# Extraction de relations en domaine de spécialité

Anne-Lyse Minard

École Doctorale d'Informatique de Paris-Sud (EDIPS)
Laboratoire d'Informatique, de Mécanique et de Sciences de l'Ingénieur (LIMSI)

Anne-Lyse Minard

# Extraction de relations en domaine de spécialité

**THÈSE DE DOCTORAT EN INFORMATIQUE**

Soutenue le *7 décembre 2012*

Jury constitué de

| | | |
|---|---|---|
| Directeur de thèse | Brigitte Grau<br>ENSIIE – LIMSI-CNRS | (Professeur) |
| Co-Encadrant | Anne-Laure Ligozat<br>ENSIIE – LIMSI-CNRS | (Maître de Conférence) |
| Rapporteur | Emmanuel Morin<br>Université de Nantes – LINA | (Professeur) |
| | Horacio Saggion<br>Université Pompeu Fabra, Barcelone | (Professeur) |
| Examinateur | Juliette Dibie-Barthélemy<br>INRA, AgroParisTech | (Maître de Conférence) |
| | Chantal Reynaud<br>Université Paris-Sud – LRI | (Professeur) |
| | Yannick Toussaint<br>INRIA, LORIA | (Chargé de Recherche) |
| Invité | Stephen Randall Thomas<br>CNRS, IGR | (Directeur de Recherche) |

# Remerciements

Je souhaite remercier dans cette partie les personnes grâce à qui cette thèse a pu avoir lieu, ainsi que celles qui ont permis qu'elle se déroule de façon agréable.

Je remercie tout d'abord Anne-Laure Ligozat et Brigitte Grau grâce à qui (et à cause de qui !!!) j'ai réalisé cette thèse. Pendant ces trois années, elles m'ont toujours encouragée, soutenue, conseillée et encadrée avec beaucoup de pédagogie, de patience et de gentillesse, et toujours dans la bonne humeur ! Un grand MERCI !

Je remercie les membres de mon jury (Juliette Dibie-Barthélemy, Chantal Reynaud, Emmanuel Morin, Horacio Saggion et Yannick Toussaint) pour la relecture de mon manuscrit et pour les pistes de recherche qu'ils ont ouvertes par leurs questions.

J'ai réalisé une partie de mes travaux en collaboration avec Randy Thomas, je le remercie pour sa gentillesse et pour le temps qu'il m'a consacré. Je remercie également Pierre Zweigenbaum qui m'a conseillé et guidé dans le domaine de l'informatique médicale.

Ich danke Ulf Leser und Philippe Thomas von der Humboldt Universität zu Berlin für den netten Empfang in die WBI-gruppe, für ihre Unterstützung und für ihre Geduld.

Je remercie Anne Vilnat pour tous ses bons conseils et pour son aide dans la préparation de mes enseignements à l'IUT d'Orsay.

J'ai pu réaliser cette thèse au LIMSI dans de bonnes conditions grâce à de nombreuses personnes. Je remercie en particulier :

- Les membres du groupe ILES (j'y inclus bien évidemment les Gestes et les TLP du rez-de-chaussée du 508 !) pour leur accueil, leur bonne humeur et tous les échanges que nous avons pu avoir à la cafet, au CESFO, dans le couloir, etc. : Delphine, Sophie, Gabriel, Michael, Annelies, Sarah, Anne, Pierre, Xavier, Aurélien, etc. ;
- Tifanie, ma collègue du bureau des doctorants, pour son dynamisme et sa gentillesse ;
- Les doctorants (ou assimilés) réguliers des activités du BDD (et du Loup-garou !) : Adrien, Florimond, Lionel, Richard, Francky, Yoren, Pierre, Sami, Jean, Max, Lyés, etc. ;
- Les doctorants des groupes ILES et TLP pour les discussions enrichissantes que nous avons eues et tous les bons moments passés ensemble (autour d'un café, en conf ou en école d'été) : Houda, Clément, Asma, Alex, Béa, Nico, Driss, Mathieu, Nadi, Souhir, Camille, etc. ;
- Asma pour toutes nos chouettes discussions ;

– Ceux qui ont partagés leur bureau avec moi et qui ont supportés mon mauvais anglais, mes pétages de cable, mes coups de stress et d'euphorie : Bruno, Alex, Driss, Cyril et Béa.

Un grand merci aux pauseurs du bureau 112 (Nico, Max et Driss) pour tous les très bons moments passés ensemble, pour leur soutien pendant les mois de rédaction et pour être venu régulièrement faire diminuer le stock de gâteaux ! Et un très grand merci à Béatrice pour son rire, ses paroles réconfortantes et sa grande gentillesse.

Le choix des différentes étapes de mon cursus universitaire jusqu'en thèse fut entre autre guidé par plusieurs enseignants passionnés par leur travail. Je remercie en particulier Antonio Balvet, Georgette Dal, Thomas Lebarbé et Olivier Kraif pour tout ce qu'ils m'ont appris en linguistique et en TAL.

Même si son intervention dans mon cursus scolaire remonte à quelques années maintenant, je souhaite remercier Sylvie Lamarre qui durant 5 ans m'a appris à lire, à écrire, à aimer autant les maths que le français, et à avoir soif de connaissances (« Merci maîtresse ! »).

Je remercie les Bons-Amis pour les week-ends et les vacances passés en leur compagnie dans la joie et la bonne humeur : Anne, Lucie, Sarah, Jean-Daniel, Pierre, Timothée et Damaris.

Mes derniers remerciements vont à ma famille. Une grande reconnaissance à mes parents qui m'ont donnés la chance de faire des études, et qui m'ont soutenus et encouragés dans mes choix d'orientation. Un grand merci à mes sœurs, mes frères, mes parents et mes grands-parents pour leur amour et leur soutien.

**Discographie**

Certains jours, je trouvais la motivation et la concentration dans la musique. Ma thèse a été rythmée entre autre par *Imany*, *les Ogres de Barback*, *Jehro*, *Ben Harper*, *Pink Martini*, *Raul Paz*, *Chupa Chuva*, *Dos soles*, *Katie Melua*, la sonate pour piano D958 de *Schubert* et la sonate pour violoncelle et piano opus 38 de *Brahms*.

# Résumé

La quantité d'information disponible dans le domaine biomédical ne cesse d'augmenter. Pour que cette information soit facilement utilisable par les experts d'un domaine, il est nécessaire de l'extraire et de la structurer. Pour avoir des données structurées, il convient de détecter les relations existantes entre les entités dans les textes. Nos recherches se sont focalisées sur la question de l'extraction de relations complexes représentant des résultats expérimentaux, et sur la détection et la catégorisation de relations binaires entre des entités biomédicales.

Nous nous sommes intéressée aux résultats expérimentaux présentés dans les articles scientifiques. Nous appelons résultat expérimental, un résultat quantitatif obtenu suite à une expérience et mis en relation avec les informations permettant de décrire cette expérience. Ces résultats sont importants pour les experts en biologie, par exemple pour faire de la modélisation. Dans le domaine de la physiologie rénale, une base de données a été créée pour centraliser ces résultats d'expérimentation, mais l'alimentation de la base est manuelle et de ce fait longue. Nous proposons une solution pour extraire automatiquement des articles scientifiques les connaissances pertinentes pour la base de données, c'est-à-dire des résultats expérimentaux que nous représentons par une relation n-aire. La méthode procède en deux étapes : extraction automatique des documents et proposition de celles-ci pour validation ou modification par l'expert via une interface.

Nous avons également proposé une méthode à base d'apprentissage automatique pour l'extraction et la classification de relations binaires en domaine de spécialité. Nous nous sommes intéressée aux caractéristiques et variétés d'expressions des relations, et à la prise en compte de ces caractéristiques dans un système à base d'apprentissage. Nous avons étudié la prise en compte de la structure syntaxique de la phrase et la simplification de phrases dirigée pour la tâche d'extraction de relations. Nous avons en particulier développé une méthode de simplification à base d'apprentissage automatique, qui utilise en cascade plusieurs classifieurs.

**Mots clés :** Extraction de relations, relation binaire, relation n-aire, domaine biomédical, SVM, information syntaxique, simplification de phrases.

# Abstract

The amount of available scientific literature is constantly growing. If the experts of a domain want to easily access this information, it must be extracted and structured. To obtain structured data, both entities and relations of the texts must be detected. Our research is about the problem of complex relation extraction which represent experimental results, and detection and classification of binary relations between biomedical entities.

We are interested in experimental results presented in scientific papers. An experimental result is a quantitative result obtained by an experimentation and linked with information that describes this experimentation. These results are important for biology experts, for example for doing modelization. In the domain of renal physiology, a database was created to centralize these experimental results, but the base is manually populated, therefore the population takes a long time. We propose a solution to automatically extract relevant knowledge for the database from the scientific papers, that is experimental results which are represented by a n-ary relation. The method proceeds in two steps: automatic extraction from documents and proposal of information extracted for approval or modification by the experts via an interface.

We also proposed a method based on machine learning for extraction and classification of binary relations in specialized domains. We focused on the variations of the expression of relations, and how to represent them in a machine learning system. We studied the way to take into account syntactic structure of the sentence and the sentence simplification guided by the task of relation extraction. In particular, we developed a simplification method based on machine learning, which uses a series of classifiers.

**Keywords:** Relation extraction, binary relation, n-ary relation, biomedical domain, SVM, syntactic information, sentence simplification.

# Table des matières

# Liste des tableaux

# Table des figures

# Introduction

Dans les domaines de spécialité, les connaissances sont principalement sous forme textuelle. Ces connaissances peuvent être par exemple les résultats de recherches dans le cas des articles scientifiques, ou les rapports des médecins dans le cas de comptes rendus cliniques. Toutes ces connaissances sont utilisées par les chercheurs pour compléter leurs résultats, les comparer, les confronter et découvrir de nouvelles informations. Le nombre des informations disponibles au format électronique ne cesse d'augmenter. Par exemple pour le domaine médical, la banque de données bibliographiques MEDLINE contenait 15 millions d'articles référencés en 2007 et en contient actuellement plus de 19,5 millions.

Une partie de ces informations est enregistrée dans des bases de données spécifiques à un domaine et à une tâche. Ces connaissances qui sont dans les bases sont structurées, accessibles et plus facilement exploitables que sous forme textuelle. Cependant les bases de données sont souvent peuplées manuellement, et elles ne sont donc pas complètes et pas nécessairement mises à jour.

Le peuplement automatique des bases de données nécessitent d'identifier dans les textes les informations pertinentes, de le typer et de les structurer. Les informations d'intérêts peuvent être de différents types ; on peut vouloir extraire des entités nommées (par exemple des noms de lieux, des noms de personnes, etc.), des entités d'un domaine (par exemple les noms de gènes, les noms de médicaments, etc.), des informations temporelles (par exemple des dates, des enchaînements de faits, etc.), etc. Pour pouvoir extraire des informations complexes dans les textes, il est nécessaire de savoir reconnaître les entités d'intérêts, à partir desquelles il est ensuite possible d'extraire des relations entre des entités, des relations entre les descripteurs d'un événement (par exemple des résultats expérimentaux), des enchaînements d'événements, etc. Les relations entre les entités permettent alors de structurer les informations extraites via les entités. Par exemple dans le domaine médical, on peut annoter les médicaments et les maladies, puis extraire les relations existantes entre ces deux types d'entités pour savoir si le médicament soigne la maladie, s'il la cause, etc.

La langue naturelle étant très riche et très complexe, un même type d'information peut ainsi être exprimé de plusieurs manières et une information peut même être implicite dans le texte. Par exemple, deux verbes sémantiquement proches (ou synonymes dans un contexte particulier) peuvent être employés pour exprimer la même idée, ou encore deux noms peuvent référer au même médicament : le nom courant du médicament et le nom de son générique. Des phrases peuvent posséder des structures différentes mais traduire une information semblable.

Dans la phrase (1), *l'aspirine* correspond à une entité de type `médicament` et *la douleur*,

*la fièvre* et *l'inflammation* sont des entités du type `problème médical`. Les entités de type `problème médical` sont liées à *l'aspirine* par des relations qui peuvent être du type `un médicament permet de soigner un problème médical`. Les relations sont exprimées respectivement par deux verbes différents : *soulager*, *faire baisser* et le groupe nominal *le traitement de*.

(1) [TREAT L'aspirine] est utilisée pour soulager [PB la douleur], faire baisser [PB la fièvre] et le traitement de [PB l'inflammation].

La reconnaissance de ces informations relève de l'extraction d'information, qui est une discipline du Traitement Automatique des Langues (TAL) qui s'intéresse à la façon dont des informations pertinentes dans un certain contexte sont exprimées dans des documents écrits et aux méthodes à utiliser pour les extraire. Les systèmes d'extraction d'information permettent ainsi d'extraire des connaissances des textes sur un domaine particulier et pour une application donnée. Notre travail de recherche s'inscrit dans ce contexte très large de l'extraction d'information dans des textes de spécialités pour disposer d'informations structurées. L'extraction des entités nommées est une tâche qui a été beaucoup étudiée et pour laquelle les méthodes proposées fournissent de bon résultats. Ainsi l'enjeu actuel porte plutôt sur la structuration de ces entités, et donc sur l'extraction de relations. Le problème sera de savoir reconnaître celles-ci dans des textes sous leurs différentes formes. Identifier les liens existants entre deux entités ou plus dans un texte permet de compléter et structurer l'information extraite par les entités d'un domaine. L'exemple précédent illustre le problème de la variation d'expression dans la langue, et ici en particulier ce qui nous a intéressé dans ce travail, la variabilité d'expression des relations.

## Problématiques et contributions

L'objectif de notre travail est d'étudier l'expression de relations en domaine de spécialité et de proposer des méthodes adaptées pour leur extraction. Les relations sont diverses et peuvent être divisées en deux catégories selon le nombre d'arguments qui sont liés : les relations n-aires et les relations binaires.

Nous avons étudié la problématique de l'extraction de résultats expérimentaux dans des articles scientifiques dans le but de peupler une base de données existante. Nous avons dans un premier temps étudié précisément ce qu'était un résultat expérimental et la façon dont il s'exprimait dans le texte. Nous avons proposé une formalisation des résultats expérimentaux par des relations n-aires, dans lesquelles le résultat numérique est l'élément pivot. Ces relations sont implicites dans le texte et étendues sur plusieurs phrases. Nous proposons une méthode d'extraction permettant d'extraire ces relations n-aires qui prend en compte les particularités d'expression de ces informations, entre autres le fait que tout l'article est nécessaire, que les résultats peuvent être dans des tableaux et qu'il n'y a pas de mentions des relations dans le texte.

La deuxième problématique auquel nous nous sommes intéressée est la détection et la classification de relations binaires entre des entités dans des textes biomédicaux. Nous nous sommes intéressée à la définition des traits à utiliser pour identifier les relations binaires :

quels types de traits doivent être utilisés et dans quel contexte ? Nous avons effectué une étude des attributs à utiliser pour caractériser les relations. Nous avons développé un système d'extraction de relations par apprentissage qui utilise ces attributs pour classer les relations.

La question de la variabilité d'expression des relations nous a intéressée et poussée à étudier la prise en compte et l'apport des informations sur la structure syntaxique des phrases. Pour extraire les relations, il est possible d'utiliser des informations sur la structure syntaxique des phrases plus ou moins précises grâce à l'arbre de constituants. Suivant les attributs définis, le contexte pris en compte (ou dans lequel sont extraits les attributs) varie, ce qui nous a amenée à nous intéresser à la définition du contexte et les façons de le faire varier. Nous avons étudié la simplification de phrases dans le but de réduire la variété d'expression des relations, de faire varier le contexte et ainsi améliorer les performances du système d'extraction de relations. Nous proposons une comparaison de méthodes de simplification qui portent sur la phrase ou sur l'arbre de constituants, et qui sont fondées sur des règles ou des classifieurs.

Nous nous sommes également intéressée à la mise à disposition des connaissances et au problème du peuplement des bases de données grâce à des systèmes d'extraction d'information, et en particulier à la manière d'aider les experts au peuplement grâce aux informations extraites par un système d'extraction d'information.

## Organisation du mémoire

Dans la première partie de notre mémoire, nous présenterons le contexte de notre travail, c'est-à-dire le domaine de l'extraction d'information en particulier dans des textes de spécialités. L'extraction de relations est le sous-domaine de l'extraction d'information que nous avons le plus étudié. Nous définirons les différents types de relations, selon leur arité et leur forme d'expression phrastique, puis nous proposons un état de l'art des méthodes d'extraction de relations n-aires et binaires. Nous présenterons ensuite le domaine de la simplification de phrases et les méthodes proposées dans l'état de l'art.

La deuxième partie sera consacrée à l'extraction de résultats expérimentaux dans des articles scientifiques pour aider au peuplement d'une base de données dans le domaine biomédical. Nous présenterons en particulier la formalisation des résultats expérimentaux par des relations n-aires et une méthode d'extraction de ces relations. Le système d'extraction a été développé pour aider les experts à peupler une base de données. Pour ce faire une interface d'aide à l'annotation d'articles scientifiques a été proposé ; nous la présenterons ainsi qu'une évaluation utilisateur.

La dernière partie s'intéressera à l'extraction de relations binaires en domaine de spécialité. La méthode d'extraction proposée est fondée sur de l'apprentissage automatique utilisant des SVM. Nous la présenterons ainsi que différentes expérimentations qui ont été faites pour évaluer les informations utilisées par le système pour détecter et classer les relations. Nous présenterons une comparaison des résultats obtenus par notre système sur plusieurs corpus. Pour améliorer l'extraction des relations, nous avons proposé plusieurs méthodes de simplification de phrases afin de ne conserver que les informations utiles pour

détecter les relations dans les phrases. Nous terminerons cette partie par une présentation de ces méthodes et des évaluations effectuées.

# Première partie

# Extraction de relations en domaine de spécialité

Dans cette partie, nous présentons les travaux existants dans le domaine de l'extraction d'information et dans une de ses sous-tâches : l'extraction de relations, et nous positionnons notre travail par rapport à l'existant. Dans un premier temps, nous faisons un point sur les recherches en extraction d'information en domaine général puis en domaine de spécialité, en particulier dans le domaine biomédical. Ensuite, nous présentons les différents types de relations entre entités qui peuvent être extraites ainsi que les méthodes utilisées. Nous nous arrêtons en particulier sur la description des traits utilisés pour apprendre à extraire les relations et sur la définition du contexte dans lequel sont extraits ces traits. Après avoir défini la notion de simplification de phrases, nous présentons des travaux pour la simplification de phrases dirigée par une tâche et en particulier dirigée par l'extraction de relations.

# Chapitre 1

# Extraction de relations

**Sommaire**



Les travaux présentés dans cette thèse s'inscrivent dans le domaine de l'extraction d'information, activité qui consiste à extraire l'information dans des documents écrits dans le but de structurer l'information disponible, par exemple en annotant le texte ou en remplissant une base de données.

**Définition 1** *Extraction d'information*

*"Information extraction may be seen as the activity of populating a structured information source (or database) from an unstructured, or free text, information source." (Gaizauskas et Wilks [1998])*

*"L'extraction d'information désigne l'activité qui consiste à remplir automatiquement une banque de données à partir de textes écrits en langue naturelle." (Poibeau [2003])*

L'extraction d'information couvre beaucoup de tâches dont les plus importantes sont la reconnaissance d'entités nommées, la résolution de coréférences, l'extraction de relations

et l'extraction d'événements. À l'heure actuelle la tâche qui a suscité le plus de travaux et qui a été la plus étudiée est l'extraction d'entités nommées.

Avant de présenter le domaine de l'extraction de relation, nous allons présenter un rapide historique du domaine de l'extraction d'information.

## 1.1 Historique de l'extraction d'information

Les premiers grands travaux en extraction d'information datent des conférences MUC (*Message Understanding Conference*). Ces conférences, initiées et financées par la DARPA (*Defense Advanced Research Projects Agency*), avaient pour objet d'évaluer et d'encourager les recherches effectuées dans le domaine de l'extraction d'information : par exemple pour MUC-3 en 1991, la tâche d'extraction avait lieu dans des dépêches journalistiques sur les activités terroristes en Amérique latine. La première conférence a eu lieu en 1987. Les deux dernières conférences MUC-6 et MUC-7, en 1995 et 1997, se sont focalisées sur l'extraction d'information à partir de schémas ou templates. Les différentes tâches d'extraction d'information définies au cours de la sixième conférence MUC étaient : la reconnaissance d'entités nommées (c'est-à-dire de noms de personnes, d'organisations, de lieux, etc.), l'extraction de coréférences, le remplissage de formulaires prédéfinis et l'extraction de scénarios. À MUC-7 en 1998 apparaît la tâche d'extraction de relations entre des entités. Les trois relations étudiées étaient : `employee_of`, `product_of` et `location_of`. Cette conférence marque le début de la recherche en extraction de relations sémantiques (non terminologiques).

Au cours des campagnes d'évaluation MUC, ont été définies deux métriques pour l'extraction d'information : le rappel et la précision. Le rappel est le nombre d'entités correctement extraites par un système divisé par le nombre d'entités à extraire. La précision est le nombre d'entités correctement extraites par un système divisé par le nombre d'entités correctes et incorrectes extraites par un système.

À la suite des conférences MUC, les conférences ACE (*Automatic Content Extraction*) ont continué à encourager les recherches en extraction d'information, en proposant des tâches de détection et caractérisation d'entités, de relations et plus tard d'événements.

En 2009 a eu lieu la première tâche KBP (*Knowledge Base Population*) dans le cadre de la campagne d'évaluation TAC (*Text Analysis Conference*). Le but de cette tâche était d'extraire de l'information à propos d'entités d'une base de connaissance. Il était proposé aux participants deux sous-tâches : relier les entités du texte avec celles de la base, et remplir les champs de la base à propos de ces entités.

## 1.2 Extraction d'information en domaine de spécialité

Nous nous sommes intéressée à l'extraction d'information en domaine de spécialité, et plus particulièrement dans le domaine biomédical. Nous avons travaillé avec des corpus de textes de spécialité, comme des articles scientifiques provenant de revues, ou des comptes rendus cliniques. Dans le domaine biomédical, la base de données bibliographiques MEDLINE (*Medical Literature Analysis and Retrieval System Online*) regroupe une grande

quantité de résumés d'articles scientifiques et d'articles entiers. En avril 2012, 19,5 millions d'articles étaient référencés sur MEDLINE [1]. L'interface PubMed [2] permet de consulter cette base de données. Beaucoup d'experts utilisent les informations et résultats contenus dans ces articles publiés, mais le nombre d'articles ne cessant d'augmenter, il est nécessaire de disposer de systèmes d'extraction d'information pour accéder plus facilement à leur contenu. Une grande partie des travaux en extraction d'information dans le domaine biomédical utilise les articles de cette base de données.

Dans des textes de spécialité, les auteurs emploient une terminologie particulière. Elle contient des termes techniques spécifiques à un domaine et des mots de la langue courante, qui peuvent avoir un sens spécialisé. Extraire de l'information dans un domaine de spécialité pose des problèmes légèrement différents de l'extraction d'information en domaine général. Il est en particulier plus aisé d'annoter les entités nommées en domaine de spécialité car la terminologie est proche d'une terminologie finie contrairement à la langue générale, mais il est plus difficile de repérer les termes qui se rapportent au même concept ou à la même entité (à cause des variations typographiques pour les noms de protéines par exemple).

Nous avons vu que dans le domaine général les campagnes d'évaluation (en particulier MUC et ACE) ont été motrices dans l'avancée des recherches sur l'extraction d'information. Il en est de même dans le domaine biomédical. Elles sont à l'origine de la majorité des corpus annotés disponibles et elles fournissent également un cadre d'évaluation permettant de comparer différentes méthodes. Nous présentons ici quelques campagnes d'évaluation (aussi appelées challenges), et dans la suite de ce chapitre nous présenterons des systèmes développés au cours de ces campagnes.

Le challenge BioCreative [3] consiste à évaluer les systèmes de fouille de texte et d'extraction d'information appliqués au domaine biologique. La première campagne d'évaluation BioCreative a eu lieu en 2004 et la dernière en date a eu lieu en 2012. Les tâches d'intérêt de cette campagne sont entre autres la détection des noms de gènes et leur liaison avec les entrées d'une base de données, l'extraction d'interactions entre protéines, la recherche de documents portant sur un gène, etc. Des systèmes développés pour la tâche d'extraction d'interactions entre protéines seront décrits dans la partie 1.6.

Le challenge LLL05 [4] (Nédellec [2005]) portait sur l'extraction des interactions entre gènes et protéines dans des résumés de MEDLINE. La campagne d'évaluation BioNLP'09 [5] concernait la reconnaissance d'événements bio-moléculaire dans la littérature.

Dans le domaine médical, le challenge de Traitement Automatique de la Langue pour les données cliniques, i2b2 (*informatics for integrating Biology & the Bedside*) a porté sur l'extraction de médication en 2009 (Uzuner *et al.* [2010b]), sur l'extraction d'entités (problèmes médicaux, tests et traitements), la classification d'assertions et l'extraction de relations en 2010 (Uzuner *et al.* [2011]), et sur la résolution de co-référence et la détection de sentiments en 2011 [6].

[1] http://www.nlm.nih.gov/bsd/revup/revup_pub.html#med_update
[2] http://www.ncbi.nlm.nih.gov/pubmed
[3] http://biocreative.sourceforge.net/
[4] http://genome.jouy.inra.fr/texte/LLLchallenge/
[5] http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/index.shtml
[6] https://www.i2b2.org/NLP/Coreference/

## 1.3 Relations : définitions

Nous nous sommes intéressée aux relations entre entités et à leur extraction. Les relatiosn entre deux entités ou plus ont pour rôle de structurer l'information extraite.

**Définition 2** *"In the context of information extraction, a* relation *represents some relationship between two entities; a* relation mention *is an expression of this relationship, and involves two entity mentions." (Grishman [2010])*

En nous appuyant sur cette définition, nous appelons *relation* une relation sémantique entre deux types d'entités (on parle aussi d'*instance de relation*) et *expression de relation* (ou *mention de la relation*) la réalisation phrastique de la relation dans une phrase ou un document. Nous utiliserons également le terme *classe de relation* (ou *catégorie de relation*) pour parler d'un type de relation. Par exemple, `auteur_de(Personne,Œuvre)` est une classe de relation, `auteur_de(Stendhal, le Rouge et le Noir)` est une instance de la relation et l'exemple (2) est l'expression de la relation dans une phrase provenant de Wikipedia.

(2) **Le Rouge et le Noir** est un roman écrit par **Stendhal**, publié à Paris chez Levasseur en 1830.

Soient $E_1, E_2, ..., E_n$ des ensembles d'entités de type $t_1, t_2, ..., t_n$. Les relations d'une classe de relation R relient $n$ entités $e_1, e_2, ..., e_n$ des ensembles d'entités $E_1, E_2, ..., E_n$. On notera la relation : $r(e_1, e_2, ..., e_n)$. Le nombre d'arguments de la relation (on parle de l'arité d'une relation), donc d'entités liées, peut varier, de 2 pour une relation binaire à n pour une relation n-aire.

Une relation peut relier des entités nommées, comme des noms de personnes, de lieux, des dates, ou des termes, comme des noms de gènes, de maladies. Dans les deux cas, nous appelerons les arguments d'une relation dans le texte *entités*. Une relation peut être restreinte à des types d'entités particuliers, comme pour la relation `auteur_de` qui n'intervient qu'entre un nom de personne et un titre, ou au contraire elle peut exister pour des types d'entités différents, comme la relation d'hyperonymie `is-a`.

L'expression de la relation dans le texte n'est pas toujours explicite, c'est-à-dire qu'elle n'est pas toujours exprimée par des mots ou des expressions dans le texte. En effet une relation peut être implicite, c'est-à-dire qu'elle peut être identifiée grâce à d'autres indices. Dans l'exemple (2) la relation `auteur_de (Stendhal, le Rouge et le Noir)` est exprimée par le verbe et la préposition « écrit par », mais la relation `fonction(Stendhal, auteur)` est implicite, et peut être identifiée grâce à la présence d'une entité de type Œuvre, au verbe « écrire », etc.

Nous avons dressé une typologie des mentions de relation suite à l'observation de différents corpus. Pour l'illustrer, nous utilisons des exemples de mentions de la classe de relation `auteur_de` issus du Web. La forme la plus courante suit la structure SUJET - VERBE - OBJET, où le sujet et l'objet du verbe sont les arguments de la relation et le verbe la mention de la relation (voir l'exemple (8)). Il est possible que le verbe ait plusieurs sujets ou plusieurs objets coordonnés. Dans l'exemple (3), le verbe *écrire* a trois objets coordonnés.

La mention de la relation peut aussi être une préposition, comme dans l'exemple (4). Le troisième type de construction est plus complexe : les arguments sont dans deux propositions différentes et la mention de la relation est une structure particulière de phrase ou un enchaînement de faits, etc. Dans l'exemple (5), la première entité est dans la proposition principale et la deuxième dans la proposition relative.

La ponctuation peut aussi être un élément de la mention de la relation. Par exemple dans la phrase (6), la parenthèse est un élément important qui marque la relation entre l'auteur et ses œuvres.

(3) Entre-temps, **Bernard Werber** a écrit **Le Livre du voyage** (1997), **Le Père de nos pères** (1998) et le scénario de deux BD : **Exit 1** (1999), **Exit 2** (2000).

(4) Après avoir commencé **La conquête de Plassans** de **Zola** il y a environ un mois, j'ai enfin terminé de lire ce magnifique roman, 4e tome de la série des Rougons Macquarts [...]

(5) En 1979, **Daniel Pennac** fait un séjour de deux ans au Brésil, qui sera la source d'un roman publié vingt-trois ans plus tard : **"Le Dictateur et le hamac"**.

(6) Cinquième roman d'**Anna Gavalda** (après **La Consolante**, **À leurs bons coeurs**, **Ensemble c'est tout** et **Je l'aimais**) **L'échappée belle** est sortie en version poche le 2 mai dernier (tiré à 300 000 exemplaires par J'ai lu).

Lors de la dernière édition de la conférence MUC : MUC-7, (voir Chinchor [1998] pour la définition des tâches de MUC-7), la tâche d'extraction de relations entre entités nommées a été "officiellement" formalisée en tant que tâche indépendante. Au cours de cette conférence, une tâche (*Template relation task*) était dédiée à l'extraction de relations avec une organisation (`employee_of`, `manufacture_of`, et `location_of`) pour remplir un schéma alors qu'une autre (*Scenario Template Task*) était construite autour d'un événement préspécifié dans lequel étaient impliqués des organisations, personnes ou artefacts particuliers. Ces deux tâches mettent en jeu les différents problèmes à résoudre : sélection de l'information pertinente, reconnaissance de termes et d'entités nommées, et reconnaissance de relations entre ceux-ci.

Par exemple pour la tâche *Template relation*, à partir de la phrase (7) il fallait trouver la relation `employee_of` entre *Dennis Gillespie* et la *NAVY*, comme suit :
<EMPLOYEE_OF-9602040136-5>:=
PERSON: <ENTITY-9602040136-11>
ORGANIZATION: <ENTITY-9602040136-1>

<ENTITY-9602040136-11>:=
ENT_NAME: "Dennis Gillespie"
ENT_TYPE: PERSON
ENT_DESCRIPTOR: "Capt."; "the commander of Carrier Air Wing 11"
ENT_CATEGORY: PER_MIL

<ENTITY-9602040136-1>:=
ENT_NAME: "NAVY"
ENT_TYPE: ORGANIZATION
ENT_CATEGORY: ORG_GOVT

(7) The officer said the decision to reassign Kilian to the Pacific headquarters of the Navy's Fighter Wing was made Saturday by the commander of Carrier Air Wing 11, Capt. Dennis Gillespie.

Cette relation n'est vraie qu'à un certain moment. Dans un document plus ancien ou plus récent, la relation `employee_of` faisant intervenir l'argument *Dennis Gillespie* pourra être différente. Nous pouvons distinguer ce genre de relations issues d'une observation ou d'une expérience qui ne sont vraies qu'à l'endroit où elles sont instanciées et les relations toujours vraies quelles que soient leurs réalisations phrastiques. Par exemple, les relations d'hyperonymies (souvent appelées relations `is-a`) sont toujours vraies, comme par exemple la relation `est-un(pomme, fruit)` qui peut être extraite de différents énoncés. La relation `auteur_de(Stendhal, le Rouge et le Noir)` est aussi une relation toujours vraie, elle peut être extraite de l'exemple (2) mais également de l'exemple (8).

(8) **Stendhal** a publié **Le Rouge et le Noir** en 1830.

Au contraire, la relation `localisation(Paul, Grenoble)` extraite de la phrase (9) n'est vraie que dans cet énoncé. Nous dirons de cette relation qu'elle est contextuelle.

(9) **Paul** est parti à **Grenoble** pour ses études.

## 1.4 Arité des relations

Les relations sont souvent divisées en deux catégories : les relations n-aires, impliquant n arguments, et les relations binaires, c'est-à-dire entre deux entités, qui sont un cas particulier des relations n-aires.

### 1.4.1 Relation n-aire

Nous allons détailler ici la typologie des relations n-aires proposée par le groupe de travail du W3C (Noy et Rector [2006]). Une relation n-aire est définie comme une relation qui fait intervenir plus de deux arguments (ou individus dans le sens ontologique). Ils en distinguent 4 cas :

1. Une relation binaire à laquelle il faut ajouter un attribut décrivant l'instance de la relation. Pour cela, il faut créer un individu qui représente l'instance de la relation, qui aura des liens avec tous les participants. La figure 1.1 est un exemple de représentation de ce type de relation n-aire.
2. Une relation entre un individu et différents aspects de l'instance de la relation (voir la figure 1.2).

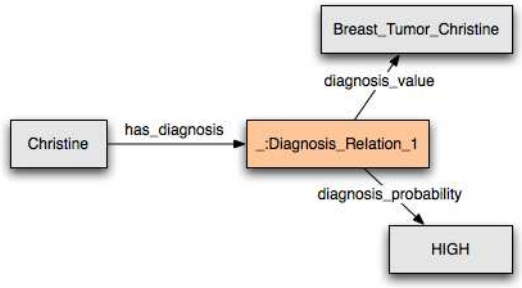



FIG. 1.1 – Représentation de la relation n-aire extraite de la phrase *Christine has breast tumor with high probability* (de Noy et Rector [2006]).

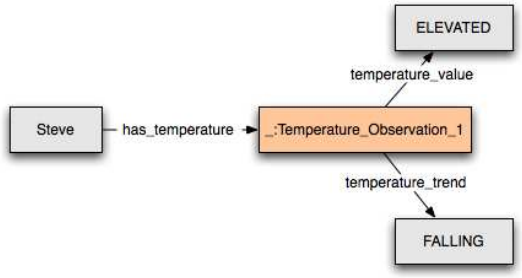



FIG. 1.2 – Représentation de la relation n-aire extraite de la phrase *Steeve has temperature, which is high, but falling* (de Noy et Rector [2006]).

3. Une relation qui fait intervenir des individus qui jouent des rôles différents et dont aucun n'est central (voir la figure 1.3).

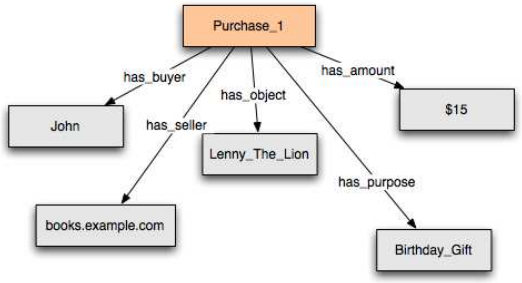



FIG. 1.3 – Représentation de la relation n-aire extraite de la phrase *John buys a "Lenny the Lion" book from books.example.com for $15 as a birthday gift* (de Noy et Rector [2006]).

4. Une relation dont l'objet est une liste d'arguments qui doivent être ordonnés (par exemple chronologiquement) (voir la figure 1.4).

Selon les tâches, le type de relation n-aire à extraire est différent. Nous allons donner quelques exemples de tâche d'extraction n-aire et le type de relation considéré.

Dans les évaluations de MUC, une tâche consistait à remplir un schéma prédéfini (Humphreys *et al.* [1998]). Le schéma peut être composé du nom d'une organisation, de sa des-

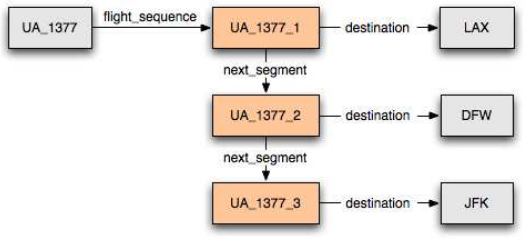



FIG. 1.4 – Représentation de la relation n-aire extraite de la phrase *United Airlines flight 1377 visits the following airports: LAX, DFW, and JFK* (de Noy et Rector [2006]).

cription, de son type, de sa nationalité, etc. Dans MUC-6, trois types de schémas à remplir ont été définis : ORGANISATION, PERSONNE et ARTEFACT (OBJET). Ce type de relation n-aire se rattache au troisième cas de relation décrit par le W3C.

Dans le domaine médical, une prescription médicale peut être représentée par une relation n-aire. L'extraction des prédictions dans des textes cliniques était l'objet du challenge i2b2 en 2009 (Uzuner *et al.* [2010b]). Dans le cadre de ce challenge, une prescription a été défini comme une relation entre un médicament et des informations concernant la prise de ce médicament par le patient (le dosage du médicament, le mode d'administration, la fréquence de la prise, la durée du traitement et les raisons pour lesquelles le médicament a été donné). Dans cet exemple de relation n-aire, un argument est central à la relation : le médicament, et les autres arguments sont en relation avec l'argument central, cela correspond au deuxième cas de relation n-aire décrit par le W3C.

Dans le domaine biomédical, McDonald *et al.* [2005] ont travaillé sur l'extraction de relations complexes représentant des événements de variations génomiques. Ils définissent une relation n-aire R par le schéma $R(t_1, t_2, ..., t_n)$ où $t_i$ sont des types d'entités. Une instance de la relation est une liste d'entités $(e_1, e_2, ..., e_n)$ qui sont du type $t_i$ ou qui sont absents de la relation. Cette définition d'une relation n-aire est proche du troisième cas de relation décrit par le W3C.

La tâche d'extraction de relations n-aires consiste à reconnaître les arguments et la relation si celle-ci est explicite, ou à reconnaître les arguments et les relier entre eux, si la relation est implicite.

### 1.4.2 Relation binaire

Une relation binaire est une relation faisant intervenir deux entités (qu'on peut appeler aussi des arguments de la relation). Ces arguments peuvent avoir un rôle particulier dans la relation, on parle alors de relation orientée (ou asymétriques), mais ils peuvent également intervenir dans une relation dont les arguments ont le même rôle (on parle alors de relation non-orientée ou symétrique). Par exemple dans la phrase (10), la relation `frêre_de(Bart, Lisa)` est non-orientée : les deux arguments ont le même rôle, alors que dans l'exemple précédent (2), les deux arguments **Stendhal** et **le Rouge et le Noir** ont des rôles différents dans la relation *auteur_ de*.

(10) **Bart** est le frêre de **Lisa** et le fils de Homer et Marge.

Plusieurs campagnes d'évaluations se sont intéressées à l'extraction des relations binaires. La première à s'y être intéressée est MUC-7, avec la tâche *Template relation Task*.

Dans le domaine biomédical, l'extraction des interactions géniques et donc la modélisation de ces interactions est une phase très importante dans la compréhension du fonctionnement cellulaire. La recherche en extraction de relations a souvent porté sur cette tâche, et plusieurs campagnes d'évaluations ont eu lieu sur ce sujet. Par exemple, le challenge LLL05 (Learning Language in Logic) (Nédellec [2005]) portait sur l'extraction des interactions entre des protéines et des gènes dans des résumés de la base MEDLINE. Un challenge un peu similaire a eu lieu en 2011 sur l'extraction des interactions entre médicaments : DDI Extraction (Segura-Bedmar *et al.* [2011b]). Le corpus était également composé de résumés d'articles scientifiques. Les relations entre des médicaments et des maladies sont aussi étudiées, entre autres pour repérer les récurrences de certains phénomènes. En 2010, le challenge i2b2 (Uzuner *et al.* [2011]) proposait une tâche d'extraction de différentes relations entre des médicaments, des examens et des problèmes médicaux.

La tâche d'extraction de relations binaires peut être posée différemment selon les données de départ (Sarawagi [2008]) :

- les entités sont annotées dans les textes, et pour une paire donnée il faut trouver le type de la relation qui existe ;
- la relation $r$ et une entité $e$ sont connues, le but est de trouver les entités avec lesquelles $e$ est en relation $r$ ;
- la relation $r$ est connue, le but est d'extraire les instances des paires d'entités reliées par cette relation.

Dans la suite de ce travail, nous nous intéresserons uniquement au premier type d'extraction de relations binaires, à savoir, les entités sont annotées dans le texte, et nous cherchons à extraire les relations existantes entre ces entités.

## 1.5 Méthodes d'extraction de relations n-aires

Une relation n-aire est une relation qui fait intervenir plus de deux arguments, souvent de types différents. La définition d'une relation n-aire est très large et de ce fait les méthodes pour extraire ces relations sont nombreuses. L'approche sera différente si la relation est dans un tableau ou s'étend sur plusieurs phrases voire tout un texte, ou si tous les attributs de la relation sont dans la même phrase. Une différence est aussi observée entre les relations s'exprimant de façon explicite dans les textes et celles qui sont implicites. Les méthodes consisteront à identifier toutes les relations binaires entre deux entités, puis regrouper les relations binaires, ou directement à regrouper les entités selon leur position dans le texte, la phrase, etc. Nous présentons dans la suite, les principales méthodes qui ont été proposées dans la littérature.

Pour extraire une relation entre un individus et différents aspects de l'instance de la relation (c'est-à-dire une relation du type 2 de la typologie du W3C), il faut annoter les entités, puis relier les entités de types différents à l'entité centrale. La difficulté majeure de cette tâche est l'annotation des entités. Souvent la mise en relation consiste principalement à relier l'élément pivot avec les entités qui le suivent dans la phrase. C'est le cas pour l'extraction de prescriptions dans des rapports cliniques, tâche proposée dans le cadre du

challenge i2b2 2009 (Uzuner *et al.* [2010b]). Une prescription, aussi appelée événement de médication, est décrite par Gold *et al.* [2008] comme un médicament administré à un patient relié à sa dose, sa fréquence, la nécessité de prendre ce traitement, etc. Les prescriptions ont souvent la forme de liste, comme le montre l'exemple (11), ce qui représente une particularité de ce type d'événement.

(11) RAISON Prophylaxis, MEDICAMENT heparin DOSAGE 5000 unit MODE subcu FREQUENCE t.i.d. - the patient has consistently refused her MEDICAMENT heparin.

Grouin *et al.* [2010] ont défini une méthode à base de règles pour mettre en relation un médicament avec les attributs situés après le médicament dans la phrase. Si un certain type d'attribut est absent, alors ils le cherchent dans le début de la phrase (avant le nom de médicament). Ils utilisent aussi des règles pour les cas de doubles entrées, par exemple si un attribut ou un médicament s'applique à deux prescriptions différentes. Ils ont obtenu une F-mesure globale de 0,773. Patrick et Li [2010] utilisent des SVM (*Support Vector Machine*) pour la mise en relation des attributs pour la même tâche. Ils sont arrivés premier du challenge avec une F-mesure globale de 0,857, mais l'utilisation des techniques d'apprentissage nécessite un corpus annoté de grande taille.

Dans le cas des prescriptions, les relations sont implicites, c'est-à-dire que la relation est exprimée par le type et la position des entités dans la phrase. Lorsque les relations sont explicites, pour les extraire il faudra s'intéresser à certains mots (comme les verbes), à la structure syntaxique de la phrase, aux relations syntaxiques entre les mots, etc., mais des critères de surface seront souvent moins pertinents.

Certaines relations explicites sont exprimées dans les phrases à l'aide de relations syntaxiques entre les verbes et leurs attributs. Les méthodes utilisées dans ce cas, s'intéressent aux dépendances entre les entités et le verbe de la phrase. De l'exemple (12) tiré de Nguyen *et al.* [2010], les relations syntaxiques suivantes peuvent être extraites : `sujet(quitter, frère)`, `objet(quitter, Pau)`, `prep(quitter,pour)` et `prep(quitter, depuis)`. Il est alors possible de construire la relation complexe `quitter (le frère de mon ami, Pau, pour une ville près de Lyon, depuis deux semaines)`.

(12) Le frère de mon ami a quitté Pau, pour une ville près de Lyon, depuis deux semaines.

Dans le cadre du projet GEONTO portant sur l'information géographique, Nguyen *et al.* [2010] ont travaillé sur l'extraction des relations dans des récits de voyage et ont proposé un modèle fondé sur l'analyse syntaxique des phrases. Ils utilisent la typologie des relations n-aires du W3C, et pour chaque type de relations, ils listent les relations syntaxiques apparaissant dans la phrase. Ensuite pour extraire les relations et les catégoriser, ils identifient les relations syntaxiques de la phrase. Les relations extraites sont ensuite utilisées pour peupler une ontologie. Nguyen *et al.* [2010] n'ont pas encore présenté les résultats obtenus avec cette méthode. Cette méthode nécessite d'avoir une analyse syntaxique des textes fiable, et que l'élément central de la relation soit un verbe.

Il existe aussi des méthodes hybrides, combinant de l'apprentissage et des règles, c'est le cas de l'approche de Björne *et al.* [2010] pour l'extraction d'événements bio-moléculaires dans le cadre de la tâche BioNLP'09. Certains événements sont des relations binaires entre

le type de l'événement et un argument, et d'autres des relations n-aires du type 1 du W3C (une relation binaire et des attributs qui décrivent l'instance de la relation). Les événements bio-moléculaires y sont définis comme le changement d'état d'une bio-molécule et ils sont composés d'un mot déclencheur et d'attributs qui lui sont reliés, souvent par des relations syntaxiques. Ils utilisent deux classifieurs multi-classes : un pour détecter le mot déclencheur et le typer, et un pour identifier les relations dans la phrase entre les entités et les déclencheurs d'événements. La construction finale des événements est réalisée grâce à des règles. Buyko *et al.* [2009] réduisent l'arbre de dépendances avec une méthode à base de règles, avant d'identifier les arguments de l'événement avec un classifieur à noyau de graphes. Ils obtiennent des résultats moins bons que Björne *et al.* [2010] (5 points de moins sur le F-score).

La tâche de détection de l'événement peut être divisée en trois : détection du mot déclencheur, typage de l'événement et mise en relation des attributs. Dans ce cas des approches à base d'apprentissage, utilisant différents classifieurs peuvent être utilisés. Ahn [2006] traite la première sous-tâche comme une tâche de classification de mots et effectue une classification binaire pour détecter les mots déclencheurs, puis une classification multi-classes pour déterminer le type de l'événement. Ils ont développé cette méthode pour la campagne d'évaluation ACE 2005. Les événements y sont décrits comme des structures complexes reliant des arguments qui sont eux mêmes complexes. 8 types d'événements sont annotés. Ainsi par exemple, l'événement LIFE peut avoir plusieurs arguments, et a cinq sous-types : be-born, marry, divorce, injure et die. Ces événements peuvent être rapprochés du type 2 de la typologie du W3C, c'est-à-dire que ce sont des relations entre des individus (ou participants de l'événement) et différents aspects de l'instance de la relation (par exemple des informations temporelles).

Une autre méthode hybride, mêlant apprentissage et règles appliquées sur un graphe, a été proposée par McDonald *et al.* [2005]. Ils travaillent sur l'extraction d'événements de variations génomiques, qu'ils modélisent comme une relation du type 3 de la typologie du W3C, c'est-à-dire que ce sont des relations qui font intervenir des individus qui ont des rôles différents. Ils décomposent en deux étapes l'extraction : détection des paires d'entités en relation en faisant une classification binaire (grâce à un classifieur basé sur le modèle d'entropie maximale), reconstruction des relations complexes en sélectionnant les cliques maximales dans le graphe des relations. Döhling et Leser [2011] se sont inspirés de cette méthode pour extraire des relations 5-aires dans des articles sur les tremblements de terre. Ils apprennent des patrons basés sur la représentation en dépendances des phrases, puis ils calculent la clique maximale dans le graphe des relations.

Jusque là nous avons parlé de relations n-aires au niveau d'une phrase, mais certaines relations peuvent s'étendre sur plusieurs phrases. Par exemple, dans un article scientifique, une expérience sera souvent décrite sur plusieurs phrases, voire sur un paragraphe ou l'article complet. Pour extraire ces relations, il sera souvent nécessaire d'utiliser des critères de positionnement, soit les positions des entités par rapport à un élément central, soit les positions des entités dans la structure du texte, etc. Feng *et al.* [2007] proposent une méthode à base de CRF pour extraire des expériences sur les projections neuronales entre des régions du cerveau. Ils abordent le problème de l'extraction d'expériences comme l'annotation des phrases selon leur rôle dans le discours (du paragraphe ou du document). Ils

définissent un classifieur pour annoter les attributs de l'expérience, et un deuxième pour regrouper les attributs en une expérience. Le deuxième classifieur annote les phrases selon qu'elles contiennent le début de la description d'une expérience, la suite de la description ou qu'elles ne contiennent pas d'information concernant l'expérience. Pour classer les phrases, ils utilisent entre autres les attributs suivants : le type des attributs dans la phrase, la disposition de la phrase dans le paragraphe et les mots indicateurs de la description d'une expérience. Le problème avec cette méthode est qu'une phrase ne peut pas être annotée pour deux expériences. Donc si une phrase contient des éléments de description de deux expériences, une seule sera extraite.

Dans la littérature scientifique, il est très fréquent que les résultats d'expériences soient détaillés dans des tableaux. Dans ce cas l'information est structurée, ce qui facilite son extraction. Touhami *et al.* [2011] se sont intéressés à ces relations. Ils ont développé pour cela une ressource termino-ontologique (RTO) dont une partie est générique pour la tâche d'annotation des relations n-aires, et une partie spécifique à un domaine d'application. Dans leur représentation, une relation est caractérisée par son étiquette et sa signature définie par un domaine (arguments d'accès) et par un co-domaine (argument résultat). Le modèle qu'ils proposent pour l'annotation des tableaux est constitué de 5 étapes. Ils distinguent les colonnes symboliques (qui contiennent des termes) des colonnes numériques. À chaque colonne, ils associent un concept de la RTO. Ils identifient ensuite quelles relations de la RTO sont représentées dans le tableau. Pour cela, ils combinent deux scores : le score d'une relation pour le tableau considérant le titre du tableau (= le maximum de la mesure cosinus entre les termes dénotant la relation dans la RTO et le titre du tableau), et le score d'une relation considérant le contenu du tableau (= le nombre de concepts reconnus dans la signature de la relation par le nombre de concepts dans la signature de la relation). La dernière étape est l'instanciation de chaque relation pour chaque ligne du tableau.

Ghersedine *et al.* [2012] ont proposé une méthode similaire pour l'extraction de relations n-aires dans le texte. Des lexiques et des patrons lexico-syntaxiques sont utilisés pour l'identification des arguments, une méthode à base de co-oocurrences est utilisée pour identifier l'argument pivot et des représentations des relations sous forme de graphes sont construites pour extraire les relations n-aires.

Dans notre travail, nous nous sommes intéressée à l'extraction de résultats d'expériences qui peuvent être vus comme des relations implicites et s'étendant sur plusieurs phrases. Dans une même phrase, nous pouvons avoir plusieurs relations ; il sera donc difficile d'utiliser des méthodes à base d'apprentissage comme Feng *et al.* [2007]. De même comme nos relations sont implicites, les méthodes qui utilisent les relations syntaxiques ne seront pas exploitables. En revanche les méthodes basées sur un élément pivot, comme le médicament dans le cas de l'extraction de prescriptions, sont proches de ce que nous avons proposé.

Maintenant que nous avons vu les méthodes d'extraction de relations n-aires (faisant intervenir plus de deux arguments), nous allons présenter les méthodes d'extraction des relations binaires qui diffèrent des précédentes.

## 1.6 Méthodes d'extraction des relations binaires

Les méthodes d'extraction de relations diffèrent selon la sous-tâche traitée : détection d'une relation entre deux entités, identification de la direction des relations, c'est-à-dire du rôle des arguments et catégorisation des relations. Nous nous intéresserons ici uniquement aux méthodes pour détecter et catégoriser une relation.

### 1.6.1 Méthodes fondées sur les co-occurrences

Les relations hiérarchiques et lexicales (par exemple les relations d'hyperonymies) sont vraies tout le temps au sein d'un domaine. Une méthode fondée sur les co-occurrences permet assez facilement de détecter ce genre de relations (par exemple Jelier *et al.* [2005]). Cette méthode est souvent utilisée comme baseline pour évaluer de nouvelles approches. Elle repose sur l'idée que deux mots qui apparaissent fréquemment dans le même contexte peuvent être sémantiquement liés. Avec cette méthode, les relations sont extraites avec un rappel élevé mais une précision très basse.

