TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Accès au contenu sémantique en langue de spécialité : extraction des prescriptions et concepts médicaux

Cyril Grouin[1] Louise Deléger[1] Bruno Cartoni[2] Sophie Rosset[1] Pierre Zweigenbaum[1]
(1) LIMSI-CNRS, BP133, 91403 Orsay Cedex, France
(2) Département de Linguistique, Université de Genève, Suisse
{cyril.grouin, louise.deleger, sophie.rosset, pierre.zweigenbaum}@limsi.fr,
bruno.cartoni@unige.ch

Résumé. Pourtant essentiel pour appréhender rapidement et globalement l'état de santé des patients, l'accès aux informations médicales liées aux prescriptions médicamenteuses et aux concepts médicaux par les outils informatiques se révèle particulièrement difficile. Ces informations sont en effet généralement rédigées en texte libre dans les comptes rendus hospitaliers et nécessitent le développement de techniques dédiées. Cet article présente les stratégies mises en œuvre pour extraire les prescriptions médicales et les concepts médicaux dans des comptes rendus hospitaliers rédigés en anglais. Nos systèmes, fondés sur des approches à base de règles et d'apprentissage automatique, obtiennent une F_1-mesure globale de 0,773 dans l'extraction des prescriptions médicales et dans le repérage et le typage des concepts médicaux.

Abstract. While essential for rapid access to patient health status, computer-based access to medical information related to prescriptions key medical expressed and concepts proves to be difficult. This information is indeed generally in free text in the clinical records and requires the development of dedicated techniques. This paper presents the strategies implemented to extract medical prescriptions and concepts in clinical records written in English language. Our systems, based upon linguistic patterns and machine-learning approaches, achieved a global F_1-measure of 0.773 for extraction of medical prescriptions, and of clinical concepts.

Mots-clés : Extraction d'information, Indexation contrôlée, Informatique médicale, Concepts médicaux, Prescriptions.

Keywords: Information extraction, Controled indexing, Medical informatics, Clinical concepts, Prescriptions.

1 Introduction

L'accès au sens présent dans les documents au moyen d'outils informatiques est indispensable, tant du point de vue de la compréhension du contenu que du développement des méthodologies informatiques facilitant cet accès. Selon le domaine de langue étudié et le format des données accessibles, la production de systèmes est loin d'être triviale. Nous avons fait le choix d'axer cette étude sur un domaine de langue particulier, le domaine médical, en travaillant sur des documents spécifiques, les comptes rendus hospitaliers. Les comptes rendus hospitaliers intègrent un nombre important d'informations sur l'état de santé des patients, tant au niveau des prescriptions médicales que des concepts médicaux utilisés. Ces informations, bien que partiellement structurées en sections (antécédents du patient, histoire de la maladie, traitement de sortie, etc.), sont rédigées en texte libre et leur appréhension par des outils informatiques, en l'absence de normalisation, se révèle difficile. Cependant, la langue employée dans les comptes rendus se caractérise par une stabilité et une formalisation élevées sur le plan syntaxique, sémantique, et même structurel (Sager, 1981; Friedman, 2000), ce qui autorise une analyse automatique.

Un accès rapide aux informations médicales contenues dans un dossier patient est essentiel pour les praticiens hospitaliers, pour résumer les antécédents du patient ou pour réaliser des études préventives. Deux types d'informations médicales émergent dans les documents cliniques : en premier lieu, les informations liées à la prise de médicaments, qu'elles concernent le médicament en lui-même ou les informations associées (dosage, fréquence, etc.) ; en second lieu, les concepts clés dans la pratique clinique, qui recouvrent les problèmes médicaux (signes, symptômes, maladies, etc.), les examens réalisés pour les diagnostiquer, et les traitements associés.

Nous présentons dans cet article un état de l'art sur l'accès au contenu sémantique dans les comptes rendus cliniques (section 2) puis les approches que nous avons développées pour accéder aux informations médicales, d'une part pour extraire les informations liées aux prescriptions médicales (section 3), d'autre part pour repérer, extraire et typer les concepts médicaux (section 4) dans le cadre de nos participations aux éditions 2009 et 2010 du challenge international i2b2 (*informatics for integrating biology to the bedside*) dont les thématiques concernaient ces aspects (Uzuner *et al.*, 2010a,b). Nous détaillons et discutons les résultats obtenus dans chacune de ces sections.

2 État de l'art

L'accès au contenu d'un document textuel peut être appréhendé de deux manières : soit par le biais d'approches à base d'apprentissage, soit par la création de patrons linguistiques faisant appel à des connaissances d'expert.

Les approches à base d'apprentissage reposent sur l'utilisation de corpus annotés avec soin, dans une volumétrie suffisante et une répartition homogène, pour permettre à un système d'apprendre les conditions dans lesquelles se rencontrent les informations à extraire. Ces approches font l'objet de nombreux travaux, en particulier dans le domaine de la reconnaissance des entités nommées médicales (Li *et al.*, 2008; Doan & Xu, 2010) ou en analyse morphologique (Claveau & Kijak, 2010), rendus possibles par la disponibilité étendue et la simplicité d'utilisation de ces outils d'apprentissage. Si ces outils permettent d'obtenir rapidement de bons résultats, ils demeurent largement dépendants des données fournies en entrée, et seules des données homogènes, de qualité et disponibles en nombre suffisant, tels les corpus des challenges médicaux i2b2, permettent d'obtenir des résultats convaincants.

À l'opposé, les techniques à base de patrons linguistiques faisant appel à des connaissances d'expert pour la production de ces patrons ne nécessitent pas de corpus annotés. Elles nécessitent une somme de travail conséquente pour produire et adapter les patrons mais proposent l'avantage de fournir de bien meilleurs résultats (Long, 2007; Hamon & Grabar, 2010), grâce aux ressources linguistiques existantes en anglais pour le domaine médical, telles que le Metathesaurus et le Specialist Lexicon de l'UMLS (Lindberg *et al.*, 1993). La généralisation de ces approches apparaît souvent délicate à opérer, du fait de la spécialisation de la langue de spécialité concernée.

