée sur la structure interne des termes a donné lieu à de nombreux travaux. Ceux-ci utilisent différents types d'informations issus de l'analyse linguistique des termes (informations morphologiques, syntaxiques et plus rarement sémantiques) et/ou des indices statistiques comme la fréquence ou la productivité. La combinaison de ces travaux s'avèrent très utiles pour la structuration d'une liste de termes (Daille, 2003).

Le partage de bases morphologiques communes peut aider à l'acquisition de relations sémantiques. Ainsi, dans le domaine biomédical, (Zweigenbaum et Grabar, 2000) ont exploité ces types d'informations morphologiques entre des termes et leur productivité pour identifier les relations de synonymie, d'hypéronymie ou inter-hiérarchiques (*acide, acido, acidité, acidurie, acidocitose, acidophile*). La dérivation est également très utile pour identifier des variantes terminologiques en corpus (Jacquemin, 1997; Grabar et Hamon, 2006) ou structurer sémantiquement un lexique dans un contexte multilingue (Namer et Baud, 2007). Ainsi, le premier travail permet d'identifier une relation entre *stenosis of the aorta* et *aortic stenosis*. Tandis que dans le second, il est possible de lier les termes *proctorrhagia* et *colorrhagia*, ou les adjectifs *bactérioïde* et *bactériforme*.

D'autres travaux utilisent des techniques d'apprentissage par analogie sur l'emploi des fonctions lexicales pour identifier des relations inter-hiérarchiques entre les termes (Claveau et L'Homme, 2005). Par exemple, les auteurs exploite la fonction lexicale connue entre *connecteur* et *connecter* (caractérisant la relation inter-hiérarchique `instrument_pour`), pour identifier la même relation ou la même fonction lexicale entre *éditeur* et *éditer*.

L'analyse syntaxique des termes permet généralement d'identifier des relations d'hypéronymie. Ainsi, de nombreux travaux exploitent l'inclusion lexicale, c'est-à-dire l'hypothèse que lorsqu'un

terme est lexicalement inclus dans un autre, une relation d'hypéronymie peut être établie (par exemple, *fatty acids **essential** / fatty acids*). Pour cela, ils s'appuient sur une décomposition en tête/expansion des termes (Bourigault, 1994; Bodenreider *et al.*, 2001; Grabar et Zweigenbaum, 2002). Une analyse fine des relations induites par inclusion lexicale montre cependant que celle-ci permet également de produire des relations inter-hiérarchiques (Ibekwe-SanJuan, 2005). Dans Faster (Jacquemin, 1997), des variantes terminologiques sont identifiées à l'aide des mécanismes syntaxiques d'insertion, de juxtaposition et de coordination. Une approche similaire mais basée uniquement sur le remplacement de chaînes de caractères est utilisée par (Verspoor *et al.*, 2003) pour acquérir des relations d'hypéronymie dans les termes de Gene Ontology. La variation verbo-nominale combinée à un apprentissage inductif peut également être pris en considération pour identifier des relations sémantiques (Bouillon *et al.*, 2000).

Bien qu'en général les informations sémantiques sont plutôt utilisées dans des méthodes basées sur le contexte des termes, quelques travaux prenant en compte la structure interne des termes exploitent ces types d'informations. Il est alors nécessaire de faire l'hypothèse de la compositionnalité des termes et de s'appuyer sur la présence d'un invariant syntaxique. Les indices sémantiques sont alors des relations entre les composants des termes. Ainsi, le processus d'acquisition de variation morpho-syntaxique proposé par (Jacquemin, 1997) peut être étendu en exploitant des relations de synonymie (Jacquemin, 1999). D'autres travaux visent à propager les relations sémantiques sur les termes complexes (Hamon et Nazarenko, 2001) en combinant différentes ressources lexicales. La qualité des relations inférées dépend de la spécialisation des relations sémantiques initiales par rapport au domaine. Un moyen d'obtenir ces relations initiales spécifiques au domaine consiste à utiliser les relations sémantiques issues d'une terminologie (Verspoor *et al.*, 2003; Hamon et Grabar, 2008). Les relations initiales induites doivent cependant être filtrées ou contextualisées, comme nous le présentons à la section 4.