Pour extraire des relations domaniales [7] et pour catégoriser des relations, cette méthode n'est pas appropriée. Par exemple, si on veut extraire les relations entre un médicament et une maladie, dans un texte scientifique on pourra trouver que le médicament est destiné à soigner la maladie, mais dans des rapports cliniques, les deux entités pourront être dans certains cas reliées par une relation du type : le médicament traite la maladie, ou le médicament ne guérit pas la maladie. Dans ce cas, le fait que les deux entités co-occcurrent un certain nombre de fois dans le corpus, ne permettra pas de correctement classer la relation.

### 1.6.2 Méthodes à base de patrons

Les relations explicites qui sont exprimées sur une seule phrase, peuvent souvent être extraites à l'aide de patrons. Les approches à base de patrons sont très nombreuses dans le domaine de l'extraction de relations binaires. Quand peu de données annotées sont disponibles, les patrons sont définis manuellement, sinon ils sont appris sur un corpus annoté. Les méthodes qui reposent sur la définition manuelle de patrons sont peu robustes, et généralement efficaces uniquement en précision. Les patrons construits ou extraits à partir de phrases dans lesquelles les entités en relation sont très éloignées, sont trop spécifiques. En effet, plus l'expression qui sépare deux entités est longue, plus la variation d'expression de la mention de la relation peut être grande.

Une relation entre deux entités pourra être exprimée de façons diverses dans une phrase ; il est nécessaire de définir des patrons qui ne soient ni trop spécifiques, ni trop génériques. Par exemple, un patron entièrement lexicalisé aura tendance à être trop spécifique, alors qu'un patron formé uniquement des catégories morpho-syntaxiques des mots sera au contraire trop générique. Il est important que les patrons capturent la variété des formes linguistiques dans lesquelles les relations peuvent être exprimées (d'un point de vue morphologique, syntaxique, lexical et sémantique).

[7] Une relation domaniale est définie dans Grabar *et al.* [2004] comme « les relations désignant les actes, les processus ou autres, spécifiques au domaine et au corpus ».

Les patrons doivent donc contenir plusieurs niveaux linguistiques. À partir des exemples (13) et (14), nous pouvons dégager les patrons du tableau 1.1.

(13) **Merci** est une oeuvre de l'écrivain français **Daniel Pennac** parue en 2004.

(14) **L'enfant** est un roman de **Jules Vallès**, premier volet de la trilogie [...].

| patron morpho-syntaxique |
|---|
| `TITRE VER:per DET:ART NOM PRP (DET:ART NOM ADJ)? AUTEUR` |
| patron lexical |
| `TITRE est (un\|une) (oeuvre\|roman) de (l'écrivain français)? AUTEUR` |
| patron lexico-syntaxique |
| `TITRE est DET:ART (oeuvre\|roman) de (DET:ART NOM ADJ)? AUTEUR` |

TAB. 1.1 – Patrons d'extraction de la relation `auteur_de`. | est l'opérateur logique OU, ? marque le caractère facultatif de l'élément, et TITRE et AUTEUR remplacent les deux entités.

Hearst [1992] a proposé une méthode pour l'acquisition d'hyponymes qui est amorcée par une liste de paires d'hyponymes (extraite par exemple d'un thesaurus ou d'une base de connaissances). Les phrases contenant ces paires sont ensuite extraites d'un corpus pour avoir le contexte lexical et syntaxique des paires. Ces contextes servent à former des schémas (ou patrons) lexico-syntaxiques pour la relation d'hyponymie qui seront appliqués sur le corpus pour extraire de nouvelles paires d'hyponymes. À partir de l'exemple (15) et des paires d'hyponymes `(Allemagne, pays européen)` et `(Royaume-Uni, pays européen)`, on pourra former le patron suivant : `NP` $\{,\}$ `en particulier à` $\{NP,\}*$ $\{ou|et\}$ `NP`.

(15) Il devra premièrement se faire connaître aux autres pays européens, en particulier à l'Allemagne et le Royaume-Uni.

Ce type de patrons (qui peuvent être lexicalisés ou pas, multi-niveaux, etc.) a également été utilisé dans des systèmes d'extraction d'information pour le domaine biomédical.

Dans le domaine médical, nous pouvons citer le système SemRep (Rindflesch *et al.* [2000]) conçu pour extraire des relations de branchement artériel dans des comptes rendus opératoires, qui a également été appliqué à des relations entre des problèmes médicaux et leurs traitements (Srinivasan et Rindflesch [2002]). Les entités en relation correspondent à des concepts dans l'UMLS. Dans un premier temps, SemRep annote les entités en les associant à un type sémantique. Les patrons d'extraction des relations proviennent également de l'UMLS et sont des patrons sémantiques. Par exemple le patron `Pharmacologic Substance TREATS Disease or Syndrome` permettra d'extraire la relation dans la phrase (16).

(16) Methotrexate therapy in systemic lupus erythematosus.

Le système MedLEE, s'intéresse aux relations dans des comptes rendus radiologiques, des interactions biomoléculaires (Friedman *et al.* [2001]) et des relations gène-phénotype (Chen et Friedman [2004]), et il utilise des patrons lexicalisés. Abacha et Zweigenbaum

[2011] utilisent également des patrons lexicalisés pour extraire des relations entre des maladies et des traitements. D'autres comme Embarek et Ferret [2008] ont utilisé des patrons multi-niveaux, c'est-à-dire contenant soit la forme des mots, soit leur catégorie morpho-syntaxique, soit leur forme lemmatisée.

Les systèmes que nous avons présentés jusqu'ici utilisent des patrons s'appliquant sur les phrases. Mais il est également possible de construire des patrons s'appliquant sur l'arbre de constituants ou l'arbre de dépendances. Ils permettront de traiter plus facilement les phrases ayant une structure complexe.

Snow *et al.* [2005] utilisent une méthode similaire à Hearst [1992] pour la tâche d'acquisition d'hyponymes mais ils apprennent des patrons s'appliquant sur les arbres de dépendances. Ils arrivent ainsi à augmenter la précision et le rappel par rapport aux patrons s'appliquant sur la forme de surface de la phrase. Fundel *et al.* [2007] extraient le chemin reliant les deux entités dans l'arbre de dépendances, puis ils utilisent trois règles reflétant les formes d'expressions des relations les plus courantes en anglais *effector-relation-effectee*, *relation-of-effectee-by-effector* et *relation-between-effector-and-effectee*. Ils s'intéressent aux relations entre des gènes et des protéines dans des résumés de MEDLINE. Ils ont testé leur méthode sur plusieurs corpus : le corpus LLL (Nédellec [2005]), le corpus hprd50 (Fundel *et al.* [2007]) qu'ils ont annoté eux-mêmes, et un corpus grande échelle contenant un million de résumés. Les résultats qu'ils obtiennent sur le corpus LLL sont meilleurs que ceux obtenus avec des méthodes par apprentissage par les participants au challenge.

Le problème des patrons est qu'ils sont souvent mauvais en couverture et très dépendants du corpus et du domaine. Les adapter à un autre domaine est une tâche longue et parfois difficile.

### 1.6.3 Méthodes fondées sur le verbe

Les relations s'expriment de différentes manières dans les textes, mais souvent le verbe principal est l'élément déclencheur de la relation. Dans l'exemple (17), la relation `auteur_de` entre un auteur et son œuvre sera exprimée par le verbe *publier*.

(17) **Stendhal** a publié **Le Rouge et le Noir** en 1830.

Sharma *et al.* [2010] ont proposé une méthode basée sur le verbe principal de la phrase pour extraire des relations entre des entités biologiques. Ils ont fait le constat que dans 97% des phrases contenant une relation dans leur corpus annoté (50 résumés de MEDLINE), la relation était basée sur un verbe. Leur approche s'appuie sur une liste de verbes du domaine exprimant les relations recherchées, qui a été construite à partir de 54 verbes contenus dans l'UMLS et complétée avec les synonymes de ces verbes présents dans WordNet et VerbNet.

### 1.6.4 Méthodes par apprentissage supervisé

De nombreuses méthodes à base d'apprentissage supervisé ou semi-supervisé ont été proposées pour l'extraction de relations. L'apprentissage automatique est une méthode permettant de combiner les informations pertinentes à utiliser pour détecter et typer les

relations. Les attributs utilisés pour la classification représentent principalement de l'information lexicale, sémantique ou syntaxique.

Les algorithmes d'apprentissage automatique à noyau utilisent un classifieur linéaire pour résoudre un problème non-linéaire, en transformant l'espace de représentation des données d'entrées. Il est alors possible de classer des données représentées sous forme d'arbres.

Les techniques d'apprentissage supervisé les plus utilisées pour l'extraction de relations sont les SVM (Machines à Vecteurs de Support) (Uzuner *et al.* [2010a]; Zhou *et al.* [2005]; Roberts *et al.* [2008]). D'autres systèmes utilisent des classifieurs basés sur des CRF (Sahay *et al.* [2008]), sur des modèles d'entropie maximale (de Bruijn *et al.* [2011]) ou sur des réseaux de neurones (Rosario et Hearst [2004]).

Les attributs utilisés par les systèmes à base d'apprentissage pour représenter les relations sous forme vectorielle, peuvent être des attributs de surface (l'ordre des concepts, la distance, etc.), des attributs lexicaux (des trigrammes lexicaux, les mots qui forment les concepts, etc.), des attributs morpho-syntaxiques (catégories des mots de la phrase, etc.) et syntaxiques (les verbes, des bigrammes syntaxiques, les dépendances syntaxiques entre les concepts, etc.).

Les informations lexicales, syntaxiques et sémantiques ne sont pas toujours extraites dans le même contexte. Il est en effet possible de prendre en compte tous les mots de la phrase ou uniquement les mots entre les deux entités, etc. Nous verrons d'abord les contextes utilisés par différentes méthodes, et nous présenterons les informations surfaciques, sémantiques puis syntaxiques prises en compte par les systèmes.

#### 1.6.4.1 Modèlisation du contexte

Dans le domaine médical sur un corpus de comptes rendus médicaux, Uzuner *et al.* [2010a] utilisent des SVM pour extraire et typer des relations factuelles entre des problèmes médicaux présents ou possibles, des tests et des traitements. Une particularité de cette tâche est le type de corpus. En effet dans les comptes rendus médicaux, il y a des énumérations, des abréviations, des phrases mal formées, etc. Les informations lexicales sont extraites dans un contexte de trois mots avant et après les deux entités, alors que les informations syntaxiques le sont dans un contexte de deux mots. D'autres informations sont extraites sur toute la phrase, comme par exemple les verbes.

Sur un corpus journalistique dans le domaine général, Zhou *et al.* [2005] ont travaillé sur l'identification des relations entre des personnes, des organisations, des lieux, etc. dans le corpus ACE. Le corpus ACE est composé d'articles de journaux, dans lesquels 5 types d'entités ont été annotées, ainsi que 24 types de relations pour le corpus d'entraînement. Ils ont développé un système d'apprentissage à base de SVM utilisant des attributs lexicaux et syntaxiques. Le contexte duquel sont extraits les attributs est composé de deux mots avant la première entité, deux mots après la deuxième entité et tous les mots entre les deux.

Roberts *et al.* [2008] ont également choisi d'utiliser des SVM pour détecter des relations entre des entités de type *investigation*, *condition*, *negation modifier*, etc. dans un corpus de comptes rendus médicaux de patients atteints d'un cancer (dans le cadre du projet

CLEF - the Clinical E-Science Framework project). Ils utilisent les mêmes attributs que Zhou *et al.* [2005] mais ils s'intéressent en plus aux relations inter-phrases (c'est-à-dire des relations qui s'étendent sur plusieurs phrases). Ils considèrent un contexte de 6 mots avant la première entité et après la deuxième, ce qui forme un contexte large. Dans leur corpus, 23% des relations sont inter-phrases. Ils essayent avec le même système de classer aussi ces relations, mais quand ils prennent en compte les relations qui s'étendent sur deux phrases adjacentes et non plus sur une seule, la F-mesure diminue de 6 points.

Également dans des comptes rendus médicaux, Dogan *et al.* [2011] cherchent à détecter et classer des relations entre des problèmes médicaux, des tests et des traitements Pour cela ils découpent le contexte des deux entités candidates en 5 blocs : l'introduction, le premier concept, la connexion entre les deux concepts, le deuxième concept, et la conclusion. Pour chaque bloc, ils extraient les mots racinisés, les mots déclencheurs d'une non-relation, les identifiants des concepts dans l'UMLS [8], les types sémantiques des concepts et les assertions. Les blocs introduction et conclusion contiennent au maximum 5 mots (dans la limite de la phrase). Ce découpage en bloc du contexte est équivalent au contexte considéré par Zhou *et al.* [2005], mis à part que la fenêtre de mots avant et après les entités est plus petite.

Giuliano *et al.* [2006] proposent l'outil jSRE [9]. Leur méthode est basée sur une combinaison de fonctions à noyaux, une pour prendre en compte le contexte global de la relation (c'est-à-dire toute la phrase dans laquelle la relation apparaît) et une pour le contexte local (c'est-à-dire le contexte autour des deux entités candidates).

Katrenko et Adiaans [2007] séparent également en deux le contexte pour la détection d'interactions entre protéines : le contexte local correspond au nœud parent et aux deux nœuds enfants des protéines, et le contexte global au nœud le plus bas dans l'arbre qui englobe les deux protéines et à la racine de l'arbre.

Dans ces différents travaux, nous avons vu que le contexte utilisé pour extraire les attributs pour le classifieur est différent. Nous allons maintenant présenter les différents attributs qui sont extraits par ces systèmes dans ces contextes pour détecter ou classer les relations.

#### 1.6.4.2 Informations surfaciques

Une grande partie des systèmes utilisent des informations surfaciques comme attributs de base pour le classifieur. Ces informations surfaciques peuvent être les mots de la phrase, la distance entre les deux entités, la taille de la phrase, etc. Par exemple, Zhou *et al.* [2005] prennent en compte le nombre de mots entre les entités, les mots qui forment les deux entités, le fait qu'il n'y a pas de mots entre les deux entités, etc.

jSRE (Giuliano *et al.* [2006]) requiert uniquement une analyse linguistique de surface (tokenisation, découpage en phrases, etc.) pour extraire des relations. Ils ont évalué leur méthode sur la tâche d'extraction d'interaction entre protéines et gènes sur deux corpus : AIMed (Bunescu *et al.* [2005]) et LLL (Nédellec [2005]). Les résultats obtenus dépassent

[8] Unified Medical Language System : méta-thésaurus développé par la US National Library of Medicine Lindberg *et al.* [1993]

[9] http://hlt.fbk.eu/en/technology/jSRE

une partie des méthodes basées sur des attributs syntaxiques et sémantiques.

Les attributs lexicaux (par exemple les mots de la phrase) et de surface (par exemple la distance entre deux entités) ne sont pas toujours suffisants pour identifier correctement une relation. L'information syntaxique ou sémantique peut améliorer la précision du système. Nous allons présenter dans la suite les attributs sémantiques et syntaxiques utilisés par les systèmes de classification.

#### 1.6.4.3 Informations sémantiques

Les informations sémantiques transposées sous forme d'attributs peuvent provenir de ressources pour la langue générale ou de ressources d'un domaine de spécialité.

Dans le domaine général, il est possible d'utiliser WordNet (Fellbaum [1998]) pour ajouter des attributs sémantiques. Culotta et Sorensen [2004] ont travaillé sur l'utilisation de la représentation en dépendance de la phrase pour extraire les relations dans le corpus ACE. Ils définissent un arbre de dépendance augmenté pour représenter les relations. Ils représentent chaque nœud de l'arbre par un vecteur d'attributs (mots, catégories morpho-syntaxiques, hyperonymes de WordNet, etc.). Pour explorer l'information disponible dans WordNet, ils ont essayé d'augmenter l'importance de deux nœuds s'ils ont le même hyperonyme dans WordNet, mais ils n'ont pas observé d'amélioration de l'extraction des relations.

Zhou *et al.* [2005] utilisent également WordNet pour ajouter des attributs sémantiques pour l'extraction des relations de parentés entre deux personnes dans le corpus d'ACE. Ils forment une liste avec tous les mots appartenant à la classe sémantique `person | ... | relative` dans WordNet, et ils la complètent avec des mots extraits des données d'entraînement. Pour les relations `citoyen_de` et `résident`, ils utilisent une liste des noms de pays. Ils augmentent leur F mesure de 1,5 points en utilisant ces deux listes.

Dans le domaine biomédical, le méta-thésaurus de l'UMLS est souvent utilisé pour typer les concepts de la phrase. Dogan *et al.* [2011] prennent en compte le type sémantique des concepts qui sont référencés dans l'UMLS. de Bruijn *et al.* [2011] annotent également les termes de comptes rendus médicaux avec les concepts de l'UMLS via l'outil MetaMap (Aronson [2001]), les négations avec l'outil ConText ou encore les entités nommées cliniques (médicaments, maladies, symptômes, etc.) avec l'outil cTAKES (Savova *et al.* [2008]). Ces annotations leur permettent d'ajouter des attributs sémantiques pour extraire des relations entre des médicaments, des maladies et des examens cliniques. Il peut être également intéressant d'utiliser des ontologies, telles que la FMA (*Fundational Model of Anatomy*) ou le MeSH (*Medical Subject Headings*), pour disposer des classes sémantiques des entités d'intérêts. Par exemple, Rosario et Hearst [2004] montrent que pour catégoriser des relations entre un traitement et une maladie, la catégorie sémantique des entités dans le MeSH leur permet d'augmenter de 13,2 points l'exactitude de leur système à base de réseaux de neurones.

#### 1.6.4.4 Informations syntaxiques

L'information syntaxique semble indispensable pour décrire et identifier une relation. Elle permet par exemple d'identifier le ou les verbe(s) de la phrase (souvent déclencheur de la relation), les groupes prépositionnels, etc. Les attributs syntaxiques les plus couramment utilisés pour l'extraction de relations, sont les catégories morpho-syntaxiques des mots du contexte des entités. Mais des travaux ont porté sur l'évaluation d'attributs syntaxiques plus riches, comme les dépendances syntaxiques.

La typologie des relations de Uzuner *et al.* [2010a] est fine ; ils ont donc dû trouver des attributs permettant de marquer les différences entre les classes de relations. Ils utilisent en plus des attributs classiques, des attributs portant des informations sur les dépendances syntaxiques reliant les entités candidates, sous forme d'attributs dans une approche vectorielle basée sur des SVM. Ils ont utilisé les chemins de dépendances entre les entités, mais pour seule une partie des paires (35% des 2471 paires analysées manuellement) il existe un lien de dépendance entre les deux entités. Cet attribut ne semble pas pertinent pour l'extraction des relations dans des comptes rendus médicaux, du fait de la nature des données. Ils évaluent leurs classes d'attributs, et montrent que les attributs qui apparaissent comme les plus utiles sont les trigrammes lexicaux et les mots qui forment les concepts. Les informations syntaxiques n'améliorent pas la classification.

Il est également possible d'utiliser des informations provenant des arbres de dépendances. Van Landeghem *et al.* [2008] définissent six classes d'attributs provenant du chemin le plus court reliant deux protéines dans l'arbre de dépendances pour extraire des interactions entre ces protéines, et ils ajoutent une classe « sac-de-mots » qui prend en compte les racines (ou stems) de tous les mots de la phrase. Les six classes d'attributs extraits du plus petit chemin entre les deux entités sont : les *v-walks* lexicaux et syntaxiques qui prennent en compte les propriétés lexicales ou syntaxiques de deux nœuds de l'arbre et le type de dépendance les reliant, les *e-walks* lexicaux et syntaxiques considèrent un nœud et les deux liens de dépendances qui lui sont associés, le nœud racine et la catégorie morpho-syntaxique du nœud racine. Ils montrent que les porformances de l'extraction sont similaires quand seules des informations lexicales ou syntaxiques sont utilisées, ou quand les deux sont prises en compte.

Il est parfois difficile de représenter les données avec des vecteurs d'attributs, comme les informations syntaxiques provenant d'arbres. Une bonne solution pour tenir compte de l'information structurelle est de calculer la similarité entre deux arbres et non plus entre deux vecteurs. La méthode à noyau d'arbres a été proposée dans ce sens.

Des travaux en domaine ouvert ont montré que l'information structurelle utilisée sous forme d'arbres grâce à des tree kernels (ou noyau d'arbres) améliore la classification (Culotta et Sorensen [2004]; Zelenko *et al.* [2003]; Zhang *et al.* [2006]). Culotta et Sorensen [2004] utilisent des arbres de dépendance sur le corpus ACE, et montrent que les tree kernels sont meilleurs que l'information structurelle mise sous forme vectorielle. Ils testent deux types de tree kernels : *contiguous kernels* qui n'apparie pas les séquences qui sont interrompues par des nœuds non appariés, et *sparse tree* qui autorise les nœuds non appariés à l'intérieur de séquences appariées. Les meilleurs résultats sont obtenus avec des *contiguous kernel* associés à des kernels « sac-de-mots ». Bunescu et Mooney [2005] tra-

vaillent également sur le même corpus et considèrent que l'information nécessaire pour modéliser une relation entre deux entités peut être capturée par le plus petit chemin entre les entités dans l'arbre de dépendance. Les résultats qu'ils obtiennent sont meilleurs que Culotta et Sorensen [2004]. Zhang *et al.* [2006] ont étudié l'apport de la structure syntaxique des phrases pour l'extraction de relations en domaine général, en s'appuyant aussi sur le corpus ACE. Ils testent différentes sélections dans les arbres syntaxiques (arbre complet englobant les deux entités en relation, plus petit arbre commun, en ne conservant que les nœuds, etc.). Ils montrent que les meilleurs résultats sont obtenus en utilisant le plus petit sous-arbre commun aux deux entités associé à des attributs pour représenter le type sémantique de l'entité, le type de la mention (nom propre, nom commun, pronom) et des attributs sémantiques (Zhou *et al.* [2005]).

Airola *et al.* [2008] proposent une approche basée sur le calcul de similarité entre des graphes de dépendances (*all-paths graph kernel*). Ils représentent la phrase sous la forme de deux graphes : un représentant la structure en dépendances de la phrase, et l'autre l'ordre linéaire des mots. Dans le premier graphe, ils associent des poids aux dépendances selon si elles font partie du plus court chemin entre les deux entités ou non. Ainsi leur méthode prend en compte à la fois les informations contenues dans le chemin entre les deux entités et toutes les informations contenues dans la phrase. Avec leur approche, sur le corpus AIMed (Bunescu *et al.* [2005]), ils obtiennent des performances inférieures à celles obtenues par Giuliano *et al.* [2006] avec l'outil jSRE, mais comparables à l'état de l'art.

Lorsque les phrases sont complexes, pour extraire les relations il semble utile de prendre en compte la structure de la phrase, et de ne pas s'arrêter à une représentation « sac-de-mots ». La prise en compte de la structure syntaxique permet de généraliser un peu l'expression des mentions des relations, mais cela ne suffit pas à prendre en compte toutes les variations d'expression. Une autre solution est de simplifier les phrases pour réduire la variabilité et ainsi mieux repérer les relations.

## 1.7 Simplification de phrases

La variété d'expression des relations est telle qu'il est difficile de la prendre en compte complètement. Pour la réduire et ainsi augmenter les performances des systèmes d'extraction d'information, on peut supprimer les éléments non indispensables, découper la phrase en plusieurs, passer de la voix passive à la voix active, etc. Nous nous sommes intéressée à cette question de la variation de formulation des relations et en particulier à la simplification de phrases en vue d'améliorer l'extraction des relations.

Dans la littérature, deux notions proches sont proposées : la simplification de phrases (ou de textes) et la compression de phrases. La première est définie comme le processus de passage d'une phrase complexe en des phrases simples, et la deuxième comme la suppression de l'information non essentielle de la phrase. Deux définitions trouvées dans la littérature sont données ci-dessous.

**Définition 3** *"Text simplification is the process of transforming complex sentences into a set of equivalent simpler sentences while preserving the original meaning. The goal is to make the resulting text easier to comprehend for human readers or to process by other*

*programs." (Damay* et al. *[2006])*

*"[...] the sentence compression task can be defined as follows: given a sentence S, consisting of words $w_1 w_2 ... w_n$, what is a subset of the words of S, such that it is grammatical and preserves essential information from S?" (Filippova et Strube [2008])*

Prenons un exemple. À partir du premier paragraphe de la page Wikipedia « Ionesco » (18), on peut vouloir produire des phrases simples ou réduites. Pour produire des phrases simples, une méthode consistera par exemple à gérer les syntagmes coordonnés, à supprimer les appositions, etc. (19). Les phrases compressées seront obtenues en supprimant des informations non essentielles dans un contexte particulier. Par exemple dans la dernière phrase de l'exemple (18), le fait que Ionesco soit un représentant du théâtre de l'absurde n'est pas une information essentielle si on cherche à extraire le nom des œuvres qu'il a écrit (20).

(18) **Eugène Ionesco**, né Eugen Ionescu le 26 novembre 1909 à Slatina (Roumanie) et mort le 28 mars 1994 à Paris, est un **dramaturge** et **écrivain** roumain et français. Il passe la majeure partie de sa vie à voyager entre la France et la Roumanie. Représentant du théâtre de l'absurde, il écrit de nombreuses œuvres dont les plus connues sont **La Cantatrice chauve**, **Les Chaises** ou bien encore **Rhinocéros**.

(19) **Eugène Ionesco**, né Eugen Ionescu le 26 novembre 19091 à Slatina (Roumanie).
**Eugène Ionesco**, mort le 28 mars 1994 à Paris.
**Eugène Ionesco** est un **dramaturge** et **écrivain** roumain et français.

(20) Il écrit de nombreuses œuvres dont **La Cantatrice chauve**, **Les Chaises** ou **Rhinocéros**.

Dans ce travail, nous utilisons uniquement le terme simplification de phrases pour parler indifféremment de simplification et de compression de phrases. Nous considérons que la simplification est l'activité qui consiste à modifier une phrase pour la rendre plus simple, soit en remplaçant les mots (ou termes) utilisés, soit en supprimant de l'information considérée inutile dans un contexte particulier, soit en découpant une phrase complexe en plusieurs phrases simples.

### 1.7.1 Simplification de phrases : pourquoi ? comment ?

La simplification de phrases est un domaine qui a intéressé les chercheurs pour plusieurs tâches du TAL. Elle peut être étudié comme une tâche à part entière ou comme un prétraitement pour d'autres tâches. Nous présentons dans cette partie principalement des travaux qui ont porté sur la simplification en tant que prétraitement.

La simplification peut porter sur le lexique employé dans la phrase ou sur la syntaxe de la phrase. Si l'objectif de la simplification est de rendre plus compréhensible un texte, par exemple pour des apprenants, alors la simplification portera principalement sur le lexique utilisé. En revanche, si l'objectif est d'améliorer les performances d'un analyseur syntaxique, alors la simplification sera syntaxique, et aura pour principale action de découper une phrase complexe en plusieurs phrases simples. Pour le résumé automatique, la compression de phrases sera souvent utilisée pour sélectionner l'information importante.

La simplification en tant que tâche indépendante, est utilisée pour améliorer la lisibilité d'un document, par exemple pour les personnes ayant des difficultés cognitives (Bott et Saggion [2012]), pour les apprenants, pour les non spécialistes (Biran *et al.* [2011]), etc. La simplification est très étudiée également en tant que prétraitement sur des textes pour améliorer les performances de tâches du domaine du TAL. Des travaux ont, par exemple, été menés pour la génération de questions (Heilman et Smith [2010]), l'annotation des rôles sémantiques d'un verbe (Vickrey et Koller [2008]), l'extraction d'interactions entre protéines (Jonnalagadda et Gonzalez [2010]), l'analyse syntaxique de phrases complexes (Chandrasekar *et al.* [1996]; Bui *et al.* [2010]), etc. Selon la tâche, les simplifications seront différentes : on dit alors que la simplification est guidée par une tâche. Pour toutes ces tâches, il est indispensable de vérifier la cohérence grammaticale des phrases après simplification. Les systèmes comporteront donc souvent un module permettant de contrôler la bonne structure des phrases, mais nous ne nous attarderons pas sur ce point.

Plusieurs méthodes ont été explorées pour simplifier une phrase, la plus courante consistant à écrire des règles s'appliquant aux phrases ou aux arbres de dépendances ou de constituants. L'analyse syntaxique de phrases complexes n'est pas toujours performante. Devant ce constat, Chandrasekar *et al.* [1996] ont proposé une méthode de simplification des phrases complexes en plusieurs phrases simples pour améliorer l'analyse syntaxique. Leur méthode consiste à repérer les points d'articulations des phrases, c'est-à-dire les points où la phrase peut être découpée. Une série de règles de transformation est ensuite appliquée. Pour la tâche d'annotation des rôles sémantiques, Vickrey et Koller [2008] ont observé que la variété d'expression syntaxique du chemin reliant un argument au verbe peut être grande. Pour réduire cette variété d'expression, ils ont écrit 154 règles s'appliquant à l'arbre de constituants pour supprimer toute l'information en dehors du verbe cible et de ses arguments. Ils proposent une méthode originale pour sélectionner les meilleures règles : ils appliquent les règles de simplification pour produire toutes les phrases simplifiées possibles, puis entraînent leur système d'annotation des rôles sémantiques. La validité de chaque règle est ensuite évaluée en fonction de l'impact de la simplification sur la tâche principale. Dans le domaine de la compression de phrases, tâche qui consiste à supprimer certains constituants d'une phrase considérés comme non essentiels, Yousfi-Monod et Prince [2006] se sont basés sur des règles linguistiques. Biran *et al.* [2011] ont également proposé une méthode à base de règles pour simplifier le lexique d'une phrase, mais celles-ci sont apprises automatiquement à partir de corpus comparables.

Les méthodes à base d'apprentissage sont plus difficiles à mettre en œuvre, entre autres parce qu'elles nécessitent de disposer d'un corpus annoté. Pour la compression de phrases, Knight et Marcu [2000] proposent deux modèles : un modèle fondé sur le modèle du canal bruité, qui pose l'hypothèse qu'une phrase non compressée était à l'origine d'une phrase compressée ; et un modèle fondé sur des arbres de décision. Waszak et Torres-Moreno [2008] proposent également une approche statistique. Ils ont défini un modèle de langage bigramme qui porte sur les lemmes et les catégories morpho-syntaxiques des mots, puis ils calculent l'entropie des phrases pour retrouver la meilleure compression possible d'une phrase. Zhu *et al.* [2010] envisagent la tâche de simplification comme de la traduction mono-langue. Ils proposent une méthode à base d'apprentissage en utilisant la Wikipedia et la Wikipedia simplifiée comme corpus d'entraînement. La simplification est divisée en 4 sous-tâches : découpage des longues phrases, suppression des parties non importantes,

réordonnancement et remplacement de mots compliqués.

Chaque tâche nécessite une simplification différente, selon si le découpage de la phrase est possible, si la cohérence grammaticale est à vérifier ainsi que la cohérence dans le discours, si la coordination est à supprimer, etc. Pour la tâche d'extraction de relations, la bonne structure de la phrase n'est pas indispensable et la cohérence du discours pas utile, mais il est nécessaire de conserver les entités susceptibles d'être en relation. Le découpage de la phrase ne sera donc pas toujours possible.

### 1.7.2 Simplification de phrases guidée par l'extraction de relations

La simplification dirigée pour l'extraction de relations binaires a aussi été étudiée, et a principalement été évaluée sur la tâche d'extraction des interactions entre protéines.

Comme nous l'avons dit précédemment, il est possible d'agir sur la phrase au niveau des mots ou des catégories morpho-syntaxiques, ou sur des représentations de la phrase : l'arbre de dépendances ou l'arbre de constituants. Pour la tâche d'extraction de relations, il est intéressant de supprimer les informations non essentielles de la phrase et celles qui peuvent gêner l'identification des relations binaires, et ainsi de réduire la variété d'expression des relations. La simplification lexicale ne nous semble pas pertinente. La simplification pour l'extraction de relations nécessite de conserver les deux entités candidates, mais conserver la grammaticalité de la phrase ne semble pas indispensable.

Les principales recherches dans le domaine biomédical ont été faites pour l'extraction d'interactions entre protéines (PPI). Les corpus desquels sont extraits ces interactions sont composés d'articles scientifiques. Coden *et al.* [2005] ont étudié la taille moyenne des phrases de plusieurs corpus, entre autres le corpus GENIA composé de résumés d'articles scientifiques et le corpus MED qui contient des rapports cliniques. Ils ont trouvé que dans le corpus GENIA, qui se rapproche des corpus utilisés pour l'extraction des PPI, la taille moyenne des phrases est de 27,18 contre 13,79 pour le corpus MED. Cette étude montre que les phrases sont plutôt longues dans les articles scientifiques et donc complexes. Pour pallier ce problème, une solution consiste à diviser les phrases complexes en phrases simples. C'est ce qu'ont fait par exemple Jonnalagadda et Gonzalez [2010] en développant l'outil bioSimplify. Ils ont écrit des règles de simplification syntaxique qui s'appliquent au niveau morpho-syntaxique. Leur système produit plusieurs phrases simples et grammaticalement correctes à partir de la phrase d'origine. Leur objectif est d'augmenter le rappel de l'extraction d'information dans le domaine biomédical. Ils ont en particulier évalué leur outil pour la tâche de PPI, et observent une légère amélioration. L'inconvénient de leur système est qu'aucune sélection de la (des) meilleure(s) phrase(s) simple(s) n'est effectuée, et que les règles n'obligent pas la conservation de la paire d'entités candidate. Segura-Bedmar *et al.* [2011a] ont proposé une méthode hybride pour la détection des interactions entre des médicaments (DDI) dans un corpus composé de documents pharmacologiques. Leur méthode combine des règles pour la résolution de construction linguistique complexe et de douze patrons écrits par un expert du domaine. Ils ont développé un algorithme pour repérer les propositions dans les phrases. Ils utilisent pour cela des informations lexicales et syntaxiques. Une fois les propositions identifiées, des règles de simplification sont appliquées pour découper les phrases complexes en phrases simples. Le découpage des propositions

n'améliore pas les performances de leur système d'extraction de DDI.

D'autres travaux se sont intéresssés à la compression des phrases et non en leur division. Miwa *et al.* [2010] ont, par exemple, utilisé des règles pour supprimer les informations inutiles et gênantes pour extraire les relations. La douzaine de règles qu'ils ont écrites s'appliquent sur la sortie d'un analyseur syntaxique. Elles sont appliquées pour chaque paire de protéines. La figure 1.5 est un exemple de l'application de la règle `Copula`. Ils ont évalué l'impact de la simplification pour l'extraction des interactions entre protéines et montrent que sur les 5 corpus différents qu'ils ont utilisés, l'extraction des relations est meilleure. Des travaux ont porté également sur la compression de phrases en simplifiant les arbres de dépendances : Thomas *et al.* [2011] pour l'extraction d'interactions entre protéines, par suppression ou modification de types de dépendances (voir figure 1.6), et Buyko *et al.* [2011] pour la tâche BioNLP'09 (extraction d'événements biologiques), par élagage de l'arbre (voir figure 1.7). Les améliorations observées sont faibles. Dans le domaine général, Garcia et Gamallo [2011] ont travaillé sur la simplification des structures des phrases par simplification de l'arbre de dépendances. Grâce à des règles, ils identifient et suppriment des constituants satellites et subordonnées de la phrase. Ainsi seuls les constituants têtes des dépendances sont conservés. À partir de la phrase simplifiée obtenue, ils peuvent extraire les relations (hasBirthPlace et hasProfession) avec des règles génériques. Ils n'évaluent pas l'apport de cette simplification.

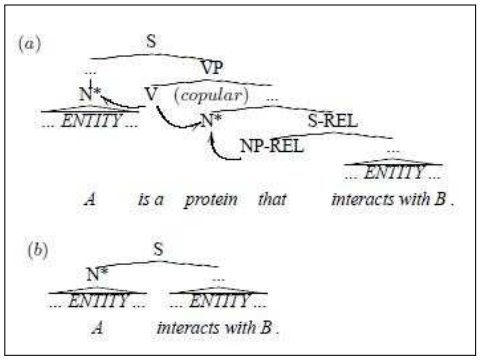



FIG. 1.5 – (a) est simplifiée en (b) (de Miwa *et al.* [2010])

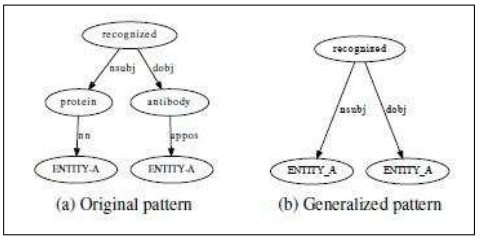



FIG. 1.6 – Arbre de dépendances de la phrase *ENTITY_A protein recognized antibody (ENTITY_A)* avant et après avoir supprimer les dépendances `nn` et `appos` (de Thomas *et al.* [2011])

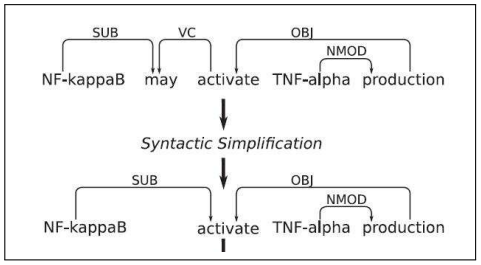



FIG. 1.7 – Élagage d'un graphe de dépendances (de Buyko *et al.* [2011])

L'information inutile peut être aussi repérée grâce à des règles s'appliquant sur l'arbre de constituants. Dans le domaine général, Qian *et al.* [2008] ont travaillé sur l'extraction

de relations dans le corpus ACE 2004. Ils modifient de trois manières l'arbre : suppression (pour la $2^e$ entité, ils ne gardent que le mot tête, etc.), compression (gestion de la coordination) et expansion (conservation de la structure possessive après la $2^e$ entité, etc.). Ces méthodes nécessitent une bonne analyse syntaxique des phrases.

La gestion de la coordination fait partie de la simplification la plus souvent effectuée avant l'extraction des relations. En effet, l'utilisation de la coordination peut rendre une phrase très complexe en particulier lorsque l'on cherche à extraire des relations. Gérer (ou résoudre) la coordination permet de simplifier la phrase pour avoir une forme simple du type SUJET VERBE OBJET. Par exemple, si on gère la coordination dans la phrase (21), on peut obtenir les phrases (22). Différents types d'éléments peuvent être coordonnés : des syntagmes nominaux (21), des syntagmes verbaux (23) ou des propositions. Dans le cas de ces dernières, il suffira souvent de diviser la phrase au niveau de la conjonction de coordination pour obtenir des phrases simplifiées.

(21) Entre temps, **Bernard Werber** a écrit **Le Livre du voyage** (1997), **Le Père de nos pères** (1998) et le scénario de deux BD : **Exit 1** (1999), **Exit 2** (2000).

(22) Entre-temps, **Bernard Werber** a écrit **Le Livre du voyage** (1997).
**Bernard Werber** a écrit **Le Père de nos pères** (1998).
**Bernard Werber** a écrit le scénario de deux BD : **Exit 1** (1999), **Exit 2** (2000).

(23) **Daniel Pennac** a publié et interprèté au théâtre **Merci**.

Dans les méthodes proposées par Qian *et al.* [2008]; Miwa *et al.* [2010]; Thomas *et al.* [2011], les problèmes de coordination sont traités. Miwa *et al.* [2010] suppriment les syntagmes en coordination avec une des deux entités. Qian *et al.* [2008] ont traité le cas de coordination de syntagmes nominaux mais également de syntagmes verbaux. Thomas *et al.* [2011] considèrent que si les deux entités ont un ancêtre commun et qu'elles sont reliées à lui par le même type de dépendance, alors les deux entités ne peuvent pas interagir. Ils ne modifient pas la phrase mais utilisent la présence de la coordination comme un indice d'absence d'interaction entre les protéines coordonnées. Segura-Bedmar *et al.* [2011a] ont utilisé des règles pour gérer les problèmes de coordination, et ils traitent dans le même temps les appositions, par exemple les structures du type `E1 such as E2 may have [...]` (E1 et E2 sont deux entités de la phrase) et les expressions corrélatives, par exemple les expressions contenant *both*. Ils augmentent ainsi le rappel de leur système à base de patrons, et obtiennent une F-mesure de 10 points supérieure.

## 1.8 Positionnement

Nous avons introduit le domaine de l'extraction d'information et présenté des travaux existants en particulier dans le sous-domaine de l'extraction de relations. Nous avons terminé cette partie en présentant la tâche de simplification de phrases et des méthodes de simplification proposées pour améliorer l'extraction des relations.

Nous avons présenté à la fois des méthodes utilisées pour l'extraction de relations n-aires et binaires. Pour l'extraction de relations n-aires, les approches développées différent un peu

des méthodes pour l'extraction des relations binaires, entre autres parce que les relations sont souvent implicites dans le texte (l'utilisation de patrons n'est donc pas pertinente), et que les arguments de la phrase peuvent être répartis sur plusieurs phrases. Les relations n-aires implicites ne sont pas exprimées par une mention de la relation dans la phrase, mais par un ensemble de phrases. Les relations entre les arguments de la relation ne sont pas forcément typées, c'est-à-dire que lors de la définition d'une relation n-aire, les types de lien entre les arguments ne sont pas connus. Une première partie de notre travail se concentre sur l'extraction de relations n-aires implicites et exprimées au niveau du discours entier. Pour extraire ces relations, nous ne disposions pas de corpus annoté permettant d'appliquer des méthodes à base d'apprentissage et nous avons vu que les méthodes à base de patrons ne sont pas utilisables pour identifier des relations implicites. Nous proposons donc une méthode qui utilise un élément pivot et des règles pour la mise en relation des arguments des relations n-aires, et des patrons et des lexiques pour l'annotation des arguments des relations. Le système ne pouvant avoir une précision et un rappel parfait, nous nous sommes intéressée à l'interaction possible avec les experts du domaine via une interface pour permettre de valider ou modifier les propositions du système.

Une seconde partie de notre travail a porté sur la détection et la classification de relations binaires. Les méthodes à base d'apprentissage, très utilisées pour l'extraction des relations binaires, obtiennent des bons résultats lorsqu'un corpus annoté de grande taille est disponible. Il ressort de la littérature que l'utilisation de ressources sémantiques améliore les résultats. Il en est de même avec la prise en compte de la structure syntaxique de la phrase dans le domaine général. Dans le domaine biomédical, les travaux existants ne montrent pas d'amélioration significative par l'intégration d'information syntaxique. Nous avons étudié l'ensemble des attributs permettant de caractériser une relation et nous nous sommes plus particulièrement intéressée à l'étude de l'intégration de connaissances sur la structure de la phrase pour la détection et la classification de relations.

Dans les corpus, les données ne sont pas toujours équilibrées et les différentes variations d'expression d'une relation pas toujours représentées. Nous nous sommes penchée sur la question de la réduction de la variabilité syntaxique dans le but d'extraire des traits caractéristiques des mentions des relations plus pertinents. Nous proposons une méthode de simplification des phrases qui permet de réduire la variabilité d'expression des relations. La difficulté du développement d'une méthode de simplification est l'identification et l'annotation de cette variabilité. La simplification de phrases guidée par une tâche n'a été que très peu étudiée jusque là. Beaucoup de travaux ont consisté en l'écriture de quelques règles permettant de supprimer certains éléments de la phrase, mais pas au développement de véritable méthode de simplification. Dans les différents travaux présentés précédemment pour la simplification destinée à améliorer l'extraction de relations binaires, nous avons vu qu'il était possible de proposer une méthode à base de règles s'appliquant sur l'arbre de dépendances, l'arbre de constituants, ou directement sur les mots de la phrase. Pour ce faire, il est nécessaire d'effectuer une étude préalable à l'écriture des règles, pour savoir si les régularités de simplification sont visibles au niveau des mots dans la phrase, ou au niveau syntaxique, dans quel cas les règles devront s'appliquer sur une représentation syntaxique de la phrase. Nous avons vu qu'il était également possible de proposer une méthode à base d'apprentissage pour effectuer ces simplifications. Les méthodes à base d'apprentissage nécessitent des corpus annotés, mais étant donné la difficulté et le temps

nécessaire pour annoter un corpus, nous avons exploré une voie permettant de définir un modèle n'impliquant pas un corpus annoté trop important.

Deuxième partie

# Extraction de relations complexes : application à des résultats expérimentaux en physiologie rénale

Une grande part des connaissances en biologie se trouve dans des articles scientifiques et c'est en particulier le cas des résultats d'expérimentations. Ces résultats sont exploités par des modèles mathématiques pour explorer des hypothèses complexes ou, dans un but plus heuristique, pour donner une représentation cohérente de l'état des connaissances. La mise en place de modèles mathématiques, leur calibration et leur validation s'appuient sur une connaissance exhaustive des mesures quantitatives expérimentales ; or, du fait de l'explosion de la quantité d'information disponible dans la littérature, la maîtrise de cette information est très difficile.

Devant ce constat, nous nous sommes intéressée au problème du peuplement de bases de données de résultats expérimentaux : nous avons mis en œuvre des méthodes d'extraction d'information, et les avons intégrées dans un assistant d'aide au peuplement de bases de données.

Notre cadre d'application est la base de données du domaine de la physiologie rénale Quantitative Kidney DataBase, ou QKDB (Dzodic *et al.* [2004]), qui a été créée dans le contexte du projet Physiome international (Hunter *et al.* [2010]). Elle s'adresse à l'ensemble de la communauté de physiologie rénale et de néphrologie, et vise à rassembler les résultats expérimentaux quantitatifs utiles pour la modélisation à partir de leurs publications dans les articles de physiologie rénale. Une version générique (QxDB, voir Ribba *et al.* [2006]), épurée de toute spécificité du domaine du rein, a aussi été développée et commence à être adoptée dans d'autres domaines tels que la croissance tumorale.

Construite en 2004, QKDB a été peuplée manuellement ; à l'heure actuelle environ 8 500 résultats ont été enregistrés dans la base, provenant de quelques 300 articles. Toutefois, et malgré l'utilité d'une telle ressource et la mise en place d'une interface web conviviale, la participation espérée des chercheurs du domaine au remplissage de la base est restée faible à cause du temps nécessaire pour entrer les valeurs et les commenter. Ce sérieux problème de *curation*, c'est-à-dire de la mise à jour manuelle de la base de données, est commun à beaucoup d'autres bases de données dans le domaine de la biologie.

Pour répondre à ce problème, nous avons conçu un système d'extraction automatique des informations qui décrivent des résultats expérimentaux, résultats que nous avons formalisés par des relations n-aires ; on parlera également d'événement. Ces relations sont souvent formulées de manière implicite dans les articles. Cependant, contrairement à la plupart des tâches existantes en extraction d'événement, les informations à extraire ne sont pas reliées à la mention de l'événement, mais à l'expression d'un résultat quantitatif. De plus, ces informations sont majoritairement exprimées par des termes du domaine (ex. par exemple « apical membrane ») et non par des entités nommées.

Pour que l'outil d'extraction puisse être utilisé facilement par les experts du domaine, nous l'avons intégré à un assistant d'aide au peuplement de la base de données. Cet assistant permet aux experts de visualiser dans les articles les informations susceptibles d'être ajoutées à la base de données, et de modifier, valider ou supprimer ces informations.

Quelques systèmes complets d'extraction d'information dans la litérature scientifique ont été développés. Beaucoup d'entre eux sont basés sur les résumés d'articles scientifiques, en particulier sur ceux référencés sur MEDLINE. Par exemple, le système PASTA (Demetriou et Gaizauskas [2002]) a pour but l'extraction d'information sur le rôle des résidus présents dans les molécules de protéines. La tâche réalisée consiste à remplir un schéma

défini par trois entités et deux relations à partir des résumés de MEDLINE. Moins de travaux ont porté sur l'extraction d'information sur l'article entier, d'une part parce que les articles ne sont pas toujours librement accessibles, et d'autre part à cause de la difficulté de conversion de PDF en un format structuré pour les articles anciens, pour ceux contenant des formules, des tableaux, etc. Nous pouvons citer BioRAT (Corney *et al.* [2004]) et Pharmspresso (Garten et Altman [2009]) qui ont été conçus pour extraire des informations sur l'expression et l'interaction des protéines, depuis des articles complets au format PDF convertis au format TEXT.

Les informations extraites peuvent être situées au sein d'une même phrase, d'un tableau, d'un paragraphe ou réparties dans tout l'article. Souvent les systèmes sont développés pour extraire l'information localisée dans une phrase (Demetriou et Gaizauskas [2002]; Corney *et al.* [2004]; Garten et Altman [2009]). Mais Swampillai et Stevenson [2010] ont montré que 28,5% des relations dans le corpus MUC6 sonr réparties sur plusieurs phrases. Il parait nécessaire de s'intéresser à l'extraction des relations inter-phrases. Dans Swampillai et Stevenson [2011], ils évaluent différentes approches d'apprentissage automatique à base de SVM pour extraire des relations binaires intra-phrases et inter-phrases. Ils montrent qu'une méthode combinant des attributs classiques (tokens, catégories morpho-syntaxiques, distance entre les entités et les verbes les plus proches) et des attributs structurels provenant des arbres de constituants permet d'obtenir des performances équivalentes pour l'extraction des relations intra- et inter-phrases.

Nous avons travaillé sur ce projet avec Stephen Randall Thomas (directeur de recherche au CNRS, spécialisé en physiologie rénale) pour les aspects biologiques, Anne-Laure Ligozat et Brigitte Grau pour le travail sur l'extraction d'information, et deux stagiaires (Adrien Dong et Corentin Limier) pour l'intégration du système d'extraction d'information à un assistant d'aide au peuplement de la base de données. Pour le développement de notre système, nous sommes repartie d'un premier système d'extraction d'information développé par un staigiaire encadré par Anne-Laure Ligozat et Brigitte Grau (Rémi Delorme).

# Chapitre 2

# Extraction d'une relation n-aire : un résultat expérimental

**Sommaire**



Nous nous sommes intéressée aux expérimentations présentées dans les articles scientifiques. Avant tout travail d'extraction, il a fallu étudier ce que les experts du domaine appellent *résultat expérimental*. Nous avons pour cela utilisé la base de données QKDB dans laquelle des résultats expérimentaux avaient été enregistrés, et nous avons travaillé avec un

expert du domaine. Dans une première partie, nous définissons ce que nous entendons par *résultat expérimental*, nous présentons ensuite la base de données et nous terminons par une description du corpus. La base de données QKDB a été conçue en 2004 pour héberger des données quantitatives, publiées dans le domaine de la physiologie rénale et épithéliale.

Nous appelons *résultat expérimental*, un résultat quantitatif obtenu suite à une expérience et mis en relation avec les informations permettant de décrire cette expérience. L'exemple 24 est un résultat expérimental extrait d'un article scientifique.

(24) Apical membrane $P_f$ averaged (in cm/s) **9.37 +- 0.77 e-4** (n = 5) at 20 ° C.

La phrase indique que l'expérience consistait à mesurer la perméabilité ($P_f$) de la membrane apicale, à une température de 20 ° C sur 5 individus (l'espèce n'est pas précisée dans la phrase). Le résultat de cette expérience est exprimé en cm/s. L'ensemble de ces informations sera appelé résultat expérimental.

Pour travailler sur la problématique de l'extraction de résultats expérimentaux, nous avons constitué un corpus avec des articles de physiologie rénale référencés dans la base de données QKDB. Nous présentons ce corpus dans une première partie. La partie suivante est consacrée à la représentation des résultats expérimentaux par des relations n-aires. Nous détaillerons ensuite la méthode mise en place pour extraire ces relations n-aires et nous terminerons par une évaluation de la méthode.

## 2.1 Corpus

### 2.1.1 Constitution du corpus

Dans la base, les articles proviennent en majorité des trois journaux prédominants dans le domaine de la physiologie rénale : *American Journal of Physiology - Renal Physiology*, *Kidney International* et *Journal of the American Society of Nephrology*. Ils sont disponibles au format PDF et parfois en XHTML sur le Web. Pour pouvoir analyser les articles, il était nécessaire d'en disposer dans un format avec une structuration simple et facilement convertible en texte, comme le XML. La conversion du format PDF au format XML n'est possible que pour des articles récents (d'une quinzaine d'années), puisque les documents plus anciens sont souvent des copies scannées de la version originale imprimée, et non une version électronique du document convertie en PDF. Pour les articles les plus récents, la conversion est possible mais pose des problèmes, principalement pour extraire les tableaux et certains caractères spéciaux (comme ±). Or les tableaux contiennent une quantité importante de résultats expérimentaux et des caractères spéciaux sont présents dans l'expression des résultats numériques (« 9.37 ± 0.77 »), les unités (« $\mu$mol/mg creatinine »), etc. De ce fait, nous n'avons gardé que les articles librement disponibles au format XHTML, soit 20 articles.