La combinaison de ces deux approches permet d'accroître sensiblement la qualité des résultats produits, soit comme approches complémentaires l'une de l'autre (une technique suivie de la seconde (Tikk & Solt, 2010)), soit comme apport de l'une pour l'autre (les patrons linguistiques utilisés pour extraire des informations réutilisées comme caractéristiques lors de la construction des modèles d'apprentissage (Wang, 2009)).

Le choix de mobiliser une approche plutôt qu'une autre est souvent dicté par le type de corpus rendu disponible : une approche à base d'apprentissage en cas de corpus annoté, une approche à base de lexiques et de règles le cas échéant. Nous avons suivi cette observation dans les choix méthodologiques décrits dans les sections suivantes.

3 Accès aux prescriptions médicales

Nous avons d'abord mis au point les méthodes d'extraction de prescriptions médicales pour l'anglais, dans le cadre de notre participation à l'édition 2009 du défi i2b2 (Deléger *et al.*, 2010). Nous les avons ensuite adaptées au français. Les données étant relatives à une langue de spécialité, les techniques décrites sont en conséquence conditionnées par cette langue de spécialité.

3.1 Présentation générale

Les prescriptions médicales recouvrent le nom du médicament (qu'il s'agisse d'un nom commercial, du nom générique, ou du principe actif) et les informations associées à ce médicament. On distingue ainsi différents types d'informations. En premier lieu, les informations relatives à la posologie (dosage, fréquence, quantité, mode d'administration, durée), à la forme galénique, etc. Ces informations se présentent sous des formes relativement stables qu'il est alors possible de décrire au moyen de patrons linguistiques. Un deuxième type d'information concerne la raison de la prise de ce médicament. Ce type d'information n'apparaît pas sous une forme régulière et doit faire l'objet d'une analyse plus complexe du texte. Enfin, un troisième type d'information se situe au niveau des événements et de la temporalité relatifs à ces prescriptions médicales et nécessite une analyse des phénomènes linguistiques entrant en jeu autour des noms de médicaments.[1] Le traitement de ce dernier type d'information a été abandonné lors du déroulement du défi 2009.

3.2 Présentation du corpus

Le corpus est composé de comptes rendus hospitaliers rédigés en anglais. Les documents proviennent d'un centre médical américain spécialisé en cardiologie. Ils ont fait l'objet d'une anonymisation où les informations personnelles (noms, prénoms, etc.) ont été remplacées par d'autres informations de même type en conservant un caractère vraisemblable. Le corpus de développement intègre 696 documents, parmi lesquels 17 ont fait l'objet d'une annotation, tandis que le corpus de test intègre 547 documents. Les documents sont structurés en sections assez générales telles que histoire de la maladie, allergies, examens de laboratoire, suivi de l'hospitalisation, et prescriptions de sortie. Les textes contiennent des abréviations qui concernent les noms de médicaments ("vanc" pour vancomycin, "levo" ou "levoflox" pour levofloxacin), les symptômes médicaux ("afib" pour *atrial fibrillation*, "abd pain" pour *abdominal pain*), les fréquences ("bid" pour *bis in diem*), et les modes d'administration ("iv" pour *intravenous*, "sub" pour *sub-lingual*).

Aucune annotation de référence n'existant préalablement au lancement du défi, la référence a été constituée en deux temps, premièrement par un vote majoritaire des sorties produites par les participants, et deuxièmement via une phase d'adjudication faisant intervenir l'ensemble des participants au défi (Uzuner *et al.*, 2010b). Au final, la référence a été constituée de manière collective pour 251 documents du corpus de test. Les résultats que nous présentons dans cet article pour la partie extraction de prescriptions médicales se fondent donc sur l'évaluation opérée sur ces documents de référence.

	Développement		**Test**	
Nombre de documents	17		251	
Médicaments	749	100,0 %	8 941	100,0 %
Dosage	397	53,0 %	4 460	49,9 %
Mode d'administration	253	33,8 %	3 389	37,9 %
Fréquence	374	49,9 %	4 042	45,2 %
Durée	66	8,8 %	550	6,2 %
Raison	150	20,0 %	1 636	18,3 %

TAB. 1 – Nombre d'éléments à extraire dans les documents annotés des corpus de développement et de test.

[1] La prescription médicale est-elle en cours, ou bien doit-elle être commencée ou arrêtée ? Où se situe la prescription médicale sur l'échelle temporelle (dans le passé, le présent ou le futur) ? Comment la prescription médicale est-elle présentée au patient (le médicament doit-il être pris obligatoirement, sous certaine condition, ou s'agit-il d'une suggestion) ?

Le tableau 1 renseigne du nombre d'informations attendues dans chaque corpus. Faute de disposer d'un corpus de développement entièrement annoté, nous donnons la volumétrie pour les 17 fichiers annotés qui nous ont été fournis par les organisateurs avec le corpus de développement. Si le nombre d'informations de chaque type reste proportionnel entre les deux corpus, il apparaît d'emblée que certaines informations sont peu présentes dans l'ensemble des corpus, rendant difficile le développement d'outils robustes pour les traiter. C'est notamment le cas des informations de durée renseignées dans moins de 10 % des prescriptions. Une prescription sur cinq seulement intègre la raison pour laquelle le médicament a été prescrit. Les autres types d'information sont davantage renseignés : le mode d'administration dans une prescription sur trois, les dosage et fréquence dans des proportions équivalentes d'une prescription sur deux.