2.3 Prise en compte du contexte des termes

Les approches exploitant le contexte des termes s'appuient sur l'hypothèse que la sémantique des termes peut être identifiée avec les contextes dans lesquels ils apparaissent. Ainsi, outre la désambiguïsation sémantique des termes, l'étude du contexte des termes fournit des indices importants et parfois indispensables pour acquérir et caractériser les relations sémantiques qu'ils entretiennent entre eux.

Nous nous intéressons ici principalement aux travaux réalisés sur les langues de spécialité et qui s'intéressent essentiellement à identifier les relations hiérarchiques et les relations inter-hiérarchiques. Les relations de synonymie sont plus rarement identifiées de cette manière.

2.3.1 Définition de patrons lexico-syntaxiques

Une des principales stratégies pour acquérir des relations sémantiques à l'aide du contexte des termes consiste à définir des patrons lexico-syntaxiques caractéristiques de la relation visée (Hearst, 1992; Auger et Barrière, 2008). Les patrons sont définis à partir d'observations en corpus et permettent d'extraire des relations d'hypéronymie. Par exemple, le patron *NP , NP * , or other NP* appliqué à la phrase "*Bruises, wounds, broken bones or other injuries...*" identifie des relations d'hypéronymie *bruise* `is_a` *Injury*, *wound* `is_a` *injury* et *broken bone* `is_a` *injury*. L'utilisation

d'une méthode d'identification automatique des patrons lexico-syntaxiques en corpus permet d'affiner les observations réalisées et d'obtenir de meilleurs résultats (Morin, 1999). Des relations transversales peuvent également être acquises grâce à cette stratégie (Hamon *et al.*, 2010; Røst *et al.*, 2010).

Si les approches basées sur les patrons exploitent en général des informations lexicales et morpho-syntaxiques dans un contexte relativement restreint, certains travaux cherchent à utiliser des relations de dépendance syntaxique (Snow *et al.*, 2005). Les patrons syntaxiques sont alors construits par apprentissage supervisé sur des chemins de dépendance syntaxique entre les termes. Un filtrage sur la longueur de ces chemins est ensuite appliqué. C'est également le cas lorsqu'il s'agit d'identifier des relations inter-hiérarchiques. Les patrons lexico-syntaxiques servent de base à la définition de modèle d'apprentissage exploitant les CRF (Yang et de Roeck, 2010) ou les SVM (Grouin *et al.*, 2010). Dans ces deux travaux, les entités mises en relations ainsi que leur types sémantiques sont connues, ce qui est généralement pas le cas dans les approches classiques visant à acquérir des relations sémantiques entre termes. Les indices trouvés en corpus dans le contexte peuvent être exploités pour inférer des patrons par programmation logique inductive (Martienne et Morin, 1999; Claveau et L'Homme, 2004). Des relations d'hypéronymie, dans le premier travail, ou des relations entre des verbes et des noms, dans le second, peuvent ainsi être identifiées.

On peut aussi remarquer que les patrons lexico-syntaxiques sont plus rarement utilisés pour l'acquisition de relations de synonymie. Cependant, il semble que certains domaines de spécialité, comme la biologie, s'y prêtent mieux (Weissenbacher, 2004; Mccrae et Collier, 2008). Ce phénomène particulier est probablement dû aux pratiques de renommage des noms de gènes lors de la découverte de leur fonction, l'explicitation de termes, mais aussi aux efforts dans ce domaine pour améliorer la recherche d'information et l'interopérabilité sémantique entre les terminologies existantes et les textes. Ainsi, par exemple le patron *also known* permet d'extraire la relation de synonymie dans l'extrait suivant : *reculatory factor (IRF) also know IRF-8*.

Des contextes plus larges peuvent être exploités pour acquérir des relations entre termes. Ainsi, des contextes riches en connaissance (*Knowledge-Rich Context*) sont définis selon (Meyer, 2001) comme étant un contexte qui contient des termes d'un domaine spécialisé et des models (patterns) de connaissaces. Des relations hiérarchiques (hypéronymie et méronymie) et des relations transversales (Schumann, 2011) entre les termes et les models sont alors obtenues. Par exemple, la phrase *tNF kappa B is a potent mediator of specific gene expression in human monocytes and has been shown to play a role in transcription of the HIV-1 genome in promonocytic leukemias* est défini comme un contexte riche en connaissance pour la relation d'hyperonymie *tNF kappa B* / *potent mediator*.

