Simplification de phrases pour l'extraction de relations

Anne-Lyse Minard[1, 2] Anne-Laure Ligozat[1, 3] Brigitte Grau[1, 3]
(1) LIMSI-CNRS, BP 133, 91403 Orsay Cedex
(2) Université Paris Sud, 91400 Orsay
(3) ENSIIE, square de la résistance, 91000 Évry
prenom.nom@limsi.fr

RÉSUMÉ
L'extraction de relations par apprentissage nécessite un corpus annoté de très grande taille pour couvrir toutes les variations d'expressions des relations. Pour contrer ce problème, nous proposons une méthode de simplification de phrases qui permet de réduire la variabilité syntaxique des relations. Elle nécessite l'annotation d'un petit corpus qui sera par la suite augmenté automatiquement. La première étape est l'annotation des simplifications grâce à un classifieur à base de CRF, puis l'extraction des relations, et ensuite une complétion automatique du corpus d'entraînement des simplifications grâce aux résultats de l'extraction des relations. Les premiers résultats que nous avons obtenus pour la tâche d'extraction de relations d'i2b2 2010 sont très encourageants.

ABSTRACT
Sentence simplification for relation extraction

Machine learning based relation extraction requires large annotated corpora to take into account the variability in the expression of relations. To deal with this problem, we propose a method for simplifying sentences, i.e. for reducing the syntactic variability of the relations. Simplification requires the annotation of a small corpus, which will be automatically augmented. The process starts with the annotation of the simplification thanks to a CRF classifier, then the relation extraction, and lastly the automatic completion of the training corpus for the simplification through the results of the relation extraction. The first results we obtained for the task of relation extraction of the i2b2 2010 challenge are encouraging.

MOTS-CLÉS : Extraction de relations, simplification de phrases, apprentissage automatique.

KEYWORDS: Relation extraction, sentence simplification, machine learning.

1 Introduction

Dans le domaine médical, de nombreux documents électroniques sont produits chaque jour, mais ces documents sont sous forme textuelle, et les informations qu'ils contiennent sont donc difficilement exploitables. L'extraction d'information consiste à structurer cette information. Pour une tâche donnée, les documents disponibles ne contiennent cependant pas nécessairement de nombreux exemples d'apprentissage et les corpus peuvent présenter une grande variabilité. Par conséquent, il est nécessaire de pouvoir apprendre à partir de peu d'exemples, éventuellement

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 1–14,
Grenoble, 4 au 8 juin 2012.

très disparates. Dans cet article, nous nous intéressons à une tâche d'extraction de relations médicales et proposons une méthode qui consiste à effectuer une simplification syntaxique préalable des phrases. Cette simplification a pour but de normaliser le corpus en ne gardant que les informations qui sont pertinentes pour l'extraction. Elle est donc guidée par la tâche, et peu d'exemples sont nécessaires pour apprendre la simplification puisque l'extraction de relation est utilisée pour augmenter le corpus annoté.

Après un état de l'art sur le domaine de l'extraction de relations et sur la simplification (section 2), nous présenterons la tâche d'extraction de relations en domaine médical et son application dans le cadre du challenge i2b2 2010 [1], ainsi que le système que nous avons développé (section 3). Ensuite, nous présenterons notre méthode pour l'annotation des simplifications (section 4). Nous détaillerons la méthode originale proposée pour améliorer la simplification grâce à la combinaison du système d'extraction de relations et du classifieur pour l'annotation des simplifications (sections 5 et 6), et terminerons par la présentation des expérimentations que nous avons menées et des résultats obtenus (section 7) [2].