Les 20 documents conservés ont été convertis dans un format XML structuré avec les balises suivantes : titre, auteurs, corps de l'article, paragraphes, tableaux (avec lignes et colonnes) et notes de bas de page. Les tableaux ont été conservés car ils contiennent de nombreux résultats numériques : dans le corpus, 73% des valeurs numériques des résultats sont dans des tableaux. Les balises XHTML `td`, `th`, `caption`, etc. ont été converties en

balises `colonne`, `ligne`, `legende` et `tableau`. Un attribut `num` a été associé aux balises `colonne` et `ligne`, il a pour valeur le numéro de la ligne ou de la colonne du tableau. Grâce à cet attribut, il est possible de récupérer toute l'information contenue dans une ligne donnée ou dans une colonne. La figure 2.1 est un exemple d'un tableau dans un article en XHTML, visualisé avec Mozilla Firefox.

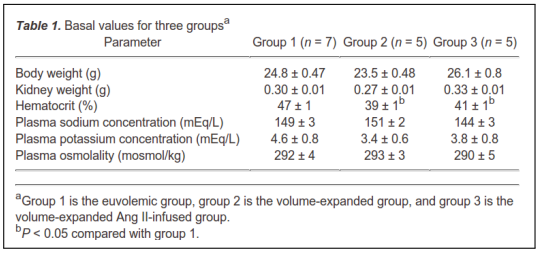


***Table 1.*** Basal values for three groups[a]

| Parameter | Group 1 ($n$ = 7) | Group 2 ($n$ = 5) | Group 3 ($n$ = 5) |
|---|---|---|---|
| Body weight (g) | 24.8 ± 0.47 | 23.5 ± 0.48 | 26.1 ± 0.8 |
| Kidney weight (g) | 0.30 ± 0.01 | 0.27 ± 0.01 | 0.33 ± 0.01 |
| Hematocrit (%) | 47 ± 1 | 39 ± 1[b] | 41 ± 1[b] |
| Plasma sodium concentration (mEq/L) | 149 ± 3 | 151 ± 2 | 144 ± 3 |
| Plasma potassium concentration (mEq/L) | 4.6 ± 0.8 | 3.4 ± 0.6 | 3.8 ± 0.8 |
| Plasma osmolality (mosmol/kg) | 292 ± 4 | 293 ± 3 | 290 ± 5 |

[a]Group 1 is the euvolemic group, group 2 is the volume-expanded group, and group 3 is the volume-expanded Ang II-infused group.
[b]$P < 0.05$ compared with group 1.

FIG. 2.1 – Exemple d'un tableau provenant d'un article en XHTML

Le corpus est composé d'environ 950 résultats expérimentaux (provenant des 20 articles). 95 résultats (issus de cinq articles) ont été utilisés pour le développement et 855 (issus des quinze autres articles) pour l'évaluation. Nous avons choisi cette répartition en considérant que la variété de 95 résultats expérimentaux était suffisante pour étudier le problème et concevoir notre système.

### 2.1.2 Annotation du corpus

Pour le développement et l'évaluation du système d'extraction d'information, il était nécessaire de disposer d'un corpus annoté en termes et relations. Pour cela, nous avons projeté les descripteurs des résultats expérimentaux contenus dans la base dans les 20 articles de notre corpus. Nous avons donc réalisé la tâche inverse à l'extraction : nous avons utilisé les données de QKDB qui avaient été extraites manuellement des articles scientifiques et nous les avons recherchées et annotées dans les textes. Une vérification et une complétion manuelle ont ensuite été faites.

Dans la figure 2.2 nous avons représenté la proportion de descripteurs par rapport au nombre de résultats expérimentaux, calculs effectués à partir de la base de données. On observe par exemple que pour tous les résultats expérimentaux, l'espèce et le paramètre sont renseignés ; en revanche le soluté l'est dans 60% des cas.

### 2.1.3 Structure des articles

Une étude des instructions données aux auteurs souhaitant soumettre un article dans les revues de physiologie, nous a permis d'avoir une vue d'ensemble sur la structure des articles. L'auteur doit suivre un plan bien défini, dans lequel le nom des parties est fixé [1]. La première partie d'un article contient le titre, le nom des auteurs, leurs institutions d'affiliations, etc. La seconde partie est le résumé où on peut trouver quelques informations

[1] Pour le journal AJP voir : http ://www.the-aps.org/publications/i4a/prep_manuscript.htm

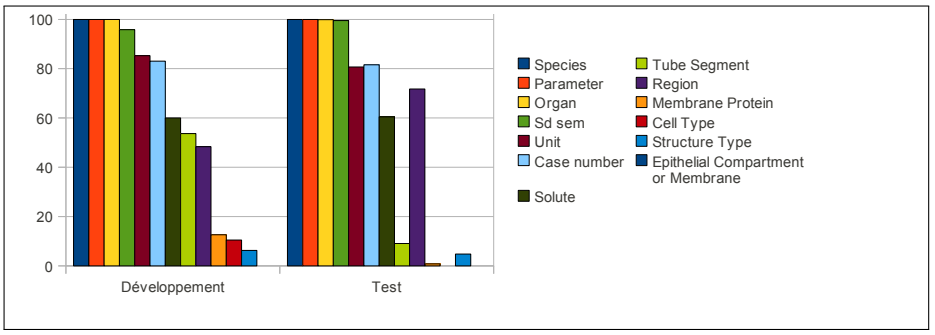



FIG. 2.2 – Proportion de descripteurs par rapport au nombre de résultats expérimentaux dans chaque corpus

générales sur les expériences réalisées. Il y a ensuite l'introduction dans laquelle l'auteur doit faire un état des recherches précédentes dans ce domaine et donner un aperçu de l'intérêt de son étude. À la suite de l'introduction, vient la partie *Materials and Methods*, qui peut s'appeler aussi *Experimental Procedures* ou *Methods*, dans laquelle est faite une description des techniques, des modéles, des produits chimiques utilisés, etc. Après quoi l'auteur peut présenter les données expérimentales et les résultats obtenus, dans la partie *Results*. Ensuite, dans la partie Discussion, l'auteur doit donner une interprétation des résultats qu'il a obtenus et les comparer aux résultats présentés dans d'autres publications. Cette partie contient ainsi des informations portant directement sur les expériences faites par l'auteur, mais également la synthèse de résultats obtenus par d'autres chercheurs faisant l'objet d'études similaires qu'il peut être également intéressant d'extraire. Les 4 dernières parties sont les remerciements, les subventions, les références et les notes de bas de page.

### 2.1.4 Étude du corpus

À partir du corpus annoté, nous avons réalisé une étude qui nous a amené à faire certains choix plutôt que d'autres lors du développement du système d'extraction. Nous présentons ci-dessus les conclusions de notre étude.

1. Localisation de l'information :
   94% des résultats quantitatifs sont donnés dans les parties *Results* et *Discussion*, les 6% restants sont des résultats provenant de figures et qui ne sont cités de façon textuelle que dans la section *Abstract*. Les informations contenues dans les champs de commentaires sont présentes majoritairement dans la section *Methods*.

   Les valeurs numériques des résultats sont souvent dans des tableaux. Dans le corpus de développement, sur 95 résultats, 33 sont dans des tableaux, ce qui représente 35% des résultats quantitatifs.
2. Silence :
   Le silence correspond aux concepts qui ne sont pas exprimés dans les articles de la même manière que dans la base.

Les différents types de variations que nous avons observés sont présentés dans le tableau 2.1. Une autre cause du silence lors de la projection, à laquelle nous ne nous intéresserons pas, provient d'erreurs dues à l'insertion manuelle des données par les experts (fautes de frappes).

| | **Exemple** | |
|---|---|---|
| **Type** | **Dans la base** | **Dans le texte** |
| **Dérivation/flexion** | urine | urinary |
| **Formule chimique** | $Na^+$ | sodium |
| **Variation d'écriture** | $0.000937 \pm 0.77$ | $9.37 \pm 0.77$ e-4 |
| **Abréviation** | permeability | $P_f$ |
| **Variation sémantique** | flow rate | excretion |

TAB. 2.1 – Types de variations observés

3. Ambiguïté :
   Nous parlons d'ambiguïté quand plusieurs termes correspondant à un même type de descripteur sont exprimés dans la même phrase ou dans le même article.
   L'ambiguïté pour le champ espèce est très faible : en effet dans 90% des expérimentations une seule espèce est citée dans l'article. Dans les autres cas, l'espèce est citée à proximité du résultat. Il en est de même pour le champ organe. Pour les autres descripteurs, l'ambiguïté, au sein d'une phrase peut être levée grâce à la ponctuation ou à des critères de proximité par rapport à la valeur numérique de l'expérience.
4. Position des descripteurs :
   90% des solutés sont dans la même phrase que la valeur numérique, ainsi que 65% des paramètres. Mais seulement 35% des commentaires sont dans la même phrase que le résultat. Extraire les descripteurs uniquement dans la phrase contenant la valeur numérique ne sera donc pas suffisant.

Après l'étude de l'expression des résultats expérimentaux dans le corpus, il nous a fallu chercher une modélisation de ces résultats.

## 2.2 Passage d'un résultat expérimental à une relation n-aire

La méthode que nous proposons pour extraire les résultats expérimentaux, s'appuie sur leur modélisation représentée par une base de données générique (QKDB). Cependant, le schéma relationnel de cette base de données n'explicite pas toutes les informations, aussi nous avons formalisé ce modèle sous forme d'une ontologie associée à une terminologie. Nous présenterons d'abord ce modèle, puis expliquerons sa représentation dans le schéma relationnel de QKDB.

### 2.2.1 La ressource termino-ontologique

Une ontologie est généralement composée d'une composante générique, représentant des concepts généraux indépendants du domaine, complétée par une ontologie du domaine,

plus éventuellement des ontologies décrivant la tâche et l'application (Guarino [1998]). Une ressource termino-ontologique (RTO) met en relation les concepts de l'ontologie avec leurs dénominations dans la langue, les termes.

Notre objectif est de définir un modèle générique pour représenter un résultat expérimental. Un résultat expérimental est défini par un résultat quantitatif et les différents descripteurs de l'expérimentation qui ont permis de l'obtenir, et peut donc être vu comme une relation n-aire. Les descripteurs forment les concepts du domaine, en l'occurrence ceux de la physiologie rénale.

Les recommandations du W3C (Noy et Rector [2006]) (voir section 1.4.1) pour représenter les relations n-aires amènent à représenter une relation par un concept, et à rattacher les éléments mis en relation par des propriétés. C'est par exemple le choix fait par Touhami *et al.* [2011], qui porte sur l'extraction de relations n-aires en microbiologie. En suivant ce type de modélisation, il faudrait créer un concept-relation par expérimentation lorsque le domaine change. Or, les descripteurs de l'expérimentation jouent tous le même rôle vis-à-vis de la relation. De ce fait, il est possible de représenter une relation par un concept générique, lié à un et un seul résultat quantitatif et à un seul type de concept représentant l'ensemble des descripteurs. Ce concept est ensuite précisé par les concepts du domaine.

La figure 2.3 illustre ce choix : un résultat expérimental est représenté par un concept-relation *ExperimentalResult*. Celui-ci est relié au concept *QuantitativeResult* qui correspond notamment à la valeur numérique du résultat, en précisant qu'il y a un et un seul concept possible, et au concept *ExperimentConcept* qui correspond aux descripteurs du résultat (espèce ou organe concernés par exemple), avec la restriction qu'il y a au moins une valeur de ce concept. Cette modélisation permet de décrire un résultat d'expérimentation, dans quelque domaine qu'il soit, et correspond donc à la partie générique de l'ontologie.

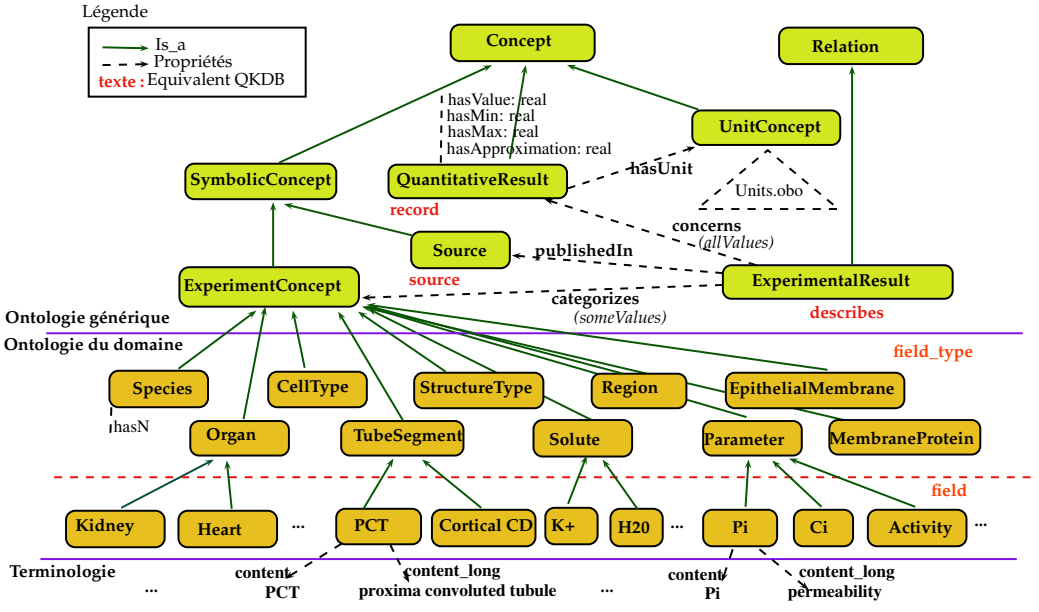



FIG. 2.3 – La ressource termino-ontologique

Un *résultat quantitatif* (*QuantitativeResult*) est quant à lui décrit par des propriétés permettant d'en donner sa valeur (*hasValue*), ses valeurs minimale et maximale (*hasMin* et *hasMax*), et l'approximation faite (*hasApproximation*).

Chaque résultat étant extrait d'un article scientifique, ce dernier est aussi décrit dans l'ontologie générique (concept *Source*). De la même manière, les résultats quantitatifs sont associés à une unité (concept *UnitConcept*), appartenant aussi à la partie générique.

Cette partie générique est complétée par une ontologie du domaine. En effet, un résultat quantitatif n'a de sens que si les données expérimentales associées sont précisées : espèce animale, organe, soluté...

QKDB est peuplée de données portant sur plusieurs niveaux depuis le niveau des canaux et transporteurs des membranes cellulaires, en passant par les niveaux de l'épithélium et du tubule rénal (le néphron), jusqu'au niveau du rein entier. Les données expérimentales particulièrement ciblées, pour l'humain et pour les principales espèces d'animaux expérimentaux, ainsi que pour les épithéliums "modèles" tels que cellules en culture, peau ou vessie d'amphibiens, etc., sont :

- les paramètres de transport (par exemple perméabilités, conductances, caractéristiques cinétiques des canaux ioniques et transporteurs membranaires...) ;
- les concentrations et débits des principaux solutés (tels que sodium, potassium, chlorure, bicarbonate, urée, glucose et lactate, les acides aminés...) mesurés le long du néphron et dans les vaisseaux et tissus du rein ;
- les mesures anatomiques (par exemple diamètres et longueurs des segments du néphron, épaisseurs des épithéliums, nombre et position des différents types de néphrons et vaisseaux, tailles des différentes régions du rein (qui diffèrent selon les espèces tels l'humain, le rat, la souris...) ;
- l'expression membranaire des protéines de transport et récepteurs membranaire dans les différents segments du néphron ;
- aspects du métabolisme importants pour la fonction du rein ;
- etc.

Ils se traduisent dans l'ontologie de domaine par des concepts à deux niveaux de hiérarchie. Le premier niveau correspond au type des descripteurs qui peuvent décrire un résultat expérimental (*Species*, *Organ*...), et le second niveau aux descripteurs eux-mêmes (*Kidney*, *Heart*...). Tous ces concepts ne possèdent pas de propriétés spécifiques, sauf le concept *Species*, auquel on associe le nombre d'individus, *hasN*, sur lesquels a été faite l'expérimentation.

La partie terminologique consiste enfin à associer à chaque concept feuille un terme préféré [2], et éventuellement des variantes, qui peuvent être des synonymes, des hyponymes, des abréviations, des acronymes ou des symboles. Les termes peuvent donc être des mots simples, des abréviations et acronymes et sont aussi souvent formés de plusieurs mots que nous appelons termes complexes.

[2] Un terme préféré est un terme choisi par les experts du domaine lors de la construction de la terminologie pour désigner le concept.

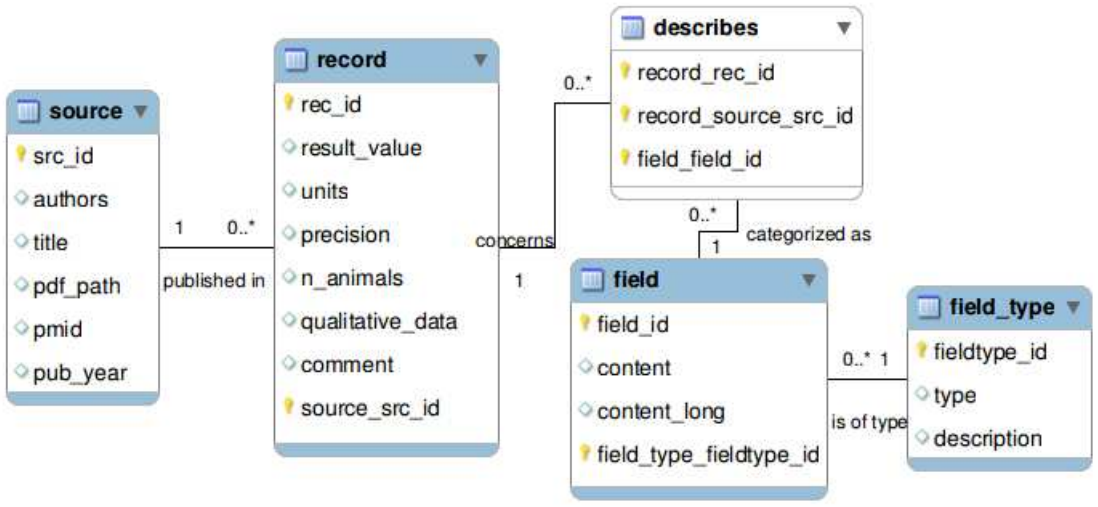



FIG. 2.4 – Schéma UML simplifié de la base de données QKDB

### 2.2.2 Représentation de l'ontologie par la base de données

Le modèle que nous venons de décrire est représenté par la base de données qui permet de stocker les instances trouvées dans les articles.

Ainsi qu'il a déjà été dit, le schéma de la base de données a été conçu pour faciliter la comparaison de données mesurées sur différentes espèces et dans des conditions expérimentales variées, mais aussi pour être facilement extensible et généralisable à d'autres domaines.

La figure 2.4 présente un schéma partiel de la base (voir la figure 2.5 pour le schéma complet). La formalisation faite sous forme d'ontologie se retrouve bien dans la base de données. Les concepts génériques se traduisent par des tables : les concepts *Experimental-Result* et *Source* sont représentés par les tables *record*, et *source*. La table *record* contient les attributs suivants :

- la valeur numérique du résultat (*result_ value*) ;
- les unités du résultat, qui qualifient la valeur numérique (*units*) ;
- une précision, qui indique généralement l'erreur standard de la mesure (*precision*) ;
- le nombre d'animaux observés (*n_ animals*) ;
- des données qui décrivent qualitativement le résultat (*qualitative_ data*) ;
- un commentaire, qui donne des informations complémentaires sur les techniques expérimentales (*comment*).

On peut noter que QKDB ne stocke pas les unités possibles sous forme de table, l'unité étant un simple attribut de la table *record*. Notons aussi que le nombre d'individus sur lesquels l'expérimentation est faite, est représenté au sein de la table record. Ce choix est cohérent avec le fait qu'il n'y a qu'un résultat quantitatif et qu'une espèce par expérimentation.

Les concepts du domaine sont représentés par deux tables : *field* et *field_ type*. *field* contient les nœuds feuille de l'ontologie, les termes correspondant étant stockés dans les

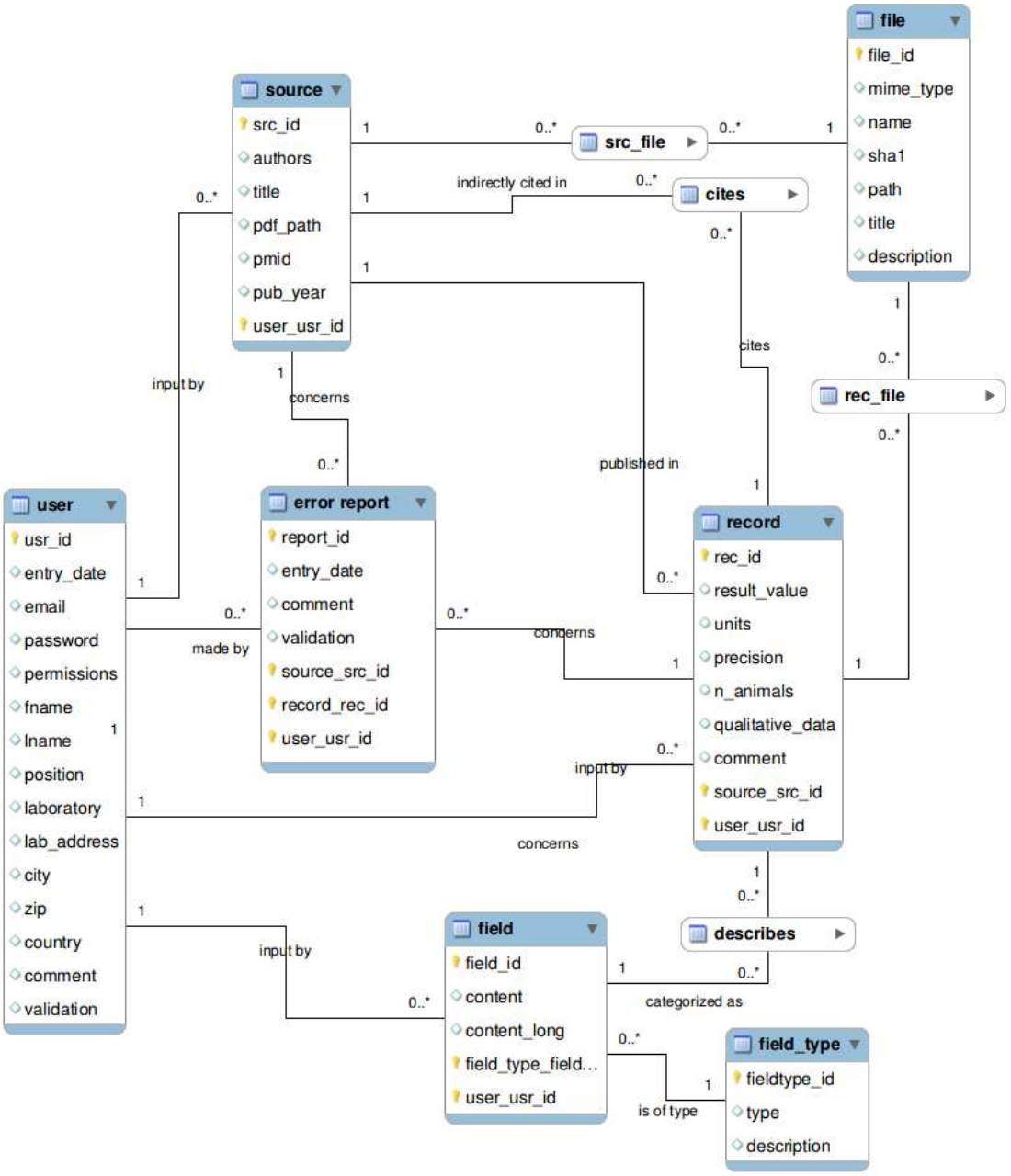



FIG. 2.5 – Schéma UML complet de la base de données QKDB

champs *content* (variante préférentielle) et *content_long* (liste de termes constituant des variantes). Chaque *field* est lié au nœud père correspondant de l'ontologie, représenté dans une table *field_type*. Ainsi, le concept « Pi » de l'ontologie correspond à une entrée de la table *field* dont l'attribut *content* est « Pi », et l'attribut *content_long* est « permeability »; cette entrée est reliée par une clé étrangère à une entrée de la table *field_type*, dont l'attribut *type* est « parameter », ce qui traduit le fait que la perméabilité est un descripteur de type paramètre. Les descripteurs généralement utilisés pour un résultat expérimental en physiologie rénale sont les suivants (correspondant donc à des entrées de la table *field_type*) :

- l'espèce sur laquelle l'expérience a été menée ;
- l'organe, la région, le segment, le compartiment et éventuellement le type de cellule, qui représentent les endroits sur lesquels l'expérience a été menée ;
- le type de paramètre, qui indique la propriété qui a été mesurée, comme le poids, la perméabilité, le diamètre ou la concentration ;
- le soluté, qui précise ce qui a été mesuré, par exemple $K^+$ si la concentration mesurée concerne ce soluté.

Seuls les deux niveaux les plus bas de l'ontologie ont été jugés nécessaires à traduire dans la base de données, mais cette structure pourrait être étendue en ajoutant un lien récursif sous la forme d'une clé étrangère dans la table *field_type*.

La relation n-aire, quant à elle, est représentée par la table *describes* qui stocke chaque occurrence de la relation trouvée dans un article par l'ensemble des couples qui lient l'occurrence de résultat (*record*) avec chaque descripteur trouvé. Ce sont ces tables (*describes* et *record*) qui seront complétées lors du processus d'extraction.

### 2.2.3 Exemple de représentation d'un résultat expérimental dans la base de données

Voici un exemple de phrase contenant un résultat expérimental stocké dans QKDB : « In controls versus PAN **rats**, **Na+/K+-ATPase activities** were (**pmol ATP/mm/h**): **proximal convoluted tubule**, **2954+-369** vs 2769+-230; thick ascending limb, 5352+-711 vs 5239+-803; and cortical collecting duct, 363+-96 vs 848+-194 (P<0.01), respectively. ».

Les informations stockées dans QKDB concernant le premier résultat cité dans la phrase (en gras) sont représentées dans les tableaux 2.2 et 2.3. On observe dans cet exemple que certains descripteurs ne sont pas dans la même phrase que la valeur numérique, comme l'organe ou le commentaire qui donne une information supplémentaire sur l'espèce étudiée. D'autres descripteurs ne sont pas exactement sous la même forme dans la base et dans le texte, par exemple le paramètre « activity » est au singulier dans la base mais au pluriel dans le texte.

Cette variation d'expression des résultats expérimentaux dans le corpus et leurs représentations dans la base de données nous a guidé dans la mise au point de la méthode d'extraction des relations n-aires.

| **table** | record | | | |
|---|---|---|---|---|
| **attribut** | result_value | precision | units | comment |
| **valeur** | 2954 | 369 | pmol ATP/mm/h | Male wistar rats (100-130g) |

TAB. 2.2 – Informations liées à un résultat expérimental dans QKDB (table *record*)

| **table** | field_type | field |
|---|---|---|
| **attribut** | type (type de descripteur) | content (descripteur) |
| **valeurs** | espèce | rat |
| | organe | kidney |
| | tube | PCT |
| | paramètre | activity |
| | protéine membranaire | Na-K-ATPase |

TAB. 2.3 – Informations liées à un résultat expérimental dans QKDB (tables *field_ type* et *field*)

## 2.3 Extraire un résultat expérimental

À partir du modèle et de l'analyse du corpus, nous avons développé un système d'extraction d'information pour extraire les résultats expérimentaux présentés dans les articles.

Comme on l'a déjà vu, l'extraction de relations n-aires implicites dans le texte nécessite l'identification des arguments, dans notre cas la valeur numérique, les attributs du résultat et les descripteurs de l'expérience, puis la mise en relation de ces arguments pour former la relation. Notre méthode s'articule autour de ces deux points : détection de la valeur numérique, de ses attributs puis des descripteurs, et ensuite mise en relation des arguments.

### 2.3.1 Méthode

Rappelons que nous avons défini un *résultat expérimental* comme étant une relation n-aire entre :

- un *résultat quantitatif* composé notamment d'une *valeur numérique*, associée à certains *attributs* comme l'unité, la précision...
- et des *descripteurs*, qui précisent les conditions expérimentales, comme l'espèce ou l'organe considéré.

Nous proposons une méthode en trois étapes pour la reconnaissance de ces résultats :

1. détection d'une valeur numérique, qui joue le rôle de déclencheur et constituera l'élément pivot du résultat. En effet, nous considérons qu'une expérience n'est pertinente que si un résultat quantitatif est fourni. La reconnaissance de la valeur numérique s'apparente à une reconnaissance d'entité numérique. Les autres attributs du résultat quantitatif comme le nombre d'animaux étudiés doivent aussi être reconnus, ce qui peut se faire assez simplement avec des expressions régulières;
2. reconnaissance de descripteurs d'expérience. Il s'agit alors d'une reconnaissance terminologique puisqu'il faut faire le lien entre les concepts de la base de données et les

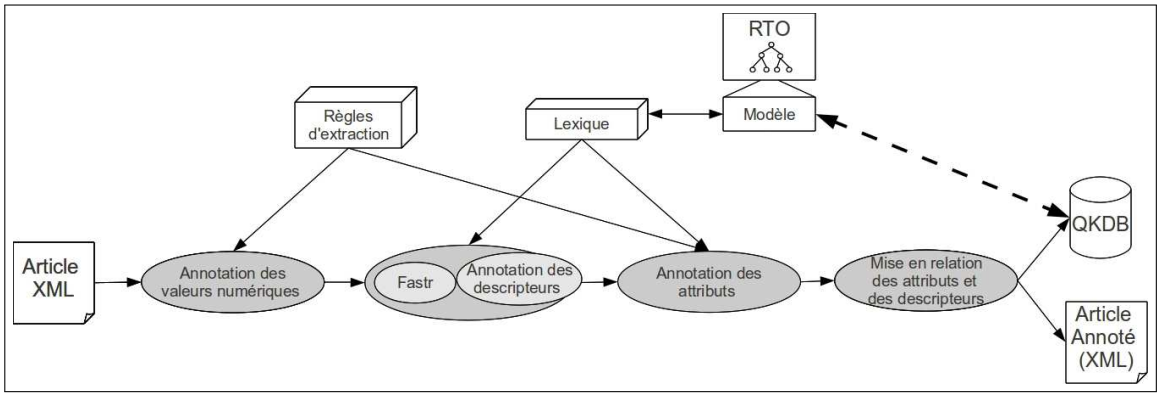



FIG. 2.6 – Schéma de fonctionnement du système d'extraction

termes de l'article. Pour cela, nous utilisons les lexiques de notre ressource termino-ontologique et identifions les variantes de ces termes (flexionnelles, dérivationnelles, etc.);

3. mise en relation des attributs ou descripteurs et de la valeur numérique. Cette mise en relation prend en compte la distance entre la valeur numérique du résultat et les attributs ou descripteurs, ainsi que des critères de fréquence.

Après avoir présenté l'architecture générale de notre système, nous présenterons le lexique puis chacune des trois étapes du système.

### 2.3.2 Architecture

Le schéma de la figure 2.6 présente les différents modules de notre système d'extraction. L'article XML est d'abord étiqueté par le TreeTagger Schmid [1994], ce qui permet de disposer des formes canoniques des mots (c'est-à-dire les lemmes) et de leur catégorie, ainsi que de procéder à un découpage en phrases selon l'étiquetage des signes de ponctuation. L'article étiqueté est ensuite fourni au module d'annotation des valeurs numériques, puis des descripteurs (qui utilise Fastr Jacquemin [1996] pour détecter les variantes des termes du lexique). Le dernier module effectue la mise en relation des résultats numériques avec leurs descripteurs. Les tables *record* et *describes* de la base de données sont ensuite complétées avec les résultats expérimentaux extraits.

### 2.3.3 Lexique

Pour extraire les descripteurs, il était nécessaire d'avoir un lexique des termes du domaine. Nous avons construit un lexique de base avec les termes associés aux concepts de la base de données, ainsi que leurs variantes, dont on donne le détail dans le tableau 2.4. L'identifiant associé à chaque concept dans la base est conservé dans le lexique pour relier les termes au concept désigné. Les entrées du lexique ont la forme suivante : « PAR_16 concentration ». *PAR* donne le type du concept, en l'occurrence paramètre, « 16 » est l'identifiant du concept de la base, et « concentration » est le terme relié à ce concept. Si plusieurs termes sont reliés au même concept, il y aura plusieurs entrées avec le même

type et le même identifiant, par exemple : « TUS_57 cortical CD » et « TUS_57 cortical collecting duct ».

| | Lexique de base | | Lexique enrichi |
|---|---|---|---|
| **Descripteur** | **# concepts** | **# termes** | **# termes** |
| Species | 24 | 25 (1,0) | 68 (2,8) |
| Organ | 12 | 15 (1,2) | 22 (1,8) |
| Structure type | 7 | 10 (1,4) | 10 (1,4) |
| Tube segment | 46 | 82 (1,8) | 94 (2,0) |
| Epithelial compartment or membrane | 14 | 28 (2,0) | 30 (2,1) |
| Cell type | 16 | 24 (1,5) | 24 (1,5) |
| Region | 30 | 47 (1,6) | 81 (2,7) |
| Parameter | 85 | 153 (1,8) | 228 (2,7) |
| Solute | 66 | 92 (1,4) | 119 (1,8) |
| Membrane Protein | 11 | 22 (2,0) | 22 (2,0) |

TAB. 2.4 – Contenu du lexique : nombre de termes associés à chaque type de concept, et nombre moyen de termes par concept, dans le lexique de base et le lexique enrichi. Entre parenthèses on indique la proportion moyenne de termes reliés à un concept.

Pour enrichir le lexique de base, nous avons ajouté des termes provenant de ressources externes. Nous avons utilisé des listes provenant de sites spécialisés que nous avons sélectionnés pour l'espèce et le soluté [3]. Nous avons également voulu ajouter des données provenant d'une ressource du domaine médical : l'UMLS (Unified Medical Language System) (Lindberg *et al.* [1993]). C'est un métathésaurus très riche, développé par la NLM (National Library of Medicine). Il rassemble plusieurs thésaurus en créant des relations entre les concepts. Il est très complet mais aussi très complexe. Pour évaluer la couverture de l'UMLS dans le domaine de la physiologie rénale, nous avons utilisé l'outil MetaMap de Aronson [2001] qui annote les concepts de l'UMLS dans des textes biomédicaux. Sur le corpus de développement, nous avons observé un grand nombre de termes annotés non pertinents, et ceux annotés par notre système ne l'étaient pas par MetaMap : seulement 6,5% des termes repérés par MetaMap étant dans notre lexique, le bruit est donc très important. Par exemple, dans la phrase « we performed a separate series of experiments in Nhe3+/+ and Nhe3-/- mice to directly evaluate $P^{sf}$ during loop of Henle perfusion », MetaMap annote « directly » comme un concept qualitatif, qui n'est pas pertinent pour notre tâche, et il n'annote pas le descripteur « loop of Henle » de type « tube segment ». Aussi, nous n'avons pas retenu cette ressource.

Pour extraire les unités nous avions besoin d'une liste des unités de base. Nous avons pour cela utilisé l'ontologie units.obo (dans le format de Gene Ontology), dans laquelle nous avons extrait les unités de base (*cm*, *mol*, etc.) et les unités composées (*ml/kg*, etc.). Nous avons ajouté manuellement quelques préfixes, comme par exemple *k* pour kilo, ce qui nous permet d'extraire des unités complexes qui étaient absentes de notre lexique.

Le lexique de base et la liste d'unités n'étaient pas assez complets pour permettre une bonne identification des descripteurs, et la complétion manuelle des listes de termes à partir de terminologies ou d'ontologies existantes ne permettait pas de disposer de tout le

[3] Exemple des sites utilisés : www.kterre.org/dossiers/atomes_liste.php ou en.wikipedia.org/wiki/Dictionary

vocabulaire nécessaire. Aussi avons nous appliqué une méthode d'acquisition de termes à partir de corpus (travail publié dans Minard *et al.* [2010]). ce qui nous a permis d'augmenter notre lexique de 5%. Le tableau 2.4 indique la taille du nouveau lexique après l'ajout de termes grâce à des ressources externes et à l'acquisition de nouveaux termes.

#### 2.3.3.1 Extraction de nouveaux termes

L'enrichissement manuel du lexique de base n'était pas assez complet pour permettre une bonne identification des descripteurs. Aussi avons nous appliqué une méthode d'acquisition de termes à partir de corpus qui exploite la connaissance du domaine dont nous disposons.

De nombreuses approches ont exploité l'existence de relations entre termes afin d'acquérir des patrons d'extraction de ces relations, et pouvoir les appliquer ensuite pour collecter de nouveaux couples de termes reliés par le même type de relation, ou collecter de nouveaux termes en fixant un terme et la relation (voir Auger et Barrière [2008] pour une synthèse).

Lin et Pantel [2001] ont travaillé sur l'extraction automatique de règles d'inférences, c'est-à-dire de phrases qui n'ont pas exactement la même signification, mais qui sont liées par une relation sémantique. Ils cherchent à découvrir ces règles d'inférences depuis un corpus. Ils basent leur algorithme sur l'hypothèse distributionnelle d'Harris, qui pose que des mots employés dans des contextes proches tendent à être similaires, d'un point de vue de leur distribution syntaxique et de leur sens. Ils appliquent cette hypothèse à des patrons construits à partir d'arbres de dépendances. Ils posent l'hypothèse que si deux patrons tendent à relier la même série de mots, alors leur sens est similaire. Ils construisent ainsi des patrons à partir de relations de dépendances. Ils sont de la forme : « X finds solution Y », où X et Y sont des champs qui sont obligatoirement des noms. Ils excluent les relations qui ne sont pas entre des mots des catégories suivantes : noms, adjectifs, verbes ou adverbes.

Embareck [2008] a travaillé sur l'extraction de patrons linguistique caractérisant une relation dans le domaine médical. La première étape du système qu'il a développé consiste en l'application des règles de reconnaissance d'entités médicales sur des textes médicaux. Les règles utilisées sont composées d'un déclencheur, du contexte précédent, du contexte suivant, et du type d'entités identifiées. Il extrait ensuite toutes les phrases contenant deux entités de la relation cible (par exemple une maladie et un traitement), et supprime à la main celles qui ne correspondent pas à la relation cible. Dans les phrases, les entités médicales sont remplacées par leur type. Et pour finir, il utilise un algorithme d'extraction de patrons multi-niveaux qui appliquent les patrons pour chaque couple de phrases.

Ces approches ascendantes à base de patrons ont été appliquées essentiellement sur les relations d'hyperonymies/hyponymies, entre mots en domaine général, en initialisant la recherche par des couples de mots qui partagent la relation que l'on veut modéliser, et généralement sur de grands corpus.

Nous utilisons une méthode similaire à Embareck [2008] pour acquérir des nouveaux termes et ainsi enrichir notre lexique. Le corpus annoté est utilisé pour extraire des patrons entre deux descripteurs de types différents. Par exemple entre une valeur numérique et un paramètre, l'expression suivante est extraite : `PAR of the anesthetized ESP was RES`, où PAR est le paramètre, RES la valeur numérique et ESP une espèce.

Les patrons ainsi extraits sont trop spécifiques pour être utiles, il faut donc les généraliser. Pour cela, nous avons utilisé l'algorithme décrit dans Pantel *et al.* [2004], qui réalise un alignement d'expressions deux à deux pour généraliser les patrons. Pantel *et al.* [2004] utilisent cet algorithme pour extraire des relations *is-a* entre un terme et son hyperonyme. À la différence de Pantel *et al.* [2004], nous utilisons l'algorithme pour extraire des relations non connues et contextuelles (c'est-à-dire que les relations existent dans le cadre d'une expérimentation) entre deux classes sémantiques.

L'algorithme est composé de deux parties, la première consiste à calculer la distance d'édition minimale (suppression, insertion ou remplacement) pour passer d'une expression à une autre (cf. Algorithme pour le calcul de la distance d'édition minimale). La seconde partie produit un alignement optimal de deux expressions, le résultat est un patron (cf. Algorithme pour la recherche optimale de patrons).

**Algorithme pour le calcul de la distance d'édition minimale**

```
D[0,0]=0
for i = 1 to n do D[i,0] = D[i-1,0] + cost(insertion)
for j = 1 to m do D[0,j] = D[0,j-1] + cost(deletion)
for i = 1 to n do
        for j = 1 to m do
                D[i,j] = min( D[i-1,j-1] + cost(substitution),
                              D[i-1,j] + cost(insertion),
                              D[i,j-1] + cost(deletion))
Print (D[n,m])
```

**Algorithme pour la recherche optimale de patrons**

```
i = n, j = m;
while i ≠ 0 and j ≠ 0
        if D[i,j] = D[i-1,j] + cost(insertion)
                print (*s*), i = i-1
        else if D[i,j] = D[i,j-1] + cost(deletion)
                print(*s*), j = j-1
        else if a_1i = b_1j
                print (a_1i), i = i -1, j = j =1
        else if a_2i = b_2j
                print (a_2i), i = i -1, j = j =1
        else
                print (*g*), i = i -1, j = j =1
```

Les règles suivantes sont appliquées :

- si les lemmes sont identiques, on les conserve dans le patron,
- sinon, si les catégories morpho-syntaxiques sont identiques, elles sont conservées,
- sinon, les éléments sont remplacés par *g*,
- si un élément a été inséré, il est remplacé par *s*.

Un exemple de l'alignement optimal entre deux expressions est présenté dans le tableau 2.5. Nous ne conservons que les patrons qui contiennent au plus deux *g* ou *s*,

comme Embareck [2008], pour éliminer les patrons trop génériques et généralement non pertinents.

| Expressions | Patron résultant |
|---|---|
| PAR of the -/- ESP (RES<br>PAR of the anesthetized ESP was RES | PAR of the *g* ESP *s* RES |

TAB. 2.5 – Exemple de l'alignement optimal entre deux expressions.

De cette façon, nous avons extrait des patrons entre les couples de descripteurs suivants : Paramètre et Résultat, Soluté et Résultat, et Région et Résultat. Nous avons obtenu respectivement 39, 18 et 16 patrons. D'autres couples de descripteurs ont été étudiés mais pas utilisés, car certains ne sont pas pertinents (par exemple entre Espèce et Soluté), pour d'autres nous n'avions pas assez d'exemples pour apprendre (par exemple entre la Structure Type et la Région), et parfois la relation est de type « complément du nom » (par exemple entre le paramètre « excretion » et le soluté « Na+ » dans la phrase « Na+ excretion was significantly increased in [...] ») et l'apprentissage de patrons est donc inutile.

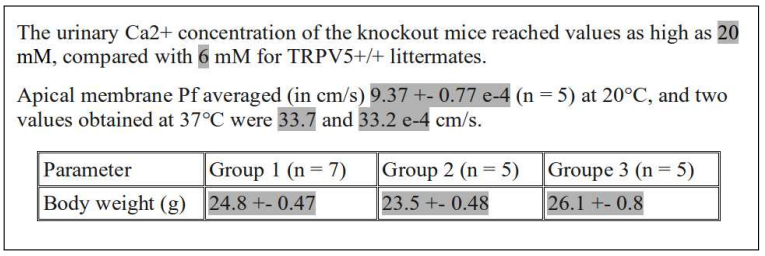
Les patrons ont été appliqués sur un corpus de 20 articles qui ne sont pas dans la base de données QKDB (173 phrases contenant des résultats numériques), dans lesquels les résultats numériques avaient été annotés avec des patrons (voir section 2.3.4.1). Les termes extraits ont été validés manuellement. 1866 termes ont été typés soluté, paramètre ou région, sachant que plusieurs patrons peuvent être appliqués sur la même phrase. 253 termes ont été correctement typés (13,5%), soit 38 termes distincts, parmi lesquels 196 sont des nouveaux termes (77,5%), soit 19 termes distincts. Le lexique est donc augmenté de ces 19 termes ce qui représente 5% d'augmentation (cf. table 2.4). Ce nombre peut sembler faible, mais les termes que nous cherchons ne donnent pas lieu à un très grand nombre de synonymes ou hyponymes chacun.

### 2.3.4 Reconnaissance des résultats quantitatifs

Les valeurs numériques des expériences sont le point de départ de la reconnaissance des expérimentations. Il est nécessaire de pouvoir tous les reconnaître, sans pour autant extraire toutes les valeurs numériques présentes dans l'article. La figure 2.7 présente différents types de résultats quantitatifs d'expériences. L'étude du corpus de développement nous a montré que 94% des résultats quantitatifs sont donnés dans les parties *Results* et *Discussion*, les 6% restants étant des résultats provenant de figures et qui ne sont cités de façon textuelle que dans la section *Abstract*. Les informations contenues dans les champs de commentaires sont présentes majoritairement dans la section *Methods* (nous ne nous intéressons pas à l'extraction de ces champs). Nous pouvons donc limiter les parties de l'article à analyser pour repérer les valeurs numériques aux sections *Results* et *Discussion*. Pour les identifier, nous avons testé une méthode à base de règles et une méthode à base d'apprentissage.

#### 2.3.4.1 Extraction à base de règles

La première méthode est fondée sur des expressions régulières permettant de repérer dans le texte les valeurs numériques qui peuvent correspondre à des résultats d'expériences.



The urinary Ca2+ concentration of the knockout mice reached values as high as 20 mM, compared with 6 mM for TRPV5+/+ littermates.

Apical membrane Pf averaged (in cm/s) 9.37 +- 0.77 e-4 (n = 5) at 20°C, and two values obtained at 37°C were 33.7 and 33.2 e-4 cm/s.

| Parameter | Group 1 (n = 7) | Group 2 (n = 5) | Groupe 3 (n = 5) |
|---|---|---|---|
| Body weight (g) | 24.8 +- 0.47 | 23.5 +- 0.48 | 26.1 +- 0.8 |

FIG. 2.7 – Exemples de résultats quantitatifs à extraire

Les patrons utilisés sont les suivants :

P1 : {nombre}%? +- {nombre}%?( e<sup>{nombre}</sup>)?
P2 : {nombre}%? ( e<sup>{nombre}</sup>)?

*{nombre}* décrit un nombre qui peut être décimal, négatif, ou contenant un séparateur de milliers. *e<sup>{nombre}</sup>* indique que le nombre peut être suivi d'une puissance. Le premier patron permet de reconnaître les valeurs qui sont suivies d'une précision, et le deuxième annote toutes les autres valeurs numériques. Ce dernier a un rappel très élevé mais une précision très basse : en effet il repère les numéros des figures, les renvois bibliographiques, etc. Toutes les valeurs n'étant pas repérées avec le premier patron, il est nécessaire d'appliquer un patron peu sélectif pour ne pas omettre un résultat potentiel à ce stade de l'extraction. Pour améliorer la précision, nous appliquons ensuite un filtre qui permet de supprimer les valeurs numériques entre parenthèses, celles qui suivent les mots *figure* et *table*, et les numérotations de références. Après cette dernière étape, la précision reste encore faible; une dernière sélection des valeurs numériques aura lieu lors de la mise en relation des descripteurs.

#### 2.3.4.2 Reconnaissance par apprentissage

Nous avons également testé une méthode à base d'apprentissage considérant la reconnaissance des valeurs numériques des résultats expérimentaux comme un problème de classification binaire. Nous avons utilisé libSVM (Chang et Lin [2001]), qui est un classifieur à base de SVM. Nous avons repéré toutes les unités numériques de l'article avec un patron simple : {nombre}%?( +- {nombre}%?)?, que nous avons appliqué dans les tableaux et dans le texte. Pour chaque valeur numérique annotée, nous avons construit un vecteur d'attributs, que nous avons fourni en entrée du classifieur. Dans cette méthode, les attributs et les descripteurs sont annotés avant l'extraction des valeurs numériques. Les attributs suivants sont utilisés :

- le caractère décimal du nombre,
- le nombre de chiffres de la partie entière de la valeur numérique (s'il n'y a qu'un chiffre et que ce n'est pas un nombre décimal, il est probable que ce soit un numéro de figure et non pas un résultat),

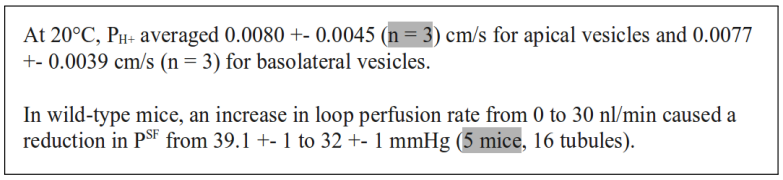



FIG. 2.8 – Exemples du nombre d'animaux étudiés à extraire

- l'inclusion de la valeur dans un tableau,
- le lemme et la catégorie morpho-syntaxique des cinq mots avant et des cinq mots après la valeur numérique (le choix de la fenêtre a été faite après plusieurs tests),
- la classe de VerbNet [4] (Kipper *et al.* [2008]) des verbes dans les cinq mots avant et après (les verbes sont importants car ils introduisent souvent les résultats, comme par exemple « average »),
- les unités dans les cinq mots avant et après (très peu de valeurs numériques de résultats sont indiquées sans unités),
- la distance entre la valeur numérique et les unités les plus proches,
- le type des descripteurs de la phrase (*Parameter*, *Solute*, etc.),
- le titre de la partie où est située la valeur, comme *Results*, *Discussion*, etc. (aucune valeur numérique de résultat ne sera donnée dans la partie *Methods* par exemple).

#### 2.3.4.3 Reconnaissance des autres attributs du résultat quantitatif

Les autres attributs des résultats quantitatifs sont repérés par des patrons. Le **nombre d'animaux** étudiés est un nombre entier qui peut être précédé de $n =$ ($1^{\text{er}}$ exemple de la figure 2.8), ou suivi du nom de l'espèce concernée ou de « males » ou « females » ($2^{\text{e}}$ exemple de la figure 2.8).

La **précision** est annotée avec le même patron que la valeur numérique associée : {nombre}%? +- {nombre}%?( e<sup>{nombre}</sup>)?.

Pour annoter les **unités**, nous utilisons un patron qui repère les chaînes de caractères composées de la combinaison d'unités de base (comme *g* ou *mol*), de préfixes (comme *k* ou $\mu$), de suffixes ($^{-1}$) et de symboles de séparation (. ou /). Nous repérons des unités comme *cm/s*, *mmHg*, *pmol ATP/mm/h* ou encore *µmol/mg creatinine*. Pour repérer cette dernière unité, nous acceptons la présence d'un soluté juste après l'unité ou entre les composants de l'unité.

### 2.3.5 Reconnaissance des descripteurs

La reconnaissance des descripteurs se heurte au problème de grande variabilité de ces termes. Les différents types de variations que nous avons observés dans le corpus de développement sont présentés dans le tableau 2.1 (page 64). Les variations les plus évidentes

[4] VerbNet est un lexique de verbes en anglais regroupés en classe (une extension des classes de Levin) http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

à détecter sont les variantes flexionnelles, comme entre « activity » et « activities ». Les variantes d'écriture nécessitent souvent l'écriture de règles, par exemple pour détecter la variante « $Ca^{++}$ » du terme « $Ca^{2+}$ ». Pour les variations que nous avons typées formule chimique, abréviation et sémantique, il est nécessaire de disposer de lexiques spécifiques.

#### 2.3.5.1 Reconnaissance de variantes avec Fastr

Pour trouver les variantes des termes, nous avons utilisé Fastr (Jacquemin [1996]). Fastr est un analyseur de surface basé sur des métarègles, qui permet de détecter des variantes morphologiques, syntaxiques ou sémantiques. Par exemple, la métarègle suivante permet de détecter des variantes syntaxiques de type insertion :
Metarule Ins( X1 -> X2 N3 ) = X1 -> X2 < N PREP ART? A? > N3:

<X2 num> ! plu

<X1 metaLabel> = 'XX'.
avec X1 le terme de départ, N pour un nom, PREP pour une préposition, ART pour un déterminant et A pour un adjectif. L'expression entre parenthèses est le modèle du terme sur lequel est appliqué la variation décrite après le signe =. Dans la métarègle donnée en exemple, si le terme de départ contient deux mots et que le deuxième est un nom, la règle autorise l'insertion d'un nom, d'une préposition, d'un déterminant et d'un adjectif entre les deux mots du terme. Par exemple, à partir de « inner medulla », la règle détecte « inner stripe of the outer medulla ». Étant donné un texte et une liste de termes multi-mots, les données sont analysées par le TreeTagger, puis Fastr compile les métarègles et les instancie en fonction des termes données en entrée, et il applique ces règles sur le texte étiqueté. Deux ressources de la langue générale sont utilisées : la base CELEX [5] pour détecter des variantes morphologiques et WordNet (Fellbaum [1998]) pour les synonymes.