2.3.2 Exploitation de la distribution contextuelle des termes

Une autre utilisation du contexte des termes consiste à réaliser une analyse distributionnelle (Harris, 1990) pour regrouper des termes partageant des contextes (à l'origine syntaxique). Par exemple, les termes *insuffisance rénale* et *détresse respiratoire* sont sémantiquement proches car ils partagent les mêmes contextes *prise en charge d'une insuffisance rénale* et *prise en charge d'une détresse respiratoire*, *apparition d'une insuffisance rénale* et *apparition d'une détresse respiratoire*. Il est ainsi possible d'identifier une relation de proximité sémantique entre des termes (Bourigault *et al.*, 2004), voire des relations de synonymie (Ferret, 2011). Par exemple, dans (Grefenstette,

1994), l'analyse distributionnelle appliquée à un corpus médical permet de repérer plusieurs types de relations : synonymie (*large* / *important* / *great*), méronymie (*patient* / *group*) ainsi que des relations d'hypéronymie *patient* / *woman*.

Cette méthode a été également utilisée par (Resnik, 1993) afin de mettre en évidence les relations sémantiques associées aux termes. Il s'agit alors de remplacer les termes présents dans les contextes par leurs classes sémantiques, issues de WordNet. Par exemple les termes *infirmier* et *docteur* sont remplacés par la classe *profession de santé* de WordNet. Il est également possible d'adopter une approche mixte qui combine l'analyse distributionnelle et les patrons lexico-syntaxiques (Caraballo, 1999). Les termes sont d'abord regroupés en fonction des contextes partagés, et des patrons sont alors appliqués pour identifier des relations d'hypéronymie.

Enfin, nous pouvons également mentionner l'approche proposée par (Cimiano *et al.*, 2000) pour acquérir des relations taxonomiques de manière semi-automatique. L'analyse syntaxique des phrases est utilisée pour construire des contextes formels (chaque nom est caractérisé par un ensemble d'attributs composé par des verbes et pour lesquels le nom apparaît comme un argument). Les noms partageant les mêmes contextes seront utilisés pour former un treillis. L'Analyse Formelle de Concepts (FCA) est alors appliquée pour repérer les appariements entre les concepts (Ganter et Wille, 1999). Cette approche vise à structurer les connaissances sous forme d'une hiérarchie à partir d'un ensemble d'entité. Les entités sont représentées sous forme d'un treillis de Galois. Le treillis est composé d'un ensemble d'individus (les objets formels, par exemple des termes), d'un ensemble de caractéristique (les attributs formels, par exemple des contextes) et d'une relation binaire entre les objets et les attributs. Il est également possible d'extraire des relations autre que des relations d'hypéronymie, notamment des relations inter-hiérarchique à l'aide de l'analyse relationnelle de concepts (Bendaoud *et al.*, 2010).

2.3.3 Discussion et analyse

Les travaux présentés ci-dessus reflètent l'importance de ce champ scientifique et montrent une certaine diversité dans les approches permettant d'acquérir des relations en termes issus de domaine de spécialité. Dans cet article, nous proposons de combiner des approches basées sur des informations externes et internes aux termes. Notre objectif se rapproche des travaux de (Resnik, 1993). Mais ici, nous nous situons dans un domaine de spécialité et nous visons à contextualiser sémantiquement les relations de synonymie acquises par une approche exploitant la structure interne des termes.

3 Matériel

Dans cette section, nous décrivons le matériel que nous avons à notre disposition pour nos expériences. Nous avons sélectionné des ressources terminologiques (UMLS et Gene Ontology) et des corpus issus du domaine biomédical (corpus Genia et BioNLP). Nous envisageons également d'utiliser des ressources générales (notamment WordNet) pour étiqueter sémantiquement nos corpus de travail.

