Apport de la syntaxe pour l'extraction de relations en domaine médical

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, 1 square de la résistance, 91000 Évry
prenom.nom@limsi.fr

Résumé. Dans cet article, nous nous intéressons à l'identification de relations entre entités en domaine de spécialité, et étudions l'apport d'informations syntaxiques. Nous nous plaçons dans le domaine médical, et analysons des relations entre concepts dans des comptes-rendus médicaux, tâche évaluée dans la campagne i2b2 en 2010. Les relations étant exprimées par des formulations très variées en langue, nous avons procédé à l'analyse des phrases en extrayant des traits qui concourent à la reconnaissance de la présence d'une relation et nous avons considéré l'identification des relations comme une tâche de classification multi-classes, chaque catégorie de relation étant considérée comme une classe. Notre système de référence est celui qui a participé à la campagne i2b2, dont la F-mesure est d'environ 0,70. Nous avons évalué l'apport de la syntaxe pour cette tâche, tout d'abord en ajoutant des attributs syntaxiques à notre classifieur, puis en utilisant un apprentissage fondé sur la structure syntaxique des phrases (apprentissage à base de tree kernels) ; cette dernière méthode améliore les résultats de la classification de 3%.

Abstract. In this paper, we study relation identification between concepts in medical reports, a task that was evaluated in the i2b2 campaign in 2010, and evaluate the usefulness of syntactic information. As relations are expressed in natural language with a great variety of forms, we proceeded to sentence analysis by extracting features that enable to identify a relation and we modeled this task as a multiclass classification task based on SVM, each category of relation representing a class. This method obtained an F-measure of 0.70 at i2b2 evaluation. We then evaluated the introduction of syntactic information in the classification process, by adding syntactic features, and by using tree kernels. This last method improves the classification up to 3%.

Mots-clés : extraction de relation, domaine médical, apprentissage multi-classes, tree kernel.

Keywords: relation identification, medical domain, multiclass learning, tree kernel.

1 Introduction

L'extraction d'information permet d'obtenir des représentations structurées du contenu d'un corpus. Le domaine médical représente en cela un domaine d'application intéressant : en effet, les documents médicaux tels que les comptes-rendus cliniques contiennent de nombreuses informations sur le suivi médical des patients, et la structuration automatique de ces informations pourrait améliorer la prise en charge de ceux-ci.

L'extraction de ces informations amène à se poser différents problèmes, liés aux types d'information cherchés : la reconnaissance des termes du domaine dans les textes, des concepts qui leur sont liés, ainsi que l'identification des types de relations qui les lient dans les documents.

Nous nous sommes intéressées à l'identification de relations dans des comptes-rendus médicaux, tâche qui a fait l'objet d'une campagne d'évaluation i2b2 en 2010 [1]. Un premier travail a été réalisé modélisant l'identification des relations comme une tâche de classification multi-classes (Minard *et al.*, 2011b). Cette approche a été choisie car elle permet de caractériser les relations par des ensembles de traits lexicaux et de surface, et a conduit à l'obtention de bons résultats à i2b2. Néanmoins, les phrases étant parfois complexes, leur représentation par des traits de surface uniquement ne permet pas de capturer des relations entre termes distants. C'est pourquoi nous nous sommes posé 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 ? Et des approches par apprentissage sur des arbres syntaxiques sont-elles meilleures que des approches «sac de mots» ?

Après avoir présenté les travaux existant dans ce domaine, nous présenterons le contexte de notre étude, puis nous expliquerons nos méthodes et évaluerons notre approche.

2 L'extraction de relations

L'extraction de relations a donné lieu à de nombreux travaux, notamment dans le domaine biomédical. Les approches actuelles se fondent sur une classification automatique plus ou moins supervisée.

(Roberts *et al.*, 2008) proposent une approche classique fondée sur des SVM (machine à vecteurs de support) pour extraire des relations dans un corpus de spécialité : le corpus du projet CLEF (the Clinical E-Science Framework project). Ils extraient des relations entre des entités (ex : condition, médicament, résultat) et des modifieurs (ex : marqueur de négation) dans des dossiers de patients atteints d'un cancer. Les relations sont de sept types. Deux types d'entités (ou une entité et un modifieur) ne peuvent être reliées que par une relation (sauf entre une investigation et une condition où la relation peut être de deux types). La tâche est donc modélisée comme une classification binaire, c'est-à-dire que les SVM sont entraînés pour une décision entre une classe et toutes les autres. Les attributs qu'ils utilisent correspondent à des attributs de notre système de base :

– des attributs lexicaux : les mots (et les radicaux des mots) dans une fenêtre de 6 mots avant et après les entités en relation, les mots formant les entités, les mots situés entre les entités ;
– des attributs morpho-syntaxiques : les catégories morpho-syntaxiques et ces mêmes catégories généralisées (par exemple toutes les catégories verbales sont regroupées en *VB*) ;
– des attributs sémantiques : le type des entités en relation et des autres entités de la phrase.