Dans l'exemple du tableau 2, nous représentons les informations à extraire en les encadrant de balises. Les deux occurrences du médicament *heparin* doivent donner lieu à deux lignes de sortie. La première ligne – relative à la première occurrence – intègrera les informations de dosage, de mode d'administration, de fréquence et de raison, alors que la seconde ligne – relative à la seconde occurrence – ne comprendra que l'information de raison, les autres informations se rapportant uniquement à la première apparition.

<raison> *Prophylaxis* </raison> , <medicament> *heparin* </medicament> <dosage> *5000 units* </dosage> <mode> *subcu* </mode> <frequence> *t.i.d.* </frequence> - *the patient has consistently refused her* <medicament> *heparin* </medicament> .

TAB. 2 – Exemple d'annotation en prescriptions médicales.

3.3 Description du système

Notre système ayant été développé dans le cadre de la participation à l'édition 2009 du défi i2b2, nous l'avons orienté vers le traitement des informations suivantes : nom du médicament, dosage, mode d'administration, fréquence, durée, raison de la prescription, et type de portion de texte dans lequel apparaît la prescription (liste ou passage narratif). Nous avons fait le choix de développer un système reposant entièrement sur des règles d'extraction et des listes, sans recourir à des outils externes tels que des étiqueteurs, lemmatiseurs ou analyseurs syntaxiques. Ce choix repose sur le fait que les informations à extraire peuvent l'être, soit par la projection de lexiques (noms de médicaments, modes d'administration), soit par l'utilisation de règles (les chiffres des dosages, fréquences, durées, etc.), ces méthodes permettant l'obtention rapide de résultats de qualité.

Les problèmes à résoudre dans cette tâche consistaient à gérer l'exhaustivité des noms de médicaments (génériques, marques, classes thérapeutiques) et l'ambiguïté intrinsèque de ces noms (distinguer la concentration du dosage, repérer les substances actives utilisées comme nom de médicament). Nous devions également calculer le rattachement des informations aux noms de médicaments, prendre en compte la factorisation des informations, et considérer les cas particuliers de reprises pronominales.

3.3.1 Lexiques

Nous avons créé trois types de lexiques. Le premier lexique concerne les noms de médicaments et existe sous deux versions : une version réduite de 8 923 noms de médicaments issus de deux sites Internet (FDA[2] et RxList[3]), et une version plus large contenant 180 089 noms correspondant aux entrées du Metathesaurus de l'UMLS[4] pour le type sémantique *Clinical drug*. Les éléments présents dans cette seconde liste sont néanmoins sujets à discussion et ne correspondent pas toujours à des noms de médicaments tels que ceux attendus (*alcool, tabac,* etc.). Le second lexique est constitué d'une liste de symptômes médicaux pour permettre l'identification de la prescription d'un médicament. Il a été créé à partir des entrées de l'UMLS classées sous le type sémantique *Sign and Symptom*. Enfin, le dernier lexique consiste en une liste d'abréviations et termes spécifiques issue des travaux de (Berman, 2004). Nous avons mis en correspondance chaque terme avec le type d'information qui lui correspond : des abréviations ou termes de types dosage (*mg, sliding scale*), mode d'administration (*iv, intramuscular*), fréquence (*qd, prn*), durée (*week*).

[2]FDA : Food and Drug Administration, `http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm`
[3]`http://www.rxlist.com/`
[4]UMLS : Unified Medical Language System.

3.3.2 Algorithme

Nous avons défini une stratégie d'extraction d'information reposant sur deux étapes principales (figure 1) : dans un premier temps, nous identifions les noms de médicaments ; à partir de cette première étape, nous recherchons les informations associées à chaque médicament.

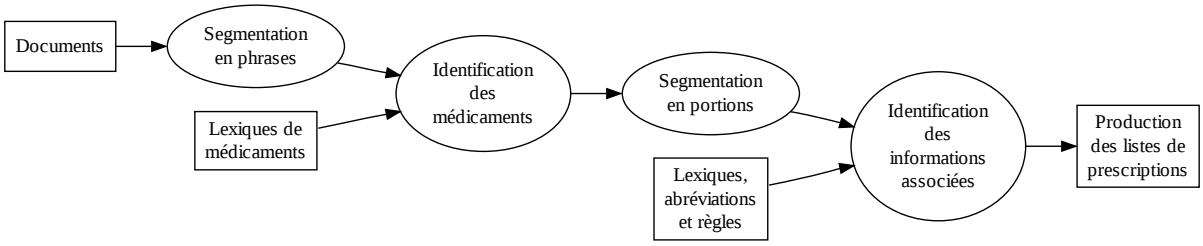

FIG. 1 – Architecture du système d'extraction de prescriptions médicales utilisé pour i2b2 2009.

L'identification des noms de médicaments repose uniquement sur un appariement exact avec le contenu des lexiques de médicaments précédemment décrits. Une fois les noms de médicaments identifiés dans un document, nous cherchons les informations qui lui sont associées. Nous avons élaboré des expressions régulières pour chaqye type d'information à traiter, à partir des guides d'annotation et d'exemples identifiés en corpus. Nous complétons l'application de ces règles par une recherche dans les listes d'abréviations et de symptômes.

Pour déterminer les informations devant être associées à chaque médicament, nous avons procédé à deux étapes de segmentation du texte. Dans un premier temps, nous segmentons le texte en phrases en nous fondant sur la mise en forme du document (lignes séparatrices et titres de section) et la ponctuation (en distinguant les points de fin de phrase des points d'abréviation ou des points mathématiques dans les décimales en anglais). Nous identifions les noms de médicaments dans ces phrases. Dans un second temps, nous procédons à une segmentation des phrases sur la base des noms de médicaments précédemment identifiés, en considérant que chaque nom de médicament constitue le début d'une portion de phrase. Nous cherchons alors les informations associées à chaque médicament à l'intérieur de ces portions, considérant que les informations associées aux prescriptions médicales suivent toujours les noms de médicaments. Pour le cas où certains types d'information n'auraient pas été trouvés à la suite du nom de médicament, nous les cherchons dans la portion qui précède.