3.1 Ressources terminologiques et lexicales

– **Gene ontology** : Gene Ontology (GO) (The Gene Ontology Consortium, 2000) est une ressource terminologique dont l'objectif est de décrire le rôle des gènes dans les organismes (prokaryotes et eukaryotes) ainsi que leurs produits géniques. Elle propose 54 453 concepts et 94 161 termes. Les termes de GO sont structurés en trois arbres hiérarchiques : processus biologiques, fonctions moléculaires et composants cellulaires. Le vocabulaire de GO est structuré à l'aide de trois types de relations : l'hypéronymie (119 430 relations), méronymie (29 573 relations) et la synonymie (101 254 relations). Actuellement, nous n'utilisons que les relations de synonymie.
– **UMLS** (Lindberg *et al.*, 1993) Unified Medical Language System (UMLS) est une ressource terminologique biomédicale. Développé par la National Library of Medicine (NLM), elle regroupe plus d'une centaine de thésaurus de différentes langues dans un méta-thesaurus. Celui-ci organise 700 000 concepts au sein d'un réseau sémantique composé de 134 types sémantiques et structuré par 54 relations sémantiques hiérarchisées par le lien is-a. Les types sémantiques associés aux termes de l'UMLS seront utilisés pour définir les contraintes sémantiques sur les mots ou les termes pour lesquels nous avons acquis des relations de synonymie.
– **WordNet** : WordNet (Fellbaum, 1998) est une base de données lexicale, développée à l'Université de Princeton en 1985. Son objectif est de structurer le contenu sémantique et lexicale de la langue anglaise. WordNet est organisé en ensembles de synonymes appelés synsets. À chaque synset, correspond un concept. Les relations lexicales présentes dans WordNet ne lient que les mots de la même catégorie morpho-syntaxique. Il existe donc quatre hiérarchies (pour les noms, les verbes, les adjectifs et les adverbes). Dans la version 3.0, WordNet contient 155 287 mots avec 117 798 noms, 11 529 verbes, 21 479 adjectifs et 4 481 adverbes organisés en 117 659 synset. Nous envisageons d'exploiter, dans un deuxième temps, les relations issues de WordNet pour étendre le processus d'acquisition de relations de synonymie, et d'exploiter les synsets dans la définition des contraintes sémantiques.

3.2 Corpus de travail

Notre approche vise à contextualiser les relations de synonymie acquises automatiquement. Nous exploiterons deux corpus de spécialité, constitués d'un ensemble d'articles issus de la base de données Medline [1].

Le corpus GENIA [2] contient, dans sa version 3.0, 2 000 résumés Medline au format XML, recueillis à l'aide des termes MESH : "Human", "Blood Cells", et "Transcription Factors" (Kim *et al.*, 2003). Il est composé de 400 000 mots. Il est annoté avec différents niveaux d'informations linguistiques et sémantiques.

Dans un premier temps, nous utilisons le corpus d'entraînement de la campagne BioNLP2011 [3] (Pyysalo *et al.*, 2011). Celui-ci est un ensemble de 800 articles de Medline issus du corpus GENIA. Il est composé de 176 146 mots. Il est également annoté avec des informations linguistiques et sémantiques et notamment des relations inter-hiérarchiques *Protein/component* et *subunit/complex*.

1. http ://www.ncbi.nlm.nih.gov/Entrez/
2. http ://www.nactem.ac.uk/genia/genia-corpus
3. http ://2011.bionlp-st.org/

4 Prise en compte du contexte des termes dans l'acquisition de relations de synonymie

Notre travail s'appuie sur la méthode implémentée dans SynoTerm pour acquérir des relations de synonymie entre termes complexes (Hamon et Nazarenko, 2001). Cette méthode se base sur l'hypothèse que des relations de synonymie peuvent être propagées à travers le principe de compositionnalité. Trois règles sont proposées pour inférer des relations de synonymie entre termes complexes à partir de relations élémentaires de synonymie entre mots [4] (étape d'inférence sur la figure 1). Ainsi, deux termes complexes sont considérés comme synonymes si au moins un de leur composant dans la même position syntaxique sont synonymes. Par exemple, étant donné la relation de synonymie entre les mots *infection* et *sepsis*, les termes *wound infection* et *wound sepsis* sont identifiés comme synonymes. Les auteurs ont également montré que la qualité des relations inférées dépend de l'origine des relations entre les termes simples : des relations de synonymie issues d'un dictionnaire de langue générale contribuent à augmenter le rappel, tandis que des relations spécialisées acquises sur un corpus du domaine permettent d'améliorer la précision.