2 État de l'art

De nombreuses méthodes ont été proposées pour l'extraction de relations, les plus courantes étant fondées sur une classification automatique plus ou moins supervisée. Les attributs utilisés pour la classification représentent en général de l'information lexicale, sémantique ou syntaxique. Par exemple (Roberts *et al.*, 2008) proposent une approche fondée sur des SVM pour extraire des relations dans des dossiers de patients atteints d'un cancer. Ils utilisent des attributs lexicaux, sémantiques et morpho-syntaxiques. (Uzuner *et al.*, 2010) utilisent des attributs syntaxiques plus riches puisqu'ils ajoutent les dépendances syntaxiques entre les concepts. Ils les utilisent dans une approche vectorielle fondée sur des SVM pour extraire des relations entre des problèmes, des tests et des traitements dans des comptes-rendus médicaux. Ces informations syntaxiques n'améliorent pas la détection des relations car dans beaucoup de cas il n'existe pas de dépendance entre les deux concepts. (Zhang *et al.*, 2006) incluent également de l'information syntaxique riche dans leur système d'extraction de relations. Pour cela, ils utilisent des arbres syntaxiques avec des tree kernels. Ils ont testé leur système sur le corpus ACE 2003, et ils montrent que les meilleurs résultats sont obtenus en utilisant le plus petit sous-arbre commun aux deux entités. Nous montrons dans (Minard *et al.*, 2011a) que pour l'extraction de relations en domaine médical (sur le corpus i2b2 2010) l'utilisation de l'arbre minimal commun aux deux entités n'est pas suffisant et qu'il est souvent nécessaire d'utiliser l'arbre complet ou tout du moins des éléments de cet arbre.

Pour améliorer l'extraction des relations, nous proposons une méthode de simplification des phrases. Simplifier les phrases consiste alors à supprimer ou à repérer les mots de la phrase qui peuvent gêner le classifieur. Dans notre cas, la simplification ne consiste pas à rendre un texte plus facile à lire, mais à ne garder que les mots permettant de classer une relation.

La simplification de textes a donné lieu à de nombreux travaux, soit en tant que tâche à part entière comme par exemple dans (Woodsend et Lapata, 2011), soit en tant que prétraitement pour d'autres tâches, comme par exemple la génération de questions (Heilman et Smith, 2010). Cette

1. https ://www.i2b2.org/NLP/Relations/
2. Ce travail a été partiellement financé par OSEO dans le cadre du programme Quæro.

simplification est généralement fondée sur des règles syntaxiques. Dans le domaine biomédical, différentes recherches sur la simplification syntaxique pour améliorer l'extraction de relations ont été menées dans le domaine des interactions entre protéines (PPI). (Jonnalagadda et Gonzalez, 2010) ont développé un outil (bioSimplify) qui produit des phrases simples à partir d'une phrase complexe. Leur objectif est d'augmenter le rappel de l'extraction d'information dans le domaine biomédical. Pour cela, 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. Aucune sélection de la (des) meilleure(s) phrase(s) simple(s) n'est effectuée, et les règles n'obligent pas la conservation de la paire d'entités candidate. L'évaluation de leur outil pour l'extraction des interactions entre protéines n'est pas assez précise pour en tirer des conclusions. (Miwa *et al.*, 2010) ont également utilisé des règles pour simplifier les phrases. 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, car leur rôle est de supprimer l'information inutile pour l'extraction des interactions. Ils ont évalué l'impact de la simplification pour l'extraction des interactions entre protéines et montrent que sur 5 corpus différents l'extraction des relations est meilleure. Deux autres travaux portent sur la simplification des arbres de dépendances pour la tâche d'extraction d'interactions entre protéines (Thomas *et al.*, 2011), par suppression ou modification de types de dépendances, et pour la tâche BioNLP'09 [3] (extraction d'événements biologiques) (Buyko *et al.*, 2011), par élagage de l'arbre.

Dans un autre domaine, l'annotation des rôles sémantiques, (Vickrey et Koller, 2008) 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.

La méthode de simplification que nous proposons dans cet article, est fondée sur un apprentissage automatique, contrairement aux travaux que nous venons de présenter. Le mode d'apprentissage semi-supervisé se rapproche de celui développé par (Vickrey et Koller, 2008) pour la tâche d'annotation des rôles sémantiques. En effet, 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. Cette méthode se rapproche ainsi des travaux en compression de phrases, qui consiste à supprimer certains constituants d'une phrase, considérés comme non essentiels. Les approches en compression de phrases peuvent se fonder sur des règles linguistiques (Yousfi-Monod et Prince, 2006) ou sur un apprentissage(Knight et Marcu, 2000; Waszak et Torres-Moreno, 2008). Cependant, notre tâche s'en distingue par deux aspects : notre objectif n'est pas de simplifier les phrases en fonction des informations saillantes, mais en fonction des informations relatives à l'extraction d'une relation, et par ailleurs, nous souhaitons développer un système qui ne nécessite pas l'annotation d'un grand corpus pour la simplification.