D'autres travaux utilisent des attributs syntaxiques plus riches ((Zhou *et al.*, 2005), (Uzuner *et al.*, 2010)). En particulier, (Uzuner *et al.*, 2010) utilisent les dépendances syntaxiques entre les concepts sous forme d'attributs dans une approche vectorielle basée sur des SVM. Ils souhaitent typer des relations entre des problèmes, des tests et des traitements dans des comptes-rendus médicaux. Ils classent les relations en six catégories, comme par exemple les relations existantes entre une maladie qu'a le patient et un traitement, ou les relations entre une éventuelle maladie et un traitement. Ils utilisent des attributs classiques (l'ordre des concepts, la distance, des trigrammes lexicaux, les mots qui forment les concepts, les verbes, des bigrammes syntaxiques, etc.) ainsi que des attributs portant des informations sur les dépendances syntaxiques reliant les entités. Pour plusieurs relations ils obtiennent des F-mesures entre 0,60 et 0,85, mais pour les relations pour lesquelles il y a peu d'exemples dans le corpus d'apprentissage les F-mesures sont nulles. 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 notamment car les dépendances syntaxiques complètes ne sont trouvées que pour une faible proportion des phrases analysées.

Des travaux en domaine ouvert ont cependant montré que l'information structurelle utilisée sous forme d'arbres grâce à des tree kernels améliore la classification ((Culotta & Sorensen, 2004), (Zelenko *et al.*, 2003), (Zhang *et al.*, 2006)). En particulier, (Zhang *et al.*, 2006) ont étudié l'apport de la structure syntaxique des phrases pour l'extraction de relation en domaine général, en s'appuyant

1. https ://www.i2b2.org/NLP/Relations/

sur le corpus ACE 2003. Ils testent différentes sélections dans les arbres syntaxiques (arbre complet englobant les deux entités en relation, plus petit arbre commun, en conservant que les chunks, etc.), et ils montrent que les meilleurs résultats sont obtenus en utilisant le plus petit sous-arbre commun aux deux entités. (Culotta & Sorensen, 2004) utilisent des arbres de dépendance sur le même corpus, et montrent que les tree kernels sont meilleurs que l'information structurelle mise sous forme vectorielle. Ils testent deux types de tree kernels : *contiguous kernel* 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».

3 Objectif et méthodes

Nous nous sommes intéressées à l'extraction de relations en domaine de spécialité. Notre objectif était d'étudier comment intégrer des informations syntaxiques dans un système de classification automatique, et quel était l'apport de ce type d'informations. Nous avons considéré deux sous-tâches : la détection de la présence d'une relation entre deux entités, en l'occurrence des concepts médicaux, et la catégorisation de cette relation éventuelle selon des catégories prédéfinies. Ces sous-tâches ont été abordées comme des problèmes de classification supervisée bi-classes ou multi-classes.

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

3.1 Classification

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. De plus ils donnent de bons résultats pour les tâches pour lesquelles il y a beaucoup d'attributs mais qui sont très épars, comme c'est souvent le cas en TAL. 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 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 2). Pour la classification binaire, c'est-à-dire la détection de relation, nous avons utilisé le logiciel SVM-Light-TK version 1.5 de (Moschitti, 2006). 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. Nous avons choisi d'utiliser l'option SST qui est plus générale et donne de meilleurs résultats selon (Moschitti, 2006).