Nous avons repris les règles fournies avec Fastr pour détecter les variantes dans notre corpus. Nous avons fait une étude des variantes trouvées par Fastr dans les articles du corpus. 3000 occurrences à partir de 194 termes multi-mots ont été trouvées, dont 520 variantes flexionnelles (16,7%), 337 variantes syntaxiques (10,8%) et 84 variantes morpho-syntaxiques (2,7%) [6]. Par exemple, « glomerular cells and capillaries » est reconnu pour le tube segment « glomerular capillary » et avec la règle NameToVerb, « fraction of filtered » est repéré pour le paramètre « filtration fraction ».

### 2.3.6 Mise en relation des informations extraites

Une fois les résultats quantitatifs et les descripteurs annotés, il est nécessaire de mettre en relation une valeur numérique avec ses attributs et les descripteurs du résultat expérimental. Pour cela, il faut sélectionner parmi tous les descripteurs annotés dans le texte ceux qui sont associés à la même expérimentation que la valeur numérique. Pour extraire des relations n-aires, il est possible d'utiliser les relations syntaxiques ou des patrons pour relier les attributs et descripteurs à la valeur numérique associée. Mais comme nous l'avons vu précédemment la relation n-aire représentant un résultat expérimental est implicite dans

[5] www.ldc.upenn.edu/readme_files/celex.readme.html
[6] 2059 occurrences ne sont pas des variantes mais le terme repéré tel quel.

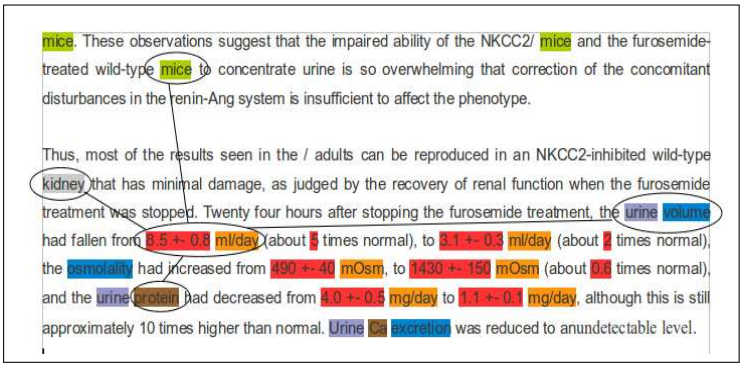



FIG. 2.9 – Exemple de la mise en relation des descripteurs

le texte et il donc difficile d'extraire une (ou des) structures syntaxiques récurrentes. De ce fait, utiliser un critère de proximité est ce qui nous semblait le plus approprié.

Les difficultés de la mise en relation vient de la présence d'ambiguïtés et du fait que les descripteurs ne sont pas toujours à proximité de la valeur numérique du résultat dans l'article (voir l'étude du corpus 2.1.4). Nous ne pouvons pas extraire les descripteurs uniquement dans la phrase de la valeur numérique. Nous avons par conséquent mis en place des heuristiques de sélection des valeurs numériques, ainsi que des attributs et descripteurs. Le principe est d'utiliser la valeur numérique comme élément pivot, et d'associer les attributs ou descripteurs en fonction de leur proximité avec cette valeur dans l'article, ou de leur fréquence dans l'article complet.

Les valeurs numériques sont ainsi sélectionnées si au moins l'une de ces conditions est vérifiée :

- la valeur numérique est dans un tableau ;
- elle est suivie d'une précision ;
- une unité est à proximité.

Pour chaque valeur numérique sélectionnée, les attributs et descripteurs sont sélectionnés en fonction des critères suivants :

- pour l'espèce et l'organe : sélection du terme le plus proche de la valeur numérique (dans la phrase) ou du plus fréquent (dans l'article) s'ils ne sont pas mentionnés dans la phrase ;
- si la valeur numérique est dans un tableau : sélection du descripteur dans l'entête de la colonne ou de la ligne ou dans la légende ;
- dans tous les autres cas :
  - sélection de l'attribut ou du descripteur le plus proche dans la phrase ou dans la proposition dans le cas des énumérations (voir l'exemple dans la section 2.2.3) ;
  - s'il n'y en a pas sélection du plus proche dans les phrases précédentes dans la limite du paragraphe.

La figure 2.9 présente un exemple de mise en relation des attributs et descripteurs avec la valeur numérique « 8.5 +- 0.5 » (tous les descripteurs repérés sont surlignés dans l'exemple) et la figure 2.10 présente la mise en relation des attributs et descripteurs dans un tableau. Dans la figure 2.9, on observe que deux paramètres sont dans la même phrase

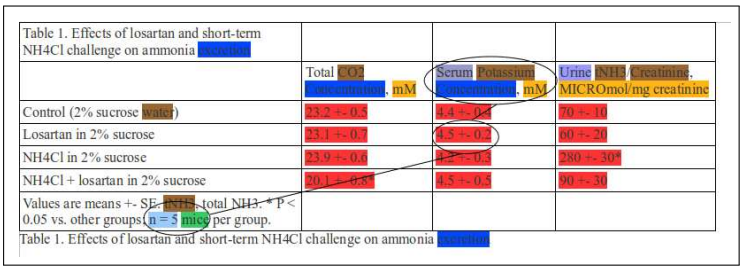



FIG. 2.10 – Exemple de la mise en relation des descripteurs dans un tableau

que la valeur numérique : « volume » et « osmolality », le plus proche est « volume ». On remarque également que le nom de l'organe n'est pas donné dans la phrase du résultat mais dans la précédente. Dans la figure 2.10, les descripteurs sélectionnés sont ceux qui sont dans le titre de la colonne de la valeur numérique ainsi que ceux qui sont dans la légende. On observe que seulement dans le titre de la première ligne un descripteur a été annoté, mais la valeur numérique est dans la deuxième ligne, ce descripteur n'est donc pas sélectionné.

Pour les résultats numériques qui sont dans des tableaux, la mise en relation des descripteurs est faite en utilisant la structure du tableau. Le numéro de la ligne et de la colonne du résultat numérique sont utilisés pour relier le résultat avec les descripteurs de la relation. Touhami *et al.* [2011] ont travaillé sur l'extraction de relations n-aires dans des tableaux et ils ont choisi de calculer des scores prenant en compte la similarité entre les termes dénotant le relation, le titre du tableau et le nombre de concepts repérés dans la signature de la relation pour identifier les relations. Nous avons jugé que de n'utiliser que la structure du tableau pour relier les éléments de la relation est aussi efficace et plus simple à mettre en œuvre.

Une fois la méthode définie, nous l'avons appliquée sur notre corpus test pour extraire les résultats expérimentaux. Nous avons ainsi pu évaluer les performances du système d'extraction.

## 2.4 Évaluations du système d'extraction d'information

L'évaluation est faite sur trois critères différents. Premièrement, nous calculons le rappel, la précision et la F-mesure pour l'extraction des résultats expérimentaux (dénommée *évaluation générale*) :

$$Rappel = \frac{\text{résultats correctement extraits}}{\text{nombre de résultats à extraire}} \tag{2.1}$$

$$Précision = \frac{\text{résultats correctement extraits}}{\text{nombre de résultats extraits}} \tag{2.2}$$

$$F - mesure = \frac{2 \cdot (\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})} \tag{2.3}$$

Nous comptons un descripteur ou un attribut correctement extrait s'il est du bon type et associé au bon résultat expérimental. Si une valeur numérique est sélectionnée à tort, tous ses attributs et descripteurs sont comptabilisés comme faux. Nous évaluons également l'extraction et la sélection des valeurs numériques uniquement, ce qui correspond à évaluer l'identification des résultats quantitatifs (dénommée *évaluation des résultats quantitatifs*). Le troisième type d'évaluation que nous faisons, concerne uniquement les résultats expérimentaux pertinents, c'est-à-dire que nous calculons le rappel, la précision et la F-mesure pour les valeurs numériques correctement sélectionnées (dénommée *évaluation des résultats expérimentaux correctement extraits*). Ce dernier type d'évaluation permet de n'évaluer que l'extraction des descripteurs et des attributs, et leur mise en relation.

Pour la tâche d'extraction de prescriptions qui a fait l'objet du challenge i2b2 2009, deux types d'évaluation ont été proposés par les organisateurs (Uzuner *et al.* [2010b]) : l'évaluation horizontale, qui consiste à calculer les f-mesures au niveau des entrées, c'est-à-dire pour chaque prescription, et l'évaluation verticale, qui consiste à évaluer l'extraction de chaque type de descripteurs (médicament, dosage, fréquence, etc.). L'évaluation générale que nous effectuons correspond à ce qu'ils ont appelé évaluation horizontale.

Rappelons que le corpus de test est composé de 855 résultats expérimentaux pertinents (issus de 15 articles).

L'évaluation du système est présentée dans le tableau 2.6. Ce système utilise des règles pour extraire les valeurs numériques, le lexique enrichi pour la détection des descripteurs et les critères de proximité et fréquence pour la mise en relation.

L'évaluation des résultats expérimentaux correctement extraits est faite sur une base de rappel de 0,63, c'est-à-dire que l'évaluation est faite sur les 63% de résultats expérimentaux correctement extraits. Nous obtenons une F-mesure de 0,78 pour les résultats pertinents, ce qui correspond à un bon résultat, au niveau de l'état de l'art si on se réfère à la tâche similaire de i2b2 2009 portant sur des termes médicaux. La F-mesure 0,61 reste bonne, ce qui signifie que même si trop de résultats sont proposés, ils ne constitueront pas un bruit trop lourd à gérer par les curateurs.

| | **Rappel** | **Précision** | **F-mesure** |
|---|---|---|---|
| Évaluation générale | 0,75 | 0,51 | 0,61 |
| Évaluation des résultats quantitatifs | 1,00 | 0,63 | 0,77 |
| Évaluation des résultats expérimentaux correctement extraits | 0,75 | 0,81 | 0,78 |

TAB. 2.6 – Évaluation du système d'extraction d'information

Dans la suite de cette partie, nous présentons l'évaluation de l'impact de certains choix que nous avons faits pour chaque étape de la méthode. Chaque évaluation est effectuée en comparant les performances du système final et du système en faisant varier un paramètre à la fois.

### 2.4.1 Évaluation de l'extraction des valeurs numériques

Le tableau 2.7 présente les résultats de l'extraction des valeurs numériques par apprentissage et par règles. La précision (P) et la F-mesure (F) sont meilleures pour l'extraction par apprentissage, mais le rappel (R) est meilleur avec les règles. Ce qui nous intéresse pour cette tâche est d'avoir un très bon rappel: en effet il paraît plus aisé pour les experts qui vérifieront les données annotées d'en enlever que d'en ajouter. De plus le système d'apprentissage n'a pas assez de données pour apprendre la pertinence des valeurs numériques. En effet, seuls les résultats suivis d'une approximation (±) et ceux dans un tableau sont extraits, mais les résultats pertinents qui ne rentrent pas dans ces deux cas ne sont pas extraits. Nous avons par conséquent conservé les règles pour cette extraction.

| Évaluation des résultats quantitatifs | **R** | **P** | **F** |
|---|---|---|---|
| Apprentissage | 0,93 | 0,78 | **0,85** |
| Règles | **1,00** | 0,63 | 0,77 |

TAB. 2.7 – Évaluation de l'extraction des résultats numériques par apprentissage et par règles

### 2.4.2 Évaluation de la complétion du lexique

Dans le tableau 2.8 nous comparons les performances du système avant l'enrichissement du lexique et du système final. Dans les deux cas, les résultats numériques sont extraits avec des règles et les descripteurs sont reliés selon des critères de proximité et de fréquence. Les précisions des deux systèmes restent très proches, alors que plus les lexiques sont complets plus le nombre de descripteurs annotés augmente, il y a donc plus de risques d'erreurs lors de la mise en relation. Comme on peut s'y attendre plus le lexique est complet et plus le rappel augmente. Nous pouvons observer que le système final a un rappel qui augmente de 12% par rapport au système de base.

| Évaluation des résultats expérimentaux correctement extraits | **R** | **P** | **F** |
|---|---|---|---|
| Système avec le lexique de base (QKDB) | 0,67 | 0,79 | 0,73 |
| Système final, avec le lexique enrichi | 0,75 | 0,81 | 0,78 |

TAB. 2.8 – Évaluation de la complétion du lexique

### 2.4.3 Évaluation de la mise en relation

La sélection de l'espèce et de l'organe se fait selon un critère de proximité puis de fréquence. Si un organe ou une espèce sont cités dans la même phrase que le résultat quantitatif de l'expérience, alors cet organe ou cette espèce est sélectionné(e). Dans le cas contraire, nous sélectionnons l'organe ou l'espèce le (la) plus fréquent(e) dans l'article. En effet la majorité des articles portent sur une seule espèce et sur un organe en particulier (dans notre cas, les articles portent sur le rein). Si ce n'est pas le cas, l'espèce et l'organe sont donnés au moment de la description du résultat et à proximité de celui-ci. Nous avons

évalué le choix de ce critère de fréquence par rapport à un critère de proximité uniquement. Dans un premier temps, nous avons évalué l'extraction de l'espèce et de l'organe avec des critères de sélection de fréquence et de proximité, et dans un second temps uniquement avec un critère de proximité. Le système utilisé est le système final avec le lexique enrichi. Les résultats de cette évaluation sont donnés dans le tableau 2.9.

| | Critère prox/freq | | | Critère prox | | |
|---|---|---|---|---|---|---|
| | **R** | **P** | **F** | **R** | **P** | **F** |
| Évaluation générale | 0,75 | 0,51 | **0,61** | 0,52 | 0,45 | 0,48 |
| Évaluation des résultats expérimentaux correctement extraits | 0,75 | 0,81 | **0,78** | 0,52 | 0,76 | 0,62 |
| Espèces | 0,97 | 0,98 | **0,97** | 0,21 | 1,00 | 0,35 |
| Organes | 0,96 | 0,98 | **0,97** | 0,07 | 0,83 | 0,13 |

TAB. 2.9 – Évaluation de l'extraction de l'espèce et de l'organe selon le critère utilisé

Nous observons que le rappel et la précision pour l'extraction des descripteurs de type espèce et de type organe avec un critère de fréquence sont proches de 1,00. Cela montre que la majorité des descripteurs de ces deux types sont extraits. En revanche, en utilisant un critère de proximité très peu de descripteurs sont extraits. En effet, les descripteurs de type espèce et organe sont cités principalement au début du document, mais de façon moins régulière dans la partie résultat de l'article. Une des raisons pour lesquelles le rappel et la précision pour l'extraction des descripteurs de type espèce ne sont pas à 1,00 est une mauvaise utilisation de la structure des tableaux. En effet, certains tableaux ont des structures assez complexes, avec par exemple des légendes communes à plusieurs lignes, et les règles de mise en relation des descripteurs dans ces tableaux ne s'appliquent pas correctement. En vu de ces résultats, nous avons utilisé un critère de proximité et de fréquence pour l'extraction de l'espèce et de l'organe dans la version finale de notre système.

Nous avons fait le choix de sélectionner les descripteurs dans le paragraphe s'ils n'étaient pas présents dans la phrase. Avec une version plus ancienne du système d'extraction, nous avions évalué le choix de ce contexte (Grau *et al.* [2009]) et nous avions observé qu'en utilisant un contexte large (c'est-à-dire le paragraphe) et non uniquement la phrase, le rappel augmentait légérement mais la précision diminuait. Pour les experts du domaine, il est important que le rappel de notre système soit bon même s'il perd un peu en précision, de ce fait nous avons choisi d'utiliser un contexte large.

### 2.4.4 Évaluation de l'extraction des résultats expérimentaux dans des tableaux

Dans le corpus de développement 35% des valeurs numériques des résultats expérimentaux étaient dans des tableaux, et dans le corpus de test 77% le sont. Les descripteurs des résultats expérimentaux dont les valeurs numériques sont dans les tableaux peuvent être dans les tableaux ou dans le texte. Dans les tableaux, les descripteurs sont dans les en-têtes des colonnes ou des lignes, ainsi que dans les légendes du tableau. Les descripteurs qui sont dans les en-têtes sont faciles à mettre en relation avec la valeur numérique ; pour cela nous prenons en compte leur position dans le tableau. Mais il est plus difficile de relier

les informations de la légende aux valeurs numériques.

Les 33 résultats du corpus de développement qui sont dans des tableaux, sont décrits par 204 descripteurs, dont 180 sont dans les tableaux et seulement 24 dans le texte (dans le texte on retrouve souvent l'espèce et l'organe). Les problèmes posés par l'extraction des informations données sous forme de tableau ou de texte étant différents, nous avons évalué séparément l'extraction des résultats expérimentaux dans des tableaux et dans le texte. Le système utilisé pour cette évaluation est le système final avec le lexique enrichi et l'utilisation des critères de proximité et de fréquence pour la mise en relation. Les résultats sont présentés dans le tableau 2.10. Nous observons que le rappel est identique pour l'extraction dans le texte et dans les tableaux, en revanche la précision est meilleure pour l'extraction dans les tableaux. Cela est dû au fait que l'information présente dans les tableaux est structurée ce qui permet de limiter les erreurs.

| | Tableaux | | | Texte | | | Système | | |
|---|---|---|---|---|---|---|---|---|---|
| | **R** | **P** | **F** | **R** | **P** | **F** | **R** | **P** | **F** |
| Évaluation des résultats quantitatifs | 1 | 0,86 | 0,92 | 1 | 0,33 | 0,50 | 1 | 0,63 | 0,77 |
| Évaluation générale | 0,75 | 0,65 | 0,70 | 0,75 | 0,39 | 0,51 | 0,75 | 0,51 | 0,61 |

TAB. 2.10 – Évaluation de l'extraction des résultats numériques dans les tableaux et dans le texte

L'originalité de notre approche réside dans le fait que les données sont extraites des articles complets et nécessitent d'être retrouvées dans différentes sections de l'article, et que nous extrayons les informations qu'elles soient présentes dans le texte ou dans des tableaux. Les principales difficultés proviennent de la présence de nombreuses variations terminologiques et de la mise en relation des descripteurs avec un résultat. Le système d'extraction atteint un résultat de très bon niveau (F-mesure de 0,78), et permet de trouver tous les résultats des articles, sans pour autant fournir trop de bruit devenant gênant pour les experts curateurs (rappel de 1 et précision de 0,63).

Le système d'extraction d'information a été intégré à un assistant d'aide à l'annotation d'article et au peuplement de QKDB par les experts.

# Chapitre 3

# Assistant d'aide à l'annotation d'article et au peuplement de la base de données

**Sommaire**



Pour répondre au besoin d'assistance des experts lors du peuplement de la base de données, nous avons développé un assistant d'aide à l'annotation d'article et au peuplement de la base. Le développement de cet assistant a été commencé par un stagiaire en 2009 : Adrien Dong. Nous avons finalisé le développement et avons réalisé une évaluation. L'intégration de l'assistant au site Web de la base de données a été fait par un autre stagiaire en 2011 : Corentin Limier.

## 3.1 Spécification de l'outil

L'outil d'extraction ne pouvant pas avoir une précision parfaite, les informations extraites par ce système doivent être vérifiées avant d'être insérées dans la base de données. Nous avons donc choisi d'intégrer le système d'extraction dans une interface d'aide à l'annotation afin de faciliter l'annotation des articles par les experts. L'objectif est de fournir un article pré-annoté aux experts, qui n'auront plus qu'à valider, modifier ou rejeter les propositions du système.

L'interface développée doit donc permettre de :
- ajouter un article dans la base d'articles;
- sélectionner un article à annoter;
- afficher les annotations extraites ou existantes de l'article;
- rajouter ou modifier manuellement les annotations de l'article.

Cette interface étant destinée à être utilisée par des experts curateurs divers, elle est conçue comme une application Web, et mise à disposition de ces experts [1].

## 3.2 Descriptif

Les différents processus utiles à l'analyse d'un texte, allant de sa conversion au format XML requis jusqu'à la proposition des enregistrements à intégrer dans la base de données, sont intégrés au travers d'une interface Web, développée en PHP et JavaScript (figure 3.1).

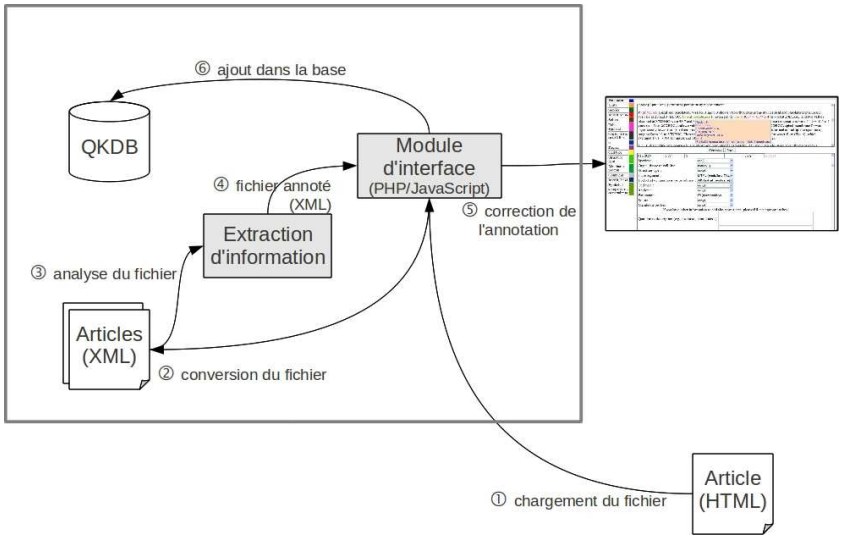



FIG. 3.1 – Architecture de l'assistant d'aide au peuplement de la base de données

L'utilisateur doit fournir à l'assistant un article HTML ainsi que des fichiers contenant les tables associées. Les articles sont transformés dans notre format XML. Pour le moment, seuls les articles HTML sont acceptés. Ensuite l'extraction est exécutée sur l'article XML. Le fichier de sortie est l'article XML dont les descripteurs ont été annotés.

Après la phase d'extraction, l'assistant propose la visualisation présentée dans la figure 3.2 [2].

L'interface permet à l'utilisateur de :
- visualiser, pour chaque résultat, les informations qui lui sont associées, surlignées en couleur (avec un code couleur par champ) (1) ;
- avoir un récapitulatif des attributs caractérisant une expérimentation ; celui-ci est affiché dans une infobulle (2) ;
- parcourir l'article en passant d'un résultat à l'autre (3) ;
- visualiser les attributs de la base de données dans un formulaire modifiable affiché en bas d'écran, qui correspond au résultat affiché dans la partie texte (4) ;
- parcourir les schémas extraits, avec un affichage de la partie texte correspondante (5) ;

[1] http://qkdb.limsi.fr/
[2] Dans ce qui suit les numéros entre parenthèses renvoient aux numéros sur la figure.

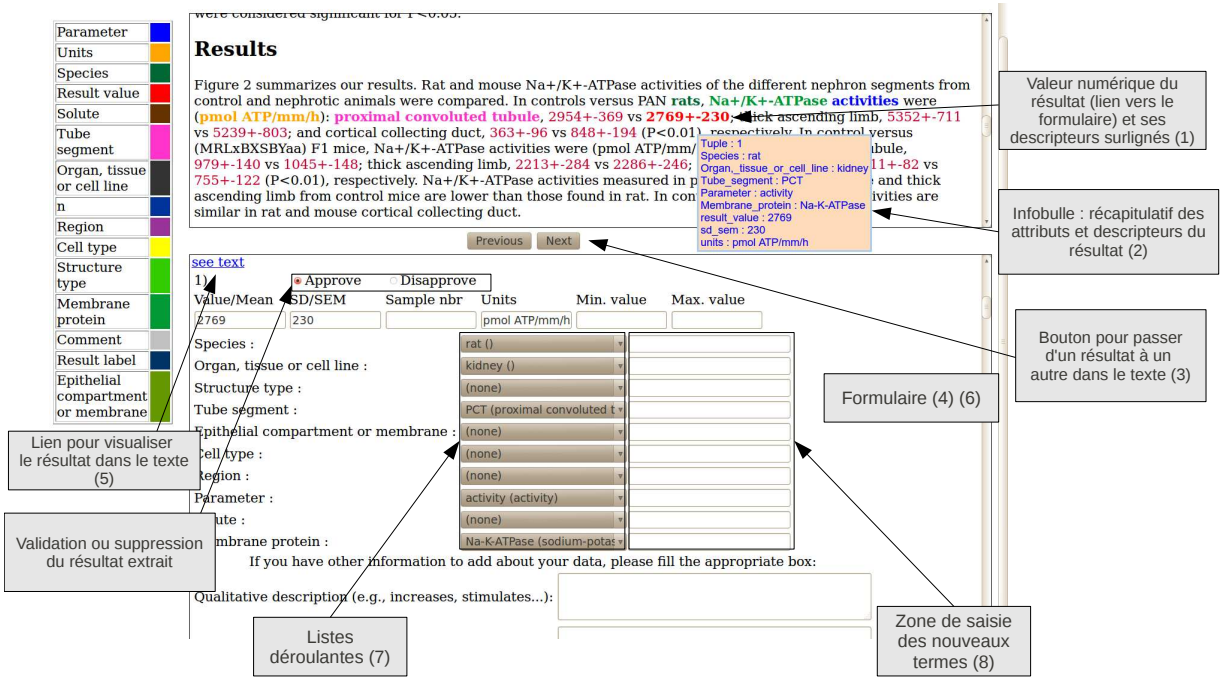



FIG. 3.2 – Présentation des fonctionnalités de l'assistant

– modifier des descripteurs via le formulaire ; les modifications sont répercutées sur l'annotation du texte de manière à conserver la cohérence entre les deux points de vue sur le texte (6).

Le formulaire est composé de listes déroulantes pour les champs pour lesquels les listes des valeurs sont fermées (7). À droite de ces listes, des zones de saisies ont été ajoutées pour permettre à l'expert d'écrire le terme utilisé dans le texte pour décrire le concept, dans le cas où un mauvais concept aurait été extrait (8).

Pour illustrer les fonctionnalités de l'assistant, nous allons présenter le processus de modification et validation, ou suppression de l'extraction du résultat expérimental surligné dans la figure 3.3 et dont le formulaire est donné dans la figure 3.4.

L'outil d'extraction a extrait un résultat expérimental dont la valeur numérique et la précision sont : $6.5 \pm 1.39$. Les attributs de cette valeur numérique et les descripteurs de l'expérience sont surlignés en couleur dans le texte, et donnés avec leur type dans le formulaire. Ce résultat expérimental semble pertinent pour la base de données QKDB et les attributs de la valeur numérique ont été correctement annotés (l'approximation et l'unité). On peut donc valider ce résultat expérimental, après avoir effectué quelques modifications au niveau des descripteurs. L'espèce est correcte mais son attribut n'a pas de valeur, il convient de rajouter la valeur 5 dans la case de l'attribut « n ». L'organe et le paramètre semblent corrects. En revanche le tube n'est pas précisé, alors qu'il est donné dans l'article : « proximal tubule ». Le terme qui a été annoté Soluté, est en fait une substance injectée à la souris, mais n'est pas en lien avec ce qui est mesuré dans l'expérience, à savoir une pression. Il est également possible, voire nécessaire, de rajouter des commentaires pour préciser que

Maximum **PSF** responses during both control and ANG II infusion periods were obtained in six nephrons of wild-type, seven nephrons of heterozygous, and nine nephrons of homozygous ACE transgenic mice. Because results were obtained in the same **nephrons**, comparison between control and angiotensin infusion periods was done by paired t-test. Data are shown in Fig. 4. It can be seen that **ANG II** infusion caused a significant increase in the TGF response magnitude in all three groups of **mice**, from **6.5 +- 1.39** to 9.8 +- 1.2 **mmHg** in +/+ (mice/tubules = 5/6; P = 0.007), from 1.14 +- 0.42 to 4.6 +- 1.3 mmHg in +/- (mice/tubules = 7/7; P = 0.016), and from 0.42 +- 0.25 to 4.02 +- 1.06 mmHg in -/- mice (mice/tubules = 4/9; P = 0.05). Increments in PSF caused by ANG II infusion were identical among the three groups of animals (3.3 +- 0.63 mmHg in +/+, 3.4 +- 1.03 mmHg in +/-, and 3.6 +- 1.1 mmHg in -/- animals), implying that the degree of increase in responsiveness was greater in the transgenic than in the control mice (1.68 +- 0.18-fold in +/+, 2.7 +- 1-fold in +/-, and 4.7 +- 1.5-fold in -/- mice). Even though TGF responses in heterozygous or homozygous mice during ANG II infusion tended to be somewhat lower than those measured during control in wild-type mice, the difference did not achieve statistical significance, suggesting that angiotensin infusion restored responsiveness to near normal levels. During ANG II infusion,

FIG. 3.3 – Exemple d'annotation d'un paragraphe extrait de "Tubuloglomerular feedback in ACE-deficient mice" de Traynor et al., 1999

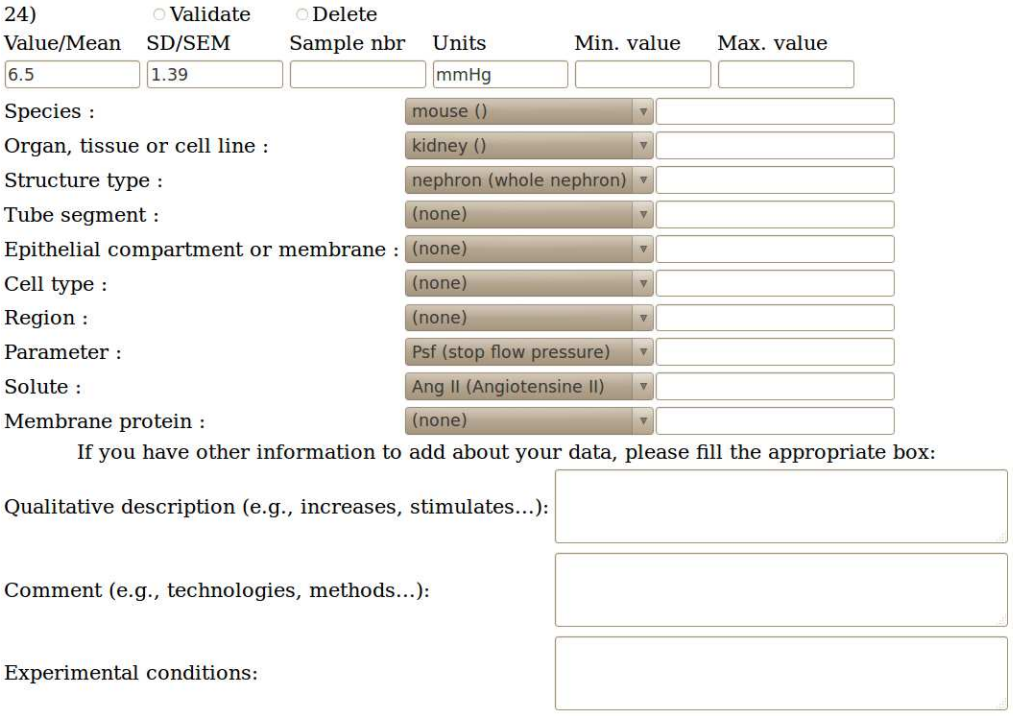



FIG. 3.4 – Formulaire d'annotation pour l'exemple de la figure 3.3

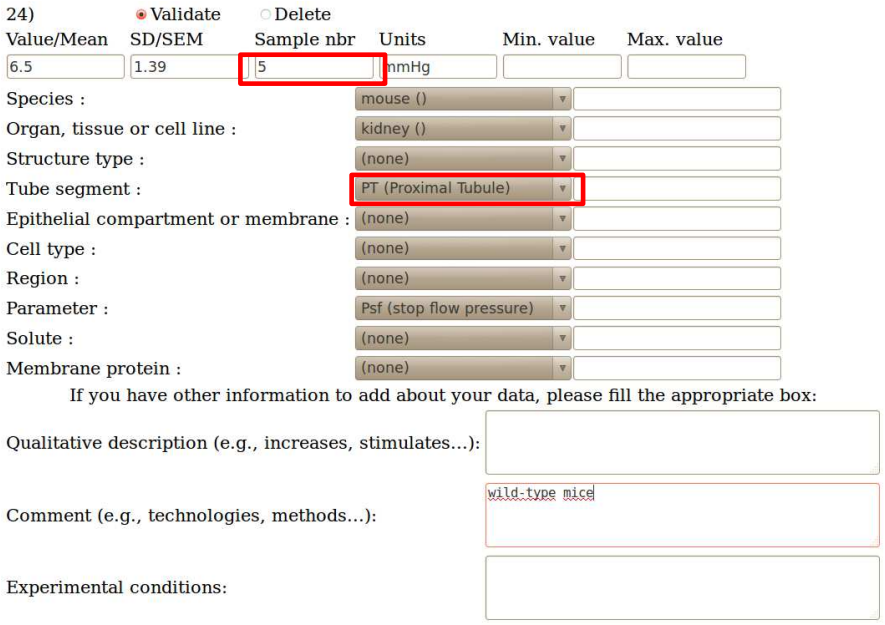



FIG. 3.5 – Formulaire d'annotation pour l'exemple de la figure 3.3 modifié

l'expérience a eu lieu sur une souris dont les gènes n'ont pas été modifiés, contrairement aux expériences suivantes. Le formulaire de la figure 3.5 a été modifié pour prendre en compte les remarques précédentes et valider le résultat expérimental. On remarque dans le texte de la figure 3.3 que la valeur numérique suivante, à savoir $9.8 \pm 1.2$, est le résultat d'une expérience pertinente pour QKDB, en revanche la suivante : 0.007 pourra être supprimée.

## 3.3 Évaluation utilisateurs

Nous nous sommes inspirée de la méthode adoptée par Alex *et al.* [2008] pour l'évaluation de l'outil. Nous avons demandé à cinq experts [3] d'annoter trois articles chacun dans trois conditions différentes (au total cinq articles différents ont été annotés) : un à partir de l'annotation de référence (c'est-à-dire à partir de l'annotation faite par projection des données de la base), un avec les annotations de notre système d'extraction et le troisième *manuellement*, c'est-à-dire sans annotation préalable de l'article. Nous avons calculé le temps qu'ils passaient sur chaque article, le nombre d'expérimentations annotées, le nombre de descripteurs et le nombre de commentaires (les commentaires ne sont pas annotés par notre système d'extraction mais très importants pour la base de données QKDB). À la suite de l'annotation, les experts ont répondu à un questionnaire.

[3] Un étudiant en bioinformatique, un post-doctorant, un ingénieur d'étude en bioinformatique, une *assistant professor* en ingénierie chimique et biologique et un directeur de recherche spécialisé en physiologie rénale (le concepteur de QKDB).

Le tableau 3.1 présente le nombre total de résultats, de descripteurs et de commentaires annotés, ainsi que le temps moyen passé pour annoter un résultat dans deux conditions : *manuellement* (le formulaire était vide) et avec l'assistant. Nous ne donnons pas les résultats obtenus dans la troisième condition (avec les annotations de référence) car c'était une évaluation de contrôle. Le temps total passé par les 5 experts sur les 5 articles est de 170 mn avec l'assistant et 150 mn sans, mais il y a presque deux fois plus de résultats annotés avec l'assistant. Nous observons que huit fois plus de commentaires sont ajoutés lorsque l'expert utilise l'assistant, par rapport à l'annotation *manuelle*.

| Condition | # résultats | # descripteurs | # commentaires | Temps moyen passé par résultat |
|---|---|---|---|---|
| **Manuellement** | 52 | 448 | 9 | 174s |
| **Avec l'assistant** | 96 | 846 | 75 | 105s |

TAB. 3.1 – Nombre de résultats expérimentaux insérés dans la base manuellement et avec l'assistant, et temps moyen passé pour annoter un résultat.

Une partie des questions posées aux experts après l'annotation est présentée dans le tableau 3.2. Pour chaque affirmation, nous leur demandions de mettre un score de 1 à 5, de d'accord avec l'affirmation à pas d'accord. Ils étaient tous d'accord pour dire que l'assistant facilite et accélère la tâche d'annotation, et que le surlignage des descripteurs permettait une meilleure visibilité, ce qui simplifiait et accélérait la tâche.

Après chaque étape, nous demandions à l'expert s'il pensait avoir annoté tous les résultats pertinents de l'article. Ils répondaient tous oui quand ils utilisaient l'assistant, et seulement la moitié répondaient oui quand ils annotaient l'article à la main, et trois experts sur cinq ont dit que de tout annoter à la main prenait trop de temps. Un expert a lu les articles en entier avant de procéder à l'annotation, les autres n'ont lu que les parties *Results* et *Methods* (et *Abstract* pour l'un d'entre eux).

| | Score |
|---|---|
| L'assistant facilite la tâche d'annotation. | 1 |
| L'assistant accélère le traitement de l'article. | 1 |
| Le surlignage des termes simplifie la tâche. | 1 |
| Le surlignage des termes accélère la tâche. | 1 |
| La visualisation des résultats un par un est u tile. | 1,4 |
| Les infobulles sont utiles. | 1,4 |
| Les liens entre le formulaire et le texte facilitent la tâche. | 1,2 |
| L'interface est simple d'utilisation. | 1,4 |

TAB. 3.2 – Moyenne des scores du questionnaire. Les scores vont de 1, si les experts sont d'accord avec l'affirmation, à 5, s'ils ne sont pas d'accord.

# Discussion et conclusion

Nous avons présenté un système complet permettant le peuplement d'une base de données. Cette base est formalisée par une ressource termino-ontologique qui représente des résultats d'expérimentation par une relation n-aire. Ce système constitue l'un des rares travaux qui permet de remplir ce type de base, plus complexe que les bases sur la génomique par exemple où l'on trouve plutôt des relations binaires à instancier. Il répond à un besoin dans le domaine, puisque la base conçue par des biologistes, et adaptée à leurs besoins, n'est pas alimentée à cause du temps que cette opération demande. Les premières présentations faites ont vivement intéressé les chercheurs du domaine.

Du point de vue de l'extraction d'information, notre approche vise à extraire des résultats expérimentaux décrits sur plusieurs phrases dans des articles complets. Nous avons étudié la détection des descripteurs et de leurs variantes, ainsi que la mise en relation des différents éléments de la relation. Notre système détecte tous les résultats expérimentaux (le rappel est de 1), ce qui facilite la tâche des curateurs, avec une précision correcte de 0,63. Une F-mesure de 0,78 est obtenue si on évalue l'extraction et la mise en relation de tous les descripteurs.

L'assistant développé pour la curation des données extraites a été évalué par différents experts. Ses performances et son ergonomie entraînent les experts à extraire tous les résultats d'un article, là où leur extraction manuelle était moins exhaustive, notamment dans le cas des tableaux dont le traitement est allégé et rendu moins fastidieux. Les résultats obtenus sur ces cas particuliers sont d'ailleurs très bons, ce qui permet de se reposer sur l'assistant.

La méthodologie mise en œuvre ainsi que les choix de développement permettent de transposer ce travail à d'autres domaines où il est important pour leur étude de représenter et décrire des résultats expérimentaux. La formalisation sous forme d'ontologie permet de voir les concepts à définir et le niveau de généralité désiré. Le lexique initial contient les termes qui désignent ces concepts, il peut ensuite être étendu dès lors que l'on dispose d'articles. Si des ontologies sont créées, elles pourront être intégrées sans problème dans notre système, puisque nos lexiques séparent bien les concepts des termes qui les désignent.

Nous n'avons pas encore pu évaluer l'adaptation de notre système à un autre domaine, mais nous souhaiterions le faire prochainement. Il est envisagé d'incorporer notre assistant dans le package QxDB générique (distribué dans la communauté du Physiome/VPH Européen), afin de faciliter l'adoption de cette approche dans d'autres domaines.

Plusieurs améliorations du système sont envisageables ; dans un premier temps, il fau-

drait travailler sur la mise en relation des descripteurs et de la valeur numérique dans les tableaux complexes. Au préalable une amélioration de la transformation des tableaux du format XHTML en XML est nécessaire, car certaines structures ne sont pas conservées lors de la conversion. Dans un deuxième temps, lors de la mise en relation des descripteurs, il serait intéressant d'étudier et de prendre en compte les liens implicites existant entre les descripteurs. Par exemple, si le descripteur de type paramètre est « concentration » alors il y a forcément un soluté associé. Ou encore, si l'unité est « g » le paramètre sera « weight ». Le dernier point pouvant permettre une amélioration de la précision du système serait d'analyser plus précisément avec un expert en quoi un résultat d'expérimentation est pertinent ou non pour la base de données. Ces améliorations nécessitent d'augmenter la taille du corpus ainsi que d'améliorer son annotation. La mise en ligne de notre outil [4] est un pas important dans ce sens, car il constitue aussi un outil d'annotation. En effet, nous avons veillé à créer et maintenir le lien et la cohérence entre QKDB et les articles d'où sont issus les informations insérées dans la base. Les utilisateurs pourront maintenant voir l'article annoté en fonction des données selectionnées.

[4] L'assistant non relié à QKDB : http://qkdb.limsi.fr ; l'assistant intégré à QKDB (en cours) : http://physiome.ibisc.fr/qkdb/

# Troisième partie

# Extraction de relations binaires dans le domaine biomédical

La littérature biomédical est riche d'observations, de résultats d'expériences, etc. Toutes ces informations sont très dispersées dans les articles et très nombreuses, et il est difficile pour les experts de disposer de toutes celles dont ils ont besoin. Pour résoudre ce problème, le domaine de l'extraction d'information s'intéresse au traitement de l'information peu structurée pour obtenir de l'information structurée.

Nous venons de présenter une méthode d'extraction de résultats expérimentaux modélisés par des relations n-aires. Nous allons maintenant nous intéresser à l'extraction de relations entre différentes entités dans des documents biomédicaux. Dans le domaine médical il est nécessaire d'analyser et d'extraire les informations des comptes rendus cliniques des patients pour aider à la prise de décision. Il peut être par exemple utile de savoir que tel médicament dans telles conditions a permis de traiter une maladie. Nous avons travaillé sur la détection et la classification (ou le typage) de relations binaires, c'est-à-dire des relations entre deux entités. Pour cela, nous avons utilisé des méthodes par apprentissage et nous nous sommes intéressée aux traits permettant de représenter les informations nécessaires pour détecter les relations (traits surfaciques, lexicaux, sémantiques et morpho-syntaxiques). Nous avons étudié l'apport de la prise en compte de la structure syntaxique pour extraire les relations. Pour améliorer la classification, nous avons proposé des méthodes de simplification de phrases, qui permettent de réduire la variation d'expression des relations.

Nous avons exploité trois corpus pour développer et évaluer nos méthodes : un corpus de comptes rendus médicaux pour la classification de relations entre des traitements, des examens cliniques et des problèmes médicaux ; un corpus d'articles scientifiques pour la détection d'interactions entre des médicaments ; et un corpus également d'articles scientifiques pour l'identification d'interactions entre des protéines.

Dans cette partie, nous présenterons dans un premier temps la méthode par apprentissage que nous avons développé conduisant au développement du système REMED qui a participé aux évaluations i2b2 en 2010 et DDI en 2011, et notre étude sur l'apport de la syntaxe pour l'extraction des relations. Le deuxième chapitre portera sur les méthodes de simplification de phrases guidée par la tâche d'extraction de relations.

# Chapitre 4

# Extraction de relations comme une tâche de classification

**Sommaire**



Nous avons considéré l'extraction de relations binaires comme une tâche de classification, c'est-à-dire que chaque couple d'entités d'une phrase sera classé soit dans la classe

des non-relations [1], soit dans une classe de relations. Pour chaque tâche, il y a au moins deux classes : non-relation et relation. Nous nous sommes intéressée aux attributs à utiliser pour représenter au mieux les caractéristiques de l'expression d'une relation et nous avons en particulier étudié l'utilisation d'attributs syntaxiques permettant de prendre en compte la structure de la phrase et les dépendances entre des groupes syntaxiques. Ces informations syntaxiques peuvent être représentées sous forme vectorielle ou sous forme d'arbre (cf. section 4.3).

Pour étudier et évaluer les informations utilisées pour la classification des relations, nous avons développé un système qui utilise des SVM pour apprendre un modèle de classification que nous avons dénommé REMED pour *Relation Extraction in bio-Medical Domain* (cf. section 4.2).

Dans un premier temps, nous introduirons les méthodes de classification, les outils, le corpus et les méthodes d'évaluation que nous avons utilisés. Nous présenterons les informations que nous avons étudiées pour apprendre à classer les relations. Une évaluation des attributs est ensuite proposée ainsi qu'une évaluation de l'apport de la prise en compte des informations sur la structure syntaxique des phrases. Nous terminerons ce chapitre par l'évaluation du système REMED sur deux autres tâches d'extraction de relations.

## 4.1 Outils et matériels

Le système que nous avons développé repose sur une méthode à base d'apprentissage. Nous avons fait ce choix car les méthodes à base d'apprentissage, contrairement aux méthodes basées sur des patrons, sont facilement adaptables à d'autres corpus et d'autres domaines, sont facilement modifiables si on veut prendre en compte des informations supplémentaires, et permettent d'obtenir les meilleurs résultats dans l'état de l'art. Le prérequis des méthodes à base d'apprentissage est de disposer d'un corpus annoté en concepts et relations. Notre travail sur l'extraction de relations binaires a été initié avec notre participation au challenge i2b2 2010, pour lequel il nous a été fourni un corpus annoté de comptes rendus cliniques. De ce fait, développer une méthode à base d'apprentissage était envisageable.

Le système utilise des classifieurs existants fondés sur des SVM (*Support Vector Machine*). Nous introduirons les SVM et les classifieurs que nous avons utilisés, puis nous présenterons le domaine d'application de notre système et les méthodes d'évaluation utilisées.

### 4.1.1 Les SVM : pourquoi ? Comment ?

Nous avons utilisé des classifieurs à base de SVM car ils sont très présents dans l'état de l'art des systèmes d'extraction de relations (Zhou *et al.* [2005]; Roberts *et al.* [2008]; Uzuner *et al.* [2010a], etc.). De plus, ils donnent de bons résultats pour les tâches pour

[1] Nous disons d'une paire d'entités qu'elle est dans la classe des non-relations s'il n'existe pas de relation entre les deux entités, ou du moins pas une relation faisant partie de la typologie de la tâche d'extraction des relations.

lesquelles il y a beaucoup d'attributs mais qui sont très épars, comme c'est souvent le cas en TAL.

Les SVM sont des méthodes à noyau, qui sont inspirées de la méthode statistique de l'apprentissage de Vladimir Vapnik Boser *et al.* [1992]. Les classifieurs à base de SVM font de la classification binaire par apprentissage supervisée.

Les classifieurs utilisent des fonctions qui permettent de faire une séparation optimale des données ; cette séparation est appelée hyperplan. Les SVM cherchent à optimiser l'hyperplan de sorte que la marge entre l'hyperplan et les données soit maximale. Dans le cas de données non linéairement séparables, l'espace de données est projeté dans un espace de plus grande dimension grâce à une fonction noyau pour avoir un espace linéairement séparable. Il est alors possible de classer des données représentées sous forme d'arbres.

Si plus de deux classes sont à séparer, il est nécessaire d'entraîner plusieurs classifieurs. Deux méthodes ont été proposées à cet effet :

- méthode « one-versus-all » (« un-contre-tous ») : construction d'autant de classifieurs binaires qu'il y a de classes, chaque classifieur attribuera le label 1 aux instances d'une classe et le label -1 à toutes les autres.
- méthode « one-versus-one » (« un-contre-un ») : construction d'autant de classifieurs binaires que de paires de classes. L'instance à classer est analysée par chaque classifieur et un vote majoritaire permet de choisir la classe.

Il est possible de paramétrer les SVM en définissant la valeur du facteur de pénalité, C. Ce facteur contrôle le compromis entre les erreurs sur les données d'entraînement et la maximisation de la marge de l'hyperplan. Si C est petit, les instances proches de la frontière entre les deux classes seront ignorées, et la marge sera plus grande.

L'utilisation d'un noyau RBF (fonction à base radiale) nécessite de définir un autre paramètre : gamma ($\gamma$). $\gamma$ augmente la flexibilité de la frontière de décision entre les deux classes. Si $\gamma$ a une petite valeur, la frontière sera presque linéaire.

### 4.1.2 Les noyaux d'arbres (tree kernels)

Pour tenir compte de l'information syntaxique, nous avons choisi d'utiliser une fonction kernel qui mesure la similarité entre deux arbres, en comptant le nombre de fragments en commun. L'arbre est découpé en fragments de différentes tailles. Deux options sont possibles, soit ST (*subtrees*) qui calcule tous les sous-arbres possibles avec tous leurs descendants (voir figure 4.1), soit SST (*subset tree*) qui autorise également les fragments d'arbres dont les feuilles ne sont pas des éléments terminaux, mais des chunks ou des étiquettes morpho-syntaxiques (voir figure 4.2).

Nous avons choisi d'utiliser l'option SST qui est plus générale et donne de meilleurs résultats selon Moschitti [2006].

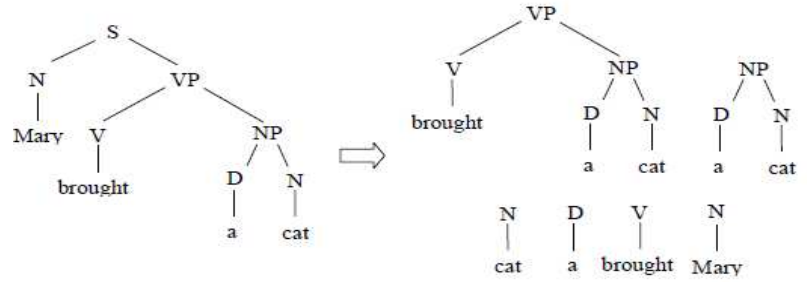



FIG. 4.1 – Exemple de subtrees (ST) de Moschitti [2006]

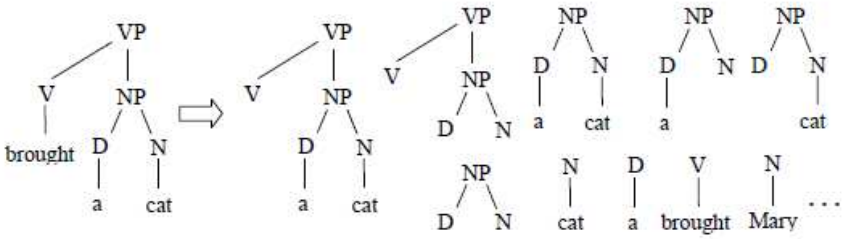



FIG. 4.2 – Exemple de subset trees (SST) de Moschitti [2006]

### 4.1.3 Outils

Pour apprendre à classer les relations, nous avons utilisé libSVM [2] qui est une bibliothèque développée par Chang et Lin [2001] en C++ et Java. Elle supporte la classification multi-classes avec la méthode « one-versus-one » ainsi que la validation croisée. Elle permet l'usage de 4 noyaux : linéaire, polynomial, sigmoïde et RBF. Nous l'utilisons avec un noyau RBF.

Nous avons également utilisé SVM-Light-TK version 1.5 développé par Moschitti (Moschitti [2006]) pour apprendre à identifier les relations en utilisant les arbres en constituants. En effet, dans SVM-Light-TK a été implémentée une fonction à noyau d'arbres, qui repose sur le calcul de similarité entre deux arbres. SVM-Light-TK nous permet donc d'utiliser une représentation sous forme d'arbres de la structure des phrases. La classification multi-classes n'est pas prise en compte dans SVM-Light-TK. Il est possible de paramétrer le classifieur pour qu'il n'utilise que les arbres, que les vecteurs ou une combinaison des deux, et pour que l'opérateur de combinaison soit la somme des contributions ou le produit des contributions.

[2] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

### 4.1.4 Domaine d'application : extraction de relations dans des comptes rendus cliniques, i2b2 2010

Pour le développement de notre système et de nos recherches sur l'apport des informations syntaxiques, nous nous sommes focalisée sur le corpus du challenge i2b2 2010 de comptes rendus cliniques (Uzuner *et al.* [2011]). Initialement le système avait été développé pour ce challenge auquel nous avons participé avec une équipe du LIMSI (Minard *et al.* [2011a]). Nous avons ensuite continué à utiliser ce corpus pour nos différentes études. Deux autres corpus ont par la suite été utilisés pour évaluer l'adaptation de nos méthodes : le corpus PPI et DDI. Ces deux corpus et les expérimentations effectuées seront présentés dans la section 4.4.

#### 4.1.4.1 Tâche

L'objectif de la tâche sur les relations d'i2b2 2010 était d'annoter dans les documents les relations existantes entre deux entités. Les types d'entités considérés sont les suivants :

- Les problèmes médicaux, définis comme les observations des patients ou des cliniciens concernant ce qui n'allait pas ou semblait être causé par une maladie. Cette catégorie comprend notamment les maladies, les syndromes, les observations sur l'état mental du patient, etc.
- Les traitements, définis comme les procédures, interventions, substances et médicaments donnés à un patient pour tenter de résoudre un problème.
- Les tests, comprenant les procédures et examens effectués sur un patient ou un fluide corporel pour vérifier ou infirmer la présence d'un problème, ou pour avoir plus d'informations sur un problème.