Le système permet également de gérer les cas de doubles entrées, lorsqu'une même information s'applique à deux prescriptions différentes (deux médicaments prescrits pour soigner la même affection), ou parce qu'une seule expression factorise deux informations de même type (un dosage différent le matin et le soir). Nous avons géré ces cas au moyen de règles définies empiriquement.

Enfin, nous avons traité quelques cas particuliers de résolution des anaphores au moyen de règles dédiées : le pronom "*this*" suivi de trois syntagmes verbaux, "*was discontinued*", "*was increased*" et "*was decreased*". Dans ces cas de reprise pronominale, nous avons créé une seconde sortie pour le médicament désigné par le pronom, éventuellement complétée par les informations suivant le syntagme verbal (en cas de modification du dosage, etc.).

3.4 Résultats et discussion

Nous donnons dans le tableau 3 les résultats obtenus par notre système sur le corpus de test composé des 251 documents annotés collectivement. Comme pour toute évaluation d'un système d'extraction d'information, deux points sont ici évalués : le typage de l'élément extrait d'une part, et la portée de l'extraction d'autre part. Les résultats présentés ici exigent que la portée ait été déterminée de façon exacte (notre système peut avoir correctement typé un élément mais l'évaluation sera incorrecte du fait d'une erreur de frontière dans la portée de l'information extraite). Les informations élémentaires de type dosage, mode d'administration, fréquence, durée, et raison ne sont considérées comme pertinentes que si elles sont associées dans la référence à un médicament. Les rangées médicament, dosage, etc., évaluent chaque type d'information séparément. La rangée « niveau horizontal » demande qu'une prescription soit complètement et exactement reconnue pour être considérée comme correcte.

	Nombre	**Rappel**	**Précision**	**F_1-mesure**
Niveau horizontal	8 941	0.725	0.827	0.773
Médicament	8 941	0.793	0.802	0.798
Dosage	4 460	0.732	0.892	0.804
Mode d'administration	3 389	0.792	0.885	0.836
Fréquence	4 042	0.770	0.893	0.827
Durée	550	0.282	0.657	0.394
Raison	1 636	0.234	0.412	0.299

TAB. 3 – Résultats obtenus par notre système au défi i2b2 2009 (recouvrement exact).

Notre système obtient globalement de bons résultats (il a été classé 8ème sur 20 participants internationaux) avec une précision toujours supérieure au rappel, notre système générant relativement peu de bruit. Certains types d'information tels que la durée et les raisons de la prescription ont produit des résultats assez bas. Concernant les durées, le nombre restreint d'exemples dans le corpus de développement ne nous a pas permis de définir de manière précise et robuste les règles appliquées pour l'identification de ce type d'information.

Nous estimons qu'un moyen d'améliorer la détections des raisons passe par l'utilisation d'outils d'analyse syntaxique, de manière à identifier précisément les syntagmes nominaux et prépositionnels. Il semble que dans une bonne partie des situations où notre méthode n'a pas pu détecter la raison d'une prescription, cette raison était exprimée dans le contexte d'une portion de phrase relativement bien formée, où les relations grammaticales ont de bonnes chances d'être analysables automatiquement et d'aider à rattacher raison et médicament. Cependant, la variation syntaxique et l'étendue des raisons annotées dans le corpus d'entraînement témoignent de la complexité de cette tâche : les raisons "*pain*" (un seul terme), "*the previous enterococcus infection*" (un syntagme nominal), et "*had a temperature to about 101*" (un syntagme verbal) ont ainsi été associées au médicament "*vancomycin*". Un autre moyen consiste à utiliser une base de connaissances faisant le lien entre médicament et symptômes traités : si le terme "*hypercholesterolemia*" (ou une variante) est trouvé dans le voisinage des médicaments "*Zocor*" et "*simvastatin*", nous pourrions extraire la raison en accordant une importance accrue à ce terme. Une autre piste permettant l'amélioration de l'identification des informations associées consiste à mobiliser des présupposés d'expert, en adoptant une approche par inférence (déduire le mode d'administration d'un médicament à partir de sa forme galénique). Le coût de constitution d'une telle base de données associé à l'absence de normalisation des textes risquent néanmoins de limiter les apports d'une telle démarche.

4 Accès aux concepts médicaux

4.1 Présentation générale

La première piste de la campagne i2b2/VA 2010 concernait la détection et le typage de concepts médicaux dans des comptes rendus médicaux, parmi trois catégories de concepts (voir tableau 4) : les *problèmes* se rapportent aux observations faites sur l'état du patient et concernent les maladies et symptômes anormaux ou liés à une maladie existante, les *traitements* décrivent les méthodes utilisées pour résoudre le problème d'un patient (procédures, médicaments, etc.), et les *examens* se rapportent aux examens prescrits pour aider à diagnostiquer ces problèmes.

Concept	**Exemples**
Problème	<problem> *C5-6 disc herniation* </problem> *with* <problem> *cord compression* </problem> PRN <problem> *Shortness of Breath* </problem>
Traitement	<treatment> *bilateral lymph node dissection* </treatment> <treatment> *LISINOPRIL* </treatment> *10 MG PO DAILY*
Examen	*If* <test> *BS* </test> *is less than 125* *He was found on* <test> *physical exam* </test> *to have an asymmetric prostate*

TAB. 4 – Exemples de concepts de chaque type pour la tâche i2b2/VA 2010.