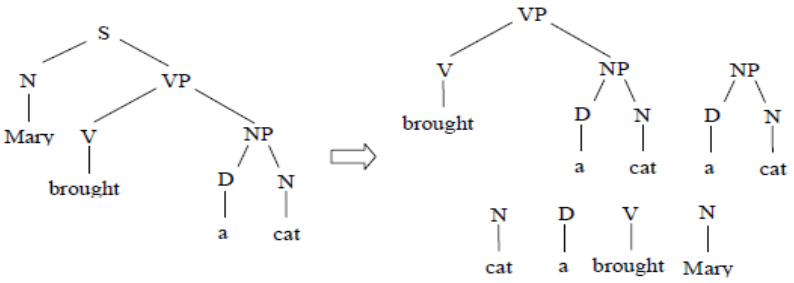

FIGURE 1 – Exemple de subtrees (ST) de (Moschitti, 2006)

Pour la classification multi-classes, c'est-à-dire la détection de relation et le typage des relations, nous avons utilisé le programme LIBSVM (Chang & Lin, 2001) avec une approche « un-contre-un ». Nous avons utilisé un noyau linéaire, ce choix est conseillé par (Hsu *et al.*, 2003) quand le nombre d'attributs est important par rapport au nombre d'instances. Nous avons fixé la valeur du paramètre *C* (*penalty parameter*) à *16* et celle du paramètre *gamma* (*kernel parameter*) à *0,03125*. Ces valeurs ont été choisies par l'outil *grid* fourni avec LIBSVM.

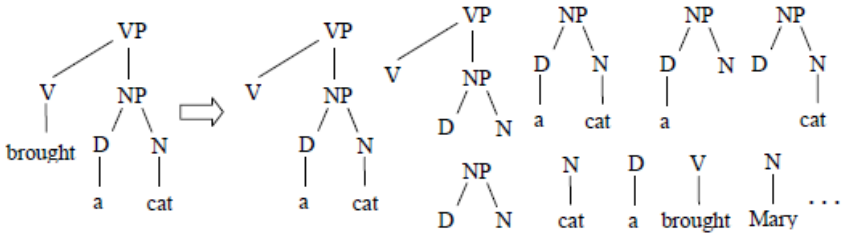

FIGURE 2 – Exemple de subset trees (SST) de (Moschitti, 2006)

3.2 Attributs de référence

Les attributs choisis pour la classification automatique 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 morpho-syntaxiques, comme les catégories morpho-syntaxiques des mots, et des informations sémantiques grâce à un typage de concepts.

Tous les attributs présentés ici constitueront les attributs de notre système de référence, auxquels seront ajoutées des informations syntaxiques. La sélection des attributs est présentée de manière plus détaillée dans (Minard *et al.*, 2011b).

3.2.1 Attributs de surface

Les attributs de surface sont relatifs à la position des concepts dans la phrase ; ce sont les suivants :
– l'ordre des concepts : celui-ci influence en effet la façon dont la relation est exprimée ;
– la distance entre les deux concepts en termes de nombre de mots [2] : deux concepts ont d'autant moins de chance d'être en relation qu'il y a de nombreux mots entre les deux ;
– la présence d'autres concepts entre les deux concepts étudiés : la présence d'un troisième concept modifie les probabilités que les deux concepts soient bien en relation.

3.2.2 Attributs lexicaux

Plusieurs attributs lexicaux sont pris en compte :
– les mots, et leurs radicaux (stems) [3], qui composent les concepts, et le mot tête de chaque concept [4]. Nous avons considéré les radicaux de manière à regrouper les variantes flexionnelles et dérivationnelles ;
– les radicaux des trois mots à gauche et à droite des concepts. La taille de la fenêtre a été choisie après plusieurs tests : avec une fenêtre plus grande ou plus petite la précision augmente très légèrement mais le rappel diminue ;
– les radicaux des mots entre les concepts, ce qui permet de tenir compte de tous les mots entre les concepts ; c'est ici qu'est située l'information la plus utile à la classification ;
– les radicaux des verbes dans une fenêtre de trois mots avant et après chaque concept, et entre eux, le verbe marquant souvent la relation ;
– les prépositions entre les concepts.

3.2.3 Attributs morpho-syntaxiques

Le TreeTagger est appliqué sur les phrases à analyser, et son étiquetage est utilisé pour les attributs suivants :

– la catégorie morpho-syntaxique des mots dans une fenêtre de trois mots à gauche et à droite des concepts ;
– la présence d'une préposition entre les concepts, peu importe la préposition ;
– la présence d'une conjonction de coordination ou d'une virgule entre les concepts ;

2. Le découpage en mots est effectué par le TreeTagger (Schmid, 1994).
3. Pour obtenir les radicaux des mots nous utilisons le module PERL lingua : :stem.
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).