3. http ://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/

3 Extraction de relations en domaine médical

Nous avons développé un système pour l'extraction de relations dans le domaine biomédical. Il utilise un classifieur à base de SVM, avec la bibliothèque LIBSVM (Chang et Lin, 2001). Le système peut être utilisé pour une classification binaire ou multi-classes (avec une approche "un-contre-un"). Il utilise des attributs qui prennent en compte des informations de surface, sur les distances entre mots par exemple, des informations lexicales, comme les mots formant les concepts, des informations syntaxiques, les catégories morpho-syntaxiques des mots, et des informations sémantiques grâce au typage des concepts. Une description détaillée du système est donnée dans (Minard *et al.*, 2011b).

Ce système a été utilisé pour le challenge i2b2 2010, pour le challenge DDI 2011 (extraction d'interactions entre médicaments (Minard *et al.*, 2011c)) et également pour l'extraction d'interactions entre protéines. Dans cet article, les tests sont effectués sur le corpus i2b2 2010 que nous présentons dans la section suivante.

3.1 Corpus de comptes-rendus médicaux

Dans le cadre du challenge i2b2 2010, un corpus annoté composé de rapports cliniques a été fourni aux participants. Les comptes rendus du corpus proviennent de 7 centres médicaux des États-Unis. Ils ont été manuellement anonymisés et annotés. Trois types de concepts ont été annotés : les problèmes médicaux (maladies, syndromes, observations sur l'état psychologique du patient, etc.), les traitements (interventions, médicaments donnés au patient, etc.) et les tests (procédures et examens). Entre ces trois types de concepts, 8 relations peuvent exister :
– un traitement améliore (TrIP), aggrave (TrWP) ou cause (TrCP) un problème médical ;
– un traitement est administré (TrAP) ou pas (TrNAP) pour un problème médical ;
– un test révèle [4] (TeRP) ou est conduit pour examiner (TeCP) un problème médical ;
– un problème médical indique un autre problème médical (PIP).
Le corpus de développement (DEV_I2B2) est composé de 349 documents (4994 relations) et le corpus d'évaluation (EVAL_I2B2) de 477 documents (9070 relations). Nous avons divisé le corpus DEV_I2B2 en deux parties afin de pouvoir entraîner notre système avant d'avoir le corpus d'évaluation : un corpus d'entraînement (TRAIN_I2B2) composé de 295 documents (4515 relations) et un corpus de test (TEST_I2B2) composé de 54 documents (479 relations). Dans le graphique 1, nous avons représenté le nombre d'instances de relations de chaque type dans les différents corpus.

3.2 Résultats obtenus

Les résultats que nous avons obtenus avec ce système sont présentés dans la figure 2. Nous avons également représenté sur le graphique l'accord inter annotateur (IAA) et le nombre d'exemples de chaque relation (nombre normalisé pour être à l'échelle du graphique). Globalement la F-mesure est d'environ 0,7 pour la classification des relations, mais cette classification est moins bonne

4. Cette relation correspond aux cas où le test indique la présence d'un problème, ou bien l'absence d'un problème. Pour ces relations, la présence de négations, très fréquentes en domaine médical, ne sera donc pas prise en compte.

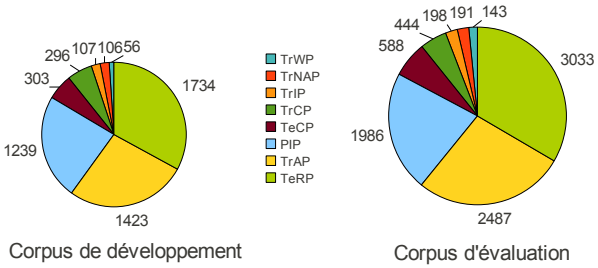

FIGURE 1 – Composition des corpus

pour les relations pour lesquelles peu d'exemples ont été annotés dans le corpus DEV_I2B2 (par exemple pour la relation TrWP ou TrIP).