La syntaxe spécifique de la langue médicale utilisée dans les comptes rendus médicaux a notamment été décrite par (Sager *et al.*, 1994, 1995; Sager & Nhàn, 2002). Nous constatons aisi que certaines phrases peuvent être constituées presque exclusivement d'énumérations, ne comprendre qu'un seul mot ou au contraire être longues et qu'il n'y a pas eu de normalisation dans la façon de noter les éléments (voir tableau 5).

Phénomène étudié	**Exemples**
Absence de normalisation	*Supprelin La* vs *Supprelin LA* *magnetic resonance imaging of ...* vs *MRI of ...* *Thaw* vs *THAUW*
Forme des phrases	*On physical examination today , his lungs are clear to auscultation and percussion . Regular rhythm .* *f / u with PCP and Dr. Pump as scheduled , return to ED with worsening sob or increased cough or sputum production*

TAB. 5 – Exemples de problèmes rencontrés en langue de spécialité.

Ces différentes considérations nous ont convaincus de ne pas procéder à une analyse syntaxique des documents comme traitement de base. Du fait de la forme très variable des expressions désignant les concepts à détecter, nous avons également décidé de ne pas chercher à modéliser complètement ces expressions par une ou plusieurs grammaires locales. Par ailleurs, disposant d'un corpus d'apprentissage de taille raisonnable, nous avons opté pour une approche s'appuyant sur des champs conditionnels aléatoires (CRF) (Lafferty *et al.*, 2001), ces derniers permettant de bonnes performances pour une tâche d'étiquetage en séquence comme celle de la détection de concepts. Nous avons pour cela utilisé l'implémentation CRF++ (Kudo, 2007). Toutefois, si ces modèles permettent de bonnes performances, des expériences (Zidouni *et al.*, 2010) ont montré qu'utiliser comme attributs des informations d'ordre linguistique (POS, informations sémantiques, etc.) permettait d'améliorer les modèles. Nous avons cherché à produire des informations et des analyses partielles des expressions concernées, et à fournir au CRF des attributs encodant ces informations. L'objectif étant de produire les analyses linguistiques que l'on peut obtenir de façon fiable et de déléguer au processus d'apprentissage les décisions finales sur les frontières et type des entités.

4.2 Description du corpus

Le corpus se compose de comptes rendus cliniques provenant à part égale de trois hôpitaux nord-américains.[5] Le corpus d'entraînement se compose de 349 documents manuellement annotés[6] tandis que le corpus de test comprend 477 documents. Il n'existe pas de type de concept sur-représenté par rapport aux autres types et la distribution des types reste équivalente entre les deux corpus (voir tableau 6). Enfin, nous observons que les concepts médicaux à identifier recouvrent des formes d'expressions assez différentes à l'intérieur de chaque type. Une abréviation ou un syntagme nominal complet peuvent tous deux constituer un concept médical (tableau 4).

	Développement		**Test**	
Nombre de documents	349		477	
Concepts	27 837	100,0 %	45 009	100,0 %
Problème	11 968	43,0 %	18 550	41,2 %
Traitement	8 500	30,5 %	13 560	30,1 %
Examen	7 369	26,5 %	12 899	28,7 %

TAB. 6 – Nombre d'éléments à extraire dans les corpus de développement et de test.

[5] Beth Israel Deaconess Medical Center (Boston, MA), Partners HealthCare (Boston, MA), University of Pittsburgh Medical Center (Pittsburgh, PA). Ces instituts ont tous trois fourni des comptes rendus cliniques ; l'Université de Pittsburgh a également fourni des notes de suivi.

[6] Les organisateurs ont également fourni 827 documents non annotés avec le corpus de développement. Nous avons fait le choix de ne travailler que sur les 349 documents annotés, notre système reposant sur la construction de modèle par apprentissage (voir sous-section 4.3).

4.3 Description du système

4.3.1 Présentation générale

L'approche que nous avons développée (Minard *et al.*, 2011) repose sur un système à base d'apprentissage. Nous avons ainsi créé des modèles d'apprentissage à base de CRF en utilisant les traits habituels pour ce genre de tâche, à savoir des n-grammes et des indices typographiques (casse, ponctuation, token alphabétique ou numérique etc.). Nous avons également ajouté des traits correspondant aux résultats d'analyses linguistiques.

Afin de procéder à différents tests lors de la construction du modèle, nous avons scindé le corpus de développement en sous-corpus d'entraînement (241 documents), de développement (54 documents) et de test à blanc (54 documents). Pour la phase de test du défi, une fois trouvée la meilleure configuration, nous avons reconstruit un modèle global fondé sur l'ensemble des 349 documents.

4.3.2 Algorithme

Notre approche reposant sur l'application d'un modèle à base d'apprentissage, nous avons mobilisé plusieurs ressources pour produire les traits nécessaires à la construction du modèle (schéma 2).

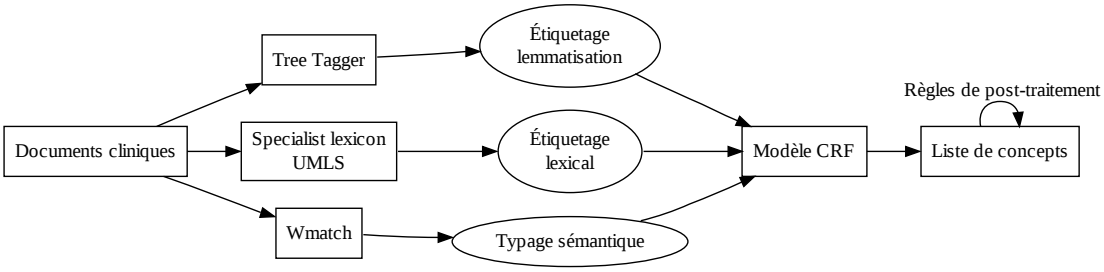

FIG. 2 – Architecture du système d'extraction de concepts médicaux utilisé pour i2b2/VA 2010.