Entre ces trois types d'entités, des relations peuvent exister : un test comme un examen peut par exemple être prescrit pour analyser un problème. Ce sont ces relations que nous avons cherché à identifier dans les documents. Afin d'étudier la reconnaissance des relations, les entités étaient annotées, et il s'agissait de déterminer si, étant donné deux entités, elles étaient en relation, et si oui, laquelle. Huit relations ont été définies par les organisateurs du challenge que nous présentons dans le tableau 4.1.

L'accord inter-annotateur pour chaque relation est donné dans le tableau 4.2. Cet accord a été calculé par Knowtator et fourni par les organisateurs du challenge. L'accord ajusté est obtenu après discussion des cas posant problème. On peut voir que l'accord inter-annotateur est faible pour la relation TrWP (*le traitement aggrave le problème*) et TrIP (*le traitement améliore le problème*).

#### 4.1.4.2 Corpus

Les corpus, fournis par les organisateurs de la tâche i2b2, sont composés de comptes rendus hospitaliers provenant de plusieurs centres médicaux aux États-Unis. Ces documents avaient été anonymisés et annotés manuellement pour constituer une référence. Un premier corpus a été fourni avant l'évaluation, composé de 350 documents. Ce premier corpus a été divisé en trois : corpus d'entraînement, de développement, et de test. Puis, les organisateurs d'i2b2 ont fourni le corpus d'évaluation, qui comporte 477 documents. Le

| Traitement - Problème | |
|---|---|
| TrIP | le traitement améliore le problème |
| | <pb> hypertension </pb> was controlled on <treat> hydrochlorothiazide </treat> |
| TrWP | le traitement aggrave le problème |
| | <pb> the tumor </pb> was growing despite the available <treat> chemotherapeutic regimen </treat> |
| TrCP | le traitement cause le problème |
| | <treat>Bactrim</treat> could be a cause of <pb>these abnormalities</pb> |
| TrAP | le traitement est administré en raison du problème |
| | <treat>antibiotic therapy</treat> for presumed <pb>right forearm phlebitis</pb> |
| TrNAP | le traitement n'est pas administré en raison du problème |
| | <treat>Relafen</treat> which is contraindicated because of <pb>ulcers</pb> |
| **Test - Problème** | |
| TeRP | le test révèle le problème |
| | <test>an echocardiogram</test> revealed <pb>a pericardial effusion</pb> |
| TeCP | le test est conduit en raison du problème |
| | <test>an VQ scan</test> was performed to investigate <pb>pulmonary embolus</pb> |
| **Problème - Problème** | |
| PIP | un problème en indique un autre |
| | <pb>Azotemia</pb> presumed secondary to <pb>sepsis</pb> |

Tab. 4.1 – Description des huit relations

nombre de relations dans chaque corpus est indiqué dans le tableau 4.2.

Sur le corpus d'entraînement (298 documents) la taille moyenne des phrases contenant au moins deux entités est de 17 mots/phrase (les signes de ponctuation ne sont pas comptés comme des mots). La phrase la plus courte contient deux mots, la phrase la plus longue 231 et la médiane est située à 15 mots/phrase. Coden *et al.* [2005] compare la taille moyenne des phrases de trois corpus, le premier est un extrait du Penn TreeBank (composé d'articles de journaux), le second le corpus GENIA (composé de résumé de MedLine) et le troisième est le corpus MED composé de rapports clinique. Les tailles moyennes des phrases de ces trois corpus ainsi que du corpus pour la campagne i2b2 2010 que nous utilisons sont répertoriées dans le tableau 4.3. Ces données montrent que le corpus sur lequel nous travaillons est composé de phrases courtes en moyenne, comparé au corpus GENIA. Les documents type rapport clinique sont composés de beaucoup de fragments de phrase (25) et d'énumérations (26).

(25) [PB C5-6 disc herniation] with [PB cord compression] and [PB myelopathy].

(26) Revealed [PB icteric sclerae], [PB the oropharynx with extensive thrush], and [PB an ulcer under his tongue].

| Relation | pré-éval | *entraî-nement* | *dévelop-pement* | *test* | évaluation | IAA strict | IAA ajusté |
|---|---|---|---|---|---|---|---|
| TrIP | 107 | 74 | 18 | 15 | 198 | 0,44 | **0,62** |
| TrWP | 56 | 39 | 10 | 7 | 143 | **0,30** | **0,58** |
| TrCP | 296 | 237 | 23 | 36 | 444 | 0,50 | 0,82 |
| TrAP | 1423 | 1013 | 229 | 181 | 2487 | 0,68 | 0,95 |
| TrNAP | 106 | 71 | 15 | 20 | 191 | 0,44 | 0,76 |
| PIP | 1239 | 855 | 224 | 160 | 1986 | **0,35** | 0,79 |
| TeRP | 1734 | 1236 | 226 | 272 | 3033 | 0,70 | 0,96 |
| TeCP | 303 | 196 | 49 | 58 | 588 | 0,43 | 0,74 |
| Toutes | 5264 | 3721 | 794 | 749 | 9070 | 0,56 | 0,94 |
| Non-relations | 11206 | 7278 | 2304 | 1624 | 17072 | | |

TAB. 4.2 – Nombre de relations par catégorie dans chaque corpus et accord inter-annotateur (IAA)

| | taille moyenne des phrases |
|---|---|
| Penn Treebank | 24 |
| MED | 14 |
| GENIA | 27 |
| i2b22010 corpus | 17 |

TAB. 4.3 – Taille moyenne des phrases de chaque corpus

### 4.1.5 Méthodes d'évaluation

Pour évaluer nos systèmes, nous utilisons les trois mesures classiques : rappel, précision et f-mesure. Nous rappelons les formules ci-dessous.

$$Rappel = \frac{\text{relations correctement classées}}{\text{nombre de relations à classer}} \tag{4.1}$$

$$Précision = \frac{\text{relations correctement classées}}{\text{nombre de relations classées}} \tag{4.2}$$

$$F - mesure = \frac{2 \cdot (\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})} \tag{4.3}$$

L'évaluation porte uniquement sur les relations et non sur toutes les paires d'entités, c'est-à-dire que nous ne comptons pas dans l'évaluation la classification des non-relations.

Giuliano *et al.* [2006] proposent deux façons d'évaluer l'extraction de relations : une réponse par occurrence dans le document (OAOD, *One Answer per Occurrence in the Document*) et une réponse par relation dans un document donné (OARD, *One Answer per Relation in a given Document*). Dans le premier cas, ils évaluent chaque occurrence des interactions ; dans le second ils évaluent les interactions, c'est-à-dire que si au moins une occurrence de l'interaction est correctement extraite, l'interaction est considérée comme correctement extraite. Toutes les évaluations que nous faisons correspondent au premier

cas proposé par Giuliano *et al.* [2006] : nous évaluons toutes les occurrences des relations dans le document.

L'évaluation proposée par les organisateurs, que nous avons choisi de suivre pour la suite de nos expérimentations, repose sur le calcul du rappel, de la précision et de la f-mesure pour chaque relation. Pour calculer le rappel et la précision d'un type de relation, nous utilisons le nombre de vrais positifs (`VP`), le nombre de faux positifs (`FN`) et le nombre de faux négatifs (`FP`) pour cette relation. Les formules sont les suivantes :

$$Rappel = \frac{VP}{VP + FN} \tag{4.4}$$

$$Précision = \frac{VP}{VP + FP} \tag{4.5}$$

Pour mesurer les performances globales du système (c'est-à-dire pour toutes les relations), nous calculons le rappel, la précision et la f-mesure en micro-moyenne. Ce type d'évaluation considère chaque classe de relations proportionnellement à son nombre d'instances. Les formules du rappel et de la précision en micro-moyenne sont données ci-dessous (`n` est le nombre de classes). La formule de calcul de la f-mesure reste la même, c'est-à-dire la moyenne harmonique du rappel et de la précision.

$$Rappel\ en\ micro - moyenne = \frac{\sum_{i=1}^{n} VP_i}{\sum_{i=1}^{n} VP_i + FN_i} \tag{4.6}$$

$$Précision\ en\ micro - moyenne = \frac{\sum_{i=1}^{n} VP_i}{\sum_{i=1}^{n} VP_i + FP_i} \tag{4.7}$$

Pour évaluer l'apport de différents paramétrages du système ou l'apport d'informations syntaxiques, etc. nous avons utilisé des tests statistiques pour savoir si les résultats obtenus étaient significativement différents. Le test le plus utilisé pour évaluer des systèmes de classification est le test de Student (ou test t) : c'est un ensemble de tests statistiques qui suivent la loi de Student quand l'hypothèse nulle est vraie. Il permet de comparer soit les moyennes observées sur deux échantillons soit deux probabilités.

Nous pouvons utiliser ce test sur les résultats de la classification binaire du système (c'est-à-dire dans le cas de l'extraction d'interactions entre protéines ou entre médicaments). Pour comparer les résultats de la classification multi-classes (sur la tâche i2b2 2010), nous ne pouvons pas utiliser nos résultats tels quels. Le test t s'utilise sur des données continues, or nos données ne sont ni continues, ni ordonnées. Pour appliquer le test à nos données, nous devons les transformer en données binaires : relations bien classées et relations mal classées, en se servant de la référence. Nous donnons un exemple dans le tableau 4.4 de la transformation des prédictions de deux classifieurs A et B en $A_{bi}$ et $B_{bi}$.

Uzuner *et al.* [2011] ont utilisé le test z pour mesurer si deux systèmes étaient significativement différents. Le test z est un ensemble de tests statistiques qui suivent une distribution normale quand l'hypothèse nulle est vraie. Le test z mesure la différence entre le taux d'erreur du système A ($p_A$) et le taux d'erreur du système B ($p_B$). La statistique

| A | B | ref | $A_{bi}$ | $B_{bi}$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 1 |
| 0 | 3 | 2 | 0 | 0 |
| 2 | 2 | 2 | 1 | 1 |
| 2 | 3 | 3 | 0 | 1 |
| 5 | 0 | 5 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 |

TAB. 4.4 – Transformation des prédictions en résultats binaires

z est calculée avec la formule suivante, où p est la moyenne des deux taux d'erreur et n le nombre d'instances :

$$z = \frac{p_A - p_B}{\sqrt{2p(1-p)/n}} \tag{4.8}$$

Nous pouvons calculer la statistique z sur nos données : pour cela nous devons compter le nombre d'instances de relations correctement classées par le système A mais pas par le système B, et inversement.

Par exemple, à partir du tableau de contingence 4.5 représentant les résultats de deux classifieurs A et B, nous pouvons calculer la statistique z de la façon suivante :
$p_A = (251 + 203)/2373 = 0,19$
$p_B = (251 + 40)/2373 = 0,12$
$p = (0,19 + 0,12)/2 = 0,155$
$z = \frac{0,19-0,12}{\sqrt{\frac{2*0,155*0,845}{2373}}} = 6,67$

z est supérieur à la valeur critique (seuil de confiance de 0,05) $Z_{0,975}$ = 1,96, donc l'hypothèse nulle est rejetée et les deux classifieurs sont significativement différents.

| | Mal classées par B | Bien classées par B |
|---|---|---|
| Mal classées par A | 251 | 203 |
| Bien classées par A | 40 | 1879 |

TAB. 4.5 – Tableau de contingence

Les conclusions que nous pouvons tirer de ces deux tests sur nos données sont identiques. De ce fait, nous n'utiliserons que le test de Student.

Une autre façon de montrer la différence entre deux systèmes est de combiner leurs résultats en donnant priorité à l'un des deux, puis de mesurer l'amélioration de la classification et voir si elle est significative. Cette méthode ne permet pas de déterminer si deux classifications sont significativement différentes, mais elle met en avant le fait que même si les rappels, précisions et f-mesures en micro-moyenne sont proches, ce ne sont pas les mêmes instances qui sont correctement classées ou pas.

Nous avons utilisé ces différentes méthodes d'évaluation et de comparaison au cours de nos expériences. Lorsque les f-mesures de deux classifieurs sont suffisamment différentes nous n'avons pas fait de tests de significativité.

## 4.2 Étude et modélisation des informations pour l'extraction des relations

Nous avons étudié quelles informations étaient pertinentes à utiliser pour classer les relations et la forme sous laquelle les représenter (sous forme vectorielle ou sous forme d'arbres). Nous avons développé un système, REMED, qui permet de détecter et de classer des relations. Nous avons fait varier les attributs représentant différentes informations pour évaluer l'apport de ces informations. Le système et les attributs utilisés ont été publiés dans Minard *et al.* [2011c,d].

L'architecture du système est présentée dans la figure 4.3. TRAIN est le corpus utilisé pour entraîner le classifieur et TEST est utilisé pour évaluer le modèle. Le système prend en entrée les textes du corpus et la référence pour l'annotation des concepts. Seules les phrases contenant au minimum deux entités sont conservées pour la suite des traitements. Elles sont prétraitées (cf. 4.2.1) et annotées en concepts à partir de la référence. Pour chaque paire d'entités susceptibles d'être en relation dans une phrase, les autres entités en coordination sont supprimées (cf. 4.2.2), puis des informations sont extraites et les vecteurs d'attributs à fournir au classifieur sont construits. À partir des vecteurs, le classifieur construit un modèle qui sera ensuite appliqué sur le corpus dans lequel on souhaite extraire des relations.

L'extraction des attributs est effectuée automatiquement et nécessite de prétraiter le corpus. Une analyse morpho-syntaxique est effectuée sur toutes les phrases du corpus avec le TreeTagger (Schmid [1994]). Nous disposons ainsi d'un découpage en tokens des phrases, des lemmes et des catégories morpho-syntaxiques de chaque token. Dans cette section, nous présentons les prétraitements effectués sur le corpus, puis le module de gestion de la coordination qui réduit les phrases contenant plusieurs entités coordonnées. Nous verrons ensuite les attributs utilisés pour représenter les informations contenues dans la phrase pouvant être utiles pour détecter et classer la relation. Nous avons mené différentes expérimentations pour mettre au point le système, comme l'évaluation des attributs. Après la mise au point du système, nous présentons son évaluation finale.

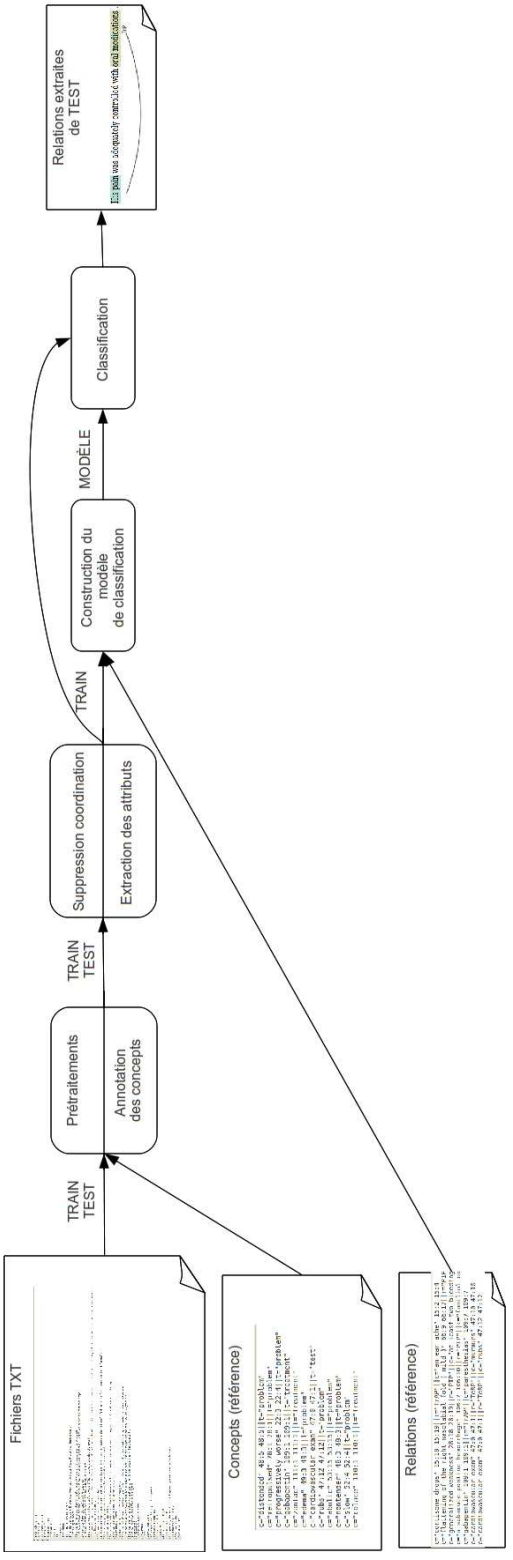



FIG. 4.3 – Schéma de fonctionnement du système d'extraction des relations

### 4.2.1 Prétraitements

Nous effectuons quelques prétraitements sur le corpus avant d'en extraire les relations. Le prétraitement commun à tous les corpus est le remplacement des abréviations par leur forme complète. Les abréviations connues ont été remplacées grâce à une liste constituée pour la campagne d'évaluation i2b2 2009 par Grouin *et al.* [2010]. Ils ont formé la liste à partir de la liste d'abréviations biomédicales construite par Berman [3] à laquelle ont été ajoutés les exemples du corpus du challenge i2b2 2009. Ainsi par exemple, *h.o.* a été converti en *history of* et *p.r.n.* en *as needed*.

Les comptes rendus médicaux avaient été anonymisés, en particulier les noms des patients et des médecins, les dates et les âges. Les données anonymisées sont différentes d'un corpus à l'autre. Par exemple les dates pouvaient être sous les formes suivantes : 2012-11-06, 06/07/2000, **DATE[Aug 16 1930], etc. Nous les avons remplacées par des balises *NAME*, *DATE* et *AGE*. La balise *NUM* est utilisée pour remplacer toutes les valeurs numériques présentes dans les comptes rendus (principalement des dosages).

### 4.2.2 Gestion de la coordination

Nous avons observé dans le corpus i2b2 2010 que beaucoup de phrases contenaient des énumérations de traitements et/ou problèmes médicaux. Du fait de ces énumérations, deux entités en relation peuvent être très éloignées dans la phrase et être séparées des mots marqueurs de la relation. Nous avons donc ajouté un prétraitement pour supprimer les entités coordonnées avec une des deux entités candidates à la relation. Pour cela, nous avons écrit des règles qui repèrent si des entités du même type sémantique (pour le corpus i2b2, les types sont examen test, traitement et problème médical) sont séparées par *and*, *or*, *,* *and* et *,*. Dans l'exemple (27), sept entités de type problème médical ont été supprimées entre *any symptoms* et *palpitations* et une après *palpitations*.

(27) The patient was instructed to contact her primary care physician and / or contact the emergency room if she had [PB **any symptoms**] including but not limited to [PB prolonged fever] , [PB nausea] , [PB vomiting] , [PB chills] , [PB night sweats] , [PB chest pain] , [PB shortness of breath] , [PB **palpitations**] or [PB any other serious complaints] .
The patient was instructed to contact her primary care physician and / or contact the emergency room if she had [PB **any symptoms**] including but not limited to [PB **palpitations**] .

Dans le corpus TRAIN issu du corpus i2b2 2010, il y a 3 745 phrases contenant plus de deux entités susceptibles d'être en relation et 14 097 paires d'entités. Pour 46% de ces paires, au moins une des entités était coordonnée avec d'autres entités du même type.

La présence d'entités en coordination dans le corpus original est une information qui pourra être utile à conserver pour la classification des relations. L'étape qui suit les pré-

[3] http://www.julesberman.info/abbtwo.htm

traitements du corpus, est la représentation sous forme vectorielle des informations disponibles pour chaque paire d'entités, comme l'indication que des entités ont été supprimées. En effet, ne fournir qu'un vecteur représentant les mots de la phrase n'est pas suffisant. Nous verrons dans la section suivante quelles sont les informations que nous avons choisi de représenter.

### 4.2.3 Les attributs ou comment représenter le contenu de l'information sous forme vectorielle

Le classifieur utilisé par notre système construit un modèle d'apprentissage à partir des données d'entraînement représentées sous forme vectorielle. Pour chaque paire d'entités, un vecteur d'attributs est construit. Nous avons défini des attributs permettant de représenter les informations décrivant les instances de relations et de non-relations. Ils peuvent être divisés en 4 classes : les attributs de surface, les attributs lexicaux, les attributs syntaxiques et les attributs sémantiques. Nous avons ajouté une 5$^{e}$ catégorie d'attributs : des attributs liés à la gestion de la coordination. Certains des attributs que nous utilisons sont repris de Zhou *et al.* [2005] et Roberts *et al.* [2008] mais l'ensemble des attributs choisis constitue un modèle original.

L'analyse morpho-syntaxique du TreeTagger permet d'avoir un découpage en mots des phrases (les mots incluent aussi les signes de ponctuation) et pour chaque mot, les lemmes et leur catégorie morpho-syntaxique. Il est ainsi possible de repérer les verbes, les prépositions, etc. dans les phrases.

Nous avons vu dans l'état de l'art qu'une définition du contexte est nécessaire. Nous avons choisi de suivre l'approche de Zhou *et al.* [2005] (repris par Roberts *et al.* [2008]) qui consiste à utiliser un contexte de deux mots (ou plus) avant la première entité (E1), deux mots (ou plus) après la deuxième entité (E2) et les mots entre E1 et E2. Après avoir fait des tests, nous avons défini le contexte local de la relation comme une fenêtre allant de 3 mots avant E1 à 3 mots après E2, en incluant les mots entre les entités (cf. figure 4.4). Avec une fenêtre plus grande ou plus petite, la précision augmente légèrement mais le rappel diminue.

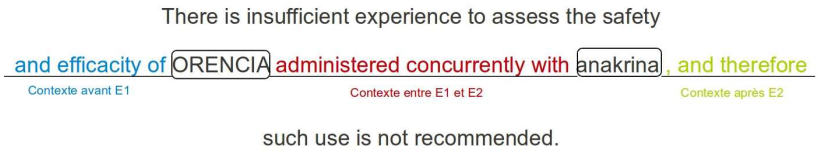



FIG. 4.4 – Contexte local de la relation

Chaque attribut a un identifiant unique, qui a pour valeur un s'il apparaît, zéro sinon. Seuls trois attributs peuvent prendre des valeurs différentes : la distance entre E1 et E2, le nombre d'entités présentes entre les deux entités candidates et le nombre d'entités supprimées lors de la suppression des entités en coordination.

#### 4.2.3.1 Attributs de surface

Les attributs surfaciques concernent la position de E1 et E2 dans la phrase, ainsi que la présence d'autres entités.

- L'ordre des concepts : permet de savoir si l'entité de type t1 est placée avant ou après l'entité de type t2 dans la phrase, ce qui influera sur la façon dont est exprimée la relation. Dans l'exemple (28) le test (*an MRI*) est donné avant les problèmes qu'il a révélés. L'exemple (29) montre un cas où le problème est cité avant le test, ce dernier étant suivi des résultats obtenus (<*NUM*>).

  (28) She had [TEST a workup] by her neurologist and [TEST an MRI] revealed [PB a C5-6 disc herniation] with [PB cord compression] and [PB a T2 signal change] at that level.

  (29) The patient was [PB thrombocytopenic] with [TEST a platelet count] of <NUM> on the <NUM>.

- La distance entre E1 et E2 en termes de nombre de mots : deux entités séparées par plus de 15 mots auront une probabilité moins élevée d'être en relation que deux entités éloignées de 3 mots.
- La présence d'autres entités entre E1 et E2 : dans l'exemple (30), cinq relations sont envisagées, chaque problème peut être en relation avec chaque traitement, et les problèmes peuvent être également en relation. Mais le fait qu'il y ait un traitement entre le premier traitement et le deuxième problème élimine la possibilité d'avoir une relation entre ces derniers.

  (30) Take with food [TREAT MILK OF MAGNESIA (MAGNESIUM HYDROXIDE)] <NUM> MILLILITERS per oral twice daily as needed [PB Constipation] [TREAT PERCOCET] <NUM> tablet per oral Q4H as needed [PB PAIN]

#### 4.2.3.2 Attributs lexicaux

Les attributs lexicaux portent les informations sur les mots dans le contexte local de E1 et E2.

- Les mots, et leurs lemmes, qui désignent E1 et E2, et le mot tête de E1 et E2 [4]. Nous avons considéré les lemmes de manière à regrouper les variantes flexionnelles. Les mots qui forment les entités peuvent être déclencheurs d'une relation : dans l'exemple (31) la relation entre le traitement et le problème est de type *le traitement aggrave le problème* (TrWP) et est déclenchée par l'adjectif *recurrent* à l'intérieur de la deuxième entité.

  (31) He has had <NUM> week courses of [TREAT antibiotics] with [PB recurrent bacteremia].

[4] La tête d'un concept correspond au mot précédant une préposition ou le dernier mot du concept, comme défini dans Zhou *et al.* [2005].

- Les lemmes dans le contexte local de la relation : c'est entre E1 et E2 qu'est située l'information la plus utile à la classification.
- Les lemmes des verbes dans le contexte local de la relation. Le verbe marque souvent la relation : par exemple dans (32) la relation de type *le test révèle le problème* (TeRP) est exprimée par le verbe *reveal*.

  (32) [TEST Chest x-rays] **revealed** evidence of [PB congestive heart failure].
- Les prépositions entre E1 et E2. Dans l'exemple (33) la préposition *for* marque une relation de type *le traitement est administré en raison du problème* (TrAP).

  (33) She was treated with [TREAT IVF] **for** [PB her ARF].

#### 4.2.3.3 Attributs morpho-syntaxiques

Les attributs syntaxiques permettent d'utiliser des informations relatives à la structure syntaxique de surface des phrases pour repérer les relations. Nous prenons en compte la succession des catégories syntaxiques et la présence d'éléments d'une certaine catégorie, indifféremment de leur réalisation lexicale. Les attributs syntaxiques considérés sont donc les suivants :

- La catégorie morpho-syntaxique des mots dans le contexte local de la relation.
- La présence d'une préposition entre E1 et E2, quelle que soit cette préposition.
- Un attribut marque la présence d'un signe de ponctuation lorsqu'il est le seul présent entre E1 et E2. Cet attribut permet de tenir compte des énumérations qui ne correspondent pas aux cas traités par le module de gestion de la coordination.

#### 4.2.3.4 Attributs sémantiques

Ces attributs permettent de généraliser l'information portée par certains mots des phrases et concernent les concepts du domaine d'une part et les classes de verbes d'autres part.

- Le type sémantique (issu de l'UMLS) des mots dans le contexte local de la relation. Dans l'exemple (28), le terme *neurologist* a pour type sémantique *professional or occupational group* ; un attribut l'indiquera pour la classification des paires contenant le terme *neurologist* dans leur contexte local.
- Les types de E1 et E2 (protéine, gène, traitement, test, problème, etc.) : cet attribut est indispensable pour une tâche d'extraction de relations entre des entités de types différents, par exemple si la tâche porte sur l'extraction de relations entre des traitements et des maladies ainsi qu'entre des examens et des maladies, etc. En effet, les classes de relations seront différentes selon les types des entités.
- Les classes de VerbNet [5] (une extension des classes de Levin) des verbes dans le contexte local de la relation. Par exemple le verbe *reveal* fait partie de la classe *indicate-78-1-1* qui contient également les verbes *show*, *prove*, *demonstrate*, etc. Nous observons que dans les exemples (32) et (34) les verbes *reveal* et *show* sont utilisés pour marquer des relations du même type.

[5] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

(34) [TEST Recent chest x-ray] **shows** [PB resolving right lower lobe pneumonia].

#### 4.2.3.5 Coordination

Les entités en coordination avec E1 et E2 sont supprimées (voir la partie 4.2.2 pour plus de détails). Suite à ce prétraitement, sept attributs ont été ajoutés : trois qui indiquent que des entités ont été supprimées avant E1, entre E1 et E2 et après E2, trois autres qui indiquent le nombre d'entités supprimées avant E1, entre E1 et E2 et après E2, et un attribut qui renseigne sur les mots déclencheurs de la suppression des entités (*or* / *and* / ,).

Dans l'exemple (27) page 109, pour la paire d'entités *any symptoms* et *palpitations*, les attributs ajoutés sont présentés dans le tableau 4.6.

| Attribut | Valeur |
|---|---|
| Nombre d'entités supprimées avant E1 | 0 |
| Nombre d'entités supprimées entre E1 et E2 | 7 |
| Nombre d'entités supprimées après E2 | 1 |
| Suppression avant E1 | 0 |
| Suppression entre E1 et E2 | 1 |
| Suppression après E2 | 1 |
| Mots déclencheurs | , / *or* |

TAB. 4.6 – Valeurs des attributs représentant les informations concernant la gestion de la coordination dans la phrase (27)

### 4.2.4 Étude de la pertinence des attributs

Les attributs et le module de gestion de la coordination ont été proposés suite à une étude du corpus. Pour vérifier leur utilité, nous avons mené différentes évaluations que nous présentons dans cette partie. Elles ont conduit à la mise au point du système dans sa version finale.

#### 4.2.4.1 Évaluation de la gestion de la coordination

Le module de gestion de la coordination modifie les phrases à partir desquelles nous construisons les vecteurs d'attributs. Nous avons donc évalué à part la suppression des coordinations et les attributs ajoutés aux vecteurs. Pour cela, nous avons utilisé le système avec tous les attributs décrits précédemment. Les résultats de l'évaluation sont présentés dans le tableau 4.7 et 4.8. Le système a été évalué sans l'utilisation du module de gestion de la coordination (*sans COORD*), avec le module mais sans les attributs spécifiques à la coordination (*avec COORD (+), sans attributs*) et avec le module et les attributs (*avec COORD (+), avec attributs*). Le module de gestion de la coordination peut supprimer les entités en coordination avec E1 et E2 qui sont situées entre E1 et E2, ou supprimer toutes les entités en coordination avec E1 et E2. Le premier cas correspond à ce que nous avons appelé *avec COORD* et le second *avec COORD +*. Dans les deux premières colonnes du

tableau 4.7, on indique la proportion de phrases dans lesquelles des entités en coordination avec E1 ou E2 ont été supprimées.

| Relation | phrases réduites | | sans COORD | avec COORD | | avec COORD + | |
|---|---|---|---|---|---|---|---|
| | TRAIN | TEST | | sans att | avec att | sans att | avec att |
| **TrIP** | 43% | 33% | 0,235 | 0,300 | **0,380** | 0,300 | 0,300 |
| **TrWP** | 41% | 29% | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| **TrCP** | 27% | 36% | 0,516 | 0,507 | 0,507 | **0,566** | 0,539 |
| **TrAP** | 27% | 26% | 0,751 | 0,750 | **0,770** | 0,725 | 0,748 |
| **TrNAP** | 36% | 40% | **0,666** | 0,645 | 0,645 | 0,645 | 0,645 |
| **PIP** | 31% | 31% | 0,695 | 0,687 | **0,703** | 0,675 | 0,699 |
| **TeRP** | 23% | 23% | 0,875 | **0,877** | **0,877** | 0,869 | 0,872 |
| **TeCP** | 20% | 28% | 0,500 | **0,530** | 0,512 | 0,461 | 0,500 |
| | **27%** | **28%** | **0,750** | **0,750** | **0,758** | **0,738** | **0,750** |

TAB. 4.7 – Variation de la f-mesure lors de l'utilisation du module de gestion de la coordination sur le corpus de test

Les meilleurs résultats sont obtenus en supprimant les entités en coordination entre E1 et E2 et en prenant en compte les attributs qui indiquent la suppression des entités. Avec COORD, 38 paires d'entités ont été mieux classées que sans COORD. 22 de ces 38 paires proviennent de phrases qui ont été réduites (dans lesquelles des entités en coordination avec E1 et E2 ont été supprimées). En revanche, 26 paires ont été mal classées avec COORD alors que ce n'était pas le cas sans COORD, et pour 4 d'entre elles la phrase avait été réduite. Dans le corpus utilisé pour l'apprentissage les phrases contenant des coordinations ont également été réduites, ce qui explique que des paires d'entités provenant de phrases non réduites puissent être moins bien classées.

Dans le tableau 4.8, nous donnons le rappel, la précision et la f-mesure pour la classification de toutes les relations. On observe que l'utilisation des attributs indiquant entre autres le nombre d'entités supprimées permet d'augmenter la précision du système.

| | sans COORD | avec COORD | | avec COORD + | |
|---|---|---|---|---|---|
| | | sans att | avec att | sans att | avec att |
| Rappel | 0,691 | 0,711 | 0,708 | 0,698 | 0,700 |
| Précision | 0,820 | 0,793 | 0,815 | 0,782 | 0,807 |
| F-mesure | 0,750 | 0,750 | 0,758 | 0,738 | 0,750 |

TAB. 4.8 – Variation du rappel, de la précision et de la f-mesure lors de l'utilisation du module de gestion de la coordination pour toutes les relations sur le corpus de test

#### 4.2.4.2 Évaluation des attributs

Pour évaluer l'utilité de chaque attribut, nous avons suivi la même méthode que Roberts *et al.* [2008]. Nous avons mesuré les performances du système sur le corpus de test en

utilisant des attributs de base et en ajoutant une série d'attributs à la fois. Nous avons divisé nos attributs de deux façons différentes : selon le type des informations (informations lexicales, surfaciques, syntaxiques, sémantiques et liées à la gestion de la coordination) et selon les informations qu'ils décrivent (des informations sur les autres entités présentes dans la phrase, sur les verbes, etc.). Les résultats sont donnés dans les tableaux 4.9 et 4.11. Dans les deux évaluations, le module de gestion de la coordination est utilisé et ne supprime que les entités en coordination entre E1 et E2.

L'ajout de certaines classes d'attributs peuvent améliorer la classification d'une relation mais détériorer celle d'une autre relation. Par exemple l'ajout des attributs sémantiques fait chuter la f-mesure pour la classe TrIP de 0,12 points, mais permet d'augmenter la f-mesure de la classe TrCP et TrNAP de 0,02 points. La plus grande amélioration que l'on observe est obtenue avec l'ajout de la classe des attributs morpho-syntaxiques qui contient les catégories morpho-syntaxiques des mots dans le contexte de E1 et E2 ainsi que la présence de signes de ponctuation, de conjonctions de coordination ou de prépositions entre E1 et E2.

| Relation | Attributs | | | | |
|---|---|---|---|---|---|
| | lexicaux | + coordination | + sémantiques | + surfaciques | + morpho-synt |
| **TrIP** | **0,571** | **0,571** | 0,454 | 0,454 | 0,380 |
| **TrWP** | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| **TrCP** | 0,508 | 0,491 | **0,516** | **0,516** | 0,507 |
| **TrAP** | 0,744 | 0,762 | **0,770** | 0,768 | **0,770** |
| **TrNAP** | 0,645 | 0,645 | **0,666** | **0,666** | 0,645 |
| **PIP** | 0,602 | 0,627 | 0,635 | 0,640 | **0,703** |
| **TeRP** | 0,860 | 0,866 | 0,860 | 0,867 | **0,877** |
| **TeCP** | 0,481 | 0,475 | 0,493 | 0,475 | **0,512** |
| | 0,728 | 0,738 | 0,740 | 0,742 | **0,758** |

TAB. 4.9 – Variation de la f-mesure selon les attributs utilisés sur le corpus de test

Dans le tableau 4.10, nous présentons l'évaluation pour toutes les relations. On observe que l'ajout des attributs morpho-syntaxiques permet d'améliorer le rappel, alors que l'ajout des attributs de coordination, comme nous l'avons vu précédemment, permet d'améliorer la précision du système.

| | lexicaux | + coordination | + sémantiques | + surfaciques | + morpho-synt |
|---|---|---|---|---|---|
| Rappel | 0,682 | 0,686 | 0,686 | 0,687 | **0,708** |
| Précision | 0,782 | 0,799 | 0,804 | 0,807 | **0,815** |
| F-mesure | 0,728 | 0,738 | 0,740 | 0,742 | **0,758** |

TAB. 4.10 – Variation du rappel, de la précision et de la f-mesure selon les attributs utilisés pour toutes les relations sur le corpus de test

Pour la deuxième évaluation dont les résultats sont présentés dans le tableau 4.11, les classes d'attributs utilisées par le système initial *base* sont les lemmes et les catégories

morpho-syntaxiques du contexte de E1 et E2. Sont ensuite ajoutées les classes d'attributs *dist* (distance entre les deux entités et position de E1 par rapport à E2), *conc* (autres entités entre E1 et E2, et entités supprimées), *verb* (lemmes des verbes et classes de Verbnet), *prep* (prépositions entre E1 et E2, conjonctions de coordination et ponctuations), *intra* (mots qui forment E1 et E2, et mot tête de E1 et E2) et *types* (types sémantiques de l'UMLS). Les résultats de la dernière colonne du tableau correspondent aux résultats du système final.

Quand nous ajoutons la série d'attributs de la classe *types*, les f-mesures de la classification des relations PIP, TeRP et TeCP augmentent alors que les f-mesures pour TrIP et TrCP diminuent. Nous gardons cette classe d'attribut car elle permet d'augmenter la f-mesure générale. De même l'ajout de *dist* augmente d'environ 0,007 la f-mesure pour la relation TrAP, mais fait chuter de 0,017 la f-mesure pour la relation TeCP. La f-mesure générale reste stable, mais si nous retirons cette classe d'attribut du système final, la f-mesure diminue légèrement (de 0,758 à 0,756).

| Relation | base | +dist | +conc | +verb | +prep | +intra | +types |
|---|---|---|---|---|---|---|---|
| **TrIP** | 0,117 | 0,117 | 0,222 | 0,380 | 0,380 | **0,571** | 0,380 |
| **TrWP** | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| **TrCP** | 0,508 | 0,500 | 0,517 | **0,548** | 0,515 | 0,545 | 0,507 |
| **TrAP** | 0,726 | 0,733 | 0,751 | 0,748 | 0,768 | **0,770** | 0,770 |
| **TrNAP** | 0,400 | 0,400 | 0,400 | 0,400 | 0,461 | **0,645** | 0,645 |
| **PIP** | 0,609 | 0,614 | 0,642 | 0,649 | 0,678 | 0,685 | **0,703** |
| **TeRP** | 0,839 | 0,836 | 0,844 | 0,851 | 0,856 | 0,868 | **0,877** |
| **TeCP** | 0,378 | 0,361 | 0,361 | 0,421 | 0,461 | 0,469 | **0,512** |
| **Toutes relations** | 0,704 | 0,705 | 0,720 | 0,727 | 0,740 | 0,753 | **0,758** |

TAB. 4.11 – Variation de la f-mesure selon les attributs utilisés sur le corpus de test

Dans le tableau 4.12, on observe que la précision du système est améliorée avec l'utilisation des attributs décrivant les deux entités E1 et E2 et des attributs portant de l'information sur les autres entités (ce qui valide ce que nous avons observé dans les deux précédentes évaluations). L'ajout des attributs *prep* fait légèrement diminuer la précision mais permet d'augmenter le rappel.

| | base | +dist | +conc | +verb | +prep | +intra | +types |
|---|---|---|---|---|---|---|---|
| Rappel | 0,655 | 0,656 | 0,668 | 0,675 | 0,699 | 0,702 | **0,708** |
| Précision | 0,761 | 0,761 | 0,780 | 0,788 | 0,786 | 0,812 | **0,815** |
| F-mesure | 0,704 | 0,705 | 0,720 | 0,727 | 0,740 | 0,753 | **0,758** |

TAB. 4.12 – Variation du rappel, de la précision et de la f-mesure selon les attributs utilisés pour toutes les relations sur le corpus de test

Cette manière d'évaluer les attributs a des limites. Par exemple, nous ne pouvons pas conclure de cette étude que les attributs *conc* sont plus utiles que les attributs *verb*, mais uniquement que les deux classes d'attributs sont importantes pour la classification des

relations. De plus l'apport de chaque classe d'attributs ne peut pas être quantifié, en effet l'ordre selon lequel les classes sont ajoutées modifie le gain en terme de f-mesure. Par exemple, l'augmentation des performances du système avec l'ajout de la classe *prep* ne sera pas aussi importante si la classe est ajoutée en dernière.

#### 4.2.4.3 Sélection d'attributs

L'évaluation précédente montre que les attributs se comportent différemment selon le type de la relation. De ce fait, nous avons effectué une autre sélection des attributs en évaluant le caractère discriminant de chaque attribut en fonction de chaque classe. Ce travail a été effectué dans le cadre du stage de Lamia Makour. Nous avons choisi d'utiliser une méthode basée sur le score de Fisher (aussi appelé F-score) comme dans Chen et Lin [2006], que nous avons étendu à un problème multi-classes. L'avantage de ce critère est qu'il est simple et efficace et qu'il a donné de bons résultats sur plusieurs corpus, avec des classifieurs à base de SVM.

Soit $x_k$ les vecteurs construits pour chaque instance du corpus d'entraînement (avec k le nombre d'instances, $k = 1, ..., m$) ; $n_+$ et $n_-$ le nombre d'instances positives et négatives. Le F-score du i$^e$ attribut est défini par l'équation 4.9.

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \tag{4.9}$$

où $\bar{x}_i$, $\bar{x}_i^{(+)}$ et $\bar{x}_i^{(-)}$ sont les moyennes du i$^e$ attribut respectivement dans tout le corpus, pour les instances positives et pour les instances négatives ; $x_{k,i}^{(+)}$ et $x_{k,i}^{(-)}$ sont le i$^e$ attribut de la k$^e$ instance positive ou négative. Plus le F-score est grand, plus l'attribut est discriminant.

Nous avons utilisé l'outil *fselect.py*, fourni avec la bibliothèque libSVM. Nous avons divisé les attributs en classes, comme présenté dans la section 4.2.3. Cette classification permet d'analyser l'utilité des attributs selon le niveau linguistique qu'ils encodent. Nous avons défini différents seuils en dessous desquels nous avons supprimé les attributs classe par classe. Nous avons fait des tests avec différentes combinaisons de seuils pour chaque classe d'attributs.

Le F-score calculé sur toutes les classes ne permet pas de trouver des seuils intéressants, parce que les attributs qui n'ont pas le même impact sur chaque classe obtiennent des poids analogues. De ce fait, nous avons étudié la sélection d'attributs pour l'apprentissage de chaque relation individuellement et pour l'apprentissage des relations par paires d'entités du même type (entre un test et un problème (Te-P), entre un traitement et un problème (Tr-P), et entre deux problèmes (PIP)). Les données sont divisées en trois selon le type des entités.

Dans le premier cas, nous avons construit un modèle d'apprentissage pour chaque relation, avec les instances positives qui sont les instances de cette relation et les instances négatives qui sont toutes les instances des autres classes de relations intervenant entre deux entités du même type (pour la relation TrAP, ce sera toutes les instances des relations TrIP, TrWP, TrNAP et TrCP) et les instances de non-relations entre deux entités du même type

(pour la relation TrAP, toutes les paires Traitement - Problème qui ne sont pas en relation). Une fois les modèles de classification appliqués sur les corpus de test, un vote est effectué pour décider de la meilleure classification pour une paire d'entités avec priorité donnée aux classes peu représentées (TeCP, TrIP, TrWP et TrNAP). Cette méthode est proche de la méthode de classification multi-classes de type « one-versus-all ».

Dans le second cas, trois modèles d'apprentissage ont été construits : un modèle multi-classes pour les paires Traitement - Problème, un modèle multi-classes pour les paires Test - Problème et un modèle bi-classes pour les paires Problème - Problème.

| Relation | Morpho-syntaxique | Lexical | Sémantique | Avant sélection F-mesure (# attributs) | Après sélection F-mesure (# attributs) |
|---|---|---|---|---|---|
| | Apprentissage bi-classes | | | | |
| **TrIP** | 0,001 | 0,000001 | 0,0001 | 0,250 (9 227) | 0,250 (8 363) |
| **TrWP** | 0 | 0 | 0 | 0,000 (9 227) | 0,000 (9 227) |
| **TrCP** | 0,001 | 0 | 0,001 | 0,551 (9 227) | **0,586** (8 318) |
| **TrAP** | 0,0001 | 0,00001 | 0,0001 | 0,790 (9 227) | **0,805** (8 746) |
| **TrNAP** | 0,0001 | 0,001 | 0,0001 | 0,620 (9 227) | **0,666** (770) |
| **PIP** | 0,000001 | 0 | 0,001 | 0,689 (10 238) | **0,705** (9 781) |
| **TeRP** | 0,0001 | 0,00001 | 0 | 0,881 (8 324) | **0,892** (7 683) |
| **TeCP** | 0,00001 | 0 | 0,001 | 0,441 (8 324) | 0,441 (7 923) |
| **Toutes les rel** | | | | 0,756 | **0,772** |
| | Apprentissage multi-classes (par type) | | | | |
| **Tr-P** | 0,000001 | 0,000001 | 0,000001 | 0,718 (9 939) | **0,721** (9 227) |
| **Te-P** | 0,0001 | 0,00001 | 0,001 | 0,829 (8 324) | **0,832** (8 070) |
| **PIP** | 0,000001 | 0 | 0,001 | 0,689 (10 139) | **0,705** (9 781) |
| **Toutes les rel** | | | | 0,757 | **0,763** |

TAB. 4.13 – F-mesures obtenues avec la sélection d'attributs sur le corpus de test

Les résultats obtenus avec ces deux méthodes et les meilleurs combinaisons de seuils sont présentés dans le tableau 4.13. Les lignes de résultats « Toutes les rel » contiennent les f-mesures en micro-moyenne (cf. 4.1.5). Tous les attributs des classes d'attributs surfaciques

et liés à la coordination sont conservés, car ces deux classes contiennent peu d'attributs (6 attributs surfaciques et 11 attributs liés à la coordination) et qu'en supprimer fait baisser les performances du système. Nous observons une amélioration générale de la f-mesure après sélection des attributs. Le seuil de suppression des attributs lexicaux est souvent bas (voire nul), ce qui montre que même les attributs les moins discriminants sont utiles. Au contraire, le seuil pour les attributs sémantiques est élevé, c'est-à-dire que seuls les attributs les plus discriminants sont conservés. Enfin, nous pouvons voir que pour les classes de relations peu représentées (TrIP, TrWP et TeCP) la sélection d'attributs ne permet pas d'améliorer la détection, sauf pour la classe TrNAP.

Nous avons ensuite utilisé les meilleurs seuils pour sélectionner les attributs pour la classification des relations du corpus d'évaluation. Dans le tableau 4.14, nous donnons la f-mesure en micro-moyenne. Une légère amélioration de la classification est observée après sélection des attributs dans les deux cas.

| | Avant sélection | Après sélection |
|---|---|---|
| Apprentissage multi-classes (par type) | 0,712 | 0,713 |
| Apprentissage bi-classes | 0,700 | 0,707 |

TAB. 4.14 – F-mesure obtenue avec la sélection d'attributs sur le corpus d'évaluation avec les meilleurs seuils trouvés sur le corpus de test

### 4.2.5 Évaluation de REMED

Le système présenté précédemment avec le module de gestion de la coordination a été utilisé pour extraire les relations du corpus d'évaluation. Tous les attributs présentés précédemment ont été conservés. Nous avons utilisé un noyau RBF, et nous avons fixé la valeur du facteur de pénalité à 16 et celle du paramètre $\gamma$ à 0,03125. Pour fixer ces valeurs, nous avons utilisé l'outil grid.py fourni avec libSVM et nous avons effectué une série de tests. Les résultats obtenus sont présentés dans le tableau 4.15. Ces résultats ont permis à notre système d'être classé en 3$^{e}$ position sur 16 au challenge i2b2 2010 (dans le tableau nous indiquons également les résultats obtenus par les meilleurs systèmes). La précision est supérieure à 0,67 pour chaque relation et est de 0,80 en micro-moyenne sur toutes les relations. Cela signifie que 80% des relations extraites sont correctement classées. En revanche, le rappel est bas pour les relations pour lesquelles il y a peu d'exemples dans le corpus (TrIP, TrCP, TrNAP et TeCP). En micro-moyenne sur toutes les relations, le rappel est de 0,63, c'est-à-dire que 63% des relations à extraire ont été extraites et correctement classées. Pour la relation TeRP, présentant le plus grand nombre d'exemples dans le corpus d'entraînement, la f-mesure atteint 0,850. En revanche, les relations moins bien dotées en exemples sont moins bien reconnues : pour la relation TrCP, la f-mesure n'est que de 0,491. Dans le corpus d'entraînement, il n'y a que 56 instances de la relation TrWP ; avec si peu d'exemples le classifieur n'apprend pas à classer correctement ces relations.

Le système qui est arrivé premier au challenge a été développé par Rink *et al.* [2011]. Il est basé sur un classifieur à base de SVM avec un noyau linéaire. Les particularités de leur système est l'utilisation de Wikipedia (est-ce qu'il existe un lien entre la page portant sur

E1 et celle portant sur E2, etc.), l'annotation des phrases en rôles sémantiques ainsi que le calcul de similarité entre des séries d'attributs. Le système de de Bruijn *et al.* [2011], arrivé deuxième, repose sur des techniques d'apprentissage fondées sur des modèles d'entropie maximale. Il utilise l'annotation des concepts de l'UMLS avec MetaMap (Aronson [2001]) ainsi que l'annotation des entités nommées cliniques faite par l'outil cTAKES (Savova *et al.* [2008]). Les auteurs ont travaillé sur le problème du déséquilibre des données et ont réduit le nombre d'exemples négatifs du corpus d'entraînement. Notre système est meilleur en précision que ces deux systèmes mais moins bon en rappel.

| Relation | Rappel | Précision | F-mesure |
|---|---|---|---|
| **TrIP** | 0,166 | 0,868 | 0,279 |
| **TrWP** | 0,000 | 0,000 | 0,000 |
| **TrCP** | 0,378 | 0,702 | 0,491 |
| **TrAP** | 0,705 | 0,752 | 0,728 |
| **TrNAP** | 0,109 | 0,677 | 0,189 |
| **PIP** | 0,544 | 0,763 | 0,635 |
| **TeRP** | 0,828 | 0,873 | 0,850 |
| **TeCP** | 0,260 | 0,805 | 0,393 |
| **Toutes relations** | **0,631** | **0,803** | **0,707** |
| Médiane | | | 0,664 |
| 1$^{e}$ système | 0,753 | 0,720 | 0,736 |
| 2$^{e}$ système | 0,693 | 0,773 | 0,731 |
| 4$^{e}$ système | 0,675 | 0,730 | 0,701 |

TAB. 4.15 – Rappel, précision et f-mesure obtenus sur le corpus d'évaluation

**Analyse des erreurs** Pour analyser les erreurs, nous avons construit des matrices de confusion. Le tableau 4.16 présente la matrice de confusion pour les relations entre un traitement et un problème, le tableau 4.17 pour les relations entre un test et un problème et le tableau 4.18 pour les relations entre deux problèmes. Les informations en ligne correspondent à la référence et celles en colonne à la classification réalisée par notre système. Par exemple dans le corpus, 200 instances de relations sont du type TrIP, 33 ont été correctement classées par le système, 58 ont été typées TrAP, 7 ont été typées TrCP et 102 n'ont pas été détectées comme une relation. Le système a extrait 38 relations de type TrIP : 33 sont correctes, une est en fait de type TrWP, trois de type TrAP et une n'est pas une relation. On observe que beaucoup de paires sont classées à tort dans la classe TrAP, la classe la mieux représentée dans le corpus pour les relations entre un problème et un traitement. Par exemple, 29% des relations TrIP sont identifiées à tort comme des relations TrAP. Une des raisons est le déséquilibrage des classes dans le corpus, mais aussi le caractère générique de la relation TrAP. En effet, la relation TrAP existe lorsqu'un traitement a été administré pour soigner un problème médical, mais si un traitement améliore un problème ou l'aggrave (relation TrIP et TrWP) cela signifie qu'il a été administré pour soigner le problème. Pour contrer ce problème, il pourrait être intéressant de modifier la façon dont le vote est fait entre les prédictions des classifieurs au sein du classifieur multi-classes. Par

exemple, un vote donnant la priorité aux classes peu représentées et ne prenant pas en compte le score de décision du classifieur, pourrait permettre de donner plus d'importance aux classes ayant peu d'instances.

| | TrIP | TrWP | TrAP | TrNAP | TrCP | Non-relation | Total référence |
|---|---|---|---|---|---|---|---|
| TrIP | **33** | 0 | 58 | 0 | 7 | 102 | 200 |
| TrWP | 1 | **0** | 41 | 0 | 7 | 96 | 145 |
| TrAP | 3 | 0 | **1760** | 3 | 17 | 678 | 2461 |
| TrNAP | 0 | 0 | 57 | **21** | 10 | 99 | 187 |
| TrCP | 0 | 0 | 84 | 3 | **170** | 189 | 446 |
| Non-relation | 1 | 0 | 331 | 4 | 28 | **3124** | 3488 |
| Total classifieur | 38 | 0 | 2331 | 31 | 239 | 4290 | 6927 |

TAB. 4.16 – Matrice de confusion pour les relations entre traitement et problème

| | TeCP | TeRP | Non-relation | Total référence |
|---|---|---|---|---|
| TeCP | **154** | 81 | 349 | 584 |
| TeRP | 20 | **2523** | 473 | 3016 |
| Non-relation | 16 | 275 | **2164** | 2455 |
| Total classifieur | 190 | 2879 | 2986 | 6055 |

TAB. 4.17 – Matrice de confusion pour les relations entre test et problème

| | PIP | Non-relation | Total référence |
|---|---|---|---|
| PIP | **1087** | 898 | 1985 |
| Non-relation | 327 | **10 848** | 11 175 |
| Total classifieur | 1414 | 11 746 | 13 160 |

TAB. 4.18 – Matrice de confusion pour les relations entre deux problèmes

Nous avons dégagé 3 types d'erreurs dans les prédictions du classifieur :

– La relation est clairement exprimée par un verbe ou une expression, mais cette construction n'est pas présente dans le corpus d'entraînement. Dans (35), la relation entre *pulmonary nodules in his RML* et *fu imaging* a été étiquetée TeRP (*le test révèle un problème*) : en effet le verbe *reveal* est déclencheur d'une relation TeRP, et le déclencheur d'une relation TeCP est *which need* mais il n'apparaît qu'une seule fois dans le corpus d'entraînement. Dans (36), la relation TeRP (*le test révèle un problème*) n'est pas reconnue entre *pressures* et *hypertension*, alors que la formulation n'est pas très complexe ; le problème doit provenir du faible nombre d'exemples de ce type.