La difficulté étant de disposer de relations entre termes simples, spécifiques au domaine, l'approche inverse a été définie pour induire des relations élémentaires à partir des relations fournies par une ressource terminologique (étape d'induction sur la figure 1) (Hamon et Grabar, 2008). L'application de la méthode sur Gene Ontology permet d'acquérir 3 707 relations de synonymie avec une précision de 0,72. Nous avons pu parfois observer des décalages sémantiques entre le type de la relation issue de la ressource terminologique et la relation induite : alors que la relation initiale est une relation de synonymie, la relation induite exprime un autre type de relation. Par exemple, à la figure 1, *cell receptor complex* et *lymphocyte receptor complex* sont des synonymes dans GO tandis que la relation induite est considérée comme une relation d'hypéronymie. On peut supposer que cette modification de type sémantique entre la relation induite et la relation entre les termes complexes est dû à un usage particulier qui peut être capturé à travers le contexte terminologique ou l'usage en corpus. De plus, la synonymie étant une relation contextuelle (Cruse, 1986), il semble important de prendre en compte le contexte lors de l'acquisition des relations élémentaires.

L'objectif de notre travail est ainsi de définir une méthode permettant de contextualiser les relations élémentaires induites. Nous souhaitons pouvoir associer des catégories sémantiques issues du contexte des termes pour contraindre l'utilisation des relations induites.

Pour cela, nous avons identifié deux pistes de travail :

1. Filtrage des relations induites par leur usage dans un corpus. Les relations élémentaires induites à partir d'une ressource terminologique (étape d'induction) sont d'abord exploitées sur un corpus pour inférer des relations entre termes complexes (étape d'inférence). L'objectif est d'identifier les relations élémentaires utiles et leur associer un poids ou une confiance plus importante. Ici nous faisons l'hypothèse que les relations incorrectes ne devraient pas être utilisées dans un corpus. Ce travail permettra également d'identifier des contextes lexicaux puis sémantiques utiles pour contextualiser les relations élémentaires.
2. Exploitation d'informations sémantiques associées aux termes. Il s'agit de contraindre le champ d'application des relations entre les termes simples en exploitant les informations

4. ou des termes moins complexes, c'est-à-dire de longueur plus petite.

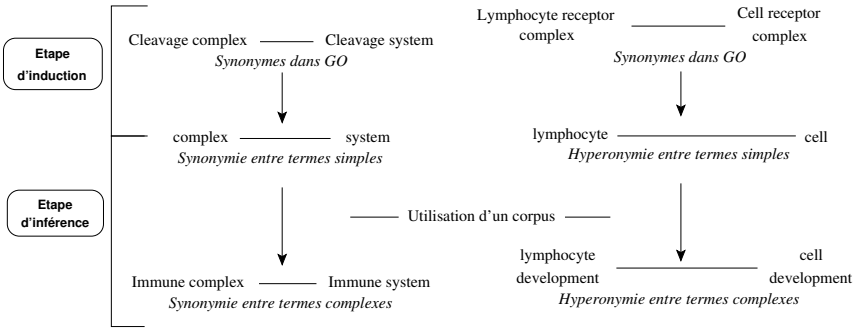

FIGURE 1 – Processus d'acquisition de relations de synonymie (inférence et induction).

sémantiques associées aux termes. Nous allons annoter les textes avec des informations sémantiques (catégories sémantiques ou rôles sémantiques). Pour cela nous allons projeter les catégories sémantiques de l'UMLS et de WordNet sur les corpus de travail ou sur les composants des termes de GO. Nous envisageons d'utiliser des méthodes d'apprentissage par analogie et par la programmation logique inductive pour en déduire les contraintes sémantiques sur les relations élémentaires.

Pour des raisons de facilité de mise en œuvre nous nous sommes pour l'instant concentrée sur la première piste, et plus particulièrement sur le filtrage des relations élémentaires induites.

5 Filtrage des relations élémentaires par l'usage en corpus

Pour filtrer les relations élémentaires en fonction de leur usage en corpus, nous avons travaillé sur le corpus BioNLP. Le corpus a d'abord été segmenté en mots et en phrases. Les mots ont été étiquetés morpho-syntaxiquement et lemmatisés avec Genia Tagger (Tsuruoka *et al.*, 2005). Nous avons ensuite extrait les termes avec YATEA (Aubin et Hamon, 2006). Les différent traitements ont été pris en charge par la plate-forme d'annotation linguistique Ogmios (Hamon et Nazarenko, 2008).