Tous les mots des corpus ont d'abord été annotés en utilisant le Tree Tagger (Schmid, 1994) et ses modèles pour l'anglais. Ainsi chaque token a été associé avec sa partie du discours et son lemme.

Nous avons ensuite effectué un étiquetage à l'aide d'informations lexicales en utilisant les ressources fournies par le Specialist Lexicon de l'UMLS (Lindberg *et al.*, 1993). Ces ressources contiennent 62 263 adjectifs et 320 013 noms, et distinguent les adjectifs relationnels des adjectifs qualificatifs, ainsi que différents types de noms (noms propres, noms comptables et non comptables). Pour les adjectifs, cette ressource contient également des informations sur la position des adjectifs dans la phrase (*attribut* ou *post-nominal*).

Nous avons également ajouté une information sémantique en nous appuyant sur les travaux de (Sager *et al.*, 1995) et sur les données d'entraînement. Nous avons ainsi construit des lexiques spécialisés (pour les noms des parties du corps, de maladie, de médicaments) et des grammaires permettant de typer des segments en fonction de différentes catégories : les parties du corps (*sternal articular facet of third costal cartilage*), les analyses de laboratoire (*blood wbc, creatinine, hematocrit*), les différents examens (*angiography, biopsy*), des pré- et post-marqueurs d'examens (*follow-up ..., physicial ..., ... levels*), les médicaments (*Abacavir Sulfate*), les mode d'administration (*inhaler, oral, pills*), les instruments et objets médicaux (*cannula, pacemaker, stent*), les procédures (*bypass, amputation, resection*), et les dosages (*100 mg, 1 dose*). Ces différentes catégories ont paru pertinentes après examen du corpus et analyse des contextes droit et gauche des concepts ainsi que de la composition des concepts eux-mêmes. Précisons que ces catégories n'avaient pas vocation à représenter directement les concepts, mais à fournir des classes permettant de regrouper des mots ou groupes de mots sous une même appellation afin de réduire l'espace de recherche. Le tableau 7 montre des exemples de ces catégories (partie gauche du tableau) et les met en rapport avec les concepts (partie droite). Nous avons par ailleurs remarqué que ces catégories fournissent des informations structurantes qui se rapportent aux concepts. Ainsi, un test se rapporte souvent à une partie de l'anatomie et une procédure alors que certains noms ou adjectifs sont fréquemment présents en partie droite (pré-marqueur) ou gauche (post-marqueur) des concepts, en particulier pour les concepts *problème* et *examen*.

Annotation sémantique + POS	**Annotation correspondante du concept**
1)_JJ Rapid_JJ <anat> atrial_JJ </anat> <diag> fibrillation_NN </diag> with_IN demand_NN <diag> ischemia_NN </diag>	1) <problem> *Rapid atrial fibrillation* </problem> *with* <problem> *demand ischemia* </problem>
<localisation> Left_VVD </localisation> <anat> heart_NN </anat> <procedure> catheterization_NN </procedure> without_IN intervention_NN (_(**DATE[Dec_NP 16_CD 07]_NN)_) ._SENT	<test> *Left heart catheterization* </test> *without* <treatment> *intervention* </treatment> *(**DATE[Dec 16 07])* .
There_EX was_VBD no_DT <diag> diplopia_NN </diag> ,_, visual_JJ <pomark-disease> loss_NN </pomark-disease> ,_, <diag> speech_NN abnormality_NN </diag> or_CC sensory_JJ change_NN in_IN her_PP$ history_NN ._SENT	*There was no* <problem> *diplopia* </problem> *,* <problem> *visual loss* </problem> *,* <problem> *speech abnormality* </problem> *or* <problem> *sensory change* </problem> *in her history* .
<premark-disease> Significant_JJ </premark-disease> for_IN non-insulin_NN <diag> diabetes_NN mellitus_NN </diag> ,_, for_IN which_WDT he_PP takes_VVZ <medoc> Diabeta_NP </medoc> ,_, one_CD QD_NNS ;_ : <anat> right_JJ eye_NN </anat> <diag> cataract_NN </diag> ,_, operated_VVN on_IN three_CD years_NNS ago_RB ._SENT	*Significant for* <problem> *non-insulin diabetes mellitus* </problem> *, for which he takes* <treatment> *Diabeta* </treatment> *, one QD ;* <problem> *right eye cataract* </problem> *, operated on three years ago* .

TAB. 7 – Lien entre information sémantique et concepts.

Enfin, nous avons cherché à voir s'il était possible de typer sémantiquement les tokens en fonction de leurs premiers ou derniers caractères que nous appelons par commodité préfixe et suffixe. Nous avons découpé les différents mots n'appartenant pas aux dictionnaires de spécialité (médicaments ou parties du corps) puis en avons extrait les successions de caractères qui permettaient à coup sûr un début de classification. Ainsi, les suffixes de type *-stomy* renvoient fréquemment à une procédure. En tout, cinq classes sémantiques ont été utilisées (position, chiffrage, procédure, examen, diagnostic).