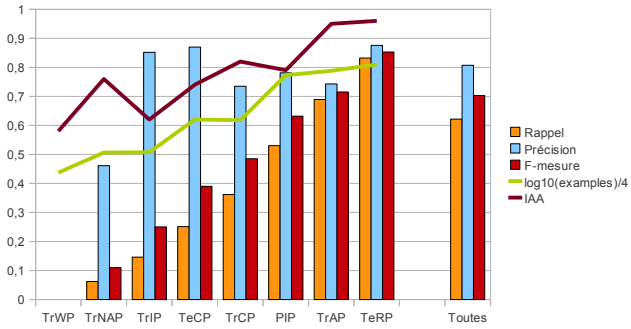

FIGURE 2 – Évaluation du système final

Les types d'erreurs que l'on peut relever proviennent du fait qu'il y a peu d'exemples pour certaines relations, associé à une grande variabilité des expressions, qu'il y a besoin de connaissances externes, et que parfois la classification est discutable car les définitions de certaines relations sont assez proches (Minard *et al.*, 2011b). Le présent travail concerne la réduction de la variabilité syntaxique des expressions par simplification. Ainsi, la relation TeCP entre *pulmonary nodules* et *fu imaging* dans la phrase de l'exemple 1 est mal classée par le système notamment à cause de la présence du verbe *reveal* qui indique généralement une relation TeRP. Dans ce type de construction, il serait intéressant de supprimer la partie de la phrase concernant les premiers tests et problèmes, pour ne conserver que la paire à étudier.

(1) [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.

Dans l'exemple 2, les deux concepts à annoter *his high right-sided filling pressure* et *his Captopril* sont séparés par la proposition *he was started on Lasix for diuresis* qui n'est pas pertinente. Si les éléments non pertinents pour cette relation étaient supprimés, la phrase pourrait devenir *Given* PROBLEM, TREATMENT *was increased.*, forme qui est fréquente dans le corpus et permet de reconnaître une relation TrAP. Une telle simplification de phrases permettrait bien de réduire la variabilité du corpus pour améliorer la classification des relations.

(2) Given [PB **his high right-sided filling pressure**], he was started on [TREAT Lasix] for [TREAT diuresis] and [TREAT **his Captopril**] was increased for [TREAT greater afterload reduction].

4 Définition du modèle de simplification

4.1 Simplification

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

Pour cela, la plupart des travaux ont défini des règles de simplification. 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 nécessiterait d'en redéfinir un grand nombre, ce qui nous a poussé à privilégier une méthode à base d'apprentissage pour annoter dans les phrases les parties à garder et celles que l'on peut supprimer.

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. Ensuite 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 3 et 4, 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 3, il s'agit de déterminer la relation TeCP entre *a magnetic resonance imaging study* et *a small vascular malformation*. Dans l'exemple 4, il s'agit d'une relation TrAP entre *the tremendous tumor burden* et *open debulking*.

(3) [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 [PBthe hemorrhage].

(4) 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.

4.2 Méthode

Nous avons choisi d'utiliser un classifieur à base de CRF («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 phrases annotées seront données en entrée du classifieur SVM. Afin de n'annoter 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. La combinaison des classifieurs est présentée dans la section 6. 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 Annotation par CRF

5.1 Constitution du corpus d'apprentissage

Nous avons sélectionné 71 phrases provenant du corpus TRAIN_I2B2. Nous avons extrait aléatoirement 14 phrases contenant des paires d'entités qui avaient été correctement classées par notre système d'extraction de relations, 37 paires mal classées et 20 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 concepts en relation sont respectivement sujet et complément de ce verbe (exemple 5).

(5) [TEST **An magnetic resonance imaging study**] **showed** [PB **basilar artery disease**], questionable aneurysm.