– si les concepts ne sont séparés que par un mot, un attribut marque la présence d'un signe de ponctuation. Cet attribut permet de considérer les énumérations différemment.

3.2.4 Attributs sémantiques

Enfin, des attributs sémantiques 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'autre part :
– le type sémantique (issu de l'UMLS [5]) des mots dans une fenêtre de trois mots à gauche et à droite de chaque concept ;
– les types des concepts : c'est l'attribut le plus important car les relations possibles dépendent des types des deux concepts considérés ;
– les classes de VerbNet [6] (une extension des classes de Levin) des verbes dans un fenêtre de trois mots à gauche et à droite des concepts, et entre les concepts.

3.3 Informations syntaxiques

Les informations syntaxiques utilisées proviennent des arbres de constituants ; ces arbres syntaxiques ont été produits par l'analyseur de Charniak/McClosky (McClosky, 2010). 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% [7]). Les phrases ont été analysées après remplacement des abréviations, normalisation des dates, âges, noms propre et nombres, et annotation des concepts. La figure 3 est un exemple d'arbre fourni par l'analyseur à partir de la phrase suivante :

2 Low back strain requiring hospitalization for pain in 2002.

avec *hospitalization* étiqueté comme un concept de type traitement et *pain* étiqueté comme un concept de type problème (la balise <NUM> remplace un nombre et la balise <DATE> une date ou une année).

À partir de cet arbre nous avons produit le sous-arbre minimal reliant les deux entités possiblement en relation. Ce sous-arbre correspondant au chemin le plus court pour aller d'un concept à l'autre, nous l'appellerons «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 la première entité ni le contexte droit de la deuxième entité.

La figure 4 représente le sous-arbre minimal complet produit à partir de l'arbre présenté dans la figure 3 et la figure 5 le sous-arbre minimal. Ces trois arbres sont utilisés comme attributs pour la classification des relations.

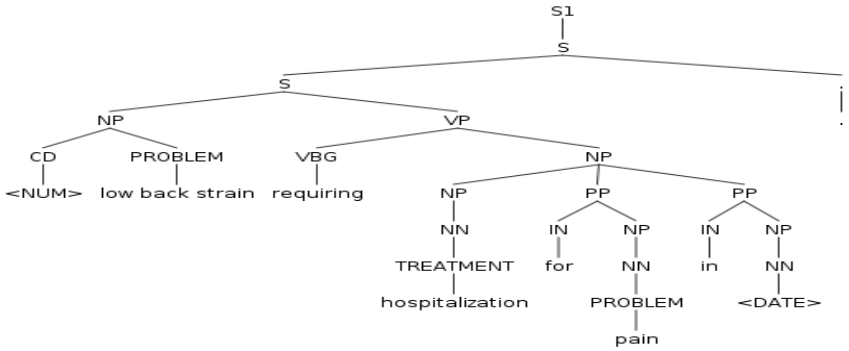

FIGURE 3 – Exemple de l'arbre complet

À partir du sous-arbre minimal nous avons calculé deux attributs : la taille du chemin reliant les deux entités (nous comptons le chemin entre les nœuds ayant pour valeur les types des concepts), et le constituant du nœud racine du sous-arbre. Pour le couple *hospitalization* et *pain*, 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 5).

5. http ://www.nlm.nih.gov/research/umls/
6. http ://verbs.colorado.edu/ mpalmer/projects/verbnet.html
7. http ://www.cs.brown.edu/ dmcc/biomedical.html

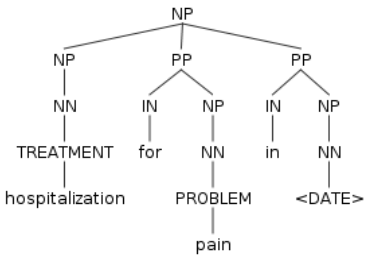

FIGURE 4 – Exemple du sous-arbre minimal complet entre les deux entités

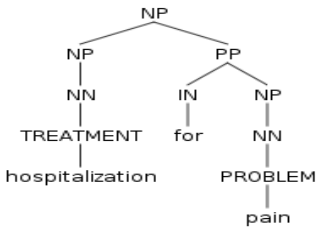

FIGURE 5 – Exemple du sous-arbre minimal entre les deux entités

4 Expérimentations

Les résultats présentés ici s'inscrivent dans le cadre de la campagne d'évaluation i2b2 en 2010. Trois tâches étaient proposées lors de cette campagne : la première consistait en l'annotation des problèmes médicaux, des traitements et des tests, la seconde en la détermination du statut d'assertion et la troisième en l'identification de relations. Nous avons travaillé sur la troisième tâche.