L'acquisition de relations de synonymie entre les termes du corpus a été réalisée à l'aide de SynoTerm. Nous avons utilisé les 3 707 relations élémentaires induites à partir de GO. 277 relations entre termes complexes ont été inférées sur le corpus BioNLP. 104 relations élémentaires ont été utilisées. Les résultats sont en cours d'analyse. La figure 2 présente quelques relations inférées. Nous sommes conscient que la taille du corpus utilisé aura une influence sur le volume de relations inférées. Un filtrage basé uniquement sur le corpus demande une réflexion sur la taille minimale pour appliquer cette approche, ou le recrutement de textes dans lesquels tous les mots ou termes simples en relation apparaissent.

Formation /Development ⟶ *Cell development /Cell formation*
⟶ *Lymphocyte development /Lymphocyte formation*
Contact /Adhesion ⟶ *Cell contact /Cell adhesion*
Embryonic /Fetal ⟶ *Development embryonic /Development fetal*

FIGURE 2 – Exemples de relations inférées

6 Conclusion

L'étude des relations élémentaires induites à partir de Gene Ontology nous conduit à nous intéresser à l'amélioration d'acquisition de relation de synonymie lorsqu'elle est basée sur l'analyse de la structure interne. Les pistes de travail que nous proposons visent à prendre en compte les contextes des termes en corpus de spécialité. Nous proposons d'abord d'utiliser la structure interne des termes pour acquérir des relations, puis de les filtrer ou les trier en contexte (en utilisant des corpus textuels ou les contextes terminologiques que constituent les termes de GO). Les relations élémentaires induites seront ainsi affinées en leur associant des contraintes.

Dans cette perspective, afin de commencer à mettre en œuvre nos pistes de travail, nous avons réalisé une première expérience sur des données textuelles issues du domaine biomédical (corpus BioNLP et Gene Ontology). Les relations élémentaires issues de l'analyse des termes de Gene Ontology ont été exploitées sur le corpus BioNLP pour inférer des relations de synonymie entre termes complexes. Ces premiers résultats vont nous permettre de définir précisément une méthode de filtrage basée sur le contexte et l'usage des termes.

Références

AUBIN, S. et HAMON, T. (2006). Improving term extraction with terminological resources. *In* SALAKOSKI, T., GINTER, F., PYYSALO, S. et PAHIKKALA, T., éditeurs : *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, numéro 4139 de LNAI, pages 380–387. Springer.

AUGER, A. et BARRIÈRE, C. (2008). Pattern-based approaches to semantic relation extraction : A state-of-the-art. *Special Issue on Pattern-based Approaches to Semantic Relation Extraction, Terminology*, 14(1):1–19.

BENDAOUD, R., TOUSSAINT, Y. et NAPOLI, A. (2010). L'analyse Formelle de Concepts au service de la construction et l'enrichissement d'une ontologie. *Revue des Nouvelles Technologies de l'Information*, E-18:133–164.

BODENREIDER, O., BURGUN, A. et RINDFLESCH, T. C. (2001). Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. *In* URI INIST CNRS, éditeur : *Actes de la conférence TIA-2001*, pages 11–21, Nancy, France.

BOUILLON, P., FABRE, C., SÉBILLOT, P. et JACQMIN, L. (2000). Apprentissage de ressources lexicales pour l'extension de requêtes. *Traitement Automatique des Langues*, 41(2):367–393.

BOURIGAULT, D. (1994). *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*. Thèse en mathématiques, informatique appliquée aux sciences de l'homme, École des Hautes Études en Sciences Sociales, Paris, France.

Bourigault, D., Aussenac-Gilles, N. et Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA) – Techniques Informatiques et structuration de terminologiques*, 18(1):87–110.

Bourigault, D. et Jacquemin, C. (2000). Constitution de ressources terminologiques. *In Ingénierie des langues*, chapitre 9, pages 215–233. Hermes Science. Sous la direction de Jean-Marie Pierrel.

Cabré, M. T. (1999). *Terminology. Theory, methods and applications*, volume 1 de *Terminology and Lexicography, Research and practice*. John Benjamins, Amsterdam/Philadelphia.