Dans 14 phrases, les deux concepts en relation sont dans deux propositions différentes (exemple 6). Sept constructions différentes ont été trouvées ; nous en présentons trois dans le tableau 1.

(6) 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 concepts sont reliés par une préposition, et la relation s'exprime au travers de la préposition et du verbe de la proposition (exemple 7).

(7) [...], she had [PB **an acute drop**] **in** [TEST **her systolic blood pressure**] to <NUM> for unclear reasons and without evidence of acute_sepsis.

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.

TABLE 1 – Phrases dans lesquelles les concepts en relation sont dans deux propositions différentes

Dans les exemples de non-relation que nous avons dans notre corpus, dans seulement deux phrases les deux concepts sont sujet et objet du même verbe. Dans 8 phrases, les deux concepts sont dans des propositions différentes et dans 9 phrases ils sont reliés par une préposition.

Cette étude fait apparaître l'existence de régularités, que la simplification pourrait dégager.

Les 71 phrases ont été annotées par 3 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 2, 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

TABLE 2 – Étude du corpus annoté

5.2 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, la catégorie morpho-syntaxique, le nombre de caractères du token, la position du token dans la phrase (position du token/nombre de tokens dans la phrase), le type sémantique si le token fait partie d'une entité et une étiquette indiquant si le token fait partie d'une des entités de la paire étudiée. Ce dernier attribut permet d'avoir une

5. http ://knowtator.sourceforge.net/

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 token courant.

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. Pour chaque lemme, nous avons compté combien de fois il apparaissait dans le corpus TRAIN_I2B2 et combien de fois il était associé à une des trois catégories. Le tableau 3 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

TABLE 3 – Exemple d'annotation de lemmes présents plus de 10 fois dans le corpus

6 Combinaison de classifieurs pour l'extraction des relations

Avec seulement 71 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 3 présente de façon simplifiée la méthode développée.

Annotation de la simplification Dans un premier temps, les 71 phrases annotées manuellement sont utilisées comme amorce pour la simplification ; elles forment le corpus d'entraînement TRAIN_SIMP. Elles sont utilisées pour apprendre les simplifications grâce à l'outil CRF++. Ensuite le modèle pour la simplification est appliqué sur la totalité du corpus DEV_I2B2.

Extraction des relations Nous utilisons ensuite ce corpus annoté pour extraire les relations grâce à notre classifieur à base de SVM. Les annotations des simplifications sont utilisées comme

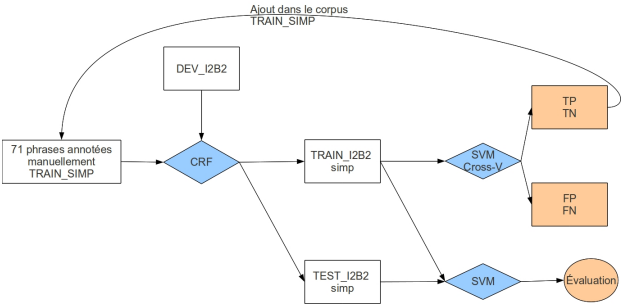

FIGURE 3 – Schéma explicatif de la méthode

des attributs supplémentaires pour la reconnaissance des relations. Un attribut identifie les tokens annotés comme «inutile», un autre pour les tokens «gênants» et un pour les «utiles».

Le corpus DEV_I2B2 a été divisé en deux : corpus TRAIN_I2B2 et corpus TEST_I2B2. Dans un premier temps, nous effectuons l'extraction des relations par validation croisée en 5 parties avec le corpus TRAIN_I2B2. 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_I2B2, le corpus DEV_I2B2 est réannoté 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_I2B2 et il est appliqué sur le corpus TEST_I2B2. Cette étape permet d'évaluer l'impact de la complétion du corpus TRAIN_SIMP sur l'extraction des relations.

7 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 simplication.