Les corpus, fournis par les organisateurs de la tâche i2b2, sont composés de comptes-rendus hospitaliers en anglais provenant de plusieurs centres médicaux aux États-Unis. Ces documents avaient été anonymisés et annotés manuellement. Un premier corpus a été fourni avant l'évaluation, composé de 350 documents. Puis, les organisateurs d'i2b2 ont fourni le corpus d'évaluation, qui comporte 477 documents.

Nous présentons dans une première partie les relations que nous devons identifier, dans une deuxième partie le corpus et nous terminons par les résultats des tests.

4.1 Relations considérées

La tâche consistait à annoter dans les comptes-rendus les relations existant entre deux concepts. Les concepts considérés étaient les suivants :

– les problèmes médicaux : les observations des patients ou cliniciens concernant ce qui n'allait pas ou semblait être causé par une maladie. Cette catégorie comprend notamment les maladies, syndromes, signes, et symptômes, les observations sur l'état mental du patient, les résultats anormaux de tests etc. ;
– les traitements : les procédures, interventions, substances et médicaments donnés à un patient pour tenter de résoudre un problème médical ;
– les tests : 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.

Ces concepts peuvent être reliés par des relations, par exemple un examen a pu être prescrit pour analyser un problème médical, il peut également révéler un problème. Nous avons cherché à extraire ces relations dans les documents du corpus. Afin d'étudier spécifiquement l'extraction des relations, les documents sur lesquels nous avons travaillé comportaient déjà l'annotation de référence des concepts [8], et il s'agissait donc de déterminer si, étant donné deux concepts, ils étaient en relation, et si oui, quel était le type de la relation.

Trois ensembles de relations ont été définis :

– relations entre problème et traitement :
  – le traitement améliore le problème (TrIP)
    Exemple : «<pb>hypertension</pb> was controlled on <treat>hydrochlorothiazide</treat>»
  – le traitement aggrave le problème (TrWP), ce qui inclut les cas où le traitement est administré pour le problème mais ne l'améliore pas
    Exemple : «<pb>the tumor</pb> was growing despite the available <treat>chemotherapeutic regimen</treat>»

8. Dans les exemples présentés dans l'article nous annotons les problèmes médicaux avec la balise <*pb*>, les tests avec la balise <*test*> et les traitements avec la balise <*treat*>

  - le traitement cause le problème (TrCP)
    Exemple : «<treat>Bactrim</treat> could be a cause of <pb>these abnormalities</pb>.»
  - le traitement est administré en raison du problème (TrAP)
    Exemple : «<treat>antibiotic therapy</treat> for presumed <pb>right forearm phlebitis</pb>»
  - le traitement n'est pas administré en raison du problème (TrNAP)
    Exemple : «<treat>Relafen</treat> which is contra-indicated because of <pb>ulcers</pb>.»
- relations entre problème et test :
  - le test révèle le problème (TeRP), et plus généralement les tests menés accompagnés de leurs résultats
    Exemple : «<test>an echocardiogram</test> revealed <pb>a pericardial effusion</pb>»
  - le test est conduit en raison du problème (TeCP)
    Exemple : «<test>an VQ scan</test> was performed to investigate <pb>pulmonary embolus</pb>»
- relations entre problème et problème
  - un problème en indique un autre (PIP), incluant les cas où des problèmes révèlent différents aspects d'un même problème médical
    Exemple : «a history of <pb>noninsulin dependent diabetes mellitus</pb>, now presenting with <pb>acute blurry vision on the left side</pb>.»

Le tableau 1 indique le nombre de relations par catégorie dans le corpus d'entraînement et d'évaluation, ainsi que l'accord inter-annotateur[9]. Nous observons que l'accord inter-annotateur est faible pour certaines catégories comme TrIP et TrWP.