Caraballo, S. (1999). Automatic acquisition of a hypernym-labeled noun hierarchy from text. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics ACL99*, pages 120–126.

Cimiano, P., Staab, S. et Tane, J. (2000). Automatic acquisition oftaxonomies from text : Fca meets nlp. *In Proceedings of the ECMLPKDD Workshop on Adaptive Text Extraction and Mining CavtatDubrovnik Croatia*, pages 10–17.

Claveau, V. et L'Homme, M.-C. (2004). Discoveringe specific semantic relationships between nouns and verbs in a specialized french corpus. *In Proceedings of the 3rd International Workshop on Computational Terminology, CompuTerm'04*, pages 39–46, Genève, Suisse.

Claveau, V. et L'Homme, M.-C. (2005). Apprentissage par analogie pour la structuration de terminologie-utilisation comparée de ressources endogènes et exogènes. *In Conférence TIA-2005, Rouen, 4 et 5 avril 2005*, Montréal, Canada.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge.

Daille, B. (2003). Conceptual structuring through term variations. *In* Bond, F., Kohonen, A., Carthy, D. M. et Villaciencio, A., éditeurs : *Proceedings of the ACL'2003 Workshop on Multiword Expressions : Analysis, Acquisition, and Treatment*, pages 9–16.

Fellbaum, C., éditeur (1998). *WordNet : An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ferret, O. (2011). Utiliser l'amorçage pour améliorer une mesure de similarité sémantique. *In Actes de TALN 2011*, pages 1–6, Montpellier.

Ganter, B. et Wille, R. (1999). *Formal concept analysis - mathematical foundations*. Springer.

Grabar, N. et Hamon, T. (2006). Terminology structuring through the derivational morphology. *In* Salakoski, T., Ginter, F., Pyysalo, S. et Pahikkala, T., éditeurs : *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, numéro 4139 de LNAI, pages 652–663. Springer.

Grabar, N. et Zweigenbaum, P. (2002). Lexically-based terminology structuring : some inherent limits. *In Proceedings of Computerm'2002 (Second Workshop on Computational Terminology)*, Taiwan.

Grefenstette, G. (1994). *Exploration in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston, USA.

Grouin, C., Abacha, A. B., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., Ligozat, A.-L., Minard, A.-L., Rosset, S. et Zweigenbaum, P. (2010). Caramba : Concept, assertion, and relation annotation using machine-learning based approaches. *In Proceedings of the workshop I2B2 2010*.

HAMON, T. (2005). Indexer les documents spécialisés : les ressources terminologiques contrôlées sont-elles suffisantes ? *In* 6ème *rencontres Terminologie et Intelligence Artificielle*, pages 71–82, Rouen, France.

HAMON, T. et GRABAR, N. (2008). Acquisition of elementary synonym relations from biological structured terminology. *In* GELBUKH, A., éditeur : *Computational Linguistics and Intelligent Text Processing - 9th International Conference, CICLing - Proceedings*, numéro 4919 de LNCS, pages 40–51, Haifa, Israel. Springer-Verlag Berlin Heidelberg.

HAMON, T., GRAÑA, M., RAGGIO, V., GRABAR, N. et NAYA, H. (2010). Identification of relations between risk factors and their pathologies or health conditions by mining scientific literature. *In MEDINFO 2010*, pages 964–968. Stud Health Technol Inform. PMID : 20841827.

HAMON, T. et NAZARENKO, A. (2001). Detection of synonymy links between terms : experiment and results. *In Recent Advances in Computational Terminology*, pages 185–208. John Benjamins.

HAMON, T. et NAZARENKO, A. (2008). Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience. *Traitement Automatique des Langues*, 49(2):127–154.

HARRIS, Z. (1990). La genèse de l'analyse des transformations et de la métalangue. *Langages*, 99:9–20. A. Daladier (resp.).

HEARST, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *In In Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545.

IBEKWE-SANJUAN, F. (2005). Inclusion lexicale et proximité sémantique entre termes. *In Actes de la conférence TIA 2005*, pages 45–57, Rouen.

JACQUEMIN, C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes.

JACQUEMIN, C. (1999). Syntagmatic and paradigmatic representations of term variation. *In 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 341–348, University of Maryland.