Les grammaires ont été construites en utilisant WMATCH, un moteur d'analyse fondé notamment sur des expressions régulières de mots (Galibert, 2009; Rosset *et al.*, 2008). L'analyseur a été construit de manière automatique à partir des données d'entraînement et des différents lexiques de spécialité à notre disposition. Ceux-ci étaient au nombre de trois : *anatomie* (145 199 mots ou expressions complexes), *médicaments* (27 518 mots ou expressions complexes), et *maladies* (175 645 mots ou expressions complexes). Nous avons d'autre part collecté les collocations des concepts et créé, en nous appuyant sur la fréquence et la distribution non ambiguë des termes, des lexiques spécifiques à la tâche (modes d'administration, procédures, outils médicaux, localisations sur le corps du patient – souvent en rapport avec une partie du corps –, examens, et pré et post-marqueurs, tant pour les examens que pour les maladies). Ces lexiques ont été utilisés pour l'acquisition des règles d'analyse au format WMATCH. Le tableau 8 présente des exemples de catégorisation de mot fondé sur le suffixe (extrait), de règle contextuelle pour la détection de la catégorie *mode* et d'appel à un lexique. La règle de catégorisation indique que les mots se terminant par "asty" sont une procédure. La règle contextuelle contient deux applications possibles (séparées par le symbole "|") : les mots détectés par la macro *&modes* (un ensemble de règles contextuelles) et suivis éventuellement de *load* sont annotés comme étant un *_mode* ; il en est de même pour le mot *release*, s'il est précédé d'un adjectif. L'application du lexique se fait en appelant la macro qui inclut le lexique de procédures.

Catégorisation	_procedure : [A-z]+ "omy" | [A-z]+ "asty" ... ;
Règle contextuelle	_mode : (&modes load ? | (<= _JJ| _VVD| _VVN) release) ;
Application de lexique	_procedure : (&procedure) ;

TAB. 8 – Exemples de règles.

Ces différentes informations ont constitué l'ensemble des traits qui ont alimenté l'apprentissage du modèle CRF. Ce modèle et les modules d'extraction de traits forment le système de base pour cette campagne d'évaluation.

Enfin, nous avons ajouté en sortie de ce système une phase de correction par l'ajout de règles de post-traitement. Nous avons supposé que l'hypothèse « *un sens par corpus* » (Fung, 1998) est vérifiée dans une langue de spécialité, à plus forte raison dans le typage de concepts médicaux : nous avons examiné les expressions étiquetées par des types de concepts différents dans le corpus et avons normalisé leur étiquette au type observé le plus fréquent (un token ayant pour trait la catégorie *médicament* qui n'aurait pas été typé ou l'aurait été typé différemment de *traitement* est corrigé avec le type *traitement*).

4.4 Résultats et discussion

Le tableau 9 présente les résultats obtenus par notre système sur l'identification et le typage des concepts médicaux. L'évaluation a été réalisée sur 477 documents. Les chiffres renseignés dans ce tableau reposent sur un appariement à l'identique des concepts ; les erreurs de frontière ont donc été pénalisantes.

	Nombre	**Rappel**	**Précision**	**F_1-mesure**
Global	27 837	0.726	0.826	0.773
Problème	11 968	0.742	0.799	0.769
Traitement	8 500	0.723	0.843	0.778
Examen	7 369	0.705	0.851	0.771

TAB. 9 – Résultats obtenus par notre système au défi i2b2 2010 (recouvrement exact).

Le système d'identification et de typage des concepts médicaux obtient une F_1-mesure générale de 0,773 (notre système s'est classé 12ème sur 22 participants internationaux). Pour cette tâche d'extraction de concepts médicaux, notre système obtient de nouveau une précision supérieure au rappel pour chaque type de concept. Nous notons que les performances du système se révèlent équivalentes sur les trois types de concepts médicaux à traiter, cette observation s'expliquant par la répartition équilibrée des concepts dans ces trois catégories. Les dix meilleurs systèmes du défi ont tous employé des méthodes d'apprentissage. Le meilleur système (De Bruijn *et al.*, 2010) a modélisé la tâche avec un CRF et s'en est servi pour définir les traits d'un modèle semi-markovien caché. Plusieurs autres systèmes bien classés ont utilisé comme traits le résultat de systèmes de reconnaissance d'entités médicales.

5 Conclusion

Dans le cadre de ce travail, nous avons constitué un ensemble de ressources nécessaires au traitement de la langue médicale. Nous avons ainsi dressé un inventaire exhaustif des noms de médicaments (génériques, marques et classes thérapeutiques) et créé des lexiques d'abréviations et de symptômes. Nous avons par ailleurs élaboré une méthodologie de détection des types d'entités de différentes sortes (par l'application d'expressions régulières et l'utilisation d'un lexique d'abréviations spécifiques) et de gestion de la factorisation d'information (coordination et duplication). Enfin, nous avons étudié les caractéristiques linguistiques à utiliser pour la construction de modèles d'apprentissage dédiés au traitement des concepts médicaux.

En Traitement Automatique des Langues, les systèmes à base de règles constituent une solution pertinente pour traiter des corpus non annotés porteurs d'informations stables syntaxiquement. L'application de patrons syntaxiques permet effectivement d'obtenir rapidement de bons résultats comme en témoignent ceux que nous avons obtenus sur l'extraction d'information dans les prescriptions médicales lors de l'édition 2009 du défi i2b2.

En revanche, la variation syntaxique des informations à extraire se révèle beaucoup plus difficile à traiter. L'utilisation seule de règles syntaxiques conduit à un manque de robustesse du système et doit faire l'objet d'une application complémentaire d'autres types de méthodes. À cet effet, l'utilisation de méthodes hybrides rassemblant un apprentissage supervisé et des informations linguistiques permet d'accroître les chances de traiter correctement ce type de données. C'est l'approche que nous avons suivie pour l'identification et le typage des concepts médicaux pour l'édition 2010 du défi i2b2 ; dans le cas présent, nous nous sommes servis d'informations d'ordre linguistique à deux reprises : en premier lieu pour constituer des traits sur chaque token de manière à construire un modèle pour l'apprentissage, puis dans un second temps, comme moyen d'affiner les résultats produits par l'application du modèle précédemment construit.