Relation	entraînement	évaluation	accord inter-annotateur
TrIP	**107**	198	**0,62**
TrWP	**56**	143	**0,58**
TrCP	296	444	0,82
TrAP	1423	2487	0,95
TrNAP	106	191	0,76
PIP	1239	1986	0,79
TeRP	1734	3033	0,96
TeCP	303	588	0,74
Toutes	5264	9070	

TABLE 1 – Nombre de relations par catégorie dans chaque corpus

4.2 Corpus

Sur le corpus d'entraînement (350 documents) la taille moyenne des phrases contenant au moins deux concepts est de 16,78 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 212 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 moyenne des phrases de ces trois corpus ainsi que du corpus pour la campagne i2b2 2010 que nous utilisons sont répertoriées dans le tableau 2. Ces données montrent que le corpus sur lequel nous travaillons est composé de phrases courtes, comparé au corpus GENIA. En effet les documents type rapport clinique sont composés de beaucoup de fragments de phrase (e.g. 1) et d'énumérations (e.g. 2) ce qui fait leur particularité.

	taille moyenne des phrases
Penn Treebank	24,16
MED	13,79
GENIA	27,18
i2b22010 corpus	16,78

TABLE 2 – Taille moyenne des phrases de différents corpus

(1) <pb>C5-6 disc herniation</pb> with <pb>cord compression</pb> and <pb>myelopathy</pb> .

9. L'accord inter-annotateur nous a été fourni par les organisateurs, il a été calculé à partir de Knowtator.

(2) Revealed <pb>icteric sclerae</pb> , <pb>the oropharynx with extensive thrush</pb> , and <pb>an ulcer under his tongue</pb> .

4.3 Prétraitement du corpus

Les fichiers du corpus ont été prétraités et normalisés. Tout d'abord, les abréviations connues ont été remplacées par leur forme complète, grâce à une liste. La liste a été constituée pour la campagne d'évaluation i2b2 2009 [10] à partir de la liste d'abréviations biomédicales formée par Berman [11] à laquelle ont été ajoutés les exemples du corpus du challenge i2b2 2009. Ainsi, *h.o.* a été converti en *history of* et *p.r.n.* en *as needed*. Puis, les données d'anonymisation (différentes pour chaque corpus) ont été remplacées par des balises *NAME*, *DATE* et *AGE*. *NUM* remplace toutes les valeurs numériques présentes dans les fichiers (principalement des dosages). Enfin, les textes ont été étiquetés par le TreeTagger et analysés par le parser de Charniak/McClosky.

4.4 Tests et résultats

4.4.1 Système : sans information syntaxique

Nous avons d'abord évalué l'approche vectorielle avec les attributs de base décrits dans la section 3.2. Les résultats sont donnés dans le tableau 3. Les résultats de la ligne *Toutes relations* sont la précision moyenne, le rappel moyen et la F-mesure moyenne pour la classification de toutes les relations. Nous avons utilisé ce système pour la campagne d'évaluation i2b2. Nous obtenons une F-mesure de 0,703, ce qui nous permet de nous placer troisième sur seize participants (Minard *et al.*, 2011a).

Relation	Précision	Rappel	F-mesure
TrIP	0,852	0,146	0,250
TrWP	0,000	0,000	0,000
TrCP	0,735	0,362	0,485
TrAP	0,743	0,689	0,715
TrNAP	0,461	0,062	0,110
PIP	0,781	0,530	0,632
TeRP	0,876	0,832	0,853
TeCP	0,870	0,251	0,390
Toutes relations	**0,807**	**0,622**	**0,703**
Médiane			0,664
1er système	0,720	0,753	0,736
2e système	0,773	0,693	0,731

TABLE 3 – Précision, rappel et F-mesure obtenus sur le corpus d'évaluation avec le système de base

4.4.2 Système : avec des informations syntaxiques sous forme vectorielle

Nous avons voulu voir 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 produit par l'analyseur Charniak/McClosky, que nous avons ajoutées aux attributs de base. Il s'agit de la taille du chemin entre le premier concept et le deuxième concept, et du nom du constituant du nœud racine du sous-arbre minimal. Ce système obtient une F-mesure de 0,700. Les résultats détaillés sont présentés dans le tableau 4.