7.1 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. Dans un premier temps, nous pouvons faire varier les attributs utilisés par les CRF pour apprendre la simplification : nous pouvons par exemple ajouter la structure syntaxique de la phrase. Nous pouvons également n'utiliser que deux classes pour la simplification, c'est-à-dire ne pas faire de distinction entre les mots inutiles et gênants, et utiles et indispensables. Deuxièmement, les informations sur la simplification peuvent être prises en compte de 2 manières par le système d'extraction de relation : les mots gênants (voire gênants et inutiles) peuvent être supprimés de la phrase ou des attributs indiquant la classe du mot peuvent être ajoutés. Finalement, nous pouvons faire varier la sélection des paires d'entités correctement classées à ajouter au corpus TRAIN_SIMP. Toutes les paires correctement classées (que les entités soient en relation ou non) peuvent être ajoutées, ou seules les paires qui n'étaient pas bien classées avec le système sans simplification et qui le sont avec la simplification, ou selon le score de décision donné par le classifieur, etc.

Nous présentons ici la configuration donnant les meilleurs résultats. Nous avons appris la simplification en ne donnant que les attributs de base (voir 5.2) et en apprenant 3 classes («utile» et «indispensable» sont regroupées en une classe, «inutile» et «gênant»). Pour prendre en compte la simplification, nous avons donné des attributs supplémentaires au classifieur. Comme il est difficile d'annoter des phrases pour des paires d'entités qui ne sont pas en relation, nous avons modifié le corpus annoté manuellement et nous avons annoté en «inutile» tous les mots de la phrase. 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 dont au moins un des mots avait été annoté «utile», 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 4 fois le système complet, après quoi nous avons obtenu 589 phrases dans le corpus TRAIN_SIMP, dont 71 qui ont été annotées manuellement. Nous avons appliqué la simplification sur le corpus d'évaluation EVAL_I2B2 afin d'évaluer la classification des relations. Dans le tableau 4, nous donnons les F-mesures obtenues sans simplification, avec la simplification apprise avec les 71 phrases annotées (avant la combinaison des deux méthodes) et avec le corpus TRAIN_SIMP obtenu après 4 itérations.

La F-mesure calculée pour toutes les relations reste stable avec ou sans la simplification même si la F-mesure pour 4 des relations diminue quand nous utilisons la simplification. La différence entre les résultats avec et sans simplification calculée avec le test T de Student est significative ($p<=0,05$), la simplification a donc un effet sur la classification mais ne permet pas encore de l'améliorer.

7.2 Évaluation manuelle de la simplification

Nous n'avons pas de corpus annoté suffisament grand pour pouvoir faire une évaluation automatique de la tâche de simplification. De ce fait, nous avons choisi d'annoter manuellement 41 relations et d'évaluer manuellement la simplification pour ces relations. Nous n'avons étudié que l'annotation des phrases portant sur une paire de concepts en relation. En effet, ainsi que nous

Relations	Sans Simplification	Simplification	
		Corpus TRAIN_SIMP non augmenté	Corpus TRAIN_SIMP augmenté
TrIP	**0,315**	0,266	0,302
TrWP	0,000	0,000	0,000
TrCP	**0,486**	0,464	0,470
TrAP	**0,732**	0,724	0,730
TrNAP	**0,195**	0,151	0,168
PIP	0,625	0,627	**0,630**
TeRP	0,852	0,852	**0,855**
TeCP	**0,452**	0,398	0,408
Toutes les relations	**0,709**	0,704	0,708

TABLE 4 – Évaluation de la classification des relations avec et sans simplification sur le corpus EVAL_I2B2

l'avons déjà mentionné, il est difficile de définir ce qui doit être annoté pour les non relations.

Nous avons donc annoté 41 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 19 cas exacts, 16 cas faux et 6 cas partiellement corrects.

Dans l'exemple 8, nous considérons que la simplification est correcte mais dans l'exemple 9 le verbe le plus utile à la détection de la relation (*revealed*) est annoté inutile, et l'annotation de la simplification est donc fausse.

(8) 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] .

(9) 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] .

7.3 Analyse des simplifications

Nous avons tenté d'établir les types de simplifications apprises. Les différentes structures de phrases qui apparaissent sont :
– *concept1 relation concept2* pour lesquelles la partie située entre les concepts doit être conservée, tout ou en partie ; cette structure est généralement bien traitée ;
– *concept1 relation (coordination de concepts) concept2* est généralement mal annotée, et la marque de la relation est souvent supprimée. Ce type de structure peut être reconnu simplement par des règles ;
– *concept1 (structure comportant des concepts) relation concept2* est généralement reconnue et la relation est gardée.