Dans le domaine médical, la langue de spécialité utilisée revêt un caractère particulièrement stable et formel, tant sur les plans syntaxique que sémantique, voire structurel. Ces caractéristiques nous autorisent à utiliser des approches hybrides lorsqu'existent des corpus annotés. Lorsque les annotations font défaut, les caractéristiques linguistiques de la langue médicale nous permettent néanmoins de travailler uniquement à base de patrons syntaxiques. Ces méthodes montrent leurs limites lorsque l'information à extraire se trouve rédigée en texte plus libre, à l'instar des raisons qui justifient une prescription médicale. Dans cette perspective, des traitements linguistiques plus complexes faisant intervenir une analyse en dépendances pourraient constituer une alternative intéressante.

Remerciements

Ce travail a été partiellement réalisé dans le cadre des projets Akenaton (ANR-07-TecSan-001) et Quæro (financement Oseo, agence française pour l'innovation et la recherche).

Les données médicales utilisées proviennent du consortium Informatics for Integrating Biology to the Bedside (i2b2) grâce aux financements numéros U54LM008748 de la National Library of Medicine, VA HSR HIR 08-374 du Consortium for Healthcare Informatics Research (CHIR), et VA HSR HIR 08-204 du VA Informatics and Computing Infrastructure (VINCI).

Références

BERMAN J. J. (2004). Pathology Abbreviated : A Long Review of Short Terms. *Archives of Pathology & Laboratory Medicine*, **128**(3), 347–352.

CLAVEAU V. & KIJAK E. (2010). Analyse morphologique en terminologie biomédicale par alignement et apprentissage non-supervisé. In *Actes de TALN 2010*.

DE BRUIJN B., CHERRY C., KIRITCHENKO S., MARTIN J. & ZHU X. (2010). MRC at i2b2 : one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. In *Proc. of i2b2/VA 2010*.

DELÉGER L., GROUIN C. & ZWEIGENBAUM P. (2010). Extracting Medical Information from Narrative Patient Records : the Case of Medication-related Information. *J Am Med Inform Assoc*, **17**(5), 555–558.

DOAN S. & XU H. (2010). Recognizing Medication related Entities in Hospital Discharge Summaries using Support Vector Machine. In *Coling2010 : Poster Volume*, p. 259–266.

FRIEDMAN C. (2000). A broad-coverage natural language processing system. In *AMIA Annu Symp Proc*, p. 270–274.

FUNG P. (1998). A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora. In *AMTA*, p. 1–17.

GALIBERT O. (2009). *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. PhD thesis, Université Paris-Sud 11, Orsay, France.

HAMON T. & GRABAR N. (2010). Linguistic approach for identification of medication names and related information in clinical narratives. *J Am Med Inform Assoc*, **17**(5), 549–554.

KUDO T. (2007). CRF++. http ://crfpp.sourceforge.net/.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional Random Fields : Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, p. 282–289.

LI D., KIPPER-SCHULER K. & SAVOVA G. (2008). Conditionnal Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts. In *BioNLP2008 : Current Trends in Biomedical Natural Language Processing*, p. 94–95.

LINDBERG D., HUMPHREYS B. & MCCRAY A. (1993). The Unified Medical Language System. *Meth Inform Med*, **32**(4), 281–291.

LONG W. (2007). Lessons Extracting Diseases from Discharge Summaries. In *AMIA Annu Symp Proc*, p. 478–482.

MINARD A.-L., LIGOZAT A.-L., BEN ABACHA A., BERNHARD D., CARTONI B., DELÉGER L., GRAU B., ROSSET S., ZWEIGENBAUM P. & GROUIN C. (2011). Hybrid Methods for improving Information Access in Clinical Documents : Concept, Assertion, and Relation Identification. *J Am Med Inform Assoc*. À paraître.

ROSSET S., GALIBERT O., BERNARD G., BILINSKI E. & ADDA G. (2008). The LIMSI participation to the QAst track. In *Working Notes of CLEF 2008 Workshop*, Aarhus, Danemark.

SAGER N. (1981). *Natural Language Processing : A Computer Grammar of English and Its Applications*. Addison Wesley.

SAGER N., LYMAN M., NHÀN N. & TICK L. (1994). Automatic Encoding into SNOMED III : A Preliminary Investigation. In *Proc. of the 18th Annual Symposium on Computer Applications in Medical Care*, p. 230–234.

SAGER N., LYMAN M., NHÀN N.-T. & TICK L. J. (1995). Medical language processing : applications to Patient Data Representation and Automatic Encoding. *Meth Inform Med*, **34**(1–2), 140–146.

SAGER N. & NHÀN N.-T. (2002). The Computability of strings, transformations, and sublanguage. In B. E. NEVIN & S. M. JOHNSON, Eds., *The legacy of Zellig Harris – Language and information into the 21st century - volume 2 : computability of language and computer applications*, volume 2, chapter 4, p. 79–120. Amsterdam/Philadelphia : John Benjamins Publishing Company.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. of the International Conference on New Methods in Language Processing*, p. 44–49.

TIKK D. & SOLT I. (2010). Improving textual medication extraction using combined conditional random fields and rule-based systems. *J Am Med Inform Assoc*, **17**(5), 540–544.

UZUNER O., SOLTI I. & CADAG E. (2010a). Extracting medication information from clinical text. *J Am Med Inform Assoc*, **17**(5), 514–518.

UZUNER O., SOLTI I., XIA F. & CADAG E. (2010b). Community annotation experiment for ground truth generation for the i2b2 medication challenge. *J Am Med Inform Assoc*, **17**(5), 519–523.

WANG Y. (2009). Annotating and Recognising Named Entities in Clinical Notes. In *Proc. of the ACL-IJCNLP 2009 Student Research Workshop*, p. 18–26, Singapore.

ZIDOUNI A., ROSSET S. & GLOTIN H. (2010). Efficient Combined Approach for Named Entity Recognition in Spoken Language. In *Proc. of InterSpeech*, Makuhari, Japon.