4.4.3 Sytème à base de tree kernels

L'ajout d'attributs n'améliorant pas la classification, nous avons testé l'ajout d'informations structurelles plus importantes que les deux attributs précédents. Pour cela nous avons utilisé le classifieur SVM-Light-TK basé sur des tree kernels. Comme certaines

10. https ://www.i2b2.org/NLP/Medication/
11. http ://www.julesberman.info/abbtwo.htm

Relation	Précision	Rappel	F-mesure
TrIP	0,875	0,141	0,243
TrWP	0,000	0,000	0,000
TrCP	0,733	0,360	0,483
TrAP	0,741	0,684	0,712
TrNAP	0,444	0,062	0,110
PIP	0,776	0,525	0,626
TeRP	0,877	0,833	0,854
TeCP	0,869	0,248	0,386
Toutes relations	**0,806**	**0,619**	**0,700**

TABLE 4 – Précision, rappel et F-mesure obtenus sur le corpus d'évaluation avec le système utilisant des informations syntaxiques sous forme vectorielle

relations ne sont pas suffisamment représentées dans le corpus, nous n'avons pas fait de classification multi-classes avec SVM-Light-TK. En effet pour 5 des 8 relations (TrIP, TrWP, TrNAP, TrCP et TeCP), le classifieur ne détectait pas de relation. Les évaluations que nous avons faites portent donc sur de la classification entre relation et non-relation, c'est-à-dire de la détection de relation.

Nous avons ajouté les arbres de constituants complets ainsi que les sous-arbres minimaux complets entre les deux entités possiblement en relation et les sous-arbres minimaux (c.f. 3.3), pour évaluer si d'autres 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'informations. Les sous-arbres minimaux contiennent en moyenne la moitié du nombre de mots de l'arbre complet. Plus précisément les arbres complets ont un nombre moyen de mots par phrase de 21 [12] (sans compter la ponctuation, et en comptant les concepts comme un seul mot), 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.

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 5 et dans la figure 6.

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

TABLE 5 – Évaluation des combinaisons des différents apprentissages à base de tree kernels pour la détection de relations

Ces résultats nous montrent que l'information contenue dans les arbres seule n'est pas suffisante pour la classification des relations quel que soit l'arbre utilisé. 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 de base 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 fournies par les arbres mais sont données par les attributs.

12. Le nombre de mots par phrase est différent de celui donné dans la section 4.2 car 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.

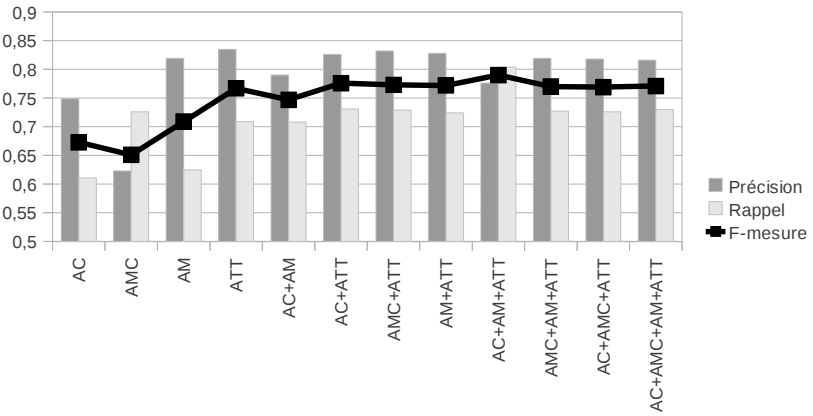

FIGURE 6 – Comparaison des différents apprentissages

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'information 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 première 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 utilisés pour détecter les relations entre des concepts éloignés dans la phrase. Les relations correctement détectées avec (AC + ATT) mais qui ne le sont pas avec (ATT) concernent deux concepts dont l'éloignement moyen est de 10,8 mots (ou ponctuations), alors que celles qui sont correctement classées par les deux systèmes concernent des concepts qui sont séparés en moyenne par 5,4 mots (ou ponctuations). Dans la figure 7 nous montrons une phrase contenant une relation de type PIP (un problème indique un autre problème) entre *increased tracer activity* et *active bleeding* ; les concepts sont séparés 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.