(35) [TEST CTS chest] was negative for [PB PE], however it did reveal [PB pulmonary nodules in his RML] which need [TEST fu imaging] in <NUM> months.

(36) [TEST His PA catheter] is showing [TREAT pressures] of about <NUM> and I suspect that he has [PB chronic pulmonary hypertension], given these numbers.

– La relation est ambiguë, aucun élément permet de typer à coup sûr la relation. Par exemple dans (37), le système reconnaît à tort une relation TrAP entre *ventilator* et *congestion* car la phrase est ambiguë. Et dans (38), le système détecte à tort une relation TrAP (*le traitement est administré en raison du problème*) entre *Mesalamine* et *flares* parce que cette relation est effectivement présente dans la phrase et partage le même contexte.

(37) He does have some signs of [PB volume overload] with [PB pulmonary vascular congestion] on [TREAT the ventilator] right now.

(38) 11. [TREAT Mesalamine] <NUM> milligram Tablet, Delayed Release (E.C.) Sig: Three (<NUM>) Tablet, Delayed Release (E.C.) PO TID (<NUM> times a day) as needed for [PB ulcerative colitis] without recent [PB severe flares].

– L'annotation de la relation est discutable, par exemple dans (39) une relation PIP (*un problème implique un autre problème*) a été annotée entre *lower abdominal pain* et *a symptom*, mais ces deux termes font référence au même concept.

(39) He'd been having [PB lower abdominal pain] for approximately the past week, [PB a symptom] for which he's been admitted in the past.

Dans le modèle que nous venons de présenter, nous n'utilisons pas d'informations sur la structure syntaxique des phrases, uniquement des informations morpho-syntaxiques. Pourtant quand on observe l'exemple (40), il semble important de prendre en compte la structure syntaxique de la phrase pour détecter la relation, par exemple la dépendance de type `nsubj` entre le problème et le verbe *resolved*.

(40) [PB **The right pleural effusion**] slowly re-accumulated after [TREAT tap] but resolved after [TREAT **continued diuresis**].

## 4.3 Étude de la prise en compte de la syntaxe

Les phrases contenant les relations sont parfois complexes, et leur représentation par des traits de surface uniquement ne permet pas de capturer des relations entre termes distants. C'est pourquoi nous nous sommes posée la question de l'utilité des traits syntaxiques pour la reconnaissance de relations en domaine de spécialité : des attributs portant de l'information sur la structure syntaxique des phrases améliorent-ils l'extraction ? Des approches par apprentissage sur des arbres syntaxiques sont-elles meilleures que des approches « sac de mots » ? Cette étude a été publié dans Minard *et al.* [2011b].

Les attributs morpho-syntaxiques que nous avons définis précédemment proviennent de l'analyse morpho-syntaxique du TreeTagger (Schmid [1994]). Une analyse morpho-syntaxique associe une étiquette aux unités lexicales de la phrase et aux autres unités comme les signes de ponctuation. L'étiquette est composée de la catégorie syntaxique de l'unité (nom, verbe, etc.) et éventuellement de ses traits morphologiques (nombre, personne, etc.). Nous avons utilisé uniquement la catégorie syntaxique des mots.

L'analyse syntaxique produit une représentation de la structure de la phrase à partir de l'exploitation des étiquettes morpho-syntaxiques. La structure de la phrase est souvent représentée sous forme d'arbre dans lequel les nœuds sont des étiquettes de partie du discours plus ou moins précises (proposition, groupe verbal, groupe prépositionnel, etc.) et les feuilles des mots. Dans le cas d'une analyse syntaxique de surface, la forme linéaire de la phrase est conservée.

Les analyses morpho-syntaxique et syntaxique sont effectuées sur les phrases dans lesquelles les entités sont annotées. Le TreeTagger ignore les balises et annote tous les mots qui forment l'entité. Pour le corpus i2b2 2010, l'analyse syntaxique est faite avec l'analyseur Charniak/McClosky (McClosky [2010]) sur les phrases dans lesquelles les entités ont été remplacées par leur type. Cette analyse produit ce que nous appelons l'arbre en constituants.

L'arbre de dépendances permet de connaitre les dépendances qui existent entre les mots de la phrase et le type de ces dépendances, par exemple la dépendance entre un verbe et son sujet. L'arbre de dépendances est construit à partir de l'arbre en constituants par le convertisseur du Stanford Parser (De Marneffe *et al.* [2006]).

L'arbre en constituants donnera des informations sur la structure de la phrase en proposition et syntagme. Il permettra de généraliser les phrases par leur structure en donnant éventuellement moins de poids aux mots. L'arbre de dépendances renseignera directement sur les fonctions des mots les uns par rapports aux autres.

Nous verrons dans un premier point l'apport de l'information syntaxique provenant de l'arbre de constituants, puis de l'information extraite de l'arbre de dépendances.

### 4.3.1 Ajout d'information provenant de l'arbre de constituants

Nous avons tout d'abord ajouté aux vecteurs linéaires des attributs portant des informations sur la structure syntaxique des phrases. Mais l'information syntaxique structurelle étant difficile à décrire par un vecteur d'attributs linéaires, nous avons ensuite utilisé les noyaux d'arbres (*tree kernels*) qui permettent d'explorer la structure des phrases en calculant la similarité des arbres deux à deux.

Les informations syntaxiques utilisées proviennent des arbres de constituants. L'analyse syntaxique des phrases a été faite après les prétraitements effectués sur le corpus (cf. 4.2.1) et après l'annotation des entités. La figure 4.5 est un exemple de l'arbre résultant de l'analyse syntaxique de la phrase de l'exemple (41).

(41) 2 $\boxed{^{\text{PB}}\text{Low back strain}}$ requiring $\boxed{^{\text{TREAT}}\text{hospitalization}}$ for $\boxed{^{\text{PB}}\text{pain}}$ in 2002.

À partir de cet arbre nous avons extrait le sous-arbre minimal reliant E1 et E2. Ce sous-arbre correspond au chemin le plus court pour aller d'une entité à l'autre ; nous l'appelons « sous-arbre minimal complet ». Nous avons également produit un sous-arbre plus restreint, qui est équivalent au sous-arbre minimal complet, sauf que nous n'avons pas gardé le contexte gauche de E1 ni le contexte droit de E2.

La figure 4.6 représente le sous-arbre minimal complet extrait de l'arbre présenté dans la figure 4.5 et la figure 4.7 le sous-arbre minimal.

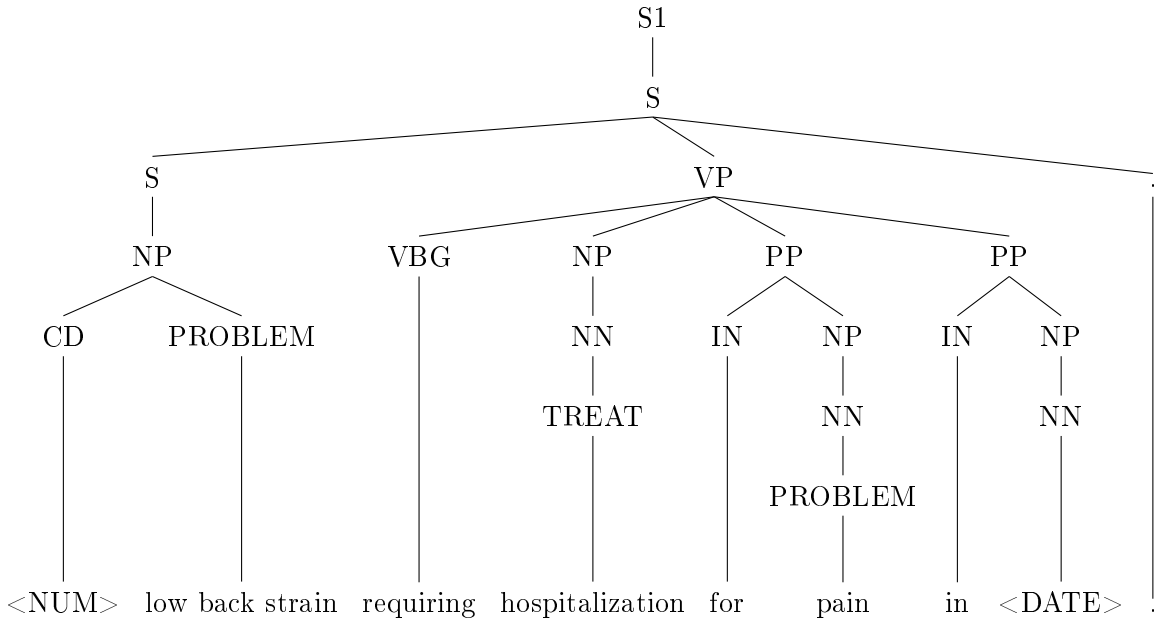



FIG. 4.5 – Exemple de l'arbre complet (la balise <NUM> remplace un nombre et la balise <DATE> une date ou une année)

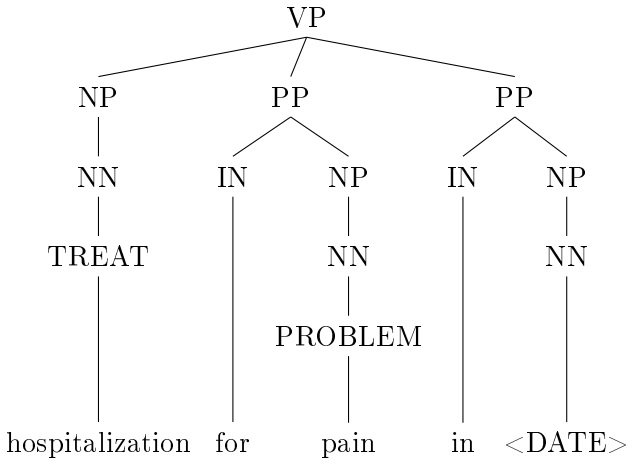



FIG. 4.6 – Exemple du sous-arbre minimal complet entre les deux entités de type traitement et problème

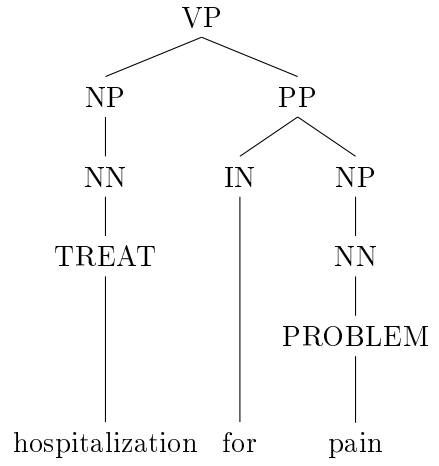



FIG. 4.7 – Exemple du sous-arbre minimal entre les deux entités de type traitement et problème

Nous avons ajouté aux attributs présentés dans la section 4.2.3 les trois types d'arbres pour évaluer si des informations contenues dans les arbres pouvaient être utilisées pour la détection des relations.

Les arbres complets contiennent des informations supplémentaires par rapport aux sous-arbres minimaux. Dans les informations supprimées, il y a du bruit qui peut gêner la classification, mais il y a également des déclencheurs des relations. Il semble donc pertinent d'utiliser les deux types d'arbres pour avoir le maximum d'information. Les sous-arbres minimaux contiennent en moyenne la moitié du nombre de mots de l'arbre complet. Dans le corpus i2b2 2010, les arbres complets [6] ont un nombre moyen de mots [7] par phrase de 21, et un nombre moyen de nœuds de 48. Alors que les sous-arbres minimaux ont un nombre moyen de mots de 8 et un nombre moyen de nœuds de 22.

À partir du sous-arbre minimal nous avons calculé deux attributs : la taille du chemin reliant les deux entités (nous comptons le nombre d'arcs entre les nœuds ayant pour valeur les types des entités), et le constituant du nœud racine du sous-arbre. Pour le couple *hospitalization* et *pain* dans la phrase 41, la taille du plus petit chemin reliant les entités est sept et le constituant du nœud racine du sous-arbre minimal est *NP* (voir figure 4.7).

### 4.3.2 Évaluations

Nous avons évalué sur le corpus i2b2 2010 ces deux façons de prendre en compte de l'information sur la structure syntaxique de la phrase. Nous évaluons dans un premier temps la prise en compte des informations extraites de l'arbre de constituants et dans un second temps l'apport des noyaux d'arbres pour l'extraction des relations.

L'analyse syntaxique du corpus i2b2 a été réalisée par l'analyseur Charniak/McClosky. Nous avons choisi d'utiliser cet analyseur car il est entraîné sur des textes biomédicaux et obtient de bons résultats dans ce domaine (f-score de 87,6% [8]).

#### 4.3.2.1 Remed avec des informations syntaxiques sous forme vectorielle

Nous avons voulu savoir si l'ajout d'informations syntaxiques améliorerait la classification des relations. Pour cela nous avons calculé des informations à partir de l'arbre de constituants que nous avons ajoutées aux attributs de base. Il s'agit de la taille du chemin entre E1 et E2, et du nom du constituant du nœud racine du sous-arbre minimal. Ce système obtient une f-mesure de 0,707. Les résultats détaillés sont présentés dans le tableau 4.19. Les résultats obtenus ne sont pas significativement différents de ceux obtenus sans l'utilisation d'informations syntaxiques selon le test de Student ($p>0,05$). 78 relations (et 47 non relations) sont mieux classées en utilisant les informations syntaxiques mais 79 relations (et 60 non relations) le sont moins bien.

[6] Nous avons un arbre par couple de concepts, c'est-à-dire que si une phrase contient trois concepts, nous avons trois arbres dans le corpus.

[7] La ponctuation n'est pas comptée et les entités comptent comme un seul mot.

[8] http://nlp.stanford.edu/~mcclosky/biomedical.html

| Relation | Rappel | Précision | F-mesure |
|---|---|---|---|
| **TrIP** | 0,161 | 0,842 | 0,271 |
| **TrWP** | 0,000 | 0,000 | 0,000 |
| **TrCP** | 0,371 | 0,717 | 0,489 |
| **TrAP** | 0,702 | 0,750 | 0,725 |
| **TrNAP** | 0,109 | 0,656 | 0,188 |
| **PIP** | 0,541 | 0,762 | 0,633 |
| **TeRP** | 0,832 | 0,874 | 0,852 |
| **TeCP** | 0,272 | 0,816 | 0,408 |
| **Toutes relations** | **0,631** | **0,803** | **0,707** |

TAB. 4.19 – Évaluation du système avec des informations syntaxiques sous forme vectorielle

#### 4.3.2.2 Système à base de tree kernels

L'ajout d'attributs n'améliorant pas la classification, nous avons testé l'ajout d'informations structurelles plus précises que les deux attributs précédents. Pour cela nous avons utilisé le classifieur SVM-Light-TK fondé sur des tree kernels. Comme certaines relations ne sont pas suffisamment représentées dans le corpus, nous avons fait à la fois de la classification bi-classes et multi-classes avec SVM-Light-TK. En effet pour 5 des 8 relations (TrIP, TrWP, TrNAP, TrCP et TeCP), le classifieur ne détectait pas ou peu de relations. Nous avons donc évalué la détection de l'existence d'une relation (modèle bi-classes) et la classification (modèle multi-classes).

Nous avons paramétré le classifieur de la façon suivante : combinaison d'arbres et de vecteurs comme type de fonction kernel, et somme des contributions des arbres et des vecteurs comme opérateur de combinaison.

**Bi-classes** Nous avons ajouté aux données fournies en entrée au classifieur les arbres de constituants complets ainsi que les sous-arbres minimaux complets entre les deux entités possiblement en relation et les sous-arbres minimaux, pour évaluer si d'autres informations contenues dans les arbres pouvaient être utilisées pour la détection des relations. Nous avons évalué plusieurs combinaisons à partir des arbres de constituants complets (AC), des sous-arbres minimaux complets (AMC), des sous-arbres minimaux (AM) et des attributs du système de base (ATT). Les résultats sont présentés dans le tableau 4.20 et dans la figure 4.8.

Ces résultats montrent que la donnée des arbres n'est pas suffisante pour la classification des relations : quel que soit l'arbre utilisé le classifieur ne sait pas en extraire l'information pertinente. De plus la combinaison des arbres minimaux et des arbres complets apportent des meilleurs résultats que les arbres complets seuls, mais la f-mesure n'atteint pas celle obtenue avec les attributs seuls. Les attributs apportent donc des informations supplémentaires par rapport aux arbres. Dans cette étude nous n'avons pas évalué l'apport de chaque attribut vectoriel mais il serait intéressant de savoir quelles données ne sont pas récupérées dans les arbres mais sont fournies par les attributs.

| Combinaison | Précision | Rappel | F-mesure |
|---|---|---|---|
| **AC** | 0,749 | 0,611 | 0,673 |
| **AMC** | 0,623 | 0,726 | 0,651 |
| **AM** | 0,819 | 0,625 | 0,709 |
| **ATT** | 0,835 | 0,709 | 0,767 |
| **AC + AM** | 0,790 | 0,708 | 0,747 |
| **AC + ATT** | 0,826 | 0,731 | 0,776 |
| **AMC + ATT** | 0,832 | 0,729 | 0,773 |
| **AM + ATT** | 0,828 | 0,724 | 0,772 |
| **AC + AM + ATT** | 0,776 | 0,804 | **0,790** |
| **AMC + AM + ATT** | 0,819 | 0,727 | 0,770 |
| **AC + AMC + ATT** | 0,818 | 0,726 | 0,769 |
| **AC + AMC + AM + ATT** | 0,816 | 0,730 | 0,771 |

TAB. 4.20 – Évaluation des combinaisons des différents apprentissages à base de tree kernels pour la détection de relations

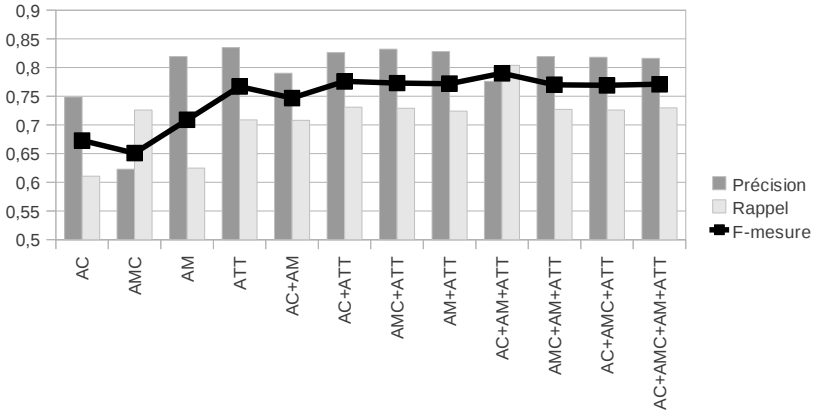



FIG. 4.8 – Comparaison des résultats des combinaisons des différents apprentissages

Nous observons également que la meilleure combinaison est celle associant les arbres complets, les sous-arbres minimaux et les attributs (AC + AM + ATT). Les sous-arbres minimaux complets n'apportent pas d'informations supplémentaires. En effet ils apportent moins d'informations que les arbres complets (la f-mesure pour AC + AM + ATT est de 0,790 et pour AMC + AM + ATT de 0,770), et plus d'informations bruitées que les arbres minimaux réduits (la f-mesure pour AC + AMC + ATT est de 0,771, contre 0,790 avec les arbres minimaux).

Nous avons effectué une étude des relations détectées avec les attributs seuls et avec les

attributs plus les arbres complets (AC + ATT). Nous avons observé que les arbres étaient effectivement utilisés pour détecter les relations entre des entités éloignées dans la phrase. Les relations correctement détectées avec (AC + ATT) mais qui ne le sont pas avec (ATT) concernent deux entités dont l'éloignement moyen est de 11 mots (ou ponctuations). Alors que celles qui sont correctement classées par les deux systèmes concernent des entités qui sont séparées en moyenne par 5 mots (ou ponctuations).

Lorsque l'arbre 4.9 est utilisé pour classer la relation entre *the treadmill exercise test* et *0.5 to 1 mm st depression in ii , iii , and avf*, celle-ci est correctement classée comme *un test qui révèle un problème médical* (TeRP), alors que sans la prise en compte de la structure syntaxique aucune relation n'avait été détectée.

Dans la figure 4.10 nous montrons une phrase contenant une relation de type *un problème médical implique un autre problème* (PIP) entre *increased tracer activity* et *active bleeding* ; les entités sont séparées par 17 mots ou ponctuations. Cette relation a été correctement détectée lorsque les arbres étaient utilisés, mais elle n'est pas repérée avec l'utilisation des attributs seuls.

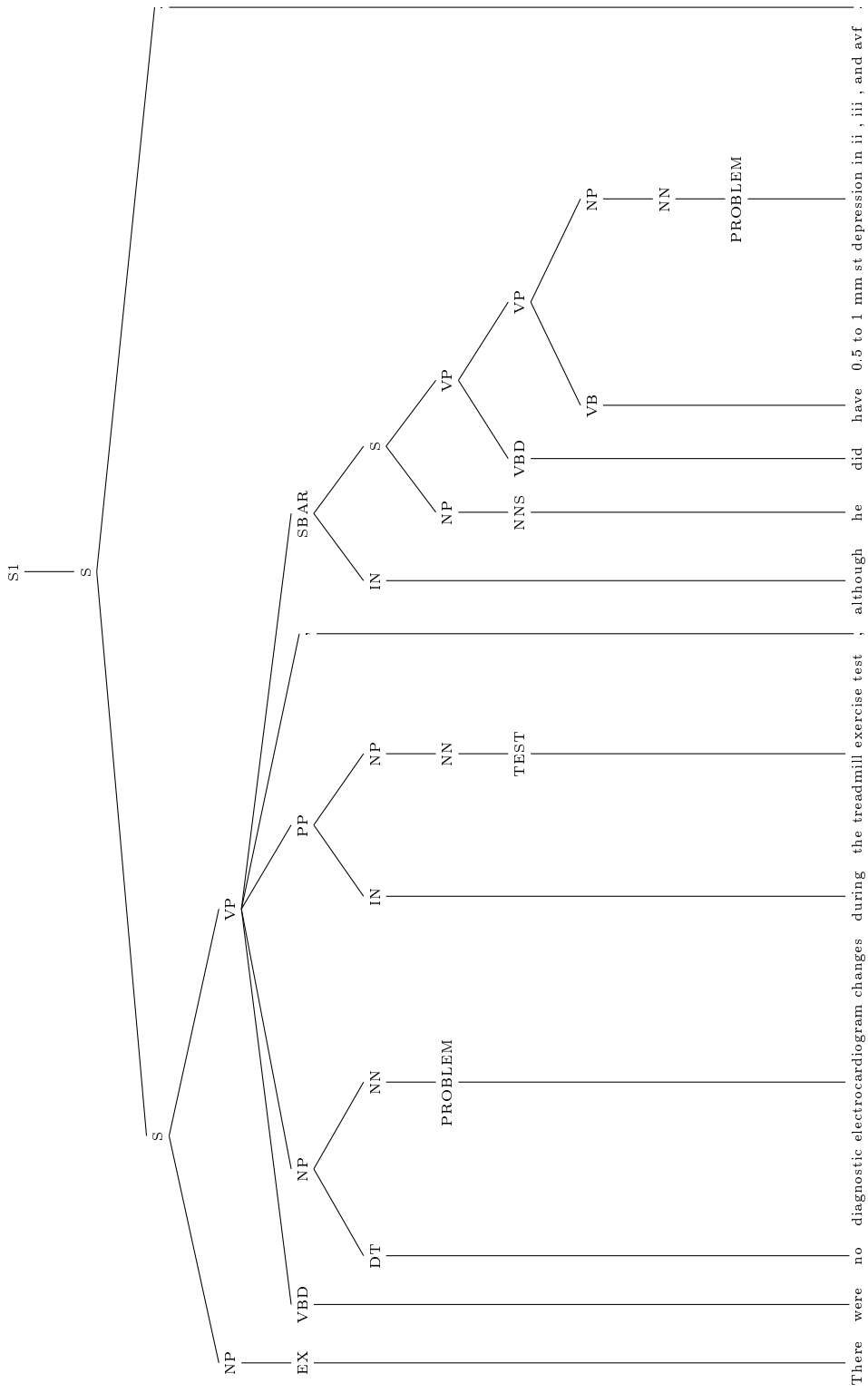



FIG. 4.9 – Exemple de phrase pour laquelle l'utilisation de la structure syntaxique a permis de classer correctement la relation entre le test et le 2$^{e}$ problème

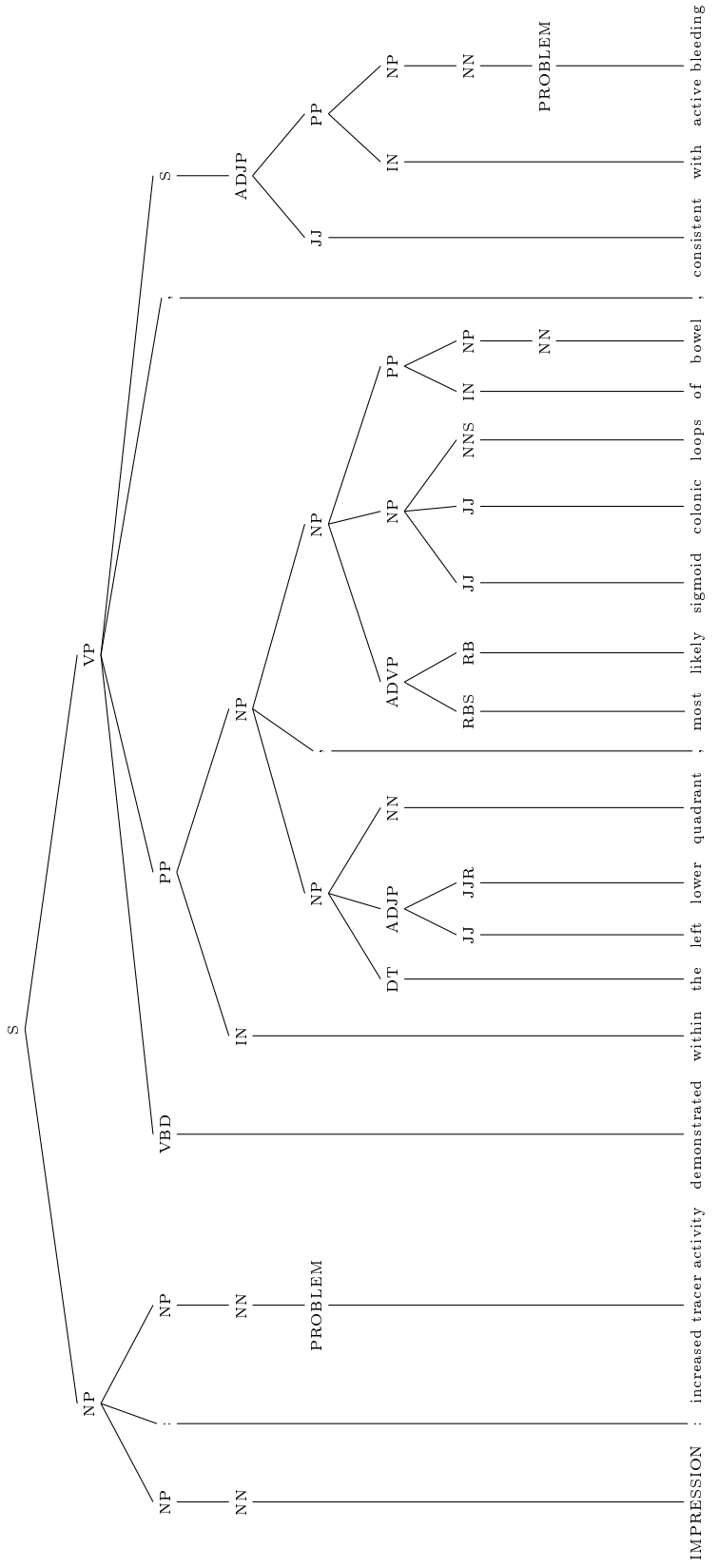



FIG. 4.10 – Exemple de l'arbre complet d'une phrase contenant deux concepts reliés par une relation de type PIP

**Multi-classes** Nous avons également évalué l'apport des arbres pour la classification des relations. Notre système construit un modèle d'apprentissage par relation, c'est-à-dire que nous utilisons une méthode de classification multi-classes de type « one-versus-all ». Dans la typologie des relations utilisée, il y a une classe pour les paires *Problème - Problème*, deux classes pour les paires *Problème - Test* et cinq classes pour les paires *Problème - Traitement*. Les données sont divisées en trois selon le type des entités et les modèles sont construits pour une classe de relation versus toutes les paires d'entités du même type. Nous disposons donc de trois corpus d'entraînement et de trois corpus de test, et de 8 modèles. Une fois que les modèles de classification ont été appliqués sur les corpus de test, un vote est effectué pour décider de la meilleure classification pour une paire d'entités avec priorité donnée aux classes peu représentées.

Les résultats de la classification multi-classes sur le corpus d'évaluation sont donnés dans le tableau 4.21. Dans les trois dernières lignes du tableau on donne le rappel, la précision et la f-mesure en micro-moyenne de chaque classifieur, c'est-à-dire les résultats de la classification pour toutes les relations. Nous avons évalué la classification produite lorsque le classifieur utilise uniquement les arbres, uniquement les attributs, une combinaison de l'arbre complet (AC), de l'arbre minimal (AM) et des attributs (ATT), et une combinaison des trois arbres et des attributs. Les meilleurs résultats sont obtenus lorsque le classifieur utilise les arbres complets, les arbres minimaux et les attributs, comme nous l'avons observé précédemment avec la classification binaire. Avec l'ajout des arbres minimaux complets les résultats restent stables.

La classification effectuée avec l'utilisation des arbres de constituants est moins bonne que celle obtenue avec libSVM et les attributs sous forme vectorielle. Ces moins bons résultats sont dus entre autres à la méthode utilisée pour la classification multi-classes : « one-versus-one » ou « one-versus-all », et à la difficulté pour SVMLight de construire des modèles avec peu de données. En effet, on remarque que pour les classes les moins représentées la f-mesure ne dépasse pas 0,200.

| | AC+AM+AMC | ATT | AC+AM+ATT | AC+AM+AMC+ATT |
|---|---|---|---|---|
| TrIP | 0,000 | 0,000 | **0,020** | **0,020** |
| TrWP | 0,000 | 0,000 | 0,000 | 0,000 |
| TrCP | 0,118 | 0,141 | **0,179** | 0,172 |
| TrAP | 0,651 | 0,647 | **0,689** | 0,682 |
| TrNAP | 0,030 | 0,040 | **0,041** | **0,041** |
| PIP | 0,617 | **0,649** | 0,601 | 0,633 |
| TeRP | 0,799 | **0,839** | 0,835 | 0,829 |
| TeCP | 0,132 | 0,036 | **0,267** | 0,215 |
| Rappel | 0,548 | 0,512 | **0,573** | 0,559 |
| Précision | 0,768 | **0,849** | 0,824 | 0,825 |
| F-mesure | 0,639 | 0,639 | **0,676** | 0,666 |

TAB. 4.21 – Classification multi-classes avec des noyaux d'arbres sur le corpus d'évaluation

Nous avons réalisé une combinaison des prédictions du meilleur classifieur (AC + AM + ATT) et des prédictions du système REMED. Les résultats de cette combinaison sont

donnés dans le tableau 4.22. Nous y avons rappelé les résultats obtenus avec les deux classifieurs séparément. La combinaison des prédictions est faite de la façon suivante pour chaque paire d'entités (le classifieur B est celui utilisant un noyau d'arbres) :

- Priorité aux relations : si le classifieur A n'a pas identifié de relation et le classifieur B a détecté une relation de type t2, alors on considère qu'il y a une relation de type t2 entre les entités.
- Priorité au classifieur A : si le classifieur A a détecté une relation de type t1 et le classifieur B une relation de type t2, alors la relation est considérée de type t1.

Nous observons que la combinaison des systèmes obtient les meilleurs résultats et le meilleur rappel, ce qui paraît normal vu que nous donnons priorité aux relations lors de la combinaison des prédictions. La meilleure précision est obtenue par le classifieur utilisant un noyau d'arbres. 432 relations ont été correctement classées par le classifieur utilisant les arbres mais pas par REMED. Ces relations ont pu être correctement classées grâce à la prise en compte de la structure syntaxique mais aussi du fait de l'utilisation d'un classifieur linéaire.

| | REMED | AC+AM+ATT | Combinaison |
|---|---|---|---|
| TrIP | 0,279 | 0,020 | 0,279 |
| TrWP | 0,000 | 0,000 | 0,000 |
| TrCP | 0,491 | 0,179 | 0,492 |
| TrAP | 0,728 | 0,689 | 0,740 |
| TrNAP | 0,189 | 0,041 | 0,189 |
| PIP | 0,635 | 0,649 | 0,680 |
| TeRP | 0,850 | 0,835 | 0,859 |
| TeCP | 0,393 | 0,267 | 0,405 |
| Rappel | 0,631 | 0,573 | **0,678** |
| Précision | 0,803 | **0,824** | 0,778 |
| F-mesure | 0,707 | 0,676 | **0,724** |

TAB. 4.22 – Combinaison des prédictions de deux classifieurs sur le corpus d'évaluation

### 4.3.3 Ajout d'informations provenant de l'arbre de dépendances

Après avoir évalué l'apport de la prise en compte de la structure syntaxique des phrases, nous avons évalué l'apport des informations provenant des arbres de dépendances, syntaxiquement plus riches. Les arbres de dépendances permettent de connaître de façon plus explicite qu'avec les arbres de constituants, la nature des liens syntaxiques entre les mots de la phrase. L'arbre de dépendances renseigne par exemple des sujets de chaque verbe de la phrase (dans 4.12, PROBLEM est le sujet des deux verbes *re-accumulated* et *resolved*). Airola *et al.* [2008] ont développé un système à noyau de graphes pour apprendre à extraire les relations à partir des arbres de dépendances pour la tâche d'extraction d'interactions entre protéines. Nous n'avons pas testé cette méthode par manque de temps, mais il nous semblerait intéressant de l'adapter pour la classification multi-classes des relations dans le corpus i2b2 2010.

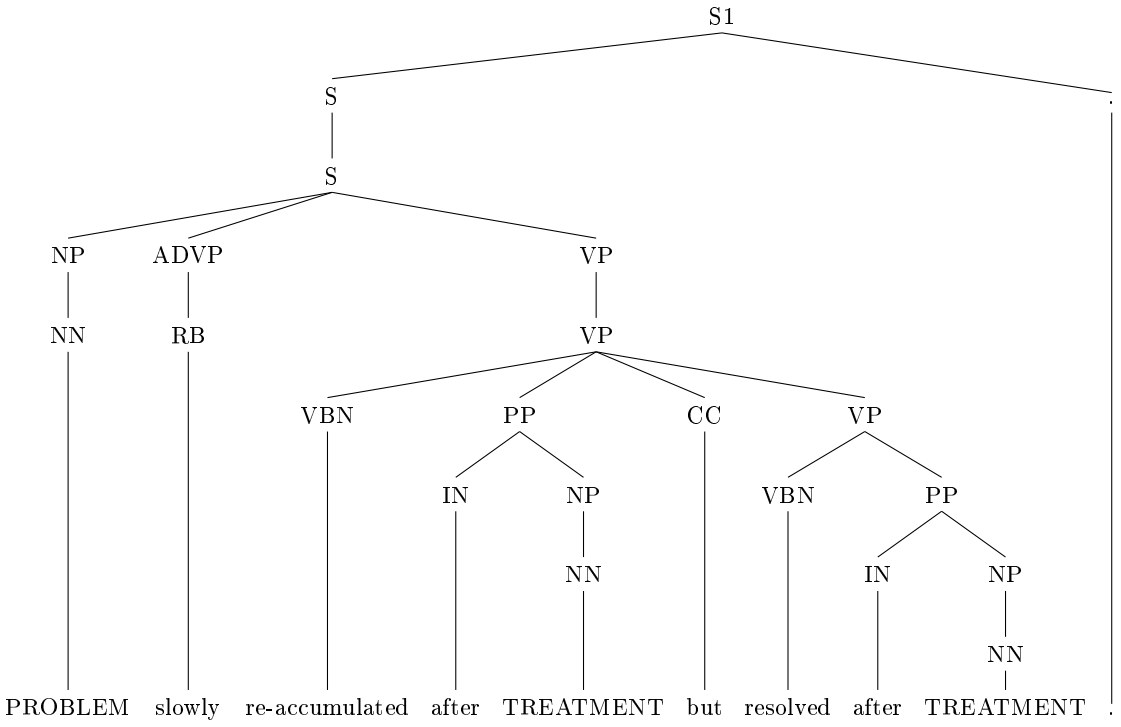



FIG. 4.11 – Exemple de l'arbre complet de la phrase (42)

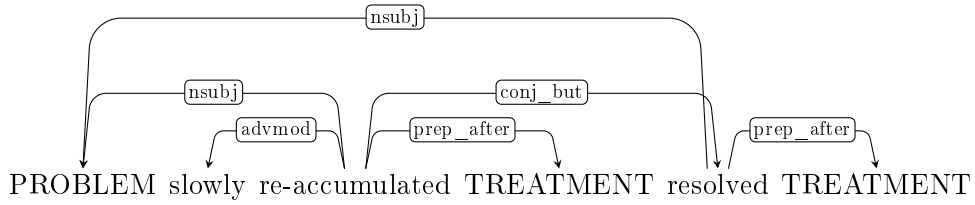



FIG. 4.12 – Arbre de dépendances de la phrase (42)

Nous avons choisi d'utiliser le convertisseur du Stanford Parser (De Marneffe *et al.* [2006]) pour produire les arbres de dépendances ; il construit les arbres de dépendances à partir des arbres de constituants. Par exemple, à partir de l'arbre de constituants de la phrase (42) présenté dans la figure 4.11, on obtient l'arbre de dépendances présenté dans la figure 4.12.

(42) [PB The right pleural effusion] slowly re-accumulated after [TREAT tap] but resolved after [TREAT continued diuresis].

Uzuner *et al.* [2010a] montre que sur un corpus de comptes rendus cliniques, les dépendances entre les deux entités E1 et E2 n'étaient pas utiles pour la classification des relations car celles-ci sont très peu nombreuses. Nous avons cherché à extraire d'autres informations plus pertinentes de ces arbres afin de les fournir au classifieur sous forme d'attributs. Les informations que nous avons extraites sont les suivantes (nous nous sommes inspirée de

| | Remed | Triplets | Triplets principaux | Triplets du sous-arbre | Tous les attributs |
|---|---|---|---|---|---|
| Rappel | 0,708 | 0,687 | **0,694** | 0,644 | 0,651 |
| Précision | 0,815 | 0,807 | **0,809** | 0,797 | 0,787 |
| F-mesure | 0,758 | 0,742 | **0,747** | 0,712 | 0,712 |

Tab. 4.23 – Évaluation de l'apport des informations extraites de l'arbre de dépendances sur le corpus de test (rappel, précision et f-mesure en micro-moyenne pour toutes les relations)

Van Landeghem *et al.* [2008] pour le choix des informations) :
- tous les triplets (mot, type de dépendance, mot) ;
- les triplets pour lesquels le type de dépendance est `subj`, `obj` ou `prep` ;
- les triplets (mot, type de dépendance, mot) du plus petit sous-arbre entre les deux entités.

### 4.3.4 Évaluation

Nous avons réalisé une petite évaluation de l'apport des informations provenant de l'arbre de dépendances sur le corpus de test. Les résultats sont présentés dans le tableau 4.23. Nous avons évalué la classification après l'ajout de chacune des séries d'attributs, puis avec tous les attributs. Dans un premier temps, nous avons utilisé tous les triplets de l'arbre, mais les résultats étaient moins bons. Ensuite, nous avons ajouté les triplets des dépendances principales de l'arbre, c'est-à-dire les dépendances de type `subj`, `obj` ou `prep`. La classification est légèrement meilleure que lorsque l'on extrait tous les triplets de l'arbre, mais elle reste moins bonne que sans l'utilisation des informations provenant de l'arbre de dépendances. Enfin nous avons utilisé uniquement les dépendances du sous-arbre, et la f-mesure a baissé de plus de 0,04 points. Nous avons également essayé de combiner les différentes informations extraites de l'arbre mais les résultats ne sont pas non plus satisfaisants (dernière colonne du tableau). Nous n'avons pas mené plus loin nos recherches sur ce point.

## 4.4 Application à deux autres corpus

Nous avons testé notre système Remed dans le cadre de deux autres tâches : l'extraction d'interactions entre médicaments et l'extraction d'interactions entre protéines. Nous allons dans cette partie présenter les tâches, les corpus et les différentes expérimentations que nous avons menées. Par ces expérimentations, nous voulons montrer que notre système donne aussi de bons résultats en classification binaire et sur des documents mieux formés que les comptes rendus cliniques à savoir des articles scientifiques.

### 4.4.1 DDI 2011 : extraction d'interactions entre médicaments

La campagne d'évaluation DDIExtraction 2011 a été organisée par Isabel Segura-Bedmar, Paloma Martínez et Daniel Sánchez-Cisneros de l'université Carlos III de Ma-

drid (Segura-Bedmar *et al.* [2011b]). Nous avons participé à ce challenge (Minard *et al.* [2011e]) avec une stagiaire (Lamia Makour) qui s'est intéressée en particulier à la sélection des attributs et au problème des classes déséquilibrées. Avec le système REMED (associé à une sélection des attributs) nous nous sommes classées $5^e$ sur 10 participants. Dans cette section nous présentons la tâche d'extraction d'interactions entre médicaments et le corpus fourni par les organisateurs du challenge, puis nous donnerons les résultats avec et sans la sélection des attributs.

#### 4.4.1.1 Tâche

L'objectif de la tâche de DDIExtraction 2011 était de détecter si deux médicaments dans la même phrase sont en interaction ou non. Deux médicaments sont considérés en interaction quand l'un influence le niveau ou l'activité de l'autre. Dans l'exemple (43), il y a une interaction entre *humorsol* et *succinylcholine*, et entre *humorsol* et *anticholinesterase agents*, mais il n'y pas d'interaction entre *succinylcholine* et *anticholinesterase agents*.

(43) Possible drug interactions of **HUMORSOL** with **succinylcholine** or with other **anticholinesterase agents**.

Les entités annotées dans le corpus sont uniquement des noms de médicaments et une seule classe de relation nous intéresse : interagit.

#### 4.4.1.2 Corpus

Le corpus est composé de textes biomédicaux collectés depuis la base de données DrugBank, annotés en noms de médicaments et en interactions entre les médicaments. Le corpus est composé d'un sous-corpus d'entraînement qui contient 2 402 interactions (dans 435 documents), et d'un sous-corpus de test qui contient 758 interactions (dans 144 documents).

| | pré-éval | évaluation |
|---|---|---|
| Documents | 435 | 144 |
| Interactions | 2402 | 755 |
| Non-relations | 21425 | 6271 |

TAB. 4.24 – Composition des corpus DDI

Nous avons identifié trois particularités dans ce corpus qui nous ont conduites à ajouter des attributs spécifiques au corpus :

- beaucoup de phrases commencent par un nom de médicament suivi de deux points, pour indiquer le médicament d'intérêt dans la phrase (voir exemple (44)) ;
- le nom de médicament peut être *drug* (c'est le cas pour 520 entités dans le corpus d'entraînement). Dans ce cas l'expression de la relation peut être différente (voir l'exemple (45)) ;
- chaque fichier du corpus traite d'un médicament en particulier.

(44) **Valproate**: **Tiagabine** causes a slight decrease (about 10%) in steady-state **valproate** concentrations.

(45) In addition to established drug interactions, there may be potential pharmacokinetic interactions between **nevirapine** and other **drug** classes that are metabolized by the **cytochrome P450** system.

Trois attributs indiquent si une des deux entités est :
- le même médicament que le premier dans la phrase ;
- le terme *drug* ;
- le médicament le plus fréquent dans l'article.

#### 4.4.1.3 Expérimentations

Nous avons utilisé le système REMED sur le corpus DDI avec les trois attributs supplémentaires décrits précédemment. La valeur du paramètre C est fixée à 2 et celle du paramètre $\gamma$ à 0,0078125. Comme nous sommes dans le cadre d'une classification bi-classes, nous pouvons ajouter des poids sur le paramètre C pour pénaliser la classe majoritaire (les non-relations). Nous fixons le poids pour la classe des non-relations à 2 et celui pour la classe des relations à 9 (ces valeurs ont été choisies après plusieurs tests).

Dans un premier temps, nous avons appliqué REMED sur le corpus d'évaluation, sans effectuer de sélection d'attributs et en utilisant les attributs spécifiques décrits précédemment. Les résultats que nous obtenons sont dans le tableau 4.25. REMED obtient une f-mesure de 0,602.

Si nous ajoutons des informations extraites de l'arbre de constituants (la taille du chemin entre E1 et E2 et le nom du constituant du nœud racine du sous-arbre minimal entre E1 et E2 (cf. section 4.3.2.1)), nous obtenons une meilleure f-mesure : 0,618. Et finalement, nous avons testé la sélection d'attributs sur ce corpus, qui nous permet d'obtenir une f-mesure légèrement meilleure : 0,622. La différence entre REMED et REMED utilisant des informations provenant de l'arbre de constituants est significative ($p<0,05$ avec le test de Student), mais elle n'est pas significative après la sélection des attributs.

Dans le tableau 4.25, nous avons aussi indiqué les résultats obtenus par les meilleurs systèmes des 5 premières équipes (Segura-Bedmar *et al.* [2011b]). L'équipe WBI a utilisé une combinaison de plusieurs noyaux et un système de raisonnement par cas utilisant une approche par vote. Les systèmes des équipes LIMSI-FBK et FBK-HLT sont fondés sur une méthode combinant différents noyaux (MEDT, PST et SL) que l'équipe LIMSI-FBK a combinée avec une méthode basée sur des attributs et un classifieur SVM. Enfin l'équipe UTurku utilise un classifieur RLS (*Regularized Least-Squares*) ainsi que la ressource DrugBank qui contient des informations sur l'interaction de paires de médicaments.

### 4.4.2 PPI : extraction d'interactions entre protéines

#### 4.4.2.1 Tâche

La tâche d'extraction d'interactions entre protéines est très similaire à celle d'extraction d'interactions entre médicaments. Dans les corpus, les protéines et/ou gènes sont annotées, et l'objectif est d'identifier les paires de protéines (ou de gènes) qui interagissent. Cette tâche a fait l'objet d'une campagne d'évaluation en 2005 (Nédellec [2005]). L'exemple 46

| | Rappel | Précision | F-mesure |
|---|---|---|---|
| REMED | 0,680 | 0,540 | 0,602 |
| REMED + info synt | 0,698 | 0,555 | 0,618 |
| REMED + info synt + sélection d'attributs | 0,707 | 0,556 | 0,622 |
| WBI | 0,719 | 0,605 | 0,657 |
| LIMSI-FBK | 0,704 | 0,585 | 0,639 |
| FBK-HLT | 0,700 | 0,583 | 0,637 |
| UTurku | 0,688 | 0,580 | 0,629 |

TAB. 4.25 – Rappel, précision et f-mesure sur le corpus d'évaluation

est extrait du corpus LLL, et exprime une relation d'interaction entre les deux protéines en gras.

(46) Here, we show that **GerE** binds near the **sigK** transcriptional start site, to act as a repressor.

#### 4.4.2.2 Corpus

Cinq corpus annotés en interactions entre protéines sont librement accessibles :
- AIMed (Bunescu *et al.* [2005]) est composé de 200 résumés de MEDLINE contenant des interactions recensées dans la base de données DIP (*the Database of Interacting Proteins*), et 30 résumés sélectionnés manuellement qui contiennent des phrases avec plus d'un nom de gène cité mais qui ne mentionnent pas d'interactions.
- BioInfer (Pyysalo *et al.* [2007]) contient également des résumés de MEDLINE (836) contenant des interactions contenues dans la base DIP.
- hprd50 (Fundel *et al.* [2007]) contient 50 résumés référencés par la base de données HPRD (*the Human Protein Reference Database*).
- IEPA (Ding *et al.* [2002]) est composé de 300 résumés de MEDLINE obtenus en utilisant 10 requêtes sur PubMed.
- LLL (Nédellec [2005]) est composé de résumés de MedLine sur la transcription et la sporulation du modèle bactérial *Bacillus subtilis*.

Le tableau 4.26 présente le nombre de phrases dans chaque corpus, le nombre de paires d'entités en relation et le nombre de paires qui ne le sont pas.

| | AIMed | BioInfer | HPRD50 | IEPA | LLL |
|---|---|---|---|---|---|
| Phrases | 1955 | 1100 | 145 | 486 | 77 |
| Interactions | 1000 | 2534 | 163 | 335 | 164 |
| Non-relations | 4834 | 7132 | 270 | 482 | 166 |

TAB. 4.26 – Composition des corpus PPI

#### 4.4.2.3 Expérimentations

Comme pour cette tâche cinq corpus sont librement disponibles, l'apprentissage de l'extraction des relations pourra être fait de trois manières différentes [9] : en validation croisée (VC) sur chaque corpus séparément, en apprentissage croisé (AC) et en corpus croisé (CC). La validation croisée consiste à diviser le corpus en k parties (ici k=10), puis à apprendre le modèle d'extraction sur (k-1) parties du corpus, ensuite à tester le modèle sur la partie restante, et à répéter cette opération k fois. Avec l'apprentissage croisé, le modèle est construit à partir de (k-1) corpus (ici k=5) et testé sur le corpus restant. Et l'apprentissage en corpus croisé consiste à apprendre le modèle avec un corpus et d'appliquer le modèle sur les 4 autres corpus. Nous avons évalué le système de cette façon en utilisant le corpus AIMed puis le corpus BioInfer comme corpus d'entraînement. Nous n'avons pas évalué en utilisant les autres corpus du fait de leur faible taille. Nous avons évalué le système REMED en utilisant ces trois méthodes.

Tikk *et al.* [2010] ont présenté les résultats obtenus avec 9 noyaux différents sur les 5 corpus. Nous comparons nos résultats avec les meilleurs résultats référencés dans Tikk *et al.* [2010].

**Système REMED** Dans les tableaux 4.27, 4.28 et 4.29, nous présentons respectivement les résultats obtenus avec REMED (avec la gestion de la coordination) en validation croisée, corpus croisé et apprentissage croisé. Les paramètres ont été fixés suite à plusieurs tests pour chaque type d'apprentissage :

- validation croisée : C vaut 32 et $\gamma$ vaut 0,0078125 ;
- corpus croisé et apprentissage croisé : C vaut 2 et $\gamma$ vaut 0,0078125, et des poids sont donnés au paramètre C pour chaque classe : 9 pour la classe de relations et 2 pour la classe de non-relations.

Dans chaque tableau, nous indiquons les meilleurs résultats référencés dans Tikk *et al.* [2010] et les méthodes utilisées. *SL* ou *Shallow Linguistic* est la méthode développée par Giuliano *et al.* [2006] qui est une combinaison de deux noyaux : un noyau pour le contexte global et un pour le contexte local. *APG* ou *all-paths graph kernel* a été proposé par Airola *et al.* [2008] et repose sur le calcul de du poids des chemins partagés dans le graphe de dépendances et dans la phrase. *kBSPS* ou *k-Band Shortest Path Spectrum Kernel* est un nouveau noyau proposé par Palaga [2009] (aussi présenté en détail par Tikk *et al.* [2010]). Le noyau compare tous les *v-walks* de taille q extraits d'un sous-graphe de dépendances entre E1 et E2 augmenté (c'est-à-dire que les nœuds et les dépendances qui sont à une distance k du sous-graphe sont ajoutés).