Certains cas nécessitent de garder la partie gauche du premier concept ; cette configuration est mal reconnue. Il en est de même pour les contextes droits du deuxième concept. Ces deux types de structure sont plus rares, et leur traitement nécessite plus d'exemples.

Cette étude nous amène à imaginer des améliorations de notre système. Il serait par exemple intéressant de pouvoir mieux sélectionner les phrases ajoutées au corpus TRAIN_SIMP pour diversifier les exemples de simplification. Une solution serait d'identifier les paires d'entités moins bien classées avec l'utilisation de la simplification que sans, et de les annoter pour apprendre de nouveaux schémas de simplification. Nous pourions également envisager d'utiliser quelques règles pour annoter les cas les plus courants (par exemple pour supprimer les concepts en coordination ou encore les indications de lieux), puis d'utiliser le système à base d'apprentissage.

8 Conclusion

Dans cet article, nous nous sommes intéressées à la simplification de phrases dans le but d'améliorer l'extraction de relations. Nous avons présenté une méthode de simplification guidée par la tâche d'extraction de relations, et nécessitant un petit corpus annoté. Les résultats que nous obtenons sur la tâche finale, à savoir l'extraction de relations, sont significativement différents des résultats de la classification sans simplification, mais la F-mesure finale reste stable. La poursuite de l'étude pourrait porter sur l'amélioration de la sélection des phrases ajoutées au corpus d'apprentissage pour la simplification, par exemple en ne gardant que celles dont le score de confiance du classifieur est élevé. Nous devons également étudier la façon dont nous traitons les cas de non-relation ; devons-nous ajouter des exemples au corpus d'apprentissage ou non, si oui, comment les annoter, etc. Pour finir, un prétraitement à base de règles sur les phrases du corpus pourrait permettre d'annoter les indications temporelles, de lieux (par exemple le nom d'une clinique), l'âge des patients, etc. et ainsi réduire d'avantage la variabilité.

Références

Buyko, E., Faessler, E., Wermter, J. et Hahn, U. (2011). Syntactic simplification and semantic enrichment—trimming dependency graphs for event extraction. 27:610–644.

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.

HEILMAN, M. et SMITH, N. A. (2010). Extracting Simplified Statements for Factual Question Generation. *In Proceedings of the 3rd Workshop on Question Generation*.

JONNALAGADDA, S. et GONZALEZ, G. (2010). Sentence simplification aids protein-protein interaction extraction. *CoRR*, abs/1001.4273.

KADO, T. (2003). CRF++ : Yet another crf toolkit. `http://crfpp.sourceforge.net/`. [consulté le 17/01/2012].

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.

MINARD, A.-L., LIGOZAT, A.-L. et GRAU, B. (2011a). Apport de la syntaxe pour l'extraction de relations en domaine médical. *In Actes TALN 2011*, pages 383–393.

MINARD, A.-L., LIGOZAT, A.-L. et GRAU, B. (2011b). Extraction de relations dans des comptes rendus hospitaliers. *In Actes des 22èmes Journées francophones d'Ingénierie des Connaissances (IC'2011)*.

MINARD, A.-L., LIGOZAT, A.-L., GRAU, B. et MAKOUR, L. (2011c). Feature selection for drug-drug interaction detection using machine-learning based approaches. *In SEPLN'11, Workshop Drug-Drug Interaction*.

MIWA, M., S, 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.

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.

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.

UZUNER, O., MAILOA, J., RYAN, R. et SIBANDA, T. (2010). Semantic relations for problem-oriented medical records. *Artificial Intelligence in Medicine*, 50:63–73.

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.

WASZAK, T. et TORRES-MORENO, J. (2008). Compression entropique de phrases contrôlée par un perceptron. *Journées internationales d'Analyse statistique des Données Textuelles*.

WOODSEND, K. et LAPATA, M. (2011). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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.

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.