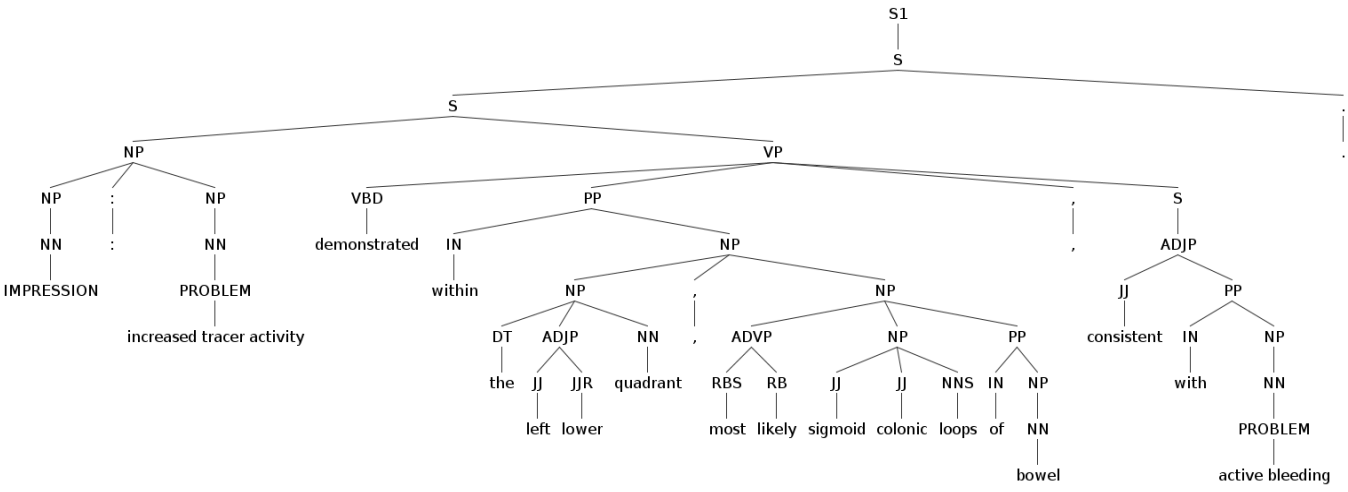

FIGURE 7 – Exemple de l'arbre complet d'une phrase contenant deux concepts reliés par une relation de type PIP

5 Conclusion

Dans cet article nous nous sommes intéressées à l'extraction de relations en domaine de spécialité. Nous avons développé un système de détection et typage de relations fondé sur une classification automatique, qui obtient une F-mesure d'environ 0,7. L'apport d'informations syntaxiques à ce système a ensuite été évalué. Les informations syntaxiques très simples ajoutées sous forme vectorielle n'ont pas amélioré la classification ; en revanche, l'utilisation des structures syntaxiques avec les tree kernels 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).

Il serait intéressant de poursuivre cette étude en faisant une classification multi-classes, pour évaluer l'apport des informations syntaxiques pour le typage des relations, ce qui nécessiterait d'augmenter le corpus d'entraînement, ou d'étendre l'étude à d'autres corpus de spécialité.

Remerciements

Ce travail a été partiellement financé par OSEO dans le cadre du programme Quæro.

Références

CHANG C.-C. & LIN C.-J. (2001). *LIBSVM : a library for support vector machines*. Software available at http ://www.csie.ntu.edu.tw/∼cjlin/libsvm.

CODEN A. R., PAKHOMOV S. V., ANDO R. K., DUFFY P. H. & CHUTE C. G. (2005). Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics*, **38**, 422–430.

CULOTTA A. & SORENSEN J. (2004). Dependency tree kernels for relation extraction. In *in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.

HSU C.-W., CHANG C.-C. & LIN C.-J. (2003). *A Practical Guide to Support Vector Classification*. Rapport interne, Department of Computer Science, National Taiwan University.

MCCLOSKY D. (2010). Any domain parsing : Automatic domain adaptation for natural language parsing. *PHD Thesis, Department of Computer Science, Brown University*.

MINARD A.-L., LIGOZAT A.-L., BEN ABACHA A., BERNHARD D., CARTONI B., DELÉGER L., GRAU B., ROSSET S., ZWEIGENBAUM P. & GROUIN C. (2011a). Hybrid methods for improving information access in clinical documents : Concept, assertion, and relation identification. *Journal of the American Medical Informatics Association*. À paraître.

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

MOSCHITTI A. (2006). Making tree kernels practical for natural language learning. In *In Proceedings of the Eleventh International Conference on European Association for Computational Linguistics (EACL), Trento, Italy, 2006*.

ROBERTS A., GAIZAUSKAS R. & HEPPLE M. (2008). Extracting clinical relationships from patient narratives. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, p. 10–18.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49.

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

ZELENKO D., AONE C. & RICHARDELLA A. (2003). Kernel methods for relation extraction. *Journal of Machine Learning Research*, **3**, 1083–1106.

ZHANG M., ZHANG J. & 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*, p. 288–295.

ZHOU G., SU J., ZHANG J. & ZHANG M. (2005). Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the ACL*, p. 427–434.