Nous observons qu'avec REMED nous obtenons des résultats inférieurs aux meilleurs résultats présentés dans Tikk *et al.* [2010]. Nos résultats sont souvent inférieurs à ceux obtenus avec les méthodes *SL*, *kBSPS* et *APG*, mais supérieurs à ceux obtenus avec les noyaux *cosine similarity*, *edit distance* ou *spectrum tree*. *kBSPS* et *APG* utilisent des informations provenant de l'arbre de dépendances, informations que nous ne prenons pas en compte dans la version initiale de REMED.

[9] Ces 3 méthodes d'évaluation ont été proposées par Tikk *et al.* [2010].

| | Corpus de test | | | | |
|---|---|---|---|---|---|
| | AIMed | BioInfer | hprd50 | IEPA | LLL |
| REMED | 0,494 | 0,546 | 0,632 | 0,696 | 0,769 |
| Tikk *et al.* [2010] | 0,562 | 0,616 | 0,710 | 0,731 | 0,797 |
| | *APG* | *APG (with SVM)* | *kBSPS* | *APG* | *APG (with SVM)* |

TAB. 4.27 – F-mesures obtenues avec le système REMED sur les corpus PPI en validation croisée (VC)

| | | Corpus de test | | | | |
|---|---|---|---|---|---|---|
| | | AIMed | BioInfer | hprd50 | IEPA | LLL |
| AIMeD | REMED | X | 0,275 | 0,596 | 0,187 | 0,269 |
| | Tikk *et al.* [2010] | X | 0,406 | 0,616 | 0,299 | 0,365 |
| | | | *SL* | *APG* | *kBSPS* | *SL* |
| BioInfer | REMED | 0,425 | X | 0,608 | 0,486 | 0,720 |
| | Tikk *et al.* [2010] | 0,415 | X | 0,698 | 0,724 | 0,806 |
| | | *SL* | | *kBSPS* | *kBSPS* | *kBSPS* |

TAB. 4.28 – F-mesures obtenues avec le système REMED sur les corpus PPI en corpus croisé (CC)

| | Corpus de test | | | | |
|---|---|---|---|---|---|
| | AIMed | BioInfer | hprd50 | IEPA | LLL |
| REMED | 0,437 | 0,347 | 0,636 | 0,430 | 0,599 |
| Tikk *et al.* [2010] | 0,438 | 0,476 | 0,697 | 0,707 | 0,791 |
| | *APG* | *kBSPS* | *APG* | *kBSPS* | *kBSPS* |

TAB. 4.29 – F-mesures obtenues avec le système REMED sur les corpus PPI en apprentissage croisé (AC)

Suite à cette observation, nous avons ajouté des attributs représentant des informations extraites de l'arbre de dépendances. Les arbres de dépendances ont été produits par le convertisseur du Stanford Parser à partir des arbres de constituants. Les arbres de constituants ont été construits par l'analyseur de Charniak-Lease (les arbres nous ont été fournis avec le corpus). Nous avons pris en compte les triplets (mot, dépendance, mot) de l'arbre complet, du sous-arbre minimal (c'est-à-dire le chemin le plus court entre E1 et E2 dans l'arbre de dépendances), ainsi que les mots contenus dans le sous-arbre minimal. Nous avons évalué la détection des relations avec ces informations supplémentaires avec la méthode d'apprentissage croisé.

Les résultats obtenus sont présentés dans le tableau 4.30. Nous observons que les résultats obtenus sont meilleurs avec les informations provenant l'arbre de dépendances, et dépassent les meilleurs résultats présentés par Tikk *et al.* [2010] pour les corpus BioInfer et AIMed.

| | Corpus de test | | | | |
|---|---|---|---|---|---|
| | AIMed | BioInfer | hprd50 | IEPA | LLL |
| REMED | 0,437 | 0,347 | 0,636 | 0,430 | 0,599 |
| REMED + dépendances | **0,461** | **0,487** | 0,687 | 0,601 | 0,777 |
| Tikk *et al.* [2010] | 0,438 | 0,476 | 0,697 | 0,707 | 0,791 |
| | *APG* | *kBSPS* | *APG* | *kBSPS* | *kBSPS* |

TAB. 4.30 – F-mesures obtenues avec la prise en compte d'informations provenant de l'arbre de dépendances avec le système REMED sur les corpus PPI en apprentissage croisé (AC)

## 4.5 Conclusion

Nous nous sommes intéressée à l'extraction de relations en domaine de spécialité et en particulier aux informations utiles pour classer correctement les relations. Nous avons défini des attributs variés qui représentent des informations permettant de détecter les relations dans la phrase, mais aussi de typer ces relations. Pour évaluer la pertinence de ces attributs, nous avons développé un système de détection et typage de relations, REMED, fondé sur une classification automatique. Dans sa version finale, REMED obtient une f-mesure d'environ 0,70. Pour une tâche s'intéressant uniquement à la détection de relations, tous les attributs ne seraient pas utiles. Les résultats que nous avons obtenus sur deux tâches d'extraction d'interactions (classification bi-classes) avec le système REMED sont bons mais ne dépassent pas les résultats obtenus avec les meilleurs systèmes de l'état de l'art.

De ce fait nous avons approfondi l'étude des informations syntaxiques. Les informations syntaxiques très simples ajoutées sous forme vectorielle n'ont pas amélioré la classification sur le corpus i2b2 2010. En revanche, l'utilisation des structures syntaxiques avec les noyaux d'arbres améliore la détection des relations, les meilleurs résultats étant obtenus avec une combinaison de l'arbre complet, le sous-arbre minimal et les attributs de base (augmentation de 3% de la f-mesure en bi-classes). Le sous-arbre minimal complet ne semble pas apporter des informations supplémentaires par rapport à l'arbre complet et au sous-arbre minimal. La combinaison des prédictions du classifieur SVM-Light utilisant l'arbre com-

plet, le sous-arbre minimal et des attributs, et des prédictions du système REMED permet d'obtenir une f-mesure de 0,72 sur toutes les relations.

L'utilisation du seul sous-arbre minimal n'est pas suffisant. Nous avons observé qu'il était aussi important de s'intéresser à l'arbre complet. La simplification qui consiste à ne conserver que ce qu'il y a entre les deux entités est soit trop drastique ou inutile (dans le cas où l'arbre minimal correspond à l'arbre complet). Dans l'exemple (47), l'arbre minimal pour la paire d'entités *the right pleural effusion* et *continued diuresis* est identique à l'arbre complet. Pour ne conserver que les informations essentielles, une simplification plus fine est nécessaire. Il faudrait par exemple supprimer le premier verbe et son complément. C'est à cette question de la simplification que nous nous intéressons dans le chapitre suivant.

(47) [PB **The right pleural effusion slowly**] re-accumulated after [TREAT tap] but resolved after [TREAT **continued diuresis**].

Suite à ce travail, nous proposons deux perspectives permettant d'approfondir l'évaluation des attributs et la prise en compte d'informations syntaxiques. L'étude des traits à utiliser pour l'extraction par apprentissage des relations binaires, a été faite pour une tâche de classification des relations. Nous pensons que les traits importants pour une tâche de détection de relations, comme la tâche d'extraction des interactions entre protéines, ne sont pas les mêmes que ceux utilisés pour caractériser le type d'une relation. Une étude pourrait être mené pour étudier les informations utilisées pour identifier la présence d'une relation entre deux entités et celles permettant de classer la relation.

Suite à l'étude de la prise en compte d'informations syntaxiques, nous avons observé que les informations extraites de l'arbre de constituants et de l'arbre de dépendances utilisées sous forme vectorielle améliorent les performances du système respectivement sur le corpus DDI et PPI, mais pas sur le corpus i2b2 2010. Les corpus DDI et PPI sont formés de résumés d'articles scientifiques, les phrases de ces corpus sont donc bien formées. Le corpus i2b2 2010 est composé de comptes rendus médicaux dans lesquels les phrases ne sont pas toujours bien construites, elles peuvent n'être composées que d'énumérations de traitements et elles contiennent beaucoup d'abréviations. De ce fait l'analyse syntaxique des textes du corpus i2b2 est moins fiable que celle des corpus DDI et PPI. Nous posons donc l'hypothèse que les informations extraites des arbres pour le corpus i2b2 ne sont pas toujours correctes. Il serait intéressant d'évaluer l'analyse syntaxique des comptes rendus médicaux et étudier la manière de l'améliorer pour permettre d'utiliser des informations syntaxiques plus riches pour la classification des relations du corpus i2b2.

# Chapitre 5

# Simplification de phrases pour l'extraction de relations

**Sommaire**



La prise en compte de la structure syntaxique n'est pas suffisante pour résoudre le problème de la variabilité d'expression des relations comme nous l'avons vu dans le chapitre précédent. Nous avons donc étudié la simplification de phrases dans le but de réduire le nombre d'expressions différentes des relations en les simplifiant. La réduction de la variabilité d'expression des relations pourrait permettre d'améliorer la classification des relations peu représentées dans le corpus. L'objectif de la simplification est de ne conserver que les mots indicateurs de la relation ou du moins supprimer les mots qui peuvent gêner la bonne classification de la relation pour améliorer l'extraction des relations.

Nous proposons quatre méthodes pour la simplification. La première est une approche à base de règles lexicalisées écrites manuellement, la deuxième adapte à la tâche d'extraction de relations la simplification faite par l'outil bioSimplify, la troisième est fondée sur des règles appliquées aux arbres de constituants, et la dernière utilise des techniques d'apprentissage automatique et est basée sur une combinaison de classifieurs.

Ces quatre méthodes appliquent des simplifications sur des formes différentes de la phrase : sur la forme linéaire de la phrase ou sur sa structure syntaxique. Les simplifications peuvent porter sur les mots de la phrase, les catégories morpho-syntaxiques, la position des mots dans la structure syntaxique de la phrase, la position des mots par rapport aux deux

entités E1 et E2, etc. Au travers de ces méthodes, nous allons explorer différents aspects de la simplification et évaluer leurs apports pour la tâche d'extraction de relations.

Avant de présenter ces méthodes, nous expliquons ce que nous entendons par simplification et les façons dont elle peut être intégrée au modèle d'apprentissage pour l'extraction de relations.

## 5.1 Définition de la simplification

Comme nous l'avons présenté dans l'introduction de la section 1.7, nous utilisons le terme simplification de phrases pour parler indistinctement de compression de phrases et de simplification de phrases.

Nous définissons ici la simplification comme l'extraction de l'information pertinente pour identifier des relations, qui consiste à garder ou annoter ce qui est nécessaire à l'identification de la relation, et à supprimer ou annoter les informations qui ne sont pas en rapport avec la relation ou qui peuvent perturber son identification. Simplifier une phrase permet donc de rassembler les entités et les marqueurs de la relation.

Les simplifications de phrases sont à effectuer pour chaque paire d'entités d'une phrase. En effet, il est nécessaire de conserver les deux entités considérées, mais les autres entités peuvent être éventuellement supprimées, et seuls les marqueurs de la relation entre E1 et E2 sont utiles. Par exemple, dans la phrase (48), pour la paire d'entités *hypotensive* et *further monitoring* (relation de type *un test est conduit en raison d'un problème*, TeCP) les mots importants sont *became*, *he was then transferred* et *for*. Alors que pour la paire d'entités *bradycardiac* et *atropine* (relation de type *un traitement améliore le problème*, TrIP) les mots les plus importants sont plutôt *which responded to*.

(48) He subsequently became [PB hypotensive] and [PB bradycardiac] in the Catheterization Laboratory which responded to [TREAT iv fluids], [TREAT dopamine], [TREAT atropine], and he was then transferred to the CCU for [TEST further monitoring].

Nous considérons que dans une phrase contenant une relation, il y a des informations indispensables pour identifier et classer la relation, d'autres inutiles, c'est-à-dire qui n'apportent pas d'informations supplémentaires pour repérer la relation et des informations qui gênent l'identification de la relation. La simplification consiste à repérer ces trois types d'informations pour ensuite les conserver, les supprimer ou leur donner des poids différents. Mais la simplification que nous proposons n'a pas pour objectif de produire une phrase grammaticalement correcte.

La simplification sera prise en compte de différentes façons par le système d'extraction de relations. Soit les informations identifiées comme inutiles sont supprimées de la phrase et le vecteur d'attributs est construit à partir de cette nouvelle phrase ; soit des attributs indiquant les mots les plus importants sont ajoutés ; soit seuls les attributs qui décrivent les mots importants sont conservés. Nous ne parlerons pas de ce dernier cas ici : ne conserver qu'une partie des attributs revient à faire une sélection d'attributs comme nous l'avons vu

dans la section 4.2.4.3. Dans les deux autres cas nous parlons de simplification et non de sélection d'attributs, puisque soit nous ajoutons des attributs pour marquer les mots les plus importants, soit nous supprimons des mots dans la phrase ; les attributs sont donc différents. Au lieu de prendre en compte les mots m3, m4 et m5 dans le contexte gauche de l'entité E1, nous prendrons en compte les mots m1, m2 et m5 par exemple. Il n'y a donc pas forcément moins d'attributs, mais ils ont des valeurs différentes. La simplification telle que nous la définissons est donc un moyen de modifier le contexte local de la paire d'entités pour ne conserver que les attributs utiles.

## 5.2 Simplification à base de règles

Nous proposons une méthode simple à base de règles Cette méthode consiste à normaliser des informations secondaires pour notre tâche d'extraction de relations, comme les indications temporelles ou les dosages des médicaments, et à supprimer les segments de la phrase qui sont facilement identifiables comme étant inutiles. Les règles ont été développées à partir du corpus d'entraînement d'i2b2 2010. Elles s'appliquent sur la forme linéaire de la phrase, c'est-à-dire que les règles prennent en compte les mots, leur position dans la phrase et par rapport à E1 et E2. Une fois les informations annotées, il est possible de les supprimer ou de les remplacer par une balise simple indiquant le type d'information (par exemple `<SERV>` pour les services hospitaliers). En utilisant des balises, on réalise une sorte de normalisation des indications de lieux, de temps, de personnes et des informations posologiques. Les balises permettent de conserver l'information sémantique sans la réalisation lexicale.

Avec cette méthode la simplification est faite en quatre étapes :

1. Annotation des indicateurs de lieux (*at home*, *in the clinic*), des services hospitaliers (*the neuro/oncology service*), des indicateurs temporels (*at the time of the discharge*), des noms de médecin (*Dr. **NAME[ZZZ]*) ou de leur fonction (*radiologist*), et des noms de personne (*Mr. Larsen*) ou des références au patient (*the patient*). Ces annotations permettent de normaliser des informations secondaires et ainsi réduire la variabilité lexicale. Un exemple de cette annotation est présenté dans la phrase (49)[1].

   (49) He then went to SERV the Operating Room TIME on <DATE> where he had TREAT resection of PB the left face squamous cell cancer by DOC Dr. <NAME> and TREAT split thickness and full thickness skin grafting of PB the defects TIME on <DATE>.

2. Annotation des informations posologiques par l'outil COKAINE développé par Grouin *et al.* [2010]. Cet outil annote les médicaments et les informations associées

[1] Les balises `<DATE>` et `<NAME>` sont issues de la normalisation des données anonymisées.

sur la prise de ce traitement : le dosage, la fréquence de prise, la durée du traitement, le mode d'administration et la raison de ce traitement. Dans le corpus i2b2 2010, les traitements étaient déjà annotés, ce qui nous intéresse est l'annotation des indications de dosage, fréquence, durée, mode et raison. Comme pour l'étape précédente, les annotations des informations posologiques constituent une normalisation d'informations non indispensables. La phrase (50)[2] montre un exemple d'annotation de ces informations. Nous balisons toutes les informations posologiques annotées par COKAINE avec une seule balise : `<DO>`.

(50) We will given [TREAT valium] [DO <NUM> milligram] [MO orally] [FREQ evry <NUM> hours] for [PB withdrawal].

3. Annotation de segments de phrases pouvant être supprimés, en fonction de la position du segment par rapport à E1 et E2, par exemple les mots qui précèdent la première virgule avant E1. Les suppressions effectuées sont les suivantes :
   - Suppression des mots situés avant la virgule précédent E1 ;
   - Suppression des mots qui précèdent *which* et *but* s'il n'y a ni E1 ni E2 ;
   - Suppression des mots situés après la virgule suivant E2 ou après *and* ;
   - Suppression des entités de type problème médical coordonnées avec E1 si c'est un problème et/ou E2 si c'est un problème ;
   - Suppression des formulations du type : *the patient had* , *he was*, etc., des auxiliaires et du déterminant *no* marqueur de négation [3] ;
   - Suppression des expressions parenthésées.

   Dans l'exemple (59), ont été supprimés le segment de phrases avant E1 et l'auxiliaire *will*.

   (51) 1. GIven patient's complaint of [PB dysphagia], will check [TEST thyroid US] to follow-up status of [PB multinodular goiter].
   2. check [TEST thyroid US] to follow-up status of [PB multinodular goiter].

   Ces règles de simplification par suppression résultent des études sur l'apport des sous-arbres minimaux versus les arbres complets (cf. section 4.3). Dans certains cas elles conduisent à ne conserver que les mots contenus dans le sous-arbre minimal, c'est-à-dire les mots entre E1 et E2 (voir exemple (52)), dans d'autres cas elles sont moins réductrices et permettent de conserver des mots dans le contexte droit et/ou gauche des entités (53)).

   (52) 1. Abdoment soft, [PB nontender] and [PB distended].
   2. [PB nontender] and [PB distended].

   (53) 1. Speech was fluent but [PB slow], and there was [PB long response latencies] before she would initiate speech.

[2] `DO` pour dosage, `MO` pour mode d'administration, `FREQ` pour fréquence et `<NUM>` pour un nombre quelconque.

[3] *no* est parfois employé devant les entités de type problème lorsqu'un examen clinique a révélé l'absence de ces problèmes, c'est-à-dire dans le cas d'une relation de type *un test révèle un problème* (TeRP). Dans la définition de cette classe de relation, il est indiqué que l'examen peut révéler la présence ou l'absence d'un problème.

2. [PB slow], and there was [PB long response latencies] before she would initiate speech.

4. Suppression des entités en coordination avec E1 et E2 grâce au module de gestion de la coordination (cf. 4.2.2).

Le tableau 5.1 présente les proportions de phrases simplifiées avec les trois premières étapes de la simplification dans le corpus d'entraînement et dans le corpus d'évaluation [4]. On observe qu'il y a peu de phrases dans lesquelles des informations posologiques ont été annotées par COKAINE, en revanche les règles de suppression s'appliquent sur plus de 70% des phrases.

| | ANNOT | COKAINE | SUP | ANNOT + COKAINE + SUP |
|---|---|---|---|---|
| TRAIN + TEST | 34% | 6% | 73% | 80% |
| EVAL | 32% | 5% | 71% | 78% |

TAB. 5.1 – Proportion de phrases simplifiées avec trois des étapes de simplification

**Résultats** Nous avons évalué les trois premières étapes de la simplification séparément sur le corpus d'évaluation, puis toutes les étapes réunies. Deux évaluations ont été faites pour chaque étape de la simplification : les informations annotées sont supprimées ou elles sont remplacées par une balise. Le système utilise le module de gestion de la coordination qui a été évalué dans la section 4.2.4.1. Les résultats sont présentés dans le tableau 5.2. Les meilleurs résultats sont obtenus avec l'annotation des posologies grâce à l'outil COKAINE. Ils sont légèrement moins bons que ceux obtenus avec REMED sans simplification mais la différence n'est pas significative ($p>0{,}05$ avec le test de Student). Peu de phrases contiennent des informations posologiques dans le corpus, l'impact de cette normalisation est donc quasiment inexistant.

Les annotations devraient permettre de normaliser et réduire certaines informations. En effet, les informations annotées sont regroupées en une seule balise ; les informations pertinentes ont donc plus de chance d'être dans le contexte local des entités. Les résultats ne confirment pas cette hypothèse, soit parce que les informations annotées ne sont pas dans le contexte local des entités (c'est-à-dire trois mots avant E1, trois mots après E2 et entre E1 et E2), soit parce que ces informations, normalisées ou pas, ne sont pas utiles.

## 5.3 Simplification avec bioSimplify

Nous avons utilisé les simplifications faites par un outil existant, bioSimplify. bioSimplify est un outil de simplification de phrases pour améliorer les systèmes d'extraction d'information dans le domaine biomédical, qui a été développé par Jonnalagadda et Gonzalez [2011]. L'outil n'a pas été développé pour la tâche d'extraction de relations. Nous avons donc ajouté une phase de post-traitement pour sélectionner les simplifications pertinentes.

bioSimplify simplifie les phrases en deux étapes :

[4] Nous considérons une phrase par paire d'entités et nous ne comptons que les phrases contenant plus de deux entités.

| Relations | REMED | ANNOT | | COKAINE | | SUP | ANNOT + COKAINE + SUP | |
|---|---|---|---|---|---|---|---|---|
| | | sup | balise | sup | balise | | sup | balise |
| **TrIP** | 0,279 | 0,271 | **0,285** | 0,279 | 0,279 | 0,279 | 0,270 | 0,279 |
| **TrWP** | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| **TrCP** | **0,491** | 0,418 | 0,447 | **0,491** | **0,491** | 0,433 | 0,438 | 0,451 |
| **TrAP** | **0,728** | 0,708 | 0,709 | 0,726 | 0,727 | 0,703 | 0,699 | 0,707 |
| **TrNAP** | **0,189** | 0,132 | 0,132 | **0,189** | **0,189** | 0,149 | 0,166 | 0,166 |
| **PIP** | **0,635** | 0,634 | 0,633 | 0,629 | 0,630 | 0,624 | 0,626 | 0,627 |
| **TeRP** | **0,850** | 0,846 | 0,846 | **0,850** | **0,850** | 0,843 | 0,842 | 0,842 |
| **TeCP** | 0,393 | 0,382 | 0,386 | **0,399** | **0,399** | 0,381 | 0,363 | 0,391 |
| | **0,707** | 0,695 | 0,697 | **0,705** | **0,706** | 0,691 | 0,691 | 0,694 |

TAB. 5.2 – Variation de la f-mesure selon les simplifications effectuées sur le corpus d'évaluation.

- suppression des pré-modifieurs dans les syntagmes nominaux ;
- découpage syntaxique de la phrase en plusieurs phrases.

Exemples de règles de simplification utilisées par bioSimplify pour découper les phrases :

- extraction des propositions de la phrase si elles sont elles-mêmes des phrases grammaticales ;
- suppression des subordonnées complétives qui ne contiennent ni des groupes verbaux, ni des groupes nominaux ;
- suppression des groupes verbaux qui post-modifient un autre groupe verbal ;
- suppression des propositions inutiles (par exemple les indications de sections).

Au fur et à mesure de l'application des règles, la grammaticalité des phrases simplifiées est vérifiée.

Pour chaque paire d'entités, nous fournissons la phrase correspondante en entrée à bioSimplify. Il produit en sortie plusieurs phrases simplifiées, mais nous n'en avons conservé qu'une seule. En premier lieu, nous ne conservons que les phrases simplifiées qui contiennent les deux entités de la paire. Ensuite nous conservons la phrase simplifiée la plus petite et dans laquelle les deux entités sont les plus proches.

Par exemple, à partir de la première phrase de (54) bioSimplify propose 6 phrases simplifiées. La phrase 1, ainsi que la 4 et la 6 ne seront pas conservées, car elles ne contiennent pas les deux entités. La phrase la plus courte contenant les deux entités est la phrase 5 : *the arteriogram did show an occluded left tibial artery.* C'est cette phrase qui sera sélectionnée.

(54) The cardiovascular surgeon was consulted, and the patient underwent [TEST the arteriogram], which did show [PB an occluded left tibial artery].

1. The surgeon was consulted.
2. The patient underwent [TEST the arteriogram], which did show [PB an occluded left tibial artery].
3. The surgeon was consulted, and the patient underwent [TEST the arteriogram], which

did show [PB an occluded left tibial artery].

4. The patient underwent [TEST the arteriogram].
5. [TEST the arteriogram] did show [PB an occluded left tibial artery].
6. The surgeon was consulted, and the patient underwent [TEST the arteriogram].

Les simplifications proposées sont parfois inutiles ou dégradent les indicateurs de la relation. Par exemple, dans la phrase (55) les éléments supprimés sont situés dans une autre proposition que celle où sont E1 et E2, et le contexte local des entités n'est donc pas modifié. Dans la phrase (56), les règles de simplification appliquées suppriment le verbe *reveal* ce qui empêche le système de classer correctement la relation existante entre *blood culture sensitivities* et *GNR*.

(55) [PB Her PE] was unlikely due to the fact the patient was on [TREAT Coumadin] as an outpatient, and had been anticoagulated for some amount of time.

1. [PB Her PE] was unlikely due to the fact the patient was on [TREAT Coumadin] as an outpatient, and had been anticoagulated for some amount of time.
2. [PB Her PE] was unlikely due to the fact the patient was on [TREAT Coumadin] as an outpatient, and had been anticoagulated.
3. [PB Her PE] was unlikely due to the fact the patient was on [TREAT Coumadin] as an outpatient.

(56) [TREAT Gent] and [TREAT Cipro] were then discontinued and changed to [TREAT meropenem] when [TEST **blood culture sensitivities**] became available revealing [PB **GNR**] sensitive to [TREAT Jacoby] but resistant to both [TREAT Gent] and [TREAT Cipro].

[TEST **blood culture sensitivities**] became [PB **GNR**] sensitive to [TREAT Jacoby] but resistant to both [TREAT Gent]

**Résultats** Nous avons appliqué bioSimplify sur le corpus i2b2 2010. Dans le tableau 5.3, nous présentons le nombre de paires d'entités dans chaque corpus, ainsi que le nombre de paires pour lesquelles une phrase simplifiée a été conservée.

| corpus | # phrases | # phrases simplifiées | % |
|---|---|---|---|
| TRAIN + TEST | 16472 | 5909 | 35,8% |
| EVAL | 26081 | 9565 | 36,6% |

TAB. 5.3 – Nombre de simplifications effectuées par bioSimplify

Nous avons remplacé dans le corpus de départ les phrases qui avaient été simplifiées, et nous avons extrait les relations de ce nouveau corpus en utilisant le système avec le module de gestion de la coordination. Les résultats que nous avons obtenus sont présentés dans

le tableau 5.4. Les résultats que nous obtenons après avoir utilisé l'outil bioSimplify sont moins bons que ceux obtenus avec REMED sans faire de simplification. Les résultats ont été comparés aux résultats obtenus sans simplification. 470 relations ont été correctement classées grâce à la simplification et 490 ont été mal classées suite à la simplification.

Nous avons ensuite réalisé une combinaison des prédictions du système sans simplification (évalué dans la section 4.2.5) et du système utilisant les simplifications de bioSimplify, ce qui permet de conserver les relations correctement classées avec et sans simplification. La combinaison a été faite par un vote avec priorité donnée aux relations par rapport aux non-relations puis aux relations classées par le système REMED (pour la méthode de combinaison voir l'évaluation de l'utilisation d'arbres pour la classification multi-classes dans la section 4.3.2.2).

| | REMED | bioSimplify | | | combinaison | | |
|---|---|---|---|---|---|---|---|
| | F-mesure | Rappel | Précision | F-mesure | Rappel | Précision | F-mesure |
| **TrIP** | 0,279 | 0,202 | 0,833 | **0,325** | 0,191 | 0,863 | 0,314 |
| **TrWP** | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| **TrCP** | 0,491 | 0,358 | 0,706 | 0,475 | 0,405 | 0,664 | **0,503** |
| **TrAP** | 0,728 | 0,695 | 0,733 | 0,713 | 0,761 | 0,707 | **0,733** |
| **TrNAP** | 0,189 | 0,094 | 0,666 | 0,165 | 0,130 | 0,694 | **0,220** |
| **PIP** | 0,635 | 0,527 | 0,708 | 0,604 | 0,610 | 0,692 | **0,649** |
| **TeRP** | 0,850 | 0,835 | 0,864 | 0,849 | 0,866 | 0,841 | **0,853** |
| **TeCP** | 0,393 | 0,316 | 0,841 | **0,459** | 0,316 | 0,812 | 0,455 |
| | 0,707 | 0,629 | 0,784 | 0,698 | 0,679 | 0,758 | **0,716** |

TAB. 5.4 – Résultats de l'extraction des relations après simplification des phrases par bioSimplify

Un exemple de relation détectée après simplification de la phrase est présenté dans (57). Il s'agit d'une relation de type *un problème implique un autre problème* (PIP) pour le couple d'entités *a malignant lymphangitic spread* et *viral bronchiolitis*. Dans l'exemple (58), il existe une relation de type *un traitement est administré en raison d'un problème* (TrAP) entre *this patient's glucose control* et *diabetic*. Cette relation est correctement détectée dans la phrase d'origine mais pas dans la phrase simplifiée. Il aurait éventuellement fallu que bioSimplify conserve *to be vigilant regarding* et pas *we need to with*.

(57) They felt it was most likely [PB **a malignant lymphangitic spread**] with [PB pericardial involvement] versus [PB infectious], [PB other cardiovascular causes], or [PB **viral bronchiolitis**].
It was [PB **a malignant lymphangitic spread**] with [PB **viral bronchiolitis**].

(58) [TREAT Steroids] are a double-edged sword, however, and we need to be vigilant regarding [TREAT infection control] and with [TREAT **this patient's glucose control**] as he is [PB **diabetic**].

We need to with [TREAT **this patient's glucose control**] as he is [PB **diabetic**].

Cette méthode ne permet pas d'améliorer l'extraction de relations pour plusieurs raisons. Premièrement parce que seul un quart du corpus a pu être simplifié par bioSimplify, ce qui limite le nombre de relations pouvant être mieux classées. Deuxièmement la simplification a pour objectif de découper les phrases complexes en plusieurs phrases simples, ce qui ne modifie pas nécessairement le contexte local de la paire d'entités. Ensuite, seules des phrases grammaticalement correctes étant construites, certaines simplifications peuvent être pertinentes pour notre tâche mais ne sont pas proposées car elles ne permettent pas d'avoir des phrases grammaticales.

bioSimplify utilise des règles non lexicalisées qui portent sur la structure de la phrase, mais les simplifications ne sont pas faites en tenant compte des deux entités. La première méthode au contraire utilise des règles lexicalisées pour normaliser et des règles qui prennent en compte les positions des deux entités. Il serait alors intéressant de développer une méthode à base de règles qui soit destinée à la simplification pour l'extraction de relations et qui porte sur la structure de phrases.

## 5.4 Simplifier des arbres de constituants

Nous avons testé une méthode à base de règles s'appliquant sur les arbres de constituants. Contrairement à la première méthode appliquant des règles au niveau des mots de la phrase, cette méthode considère uniquement la structure syntaxique de la phrase et la position des deux entités E1 et E2 dans cette structure. Les règles ne sont pas lexicalisées, elles ne prennent donc pas en compte les mots de la phrase. Cette méthode permet de simplifier la structure de la phrase, par exemple en ne conservant que la proposition incluant les deux entités ou en supprimant les expressions parenthésées.

Les arbres simplifiés peuvent être utilisés de deux façons : soit tels quels avec un classifieur à noyau d'arbres, soit en engendrant la phrase résultante de ces simplifications. Nous n'évaluerons ici que la prise en compte de la phrase résultante de l'arbre simplifié, car nous pourrons utiliser le système Remed et comparer les résultats avec ceux obtenus avec d'autres méthodes.

Dans un premier temps, nous détaillons les règles que nous avons écrites puis nous présentons les expérimentations et les résultats obtenus.

Nous avons repris certaines des règles proposées par Miwa *et al.* [2010]. Ils ont défini deux groupes de règles : les règles qui s'appliquent sur les propositions de la phrase et celles qui agissent sur le contexte proche des entités. Nous avons repris deux des trois règles du premier groupe :

- on ne conserve que la plus petite proposition qui inclut E1 et E2 ;
- si une proposition relative contient E1 et E2, alors on construit une phrase simple à partir de la relative et de son antécédent.

La troisième règle, que nous n'avons pas utilisé, porte sur les verbes attributifs. Au vue de notre corpus nous ne l'avons pas jugé pertinente.

Nous avons ensuite écrit des règles pour la gestion de la coordination au niveau de

l'arbre de constituants qui permettent de ne conserver que E1 et/ou E2 lorsqu'elles sont coordonnées avec d'autres syntagmes nominaux.

Par exemple en appliquant ces règles sur l'arbre de constituants (figure 5.1) de la phrase (59), nous obtenons l'arbre simplifié (figure 5.2) qui correspond à la phrase simplifiée (60).

(59) On admission, the patient was found to have [PB a mild fever], [PB **myalgias**], and [PB arthralgias] that were relieved by [TREAT **Tylenol**].

(60) to have [PB **myalgias**] that were relieved by [TREAT **Tylenol**]

Nous avons également écrit des règles qui suppriment des syntagmes que nous ne jugeons pas utiles ou qui remplacent des syntagmes par une balise indiquant leur type.

- Suppression des syntagmes adverbiaux quand ils sont en tête d'une proposition ;
- Suppression des syntagmes adjectivaux s'ils ne contiennent ni E1 ni E2 ;
- Suppression des adjectifs, des noms ou des nombres qui suivent une entité ;
- Remplacement des syntagmes adjectivaux ou nominaux qui contiennent une balise <AGE> suivie du mot *year* par la balise <INFO> ;
- Si des syntagmes nominaux sont coordonnés et ne contiennent pas d'entités, on ne conserve que le dernier syntagme nominal ;
- Suppression des syntagmes prépositionnels s'ils contiennent la préposition *on* suivie d'une balise <DATE> ;
- Remplacement des adjectifs contenus dans un syntagme nominal mais ne contenant pas d'entités, par la balise <JJ> ;
- Suppression des syntagmes nominaux contenant une balise <NAME> mais pas d'entités ;
- Suppression des syntagmes nominaux composés au moins d'un nombre et d'un nom mais pas d'entités ;
- Remplacement des syntagmes prépositionnels contenant la préposition *of* suivie d'un nombre par <PP> ;
- Suppression des parenthèses, de leur contenu et de ce qui les précèdent si c'est un syntagme nominal, un syntagme adjectival, un nom ou un nombre ;
- Remplacement des syntagmes prépositionnels contenus par le même syntagme nominal qu'une entité, par <PP> ;
- Suppression des syntagmes prépositionnels situés au début d'une proposition ;
- Remplacement des syntagmes prépositionnels contenant *by* ou *to* par <PP>.

Par exemple, si on applique les règles sur l'arbre de constituants de la phrase (61), on obtient la phrase simplifiée (62).

(61) [TEST **Cath**] at Kindred/North Shore today showed [PB **80% LM lesion**] with normal LAD, CX, RCA.

(62) [TEST **Cath**] <PP> showed [PB **80% LM lesion**] <PP>.

Dans le tableau 5.5, nous avons indiqué le pourcentage de phrases de chaque corpus qui ont été simplifiées en fonction des règles utilisées. Avec l'utilisation des règles inspirées

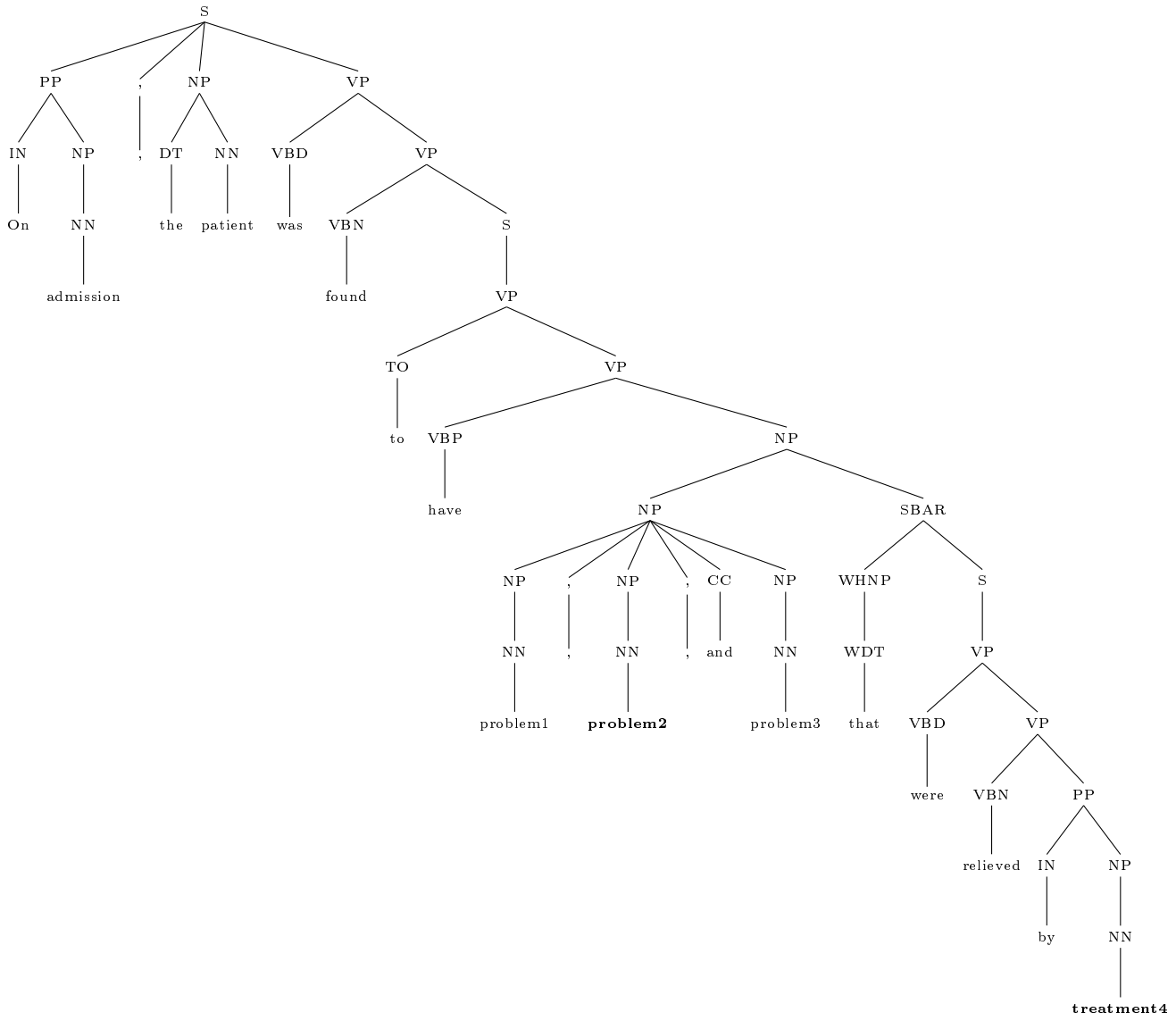



FIG. 5.1 – Arbre en constituants de la phrase (59)

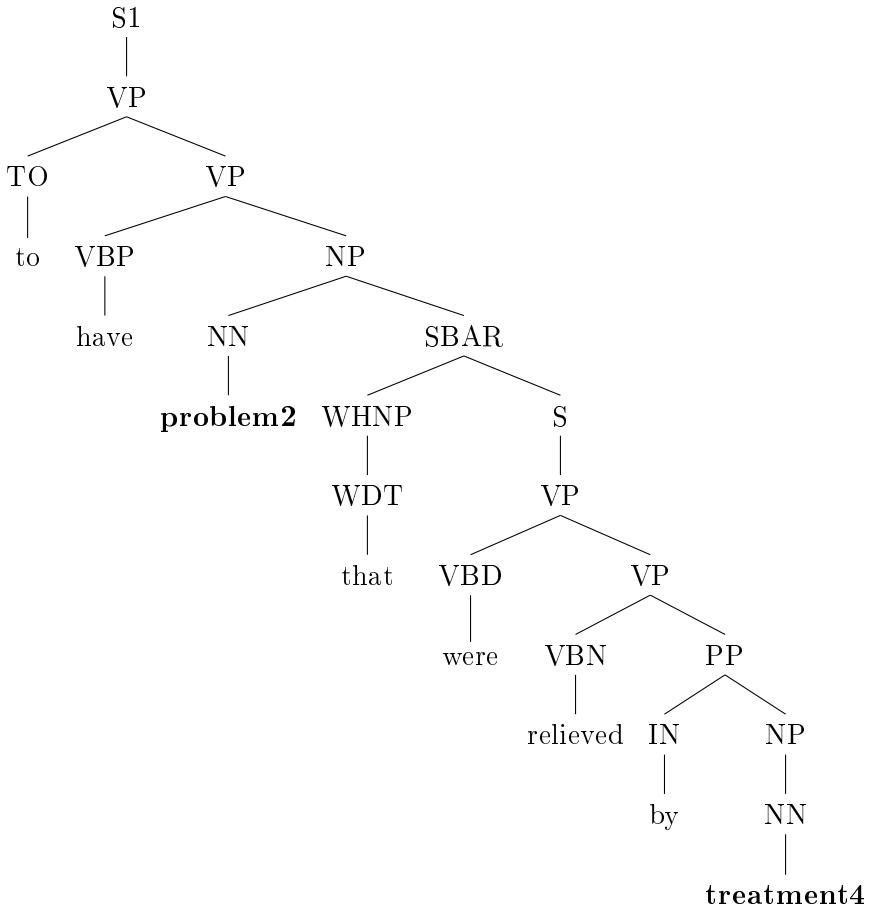



FIG. 5.2 – Arbre en constituants de la phrase (59) simplifiée

de Miwa *et al.* [2010], seul un quart des phrases du corpus d'entraînement ont été réduites et 19% des phrases du corpus d'évaluation. Les autres règles permettent de simplifier plus de la moitié des phrases.

| | RMiwa | Rcoord | Rsynt | RMiwa + Rcoord + Rsynt |
|---|---|---|---|---|
| TRAIN+TEST | 25% | 65% | 55% | 83% |
| EVAL | 19% | 65% | 56% | 84% |

TAB. 5.5 – Proportion de phrases simplifiées avec les trois séries de règles de simplification

**Résultats** Nous avons évalué indépendamment les règles que nous avons repris de Miwa *et al.* [2010] (RMiwa), les règles que nous avons écrites pour gérer la coordination (Rcoord) et les règles permettant de supprimer des syntagmes inutiles (Rsynt). Les résultats sont présentés dans le tableau 5.6. Nous observons que la classification n'est pas améliorée suite aux simplifications. Les résultats obtenus en utilisant les deux règles de Miwa *et al.* [2010] et les règles Rsynt ne sont pas significativement différents des résultats obtenus sans simplification (avec le test de Student, $p > 0{,}05$). En revanche, les prédictions du système utilisant les simplifications faites sur les coordinations sont significativement différentes de celles de REMED ($p < 0{,}05$) mais moins bonnes.

| | REMED | RMiwa | | | Rcoord | | | Rsynt | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F | R | P | F | R | P | F | R | P | F |
| **TrIP** | 0,279 | 0,166 | 0,846 | 0,278 | 0,196 | 0,886 | **0,322** | 0,191 | 0,863 | 0,314 |
| **TrWP** | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| **TrCP** | **0,491** | 0,349 | 0,679 | 0,461 | 0,362 | 0,745 | 0,487 | 0,344 | 0,701 | 0,462 |
| **TrAP** | 0,728 | 0,694 | 0,752 | 0,722 | 0,715 | 0,734 | 0,724 | 0,699 | 0,752 | **0,725** |
| **TrNAP** | **0,189** | 0,104 | 0,512 | 0,173 | 0,078 | 0,600 | 0,138 | 0,083 | 0,727 | 0,150 |
| **PIP** | **0,635** | 0,547 | 0,758 | **0,635** | 0,537 | 0,701 | 0,608 | 0,547 | 0,757 | **0,635** |
| **TeRP** | 0,850 | 0,828 | 0,878 | 0,852 | 0,830 | 0,877 | 0,853 | 0,842 | 0,867 | **0,854** |
| **TeCP** | 0,393 | 0,284 | 0,852 | **0,426** | 0,278 | 0,815 | 0,415 | 0,260 | 0,845 | 0,397 |
| | **0,707** | 0,628 | 0,804 | 0,705 | 0,633 | 0,786 | 0,701 | 0,633 | 0,802 | **0,707** |

TAB. 5.6 – Résultats de l'extraction des relations après simplification des arbres de constituants

Nous avons ensuite évalué l'impact de l'utilisation des trois groupes de règles (voir tableau 5.7). Les résultats sont moins bons qu'en appliquant une série de règles à la fois. Pour évaluer la différence dans la classification du système utilisant les phrases simplifiées et de REMED, nous avons combiné les prédictions des deux systèmes en donnant priorité aux prédictions sans simplification (cf. section 4.3.2.2 pour plus de détails sur la méthode de combinaison). Les résultats sont présentés dans les trois dernières colonnes du tableau 5.7. Grâce à la simplification, la f-mesure augment de 0,01. Nous pouvons donc conclure que la simplification permet de correctement classer des nouvelles relations par rapport à REMED, mais n'est pas suffisante pour améliorer les performances d'un classifieur.

(63) est un exemple de la simplification d'une phrase avec la combinaison de toutes les règles. Cette simplification a permis au système de correctement classer la relation de type *un traitement est administré en raison d'un problème* (TrAP). Une relation du même type

| | REMED | RMiwa + Rcoord + Rsynt | | | Combinaison | | |
|---|---|---|---|---|---|---|---|
| | F | R | P | F | R | P | F |
| **TrIP** | 0,279 | 0,196 | 0,866 | 0,230 | 0,186 | 0,860 | **0,307** |
| **TrWP** | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| **TrCP** | **0,491** | 0,351 | 0,699 | 0,467 | 0,380 | 0,681 | 0,488 |
| **TrAP** | 0,728 | 0,702 | 0,733 | 0,717 | 0,739 | 0,732 | **0,736** |
| **TrNAP** | 0,189 | 0,083 | 0,571 | 0,146 | 0,109 | 0,677 | **0,189** |
| **PIP** | 0,635 | 0,534 | 0,690 | 0,602 | 0,588 | 0,736 | **0,654** |
| **TeRP** | 0,850 | 0,832 | 0,879 | 0,855 | 0,863 | 0,854 | **0,858** |
| **TeCP** | 0,393 | 0,282 | 0,813 | **0,419** | 0,278 | 0,811 | 0,415 |
| | 0,707 | 0,629 | 0,783 | 0,698 | 0,663 | 0,782 | **0,718** |

TAB. 5.7 – Résultats de l'extraction des relations après simplification des arbres de constituants

existe entre *a cardiopulmonary arrest* et *a cardiopulmonary arrest* dans la phrase (64) mais suite à la simplification, la relation n'est plus correctement classée.

(63) Since presentation, patient had multiple episodes of [PB **cholangitis**] (<NUM> in past <AGE> years with last on <NUM>), always short lived and treated with [TREAT **antimicrobial therapy**].

Simp : patient had multiple episodes of [PB **cholangitis**], always short lived and treated with [TREAT **antimicrobial therapy**].

(64) The family realizes that the long term outlook for this patient is not very good, but they would like to currently continue to provide him with [TREAT all support necessary], short of [TREAT resuscitation measures] or [TREAT **heroic measures**] should he have [PB **a cardiopulmonary arrest**].

Simp : [TREAT **heroic measures**] should he have [PB **a cardiopulmonary arrest**]

La suite de cette étude pourrait être d'évaluer l'impact de la simplification des arbres en utilisant un classifieur à noyau d'arbres (cf. section 4.3.2.2). Il serait alors possible de comparer l'apport de l'utilisation de l'arbre complet, du sous-arbre minimal et de l'arbre simplifié pour la détection et la classification des relations.

Suite à ces conclusions, nous avons voulu proposer une méthode de simplification dirigée par la tâche d'extraction de relations, qui cherche à simplifier le contexte local de la paire d'entités et sans vouloir conserver la grammaticalité de la phrase.

## 5.5 Apprendre la simplification

Les trois méthodes précédentes ne permettent pas d'améliorer significativement les performances du système d'extraction de relations, entre autres parce qu'elles ne modifient pas nécessairement le contexte local de la paire d'entités et pour bioSimplify, parce qu'elle n'est pas dirigée par la tâche d'extraction de relations.

La plupart des travaux existants ont défini des règles de simplification. Nous avons présenté trois méthodes qui utilisent des règles dans les sections précédentes (cf. 5.2, 5.3 et 5.4). Les exemples montrent que celles-ci sont très contextuelles et dépendantes de la tâche, et reposent sur une étude de corpus plutôt que sur une connaissance a priori de la langue. Ainsi, des règles usuellement définies pour la simplification comme la suppression de relatives ne s'appliquent pas dans notre contexte. Une modélisation sous forme de règles nécessite d'en redéfinir un grand nombre sans être sûr d'obtenir une amélioration comme nous le montrent les résultats présentés dans la section 5.2.

Nous avons conçu une méthode originale fondée sur un apprentissage automatique de la simplification pour annoter dans les phrases les parties à conserver et celles que l'on peut supprimer. Cette méthode de simplification est dirigée par la tâche d'extraction de relations. Nous proposons d'annoter la simplification en apprenant sur un petit corpus annoté, puis d'évaluer l'annotation selon son impact sur l'extraction des relations, et enfin nous complétons le corpus annoté grâce aux résultats de l'extraction des relations. Notre méthode s'appuie sur un corpus de développement annoté en relations (DEV). Nous l'avons divisé en deux : un corpus d'entraînement TRAIN et un corpus de test TEST.

Cette méthode et les premiers résultats obtenus sont présentés dans Minard *et al.* [2012].

### 5.5.1 Choix du schéma d'annotation

Nous avons choisi d'annoter la simplification au niveau des mots, c'est-à-dire d'associer à chaque mot une étiquette indiquant l'importance du mot pour détecter la relation entre la paire d'entités. Quatre types d'annotation ont été définis. L'annotation « indispensable » permet de caractériser les mots qui portent l'expression de la relation. L'annotation « utile », très proche de « indispensable », indique les mots qui renforcent la relation. L'annotation « inutile » est associée aux mots n'apportant pas d'indices pour la classification de la relation, par exemple l'indication du service dans lequel est le patient. L'annotation « gênant » sert à repérer les mots pouvant gêner la bonne classification de la relation. Dans les exemples (65) et (66), les parties de phrase « indispensables » sont soulignées, les parties « inutiles » sont normales, les parties « gênantes » sont barrées et les concepts à mettre en relation sont en gras. Dans l'exemple (65), une relation du type *l'examen est conduit en raison du problème médical* (TeCP) doit être repérée entre *a magnetic resonance imaging study* et *a small vascular malformation.* Dans l'exemple (66), il s'agit d'une relation de type *un traitement est administré en raison du problème médical* (TrAP) entre *the tremendous tumor burden* et *open debulking.*

(65) [TEST **A magnetic resonance imaging study**] will be scheduled as an outpatient in three months to rule out [PB**a small vascular malformation**] if responsible for

[PB the hemorrhage].

(66) The neuro-oncologist felt that because of [PB **the tremendous tumor burden** ~~that was likely causing his symptoms~~ the patient will require [TREAT **open debulking**] as well as obtaining issue for a pathologic diagnosis.

### 5.5.2 Méthode

Nous avons choisi d'utiliser un classifieur à base de CRF (*Conditional Random Field*) ou « Champs Aléatoires Conditionnels » pour effectuer l'annotation des phrases. Les CRF sont des modèles statistiques qui ont la particularité de modéliser des dépendances entre annotations. Les CRF considèrent les probabilités des séquences d'annotations possibles pour une séquence d'observations donnée (les observations peuvent être passées et futures).

Les phrases annotées seront données en entrée du classifieur SVM pour la classification des relations. Afin de n'annoter manuellement que quelques phrases, nous proposons une architecture où la simplification est guidée par la tâche d'extraction de relations, et le corpus d'apprentissage de la simplification est augmenté itérativement en fonction des résultats de la tâche finale. Cette méthode est donc facilement adaptable à un autre domaine, contrairement aux méthodes à base de règles qui ne permettent pas toujours une adaptation simple et rapide.

#### 5.5.2.1 Annotation par CRF

**Constitution du corpus d'apprentissage** Dans un premier temps, nous avons sélectionné 71 phrases provenant du corpus TRAIN : 14 phrases contiennent des paires d'entités qui avaient été correctement classées par notre système d'extraction de relations, 37 phrases contiennent des paires mal classées et 20 des paires qui ne sont pas en relation mais qui avaient été classées comme étant en relation. Une étude de leurs caractéristiques a été menée préalablement à l'annotation.

Cette étude a montré que dans 14 phrases du corpus, la relation est exprimée par un verbe et les deux entités en relation sont respectivement sujet et complément de ce verbe (67).

(67) [TEST **An magnetic resonance imaging study**] **showed** [PB **basilar artery disease**], questionable aneurysm.