KIM, J.-D., OHTA, T., TATEISI, Y. et TSUJII, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *In ISMB (Supplement of Bioinformatics)*, pages 180–182.

LINDBERG, D., HUMPHREYS, B. et MCCRAY, A. (1993). The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291.

MARTIENNE, E. et MORIN, E. (1999). Using s symbolic machine learning tool to refine lexico-syntactic patterns. Rapport de Recherche 183, Institut de Recherche en Informatique de Nantes (IRIN).

MCCRAE, J. et COLLIER, N. (2008). Synonym set extraction from the biomedical literature by lexical pattern discovery. *BMC Bioinformatics*, 9:159+.

MCINTOSH, T. et CURRAN, J. R. (2009). Challenges for automatically extracting molecular interactions from full-text articles. *BMC Bioinformatics*, 10:311+.

MEYER, I. (2001). Extracting Knowledge-rich Contexts for Terminography. *In* BOURIGAULT, D., JACQUEMIN, C. et L'HOMME, M., éditeurs : *Recent Advances in Computational Terminology*, pages 279–302. John Benjamins, Amsterdam/Philadelphia.

MORIN, E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse de doctorat, Université de Nantes, FRANCE, Université de Nantes, Nantes, FRANCE.

NAMER, F. et BAUD, R. (2007). Defining and relating biomedical terms : towards a cross-language morphosemantics-based system. *Int J Med Inform*, 76(2-3):226–233.

PYYSALO, S., OHTA, T. et TSUJII, J. (2011). Overview of the entity relations (rel) supporting task of bionlp shared task 2011. *In Proceedings of BioNLP Shared Task 2011 Workshop*, pages 83–88, Portland, Oregon, USA. Association for Computational Linguistics.

RESNIK, P. (1993). *Selection and Information : A Class-Based Approach to Lexical Relationships*. Thèse de doctorat, University of Pennsylvania.

RØST, T. B., AKBAR, S., Øystein NYTRØ et BASGALUPP, M. (2010). Medical relation extraction with semantic grammars. *In Proceedings of the workshop I2B2 2010*.

SAGER, J. C. (1990). *A Practical Course in Terminology Processing*. John Benjamins, Amsterdam.

SCHUMANN, A.-K. (2011). A case study of knowledge-rich context extraction in russian. *In* KAGEURA, K. et ZWEIGENBAUM, P., éditeurs : *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence (TIA 2011)*, pages 143–146, INaLCO.

SNOW, R., JURAFSKY, D. et NG, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. *In* SAUL, L. K., WEISS, Y. et BOTTOU, L., éditeurs : *Advances in Neural Information Processing Systems 17*, pages 1297–1304. MIT Press, Cambridge, MA.

SPASIC, I., ANANIADOU, S., MCNAUGHT, J. et KUMAR, A. (2005). Text mining and ontologies in biomedicine : making sense of raw text. *Briefings in Bioinformatics*, 6(3):239–251.

THE GENE ONTOLOGY CONSORTIUM (2000). Gene ontology : tool for the unification of biology. *Nature genetics*, 25:25–29.

TSURUOKA, Y., TATEISHI, Y., KIM, J.-D., OHTA, T., MCNAUGHT, J., ANANIADOU, S. et TSUJII, J. (2005). Developing a robust part-of-speech tagger for biomedical text. *In Proceedings of Advances in Informatics - 10th Panhellenic Conference on Informatics*, LNCS 3746, pages 382–392.

VERSPOOR, C. M., JOSLYN, C. et PAPCUN, G. J. (2003). The gene ontology as a source of lexical semantic knowledge for a biological natural language processing application. *In SIGIR workshop on Text Analysis and Search for Bioinformatics*, pages 51–56.

WEISSENBACHER, D. (2004). La relation de sysnonymie en génomique. *In Actes de la conférence RECITAL'2004*, Fès, Maroc.

YANG, H. et de ROECK, A. (2010). Extraction of medical information using crfs, context patterns, and dependency parse trees. *In Proceedings of the workshop I2B2 2010*.

ZWEIGENBAUM, P. et GRABAR, N. (2000). Liens morphologiques et structuration de terminologie. *In Actes IC'2000*, pages 325–334, Toulouse, France.