Dans 14 phrases, les deux entités en relation sont dans deux propositions différentes (68). Sept constructions différentes ont été trouvées; nous en présentons trois dans le tableau 5.8.

(68) Finger tapping and [TEST **rapid alternating movements**] were slow on the left and she had [PB **trouble isolating individual finger movements**].

Dans 18 phrases, les deux entités sont reliées par une préposition, et la relation s'exprime au travers de la préposition et du verbe de la proposition (69).

| | |
|---|---|
| Prop Conj \| Prop Princ | Although **TREAT** were adjusted he continued to be **PB** and there was [...] |
| Prop Indep \| CC \| Prop Indep | TEST became **PB** and **TREAT** was held. |
| Prop Princ \| Prop Rel \| CC \| Prop Indep | He subsequently became PB and **PB** in the Catheterization Laboratory which responded to TREAT, TREAT, **TREAT**, and he was then transferred to the CCU for TEST. |

TAB. 5.8 – Phrases dans lesquelles les concepts en relation sont dans deux propositions différentes

(69) [...], she had [PB**an acute drop**] **in** [TEST**her systolic blood pressure**] to <NUM> for unclear reasons and without evidence of [PBacute sepsis].

Dans les 20 exemples de non-relations que nous avons dans notre corpus, dans seulement deux phrases les deux entités sont sujet et objet du même verbe. Dans huit phrases, les deux entités sont dans des propositions différentes et dans neuf phrases elles sont reliées par une préposition.

Cette étude fait apparaître l'existence de régularités, que la simplification pourrait dégager.

Les 71 phrases ont été annotées par trois annotateurs grâce au logiciel Knowtator de Protégé [5]. Les différences ont donné lieu à discussion et accord.

Une phrase pouvant contenir plus d'une paire d'entités, nous les avons annotées pour une paire d'entités définie ; de ce fait certaines phrases sont en double dans le corpus, mais à chaque fois pour une paire d'entités différente.

Dans le tableau 5.9, nous donnons pour chaque classe de simplification le nombre de mots associés à cette classe dans le corpus TRAIN_SIMP (le corpus annoté obtenu). On remarque que très peu de mots sont annotés « utile », la raison étant la difficulté de distinction entre les classes « indispensable » et « utile ». Ces deux classes seront donc regroupées ultérieurement.

| étiquette | nombre de mots |
|---|---|
| indispensable | 287 |
| utile | 52 |
| inutile | 608 |
| gênant | 177 |

TAB. 5.9 – Étude du corpus annoté

Après avoir effectué les premiers tests, nous avons augmenté le corpus TRAIN_SIMP en annotant 55 phrases supplémentaires pour des paires d'entités en relation. Le corpus TRAIN_SIMP est donc composé de 126 phrases annotées.

[5] http://knowtator.sourceforge.net/

**Application du CRF** Nous avons utilisé le classifieur CRF++ (Kado [2003]) pour apprendre à annoter les simplifications : à chaque mot il attribue une étiquette en fonction de la valeur des attributs pour ce mot.

Les attributs fournis au classifieur sont : le lemme fourni par le TreeTagger, la catégorie morpho-syntaxique provenant de l'analyse du TreeTagger, le nombre de caractères du mot, la position du mot dans la phrase (position du mot/nombre de mots dans la phrase), la classe de la relation de la paire étudiée (elle provient de la référence pour le corpus d'entraînement, et des prédictions du classifieur sans simplification pour le corpus de test), le type sémantique si le mot fait partie d'une entité et une étiquette indiquant si le mot fait partie d'une des entités de la paire étudiée. Ce dernier attribut permet d'avoir une annotation dépendante d'un couple particulier de concepts. Les dépendances séquentielles sont calculées, pour chaque type d'attributs, avec un contexte de trois mots avant et trois mots après le mot courant.

#### 5.5.2.2 Combinaison de classifieurs pour l'extraction des relations

Avec seulement une centaine de phrases annotées, la simplification obtenue ne permet pas d'améliorer l'extraction des relations. Pour augmenter le corpus TRAIN_SIMP et améliorer la simplification, nous avons combiné les deux classifieurs, et utilisé les résultats de la classification des relations pour augmenter le corpus TRAIN_SIMP. La figure 5.3 présente un schéma de la méthode.

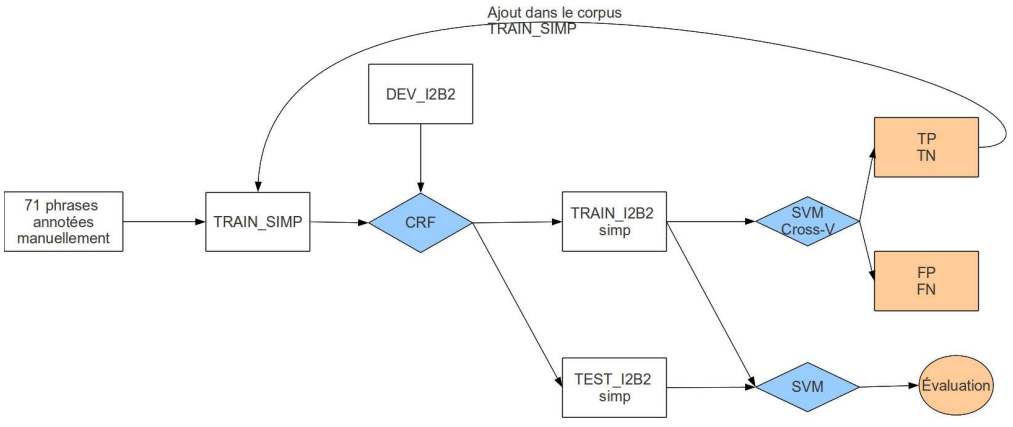



FIG. 5.3 – Schéma explicatif de la méthode

**Annotation de la simplification** Dans un premier temps, le corpus annoté manuellement est utilisé comme amorce pour la simplification; il forme le corpus d'entraînement TRAIN_SIMP. Il est utilisé pour apprendre les simplifications grâce à l'outil CRF++. Ensuite le modèle pour la simplification est appliqué sur la totalité du corpus DEV.

**Extraction des relations** Nous utilisons ensuite ce corpus annoté pour extraire les relations grâce au système REMED. Les annotations des simplifications sont utilisées comme

des attributs supplémentaires pour la reconnaissance des relations. Un attribut identifie les mots annotés comme « inutile », un autre pour les mots « gênants » et un pour les « utiles ».

Le corpus DEV a été divisé en deux : corpus TRAIN et corpus TEST. Dans un premier temps, nous effectuons l'extraction des relations par validation croisée en 10 parties avec le corpus TRAIN. À chaque itération, les phrases contenant des relations correctement extraites (les vrais positifs et les vrais négatifs) et leurs annotations pour la simplification sont ajoutées au corpus TRAIN_SIMP.

**Validation de la simplification** Une fois la validation croisée terminée sur le corpus TRAIN, le corpus DEV est de nouveau annoté pour la simplification à l'aide du classifieur à base de CRF et du corpus TRAIN_SIMP augmenté. Ensuite le modèle pour l'extraction des relations est construit à partir du corpus TRAIN et il est appliqué sur le corpus TEST. Cette étape permet d'évaluer l'impact de la complétion du corpus TRAIN_SIMP sur l'extraction des relations.

**Récapitulatifs des étapes de la méthode**

1. Construction d'un modèle de simplification à partir des 126 phrases annotées ;
2. Application du modèle de simplification sur le corpus d'entraînement ;
3. Classification des relations par validation croisée sur le corpus d'entraînement annoté en simplification ;
   - Comparaison des prédictions avec la référence : si la prédiction est bonne (c'est-à-dire qu'elle est équivalente à la référence) et qu'elle respecte un autre critère (par exemple, si la prédiction est différente de celle faite par le système REMED), alors la phrase annotée en simplification est ajoutée au corpus d'apprentissage de la simplification ;
4. Construction d'un modèle de simplification à partir du corpus d'apprentissage augmenté ;
5. Application du modèle de simplification sur le corpus d'entraînement et de test ;
6. Classification des relations du corpus de test ;
7. Si les résultats de la classification sont différents/meilleurs que ceux obtenus précédemment, alors recommencer à l'étape 2, sinon appliquer le modèle de simplification construit à l'étape 1 ou 4 sur le corpus d'évaluation et classer les relations de ce corpus.

### 5.5.3 Évaluations

Afin d'évaluer les résultats de la simplification, nous avons défini plusieurs protocoles qui permettent de :

- mesurer l'impact de la simplification sur le processus de classification;
- noter le résultat de la simplification;
- analyser de manière qualitative les annotations effectuées par le module de simplification.

#### 5.5.3.1 Évaluation de la pertinence de l'annotation manuelle

Dans un premier temps, nous avons testé la validité de notre hypothèse par vérifier manuellement que les phrases simplifiées étaient mieux classées. Nous avons extrait les relations du corpus TRAIN_SIMP avec le système REMED (après avoir enlevé du corpus TRAIN_I2B2 les paires d'entités contenues dans le corpus TRAIN_SIMP). Nous avons évalué l'extraction des relations sans modifier les phrases. Nous avons d'abord supprimé les mots gênants puis les mots gênants et inutiles. Les résultats sont présentés dans le tableau 5.10.

Nous observons que avec la suppression des mots gênants, 7 nouvelles relations sont correctement classées, et avec la suppression des mots gênants et inutiles, 9 nouvelles relations sont correctement classées mais 4 relations ne sont plus bien classées. Cette évaluation montre que la suppression des mots inutiles a un impact sur l'extraction des relations et améliore globalement la classification.

| | CC | | IC | |
|---|---|---|---|---|
| Avant simplification | 34 | | 37 | |
| | Identique | Nouveau | Identique | Nouveau |
| Suppression des mots gênants | 34 | 7 | 30 | 0 |
| Suppression des mots gênants et inutiles | 30 | 9 | 27 | 5 |

TAB. 5.10 – Évaluation de l'annotation (CC : correctement classé, IC : incorrectement classée, Identique : mêmes relations qu'avant la simplification, Nouveau : nouvelles relations correctement (ou incorrectement) classées avec la simplification

#### 5.5.3.2 Évaluation de la pertinence de l'annotation avec CRF++

Nous n'avons pas de corpus annoté pour évaluer la simplification. Pour vérifier que l'annotation de la simplification avec CRF++ était cohérente, nous avons étudié les mots classés dans chacune des trois catégories (nous avons regroupé « utile » et « indispensable » dans une même catégorie « utile »). Pour chaque lemme, nous avons compté combien de fois il apparaissait dans le corpus TRAIN et combien de fois il était associé à une des trois catégories. Le tableau 5.11 contient les lemmes les plus fréquents dans chaque catégorie et qui apparaissent au moins 10 fois dans cette catégorie. Nous observons que les lemmes les plus fréquemment étiquetés « gênant » font partie de concepts; par exemple on retrouve *fluticasone* dans le traitement *fluticasone propionate* ou *fluticasone-salmeterol*. Les lemmes les plus souvent étiquetés « utile » sont principalement des verbes, et ceux étiquetés « inutile » sont des unités (reliées à des dosages), des informations sur le patient (son nom, son âge), etc. Nous avons conclu de cette étude que le classifieur se comporte de manière cohérente pour annoter la simplification.

| UTILE | | INUTILE | | GENANT | |
|---|---|---|---|---|---|
| attribute | 50 / 56 | ml | 582 / 582 | neutropenia | 10 / 10 |
| presence | 12 / 17 | before | 260 / 260 | ph | 19 / 21 |
| questionable | 16 / 23 | yo (*year-old*) | 219 / 219 | thromboplastin | 23 / 27 |
| vs | 21 / 31 | microgram | 211 / 211 | fluticasone | 11 / 13 |
| identify | 27 / 40 | caution | 201 / 201 | diskus | 11 / 13 |
| demonstrate | 130 / 194 | mr. | 184 / 184 | migraine | 52 / 62 |
| inaccurate | 12 / 18 | ask | 177 / 177 | spiriva | 22 / 27 |
| due | 314 / 488 | asacol | 172 / 172 | panic | 42 / 52 |

TAB. 5.11 – Exemple d'annotation de lemmes présents plus de 10 fois dans le corpus

#### 5.5.3.3 Mesure de l'impact de la simplification sur la tâche d'extraction de relations

Pour évaluer l'impact de la simplification sur la classification des relations, nous avons mené plusieurs expérimentations en faisant varier un grand nombre de paramètres. Les différents paramètres sont les suivants :

- CRF (annotation de la simplification)
  - Classes : 4, 3 ou 2 étiquettes (« indispensable », « utile », « inutile », « gênant »)
  - Attributs : lemmes, catégories morpho-syntaxiques, structures syntaxique, nombre de lettres du mot, position du mot, classes de VerbNet, étiquettes de normalisation (indication de lieu, de nom de personne, date, etc.), types sémantiques, etc.
  - Paires d'entités utilisées : seulement les relations, ou les relations et les non-relations
- REMED (extraction des relations)
  - Classes de simplification : 4, 3 ou 2 étiquettes (« indispensable », « utile », « inutile », « gênant »)
  - Prise en compte de la simplification : suppression des mots gênants et/ou inutiles, ou ajout d'attributs indiquant la classe de simplification des mots
- Ajout de nouveaux exemples annotés pour l'apprentissage de la simplification
  - Ajouter toutes les paires d'entités correctement classées, ou seulement les paires qui n'ont pas été correctement classées sans simplification mais bien avec simplification
  - Sélection des paires selon le score de décision du classifieur

Nous présentons ici la configuration donnant les meilleurs résultats. Nous avons appris la simplification en donnant les attributs suivants pour chaque mot : catégorie morpho-syntaxique, position dans la phrase, relation détectée sans simplification pour cette paire d'entités, type de l'entité si le mot fait partie d'une entité et si l'entité est E1 ou E2. Nous n'avons appris que deux classes : « utile » et « indispensable » sont regroupées en une classe, et « inutile » et « gênant » aussi. Pour prendre en compte la simplification, nous avons supprimé les informations inutiles et gênantes pour l'étape de validation croisée. Pour les phases de test et d'évaluation, nous avons donné des attributs supplémentaires au classifieur : tous les mots utiles de la phrase sont identifiés par un attribut en plus de tous les mots du contexte local des entités. Ensuite, après avoir classé les relations par validation croisée sur le corpus TRAIN_I2B2, nous avons ajouté dans le corpus pour la

simplification TRAIN_SIMP les phrases contenant des relations correctement classées, et les phrases contenant des paires qui ne sont pas en relation et qui ont été correctement classées uniquement avec la simplification (elles étaient mal classées par le système n'utilisant pas la simplification). Nous avons exécuté cinq fois le système complet, après quoi nous avons obtenu 2 334 phrases dans le corpus TRAIN_SIMP, dont 126 qui ont été annotées manuellement. Dans le tableau 5.12, nous donnons les f-mesures obtenues sur le corpus de test TEST_I2B2 sans simplification, avec la simplification apprise avec les 126 phrases annotées (avant la combinaison des deux méthodes) et avec le corpus TRAIN_SIMP obtenu après cinq itérations.

| Relation | Sans Simplification | Simplification | |
|---|---|---|---|
| | | Corpus TRAIN_SIMP non augmenté | Corpus TRAIN_SIMP augmenté |
| **TrIP** | 0,380 | **0,476** | **0,476** |
| **TrWP** | 0,000 | 0,000 | 0,000 |
| **TrCP** | 0,507 | **0,561** | 0,551 |
| **TrAP** | 0,770 | 0,777 | **0,790** |
| **TrNAP** | 0,645 | **0,666** | **0,666** |
| **PIP** | **0,703** | 0,676 | 0,688 |
| **TeRP** | **0,877** | 0,873 | 0,869 |
| **TeCP** | 0,512 | 0,541 | **0,564** |
| | 0,758 | 0,760 | **0,765** |

TAB. 5.12 – Évaluation de la classification des relations avec et sans simplification sur le corpus TEST_I2B2

Nous avons appliqué la simplification sur le corpus d'évaluation EVAL_I2B2 afin d'évaluer la classification des relations. Les résultats que nous obtenons sont moins bons que sans simplification. La phrase (70) est un exemple de l'annotation de la simplification qui a permis de bien classer la relation de type *un traitement est administré en raison d'un problème* (TrAP) qui existe entre *inability to wean* et *a tracheostomy*.

(70) ~~Because of~~ [TREAT her prolonged intubation] and [PB **inability to wean**] ,~~on February <NUM>~~ , she had [TREAT **a tracheostomy**] ~~placed~~.

Nous avons comparé les relations correctement classées après simplification mais pas avec REMED et les relations correctement classées avec l'utilisation des arbres de constituants (cf. 4.3.2.2) mais pas avec REMED. Nous avons observé que presque la moitié (93 relations sur 190) des relations correctement classées après simplification l'était aussi avec l'utilisation des arbres. Mais uniquement 21% des relations correctement classées avec l'utilisation des arbres l'étaient aussi après simplification. Nous en concluons que l'utilisation des arbres de constituants permet de correctement classer plus de nouvelles relations que l'utilisation de la simplification.

#### 5.5.3.4 Évaluation manuelle de la simplification

Nous n'avons pas de corpus annoté suffisamment grand pour pouvoir faire une évaluation automatique de la tâche de simplification. De ce fait, nous avons choisi d'étudier l'annotation manuellement de 48 relations et d'évaluer manuellement la simplification pour ces relations. Nous n'avons étudié que l'annotation des phrases portant sur une paire d'entités en relation. En effet, ainsi que nous l'avons déjà mentionné, il est difficile de définir ce qui doit être annoté pour les non-relations.

Nous avons donc étudié 48 relations du corpus de test et avons comptabilisé le nombre de relations correctement simplifiées, simplifiées à tort ou partiellement simplifiées à raison. Peu d'informations sont annotées comme gênantes. Aussi, lors de l'évaluation, nous considérons que des informations annotées indispensables sont des informations à garder et que les autres sont des informations à supprimer. Nous avons considéré exacts les cas où le module garde toutes les informations pertinentes, même s'il garde aussi quelques informations que nous jugeons inutiles. Nous avons considéré comme faux les simplifications qui suppriment des informations que nous jugeons indispensables, et partiellement corrects les cas où le module aurait dû garder plus d'informations utiles, mais a gardé quand même les informations indispensables, ou lorsque trop d'informations qu'il aurait dû considérer comme inutiles sont gardées. Avec cette répartition en trois classes, nous obtenons 31 cas exacts, 11 cas faux et 6 cas partiellement corrects.

Dans l'exemple (71), nous considérons que la simplification est correcte mais dans l'exemple (72) le verbe le plus utile à la détection de la relation (*revealed*) est annoté inutile, et l'annotation de la simplification est donc fausse.

(71) He had [TEST **a cardiac catheterization**] performed which revealed [PB **a three vessel coronary artery disease**] with [PB an occluded RCA] , [PB 70%-80% proximal LAD] , and [PB a high grade left circumflex lesion after the OM] with [PB distal left circumflex occlusion] .

(72) He had [TEST **a cardiac catheterization**] performed which revealed [PB a three vessel coronary artery disease] with [PB an occluded RCA] , [PB 70%-80% proximal LAD] , and [PB **a high grade left circumflex lesion after the OM**] with [PB distal left circumflex occlusion] .

#### 5.5.3.5 Analyse des simplifications

Nous avons tenté d'établir les types de simplifications apprises. Les différentes structures de phrases qui apparaissent sont :

- *E1 relation E2* pour lesquelles la partie située entre les entités doit être conservée, tout ou en partie; cette structure est généralement bien traitée;

- *E1 relation (coordination d'entités) E2* est généralement mal annotée, et la marque de la relation est souvent supprimée. Ce type de structure peut être reconnu simplement par des règles;
- *E1 (structure comportant des entités) relation E2* est généralement reconnue et la relation est gardée.

Certains cas nécessitent de garder la partie gauche de la première entité ; cette configuration est mal reconnue. Il en est de même pour les contextes droits de la deuxième entité. Ces deux types de structure sont plus rares, et leur traitement nécessite plus d'exemples.

## 5.6 Conclusion

Nous avons évalué quatre méthodes de simplification de phrases dans le but d'améliorer les performances du système d'extraction de relations. Les trois premières méthodes sont basées sur des règles qui s'appliquent au niveau de la phrase ou de l'arbre de constituants. La première et la troisième méthodes appliquent des règles que nous avons écrites pour simplifier les phrases ou les arbres de constituants. La deuxième méthode utilise un outil existant, bioSimplify, qui n'a pas été créé spécialement pour la tâche d'extraction de relations. De ce fait, les phrases simplifiées proposées ne sont pas forcément utilisables pour extraire les relations et certaines des simplifications ne sont pas pertinentes. Ces trois méthodes n'améliorent pas les performances du système d'extraction mais ne les détériorent pas non plus. Les relations correctement classées diffèrent un peu entre les méthodes avec et sans simplification. De ce fait, la combinaison des prédictions du système avec et sans simplification permet d'améliorer la classification des relations (la f-mesure augmente de 1,5%).

L'objectif de la simplification était de réduire la variabilité d'expression des relations pour améliorer les performances du système en particulier pour les relations peu représentées dans le corpus. Nous observons une augmentation de la f-mesure pour deux des classes peu représentées : TrIP et TeCP. Les relations de la classe TrWP ne sont pas détectées par REMED, et la simplification ne permet pas non plus de les repérer.

Nous avons présenté une nouvelle méthode de simplification guidée par la tâche d'extraction de relations qui permet de faire une simplification qui n'est pas facilement représentable par des règles. Cette méthode nécessite un petit corpus annoté. Les résultats que nous obtenons sur la tâche finale, à savoir l'extraction de relations, sont meilleurs que les résultats de la classification sans simplification sur le corpus de test mais légèrement moins bons sur le corpus d'évaluation.

La poursuite de l'étude pourrait porter sur les non-relations : devons-nous ajouter plus d'exemples au corpus d'apprentissage ou non, si oui, comment les annoter, etc. En effet, nous n'avons annoté que 20 exemples de non-relations car il était plus difficile de décider si une information était utile pour ne pas classer la paire d'entités comme étant en relation que de décider si elle était utile pour détecter une relation. Les informations qui permettent de savoir que deux entités ne sont pas en relation sont souvent des informations structurelles ou surfaciques, il est donc plus difficile de typer des mots dans ces cas là. Annoter des nouveaux exemples de phrases contenant des paires d'entités qui ne sont pas en relation mais aussi annoter des relations permettrait sûrement d'améliorer les résultats.

Il serait également intéressant d'effectuer un prétraitement à base de règles sur les phrases du corpus pour faire une première simplification grossière. En effet nous pourrions utiliser des règles de la première méthode (cf. section 5.2) ou de la troisième (cf. section 5.4) pour supprimer des informations inutiles facilement identifiables, puis apprendre la simplification sur ces phrases pré-simplifiées. De cette façon, le classifieur à base de CRF apprendrait peut-être mieux les simplifications que nous ne pouvons pas représenter facilement avec des règles.

Nous pourrions envisager la simplification comme le moyen d'augmenter le corpus d'apprentissage pour l'extraction des relations. Les phrases du corpus d'entraînement simplifiées par une des méthodes que nous avons présentées peuvent être ajoutées au corpus d'entraînement non simplifié, ce qui aurait pour effet d'augmenter la couverture du corpus sur les expressions des relations. Ceci nécessiterait éventuellement de sélectionner les phrases simplifiées ajoutées au corpus d'entraînement pour ne pas insérer trop de bruit.

Pour finir, nous avons étudié de façon automatique les structures des relations, c'est-à-dire la position de E1 et E2 dans la phrase par rapport aux verbes, aux syntagmes verbaux, aux propositions, etc. Nous avons observé entre autres que pour 40% des instances de la relation TeRP (*un test révèle un problème*) dans le corpus d'entraînement, E1 est sujet d'un verbe et E2 est son objet, alors que pour plus de 40% des instances de la relation TeCP (*un test est conduit en raison d'un problème*), les deux entités sont dans des propositions différentes. Il pourrait donc être intéressant de proposer un modèle de simplification pour chaque classe de relations.

# Conclusion

L'extraction de relations est une tâche importante de l'extraction d'information car elle permet de structurer les connaissances extraites via les entités nommées ou les entités du domaine. Nous avons présenté deux méthodes pour l'extraction de relations, selon que ce sont des relations n-aires ou binaires. Nous avons étudié quelles étaient les informations qui permettaient de caractériser les relations et dans quel contexte les extraire.

Dans le domaine biomédical, beaucoup des résultats des expériences effectuées sont publiés dans des articles scientifiques. Ces résultats sont utiles entre autres pour modéliser le fonctionnement de l'organisme. Nous avons proposé une méthode originale à base de règles pour l'extraction de ces résultats expérimentaux dans des articles en physiologie rénale. Nous avons proposé une formalisation des résultats par des relations n-aires qui ont pour particularités d'être non implicites dans le texte, de pouvoir avoir des arguments dans des phrases différentes, de pouvoir être dans des tableaux et d'être décrits tout le long des articles. La méthode mise en place prend en compte toutes ces particularités et repose sur le choix du résultat quantitatif comme élément pivot de la relation. Elle permet d'extraire tous les résultats expérimentaux avec une précision de 0,6, et d'extraire les trois-quarts des descripteurs du résultat avec une précision de 0,5.

La précision du système d'extraction n'étant pas parfaite, nous avons inclus notre système d'extraction dans une interface Web pour permettre aux experts de vérifier les informations extraites et de les modifier si nécessaire. Il est en effet indispensable que les informations enregistrées dans la base de données soient exactes. Cette interface est donc un assistant d'aide à l'annotation des articles scientifiques pour le peuplement d'une base de données. Une évaluation de l'assistant par des experts du domaine a été effectuée et montre que l'assistant permet d'accélérer la tâche de peuplement de la base et de ne pas manquer des résultats expérimentaux.

Nous avons également proposé une méthode d'extraction de relations binaires dans le domaine biomédical. Cette méthode par apprentissage utilise des traits de différents types pour détecter et classer les relations. Nous avons montré qu'il était important de définir des attributs qui portent de l'information lexicale, syntaxique, sémantique et surfacique. Sur la tâche de classification de relations entre des traitements, des tests et des problèmes médicaux dans des comptes rendus cliniques, notre système obtient une f-mesure de 0,70. La variation d'expression des relations nous a conduit à étudier la prise en compte de la structure syntaxique des phrases. Dans le domaine général, des travaux ont montré l'intérêt de prendre en compte des informations syntaxiques sous forme d'arbres plutôt que sous forme vectorielle. Dans le domaine bio-médical, des systèmes utilisant des informa-

tions syntaxiques pour l'extraction de relations ont été évalués, mais ne permettent pas ou peu d'amélioration de la classification des relations. Nous avons utilisé un classifieur à noyau d'arbres qui calcule la similarité entre deux arbres et donc entre deux structures syntaxiques. Nous avons montré que la classification était meilleure lorsque l'on utilisait à la fois les attributs sous forme vectorielle, les arbres de constituants complets et les sous-arbres minimaux entre les deux entités, c'est-à-dire avec une combinaison des attributs lexicaux, syntaxiques, sémantiques et surfaciques extraits du contexte local des entités, de la structure de la phrase et de la structure du segment de phrase contenu entre les deux entités. En revanche l'ajout d'attributs extraits de l'arbre de constituants ou de dépendances n'améliore pas la classification des relations dans le corpus i2b2 2010. Le problème peut venir du type des textes contenus dans le corpus : les phrases ne sont pas toujours bien construites, contiennent beaucoup d'abréviations, d'énumérations, etc. L'analyseur syntaxique utilisé a été entraîné sur un corpus d'articles scientifiques dans le domaine biomédical et n'analyse pas correctement les textes du corpus i2b2. Pour le corpus DDI et PPI, une amélioration des performances du système est observée avec la prise en compte d'informations extraites respectivement de l'arbre de dépendances et de l'arbre de constituants. Nous avons montré par nos études que les informations sur la structure de la phrase représentées par des attributs amélioraient la classification des relations dans des corpus dont les phrases étaient bien formées (le corpus DDI et PPI) mais pas dans le corpus de comptes rendus médicaux. En revanche, l'utilisation des arbres de constituants et des sous-arbres améliorent la classification des relations dans le corpus pour lequel l'analyse syntaxique est moins fiable.

Dans le même objectif de réduire la variabilité dans l'expression des relations, nous avons comparé quatre méthodes de simplification des phrases. Les trois premières utilisent des règles s'appliquant sur la phrase ou sur l'arbre de constituants et ne permettent pas d'améliorer les performances du classifieur. Nous avons montré, via l'outil bioSimplify, que les méthodes de simplification qui ne sont pas guidées par la tâche d'extraction de relations ne permettent pas d'améliorer l'extraction des relations puisqu'elles ne modifient pas nécessairement le contexte local des entités et ne conservent pas toujours les entités. La quatrième méthode est une méthode basée sur une cascade de classifieurs. La simplification est apprise à partir d'un petit corpus annoté en simplification avec des CRF, puis le système d'extraction de relations est utilisé pour évaluer la pertinence des simplifications apprises. Avec cette méthode, nous obtenons une amélioration des performances du système sur le corpus de test mais pas sur le corpus d'évaluation.

## Perspectives

Nous allons présenter ici quelques unes des perspectives que nous envisageons pour continuer ce travail sur l'extraction des relations en domaine de spécialité.

À la suite du travail sur l'extraction de relations n-aires, il serait intéressant d'étudier la question des liens implicites entre des descripteurs des résultats expérimentaux, par exemple il existe un lien implicite entre le paramètre étudié dans l'expérience et l'unité du résultat quantitatif. La base de données contient environ 8500 résultats d'expérimentations, nous pouvons envisager d'utiliser ces données pour extraire des informations implicites dans les nouveaux articles traités. Nous avons développé notre système sur un corpus

d'article en physiologie rénale avec pour référence les données enregistrées dans QKDB, mais l'intérêt d'un assistant de ce type est qu'il soit générique et puisse s'adapter à d'autres domaines. Il faudrait former un corpus d'articles scientifiques dans un domaine dans lequel les résultats expérimentaux sont utiles et évaluer les performances de notre système. Une version générique de la base de données en physiologie a été créé, elle s'appelle QxDB et pourrait être associé à une version générique de notre système. Une dernière piste de recherche envisagée est d'étudier la façon de prendre en compte les modifications effectuées par les experts dans le but d'améliorer les extractions futures de résultats expérimentaux.

Nous avons présenté plusieurs perspectives à court terme pour l'extraction de relations dans les conclusions des chapitres 4 et 5. Entre autres, nous proposons d'étudier la différence dans les attributs utilisés pour une tâche de détection de relations et pour une tâche de classification de relations, d'évaluer les erreurs provenant de la mauvaise analyse syntaxique des phrases du corpus i2b2, de proposer une méthode hybride pour la simplification de phrases ou encore de proposer des modèles de simplification différents selon les relations.

Nous envisageons plusieurs perspectives à plus long terme sur le travail d'extraction de relations et de simplification de phrases. Pour continuer le travail sur la réduction de la variabilité, il serait intéressant d'utiliser des connaissances sémantiques plus riches dans le but de réduire la variabilité lexicale. Par exemple, l'utilisation de connaissances sur les voisins sémantiques calculées sur un corpus du domaine biomédical, permettrait de rapprocher des mots apportant des informations similaires pour classer la relation. Pour réduire la variabilité d'expression des relations, nous nous sommes surtout concentrée sur la syntaxe mais il serait aussi pertinent d'étudier la variation sémantique. En utilisant les classes de VerbNet pour les verbes, nous avons commencé à travailler dans ce sens.

Si la réduction de la variabilité ne permet pas d'améliorer la détection des relations peu représentées dans le corpus, nous pouvons nous poser la question du ré-équilibrage de nos données. Il est possible de diminuer le nombre d'exemples de la classe majoritaire, c'est-à-dire la classe des non-relations, soit aléatoirement soit en sélectionnant les exemples à conserver, c'est ce qu'on appelle le ré-échantillonnage des données. Nous avons testé cette méthode sur le corpus DDI mais les performances du système n'étaient pas meilleures. Une autre méthode consiste à augmenter le nombre d'exemples des classes minoritaires. Par exemple Chawla *et al.* [2002] proposent une technique pour effectuer du sur-échantillonnage, c'est-à-dire augmenter le nombre d'exemples des classes minoritaires en générant des nouveaux exemples à partir de ceux existants.

Les travaux futures sur la simplification pourraient porter sur la simplification des arbres de dépendances. Nous avons vu que les informations provenant de l'arbre de dépendances permettent d'améliorer l'extraction des relations sur le corpus PPI. De plus des classifieurs à noyau de graphes ont été proposés dans la littérature pour détecter les relations à partir des arbres de dépendances (par exemple Airola *et al.* [2008]). Thomas *et al.* [2011] ont proposé quelques règles pour simplifier l'arbre de dépendances, mais n'observent pas d'amélioration des résultats. Nous pourrions développer une méthode similaire à celle proposée pour la simplification des arbres de constituants (cf. section 5.4), qui consisterait à normaliser certaines dépendances ou à en supprimer. Le modèle de classification des relations pourrait être construit en utilisant soit les arbres de dépendances soit les phrases résultantes des arbres de dépendances simplifiés.

# Bibliographie


Abacha, A. B. et Zweigenbaum, P. (2011). Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of Biomedical Semantics*, 2(Suppl 5):S4+. 41

Ahn, D. (2006). The stages of event extraction. *In Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, ARTE '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics. 38

Airola, A., Pyysalo, S., Bjorne, J., Pahikkala, T., Ginter, F. et Salakoski, T. (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9(Suppl 11):S2. 47, 132, 138, 170

Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R. et Wang, X. (2008). Assisted Curation: does Text Mining Really Help? *In Proceedings of the Pacific Symposium on Biocomputing*. 90

Aronson, A. R. (2001). Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *In AMIA 2001 Symposium Proceedings*. 45, 72, 120

Auger, A. et Barrière, C. (2008). Pattern-based approaches to semantic relations-State of the art. *Special Issue of Terminology Journal on Patter-based approaches to semantic relations.*, 14(1). 73

Biran, O., Brody, S. et Elhadad, N. (2011). Putting it Simply: a Context-Aware Approach to Lexical Simplification. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA. Association for Computational Linguistics. 49

Björne, J., Ginter, F., Pyysalo, S., Tsujii, J. et Salakoski, T. (2010). Complex event extraction at PubMed scale. *Bioinformatics*, 26:i382–i390. 37, 38

Boser, B. E., Guyon, I. M. et Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *In Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 144–152, New York, NY, USA. ACM. 100

Bott, S. et Saggion, H. (2012). Automatic Simplification of Spanish Text for e-Accessibility. *In Computers Helping People with Special Needs - 13th International Conference, ICCHP 2012*, pages 527–534. 49

Bui, Q.-C., Nualláin, B. Ó., Boucher, C. A. et Sloot, P. M. A. (2010). Extracting causal relations on HIV drug resistance from literature. *BMC Bioinformatics*, 11:101. 49

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K. et Wong, Y. W. (2005). Comparative Experiments on Learning Information Extractors for Proteins and their Interactions. *Artificial Intelligence in Medicine (special issue on Summarization and Information Extraction from Medical Documents)*, (2):139–155. 44, 47, 137

Bunescu, R. C. et Mooney, R. J. (2005). A shortest path dependency kernel for relation extraction. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing HLT 05*, (October):724–731. 46

Buyko, E., Faessler, E., Wermter, J. et Hahn, U. (2009). Event extraction from trimmed dependency graphs. *In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics. 38

Buyko, E., Faessler, E., Wermter, J. et Hahn, U. (2011). Syntactic Simplification and Semantic Enrichment-Trimming Dependency Graphs for Event Extraction. *Computational Intelligence*, 27:610–644. 18, 51

Chandrasekar, R., Doran, C. et Srinivas, B. (1996). Motivations and Methods for Text Simplification. *In Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING '96)*, pages 1041–1044. 49

Chang, C.-C. et Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 76, 101

Chawla, N. V., Bowyer, K. W., Hall, L. O. et Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357. 170

Chen, L. et Friedman, C. (2004). Extracting phenotypic information from the literature via natural language processing. *In Medinfo 2004: Proceedings Of THe 11th World Congress On Medical Informatics*. 41

Chen, Y. et Lin, C. (2006). Combining SVMs with various feature selection strategies. *In Feature Extraction*, pages 315–324. Springer. 117

Chinchor, N. A. (1998). Overview of MUC-7/MET-2. *In Proceedings of the 7th Message Understanding Conference*. 32

Coden, A. R., Pakhomov, S. V., Ando, R. K., Duffy, P. H. et Chute, C. G. (2005). Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics*, 38:422–430. 50, 103

Corney, D., Buxton, B. F., Langdon, W. B. et Jones, D. T. (2004). BioRAT: extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206. 59

Culotta, A. et Sorensen, J. (2004). Dependency Tree Kernels for Relation Extraction. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. 45, 46, 47

Damay, J. J. S., Lojico, G. J. D., Lu, K. A. L., Tarantan, D. B. et Ong, E. C. (2006). SIMTEXT Text Simplification of Medical Literature. *In Proceedings of the 3rd National Natural Language Processing Symposium - Building Language Tools and Ressources*. 48

de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J. D. et Zhu, X. (2011). Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *JAMIA*, 18(5):557–562. 43, 45, 120

De Marneffe, M.-c., Maccartney, B. et D. Manning, C. (2006). Generating typed dependency parses from phrase structure parses. *In In LREC 2006*. 123, 133

Demetriou, G. C. et Gaizauskas, R. J. (2002). Utilizing text mining results: The PastaWeb system. *In Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, pages 77–84. 58, 59

Ding, J., Berleant, D., Nettleton, D. et Wurtele, E. S. (2002). Mining MEDLINE: Abstracts, Sentences, or Phrases? *In Pacific Symposium on Biocomputing*, pages 326–337. 137

Dogan, R. I., Névéol, A. et Lu, Z. (2011). A context-blocks model for identifying clinical relationships in patient records. *BMC Bioinformatics*. 44, 45

Dzodic, V., Hervy, S., Fritsch, D., Khalfallah, H., Thereau, M. et Thomas, S. R. (2004). Web-based tools for quantitative renal physiology. *Cellular and Molecular Biology*, 50(7):795–800. 58

Döhling, L. et Leser, U. (2011). EquatorNLP : Pattern-based Information Extraction for Disaster Response. *In Terra Cognita 2011 Workshop, Foundations, Technologies and Applications of the Geospatial Web*. 38

Embareck, M. (2008). *Un système de question-réponse dans le domaine médical - Le système Esculape*. Thèse de doctorat, Université Paris-Est. 73, 75

Embarek, M. et Ferret, O. (2008). Learning Patterns for Building Resources about Semantic Relations in the Medical Domain. *In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. 42

Fellbaum, C., éditeur (1998). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London. 45, 78

Feng, D., Burns, G. et Hovy, E. (2007). Extracting Data Records from Unstructured Biomedical Full Text. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. 38, 39

Filippova, K. et Strube, M. (2008). Dependency tree based sentence compression. *In Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics. 48

Friedman, C., Kra, P., Yu, H., Krauthammer, M. et Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17. 41

Fundel, K., Küffner, R. et Zimmer, R. (2007). RelEx–Relation extraction using dependency parse trees. *Bioinformatics*, 23:365–371. 42, 137

Gaizauskas, R. et Wilks, Y. (1998). Information Extraction: Beyond Document Retrieval. *Journal of Documentation*, 54(1):70–105. 28

Garcia, M. et Gamallo, P. (2011). Dependency-Based Text Compression for Semantic Relation Extraction. *In Proceedings of the Workshop on Information Extraction and Knowledge Acquisition, International Conference RANLP (Recent Advances in Natural Language Processing)*. 51

Garten, Y. et Altman, R. (2009). Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC bioinformatics*, 10(Suppl 2):S6. 59

Ghersedine, A., Buche, P., Dibie-Barthélemy, J., Hernandez, N. et Kamel, M. (2012). Extraction de relations n-aires interphrastiques guidée par une RTO. *In CORIA 2012*, pages 179–190. 39

Giuliano, C., Lavelli, A. et Romano, L. (2006). Exploiting Shallow Linguistic Information for Relation Extraction From Biomedical Literature. *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*. 44, 47, 104, 105, 138

Gold, S., Elhadad, N., Zhu, X., Cimino, J. J. et George, H. (2008). Extracting Structured Medication Event Information from Discharge Summaries. *In AMIA 2008 Symposium Proceedings*, pages 237–241. 37

Grabar, N., Malaisé, V., Marcus, A. et Krul, A. (2004). Repérage de relations terminologiques transversales en corpus. *In Actes de TALN 2004*. 40

Grau, B., Ligozat, A.-L. et Minard, A.-L. (2009). Corpus study of kindey-related experimental data in scientific papers. *In Proceedings of the Biomedical Information Extraction Workshop, International Conference RANLP (Recent Advances in Natural Language Processing)*. 83

Grishman, R. (2010). Information Extraction. *In* Clark, A., Fox, C. et Lappin, S., éditeurs : *The Handbook of Computational Linguistics and Natural Language Processing*, pages 515–530. Wiley-Blackwell. 31

Grouin, C., Deléger, L. et Zweigenbaum, P. (2010). Extracting medical information from narrative patient records: the case of medication-related information. *JAMIA*, 17:555–558. 37, 109, 144

Guarino, N. (1998). *Formal Ontology in Information Systems: Proceedings of the 1sr International Conference (FOIS'98), June 6-8, Trento, Italy*, volume 46. Ios Pr Inc. 65

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *In Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545. 41, 42

Heilman, M. et Smith, N. A. (2010). Extracting Simplified Statements for Factual Question Generation. *In Proceedings of the 3rd Workshop on Question Generation*. 49

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H. et Wilks, Y. (1998). University of Sheffield: Description of the LaSIE-II system as used for MUC-7. *In Proceedings of the Seventh Message Understanding Conferences (MUC-7)*. 34

Hunter, P., Coveney, P. V., Bono, B. d., Diaz, V., Fenner, J., Frangi, A. F., Harris, P., Hose, R., Kohl, P., Lawford, P., McCormack, K., Mendes, M., Omholt, S., Quarteroni, A., Skår, J., Tegner, J., Thomas, S. R., Tollis, I., Tsamardinos, I., Beek, J. H. G. M. v. et Viceconti, M. (2010). A vision and strategy for the virtual physiological human in 2010 and beyond. *Philosophical Transactions of the Royal Society A*, 368:2595–2614. 58

Jacquemin, C. (1996). A symbolic and surgical acquisition of terms through variation. *In Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438, London, UK, UK. Springer-Verlag. 71, 78

Jelier, R., Jenster, G., Dorssers, L. C. J., van der Eijk, C. C., van Mulligen, E. M., Mons, B. et Kors, J. A. (2005). Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics*, 21(9):2049–2058. 40

Jonnalagadda, S. et Gonzalez, G. (2010). Sentence Simplification Aids Protein-Protein Interaction Extraction. *CoRR*, abs/1001.4273. 49, 50

Jonnalagadda, S. et Gonzalez, G. (2011). BioSimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. *CoRR*, abs/1107.5744. 146

Kado, T. (2003). CRF++: Yet Another CRF toolkit. `http://crfpp.sourceforge.net/`. 159

Katrenko, S. et Adiaans, P. (2007). Learning Relations from Biomedical Corpora Using Dependency Trees. *In Knowledge Discovery and Emergent Complexity in Bioinformatics*, pages 61–80. 44

Kipper, K., Korhonen, A., Ryant, N. et Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40. 77

Knight, K. et Marcu, D. (2000). Statistics-based summarization-step one: Sentence compression. *In Proceedings of the National Conference on Artificial Intelligence*, pages 703–710. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. 49

Lin, D. et Pantel, P. (2001). DIRT - Discovery of Inference Rules from Text. *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–328. 73

Lindberg, D., Humphreys, B. et McCray, A. (1993). The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291. 44, 72

McClosky, D. (2010). Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing. *PHD Thesis, Department of Computer Science, Brown University.* 123

McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y. et White, P. (2005). Simple algorithms for complex relation extraction with applications to biomedical IE. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 491–498, Stroudsburg, PA, USA. Association for Computational Linguistics. 35, 38

Minard, A.-L., Ligozat, A.-L., Abacha, A. B., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., Rosset, S., Zweigenbaum, P. et Grouin, C. (2011a). Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. *JAMIA*, 18(5):588–593. 102

Minard, A.-L., Ligozat, A.-L. et Grau, B. (2010). Extraction de résultats expérimentaux d'articles scientifiques pour le peuplement d'une base de données. *In 10th International Conference on statistical analysis of textual data (JADT).* 73

Minard, A.-L., Ligozat, A.-L. et Grau, B. (2011b). Apport de la syntaxe pour l'extraction de relations en domaine médical. *In Actes TALN 2011*, pages 383–393. 122

Minard, A.-L., Ligozat, A.-L. et Grau, B. (2011c). Extraction de relations dans des comptes rendus hospitaliers. *In Actes des 22èmes Journées francophones d'Ingénierie des Connaissances (IC'2011).* 107

Minard, A.-L., Ligozat, A.-L. et Grau, B. (2011d). Multi-class SVM for Relation Extraction from Clinical Reports. *In Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria*, pages 604–609. 107

Minard, A.-L., Ligozat, A.-L. et Grau, B. (2012). Simplification de phrases pour l'extraction de relations. *In Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 1–14, Grenoble, France. ATALA/AFCP. 156

Minard, A.-L., Ligozat, A.-L., Grau, B. et Makour, L. (2011e). Feature selection for Drug-Drug Interaction detection using machine-learning based approaches. *In SEPLN'11, Workshop Drug-Drug Interaction.* 135

Miwa, M., Sætre, R., Miyao, Y. et Tsujii, J. (2010). Entity-focused sentence simplification for relation extraction. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 788–796, Stroudsburg, PA, USA. Association for Computational Linguistics. 18, 51, 52, 150, 154

Moschitti, A. (2006). Making tree kernels practical for natural language learning. *In Proceedings of the Eleventh International Conference on European Association for Computational Linguistics (EACL), Trento, Italy, 2006.* 19, 100, 101

Nguyen, V. T., Gaio, M. et Sallaberry, C. (2010). Recherche de relations spatio-temporelles : une méthode basée sur l'analyse de corpus textuels. *CoRR*, abs/1002.0577. 37

Noy, N. et Rector, A. (2006). *Defining N-ary Relations on the Semantic Web.* 18, 33, 34, 35, 65

Nédellec, C. (2005). Learning Language in Logic - Genic Interaction Extraction Challenge. *In Proceedings of the Learning Language in Logic 2005 Workshop at the International Conference on Machine Learning.* 30, 36, 42, 44, 136, 137

Palaga, P. (2009). Extracting Relations from Biomedical Texts Using Syntactic Information. Mémoire de D.E.A., Technische Universität Berlin. 138

Pantel, P., Ravichandran, D. et Hovy, E. (2004). Towards terascale knowledge acquisition. *In International Conference on Computational Linguistics (COLING'04)*, pages 771–777, Geneva, Switzerland. 74

Patrick, J. et Li, M. (2010). High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *JAMIA*, 17:524–527. 37

Poibeau, T. (2003). *Extraction automatique d'information. Du texte brut au web sémantique.* Hermès, Paris. 28

Pyysalo, S., Ginter, F., Heimonen, J., Bjorne, J., Boberg, J., Jarvinen, J. et Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1):50+. 137

Qian, L., Zhou, G., Kong, F., Zhu, Q. et Qian, P. (2008). Tree Kernel-Based Semantic Relation Extraction using Unified Dynamic Relation Tree. *ALPIT '08: Proceedings of the 2008 International Conference on Advanced Language Processing and Web Information Technology*, pages 64–69. 51, 52

Ribba, B., Tracqui, P., Boix, J.-L., Boissel, J.-P. et Thomas, S. R. (2006). QxDB: a generic database to support mathematical modelling in biology. *Philos Transact A Math Phys Eng Sci*, 364(1843):1517–32. 58

Rindflesch, T. C., Bean, C. A. et Sneiderman, C. A. (2000). Argument Identification for Arterial Branching Predications Asserted in Cardiac Catheterization Reports. *In AMIA Annu Symp Proc*, pages 704–708. 41

Rink, B., Harabagiu, S. et Roberts, K. (2011). Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):594–600. 119

Roberts, A., Gaizauskas, R. et Hepple, M. (2008). Extracting Clinical Relationships from Patient Narratives. *In BioNLP2008: Current Trends in Biomedical Natural Language Processing*, pages 10–18. 43, 99, 110, 114

Rosario, B. et Hearst, M. A. (2004). Classifying semantic relations in bioscience texts. *In Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics*, pages 431–438. 43, 45

Sahay, S., Lee, J. et Krishnamurthi, N. (2008). Relationship Extraction from Biomedical Documents using Conditional Random Fields. 43

Sarawagi, S. (2008). Information Extraction. *Foundations and Trends in Databases*, 1(3):261–377. 36

Savova, G. K., Kipper-Schuler, K., Buntrock, J. D. et Chute, C. G. (2008). UIMA-based Clinical Information Extraction System. *In Proceedings of The Sixth International Conference on Language Resources and Evaluation: Workshop on Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP*. 45, 120

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49. 71, 107, 122

Segura-Bedmar, I., Martinez, P. et de Pablo-Sanchez, C. (2011a). A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. *BMC Bioinformatics*, 12(Suppl 2):S1. 50, 52

Segura-Bedmar, I., Martínez, P. et Sánchez-Cisneros, D. (2011b). The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from Biomedical texts. *In Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*. 36, 135, 136

Sharma, A., Swaminathan, R. et Yang, H. (2010). A Verb-Centric Approach for Relationship Extraction in Biomedical Text. *International Conference on Semantic Computing*, 0:377–385. 42

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. 42

Srinivasan, P. et Rindflesch, T. (2002). Exploring text mining from MEDLINE. *In Proc AMIA Symp*, pages 722–726. 41

Swampillai, K. et Stevenson, M. (2010). Inter-sentential Relations in Information Extraction Corpora. *In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. 59

Swampillai, K. et Stevenson, M. (2011). Extracting Relations Within and Across Sentences. *In Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria*, pages 25–32. 59

Thomas, P., Pietschmann, S., Solt, I., Tikk, D. et Leser, U. (2011). Not all links are equal: Exploiting Dependency Types for the Extraction of Protein-Protein Interactions from Text. *In Proceedings of BioNLP 2011 Workshop*, pages 1–9, Portland, Oregen, USA. Association for Computational Linguistics. 18, 51, 52, 170

Tikk, D., Thomas, P., Palaga, P., Hakenberg, J. et Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol*, 6:e1000837. 138, 139, 140

Touhami, R., Buche, P., Dibie-Barthélemy, J. et Ibanescu, L. (2011). An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables. *In OTM Conferences (2)*, pages 662–679. 39, 65, 80

Uzuner, O., Mailoa, J., Ryan, R. et Sibanda, T. (2010a). Semantic relations for problem-oriented medical records. *Artificial Intelligence in Medicine*, 50:63–73. 43, 46, 99, 133

Uzuner, Ö., Solti, I. et Cadag, E. (2010b). Extracting medication information from clinical text. *JAMIA*, 17(5):514–518. 30, 35, 37, 81

Uzuner, Ö., South, B. R., Shen, S. et DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *JAMIA*, 18(5):552–556. 30, 36, 102, 105

Van Landeghem, S., Saeys, Y., De Baets, B. et Van de Peer, Y. (2008). Extracting protein-protein interactions from text using rich feature vectors and feature selection. *In* Salakoski, T., Rebholz-Schuhmann, D. et Pyysalo, S., éditeurs : *SMBM '08 : proceedings of the third symposium on semantic mining in biomedicine*, pages 77–84. Turku Centre for Computer Sciences (TUCS). 46, 134

Vickrey, D. et Koller, D. (2008). Sentence Simplification for Semantic Role Labeling. *In Proceedings of ACL-08: HLT*, pages 344–352, Columbus, Ohio. Association for Computational Linguistics. 49

Waszak, T. et Torres-Moreno, J.-M. (2008). Compression entropique de phrases contrôlée par un perceptron. *Journées internationales d'Analyse statistique des Données Textuelles*. 49

Yousfi-Monod, M. et Prince, V. (2006). Compression de phrases par élagage de leur arbre morpho-syntaxique. Une première application sur les phrases narratives. *TSI : Revue Technique et Science Informatiques*, 25(4):437–468. 49

Zelenko, D., Aone, C. et Richardella, A. (2003). Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106. 46

Zhang, M., Zhang, J. et Su, J. (2006). Exploring Syntactic Features for Relation Extraction using a Convolution Tree Kernel. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 288–295. 46, 47

Zhou, G., Su, J., Zhang, J. et Zhang, M. (2005). Exploring Various Knowledge in Relation Extraction. *In Proceedings of the 43rd Annual Meeting of the ACL*, pages 427–434. 43, 44, 45, 47, 99, 110, 111

Zhu, Z., Bernhard, D. et Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1353–1361, Stroudsburg, PA, USA. Association for Computational Linguistics. 49

# Index