Une étude comparative empirique sur la reconnaissance des entités médicales

Asma Ben Abacha — Pierre Zweigenbaum

LIMSI-CNRS, BP 133 91403 Orsay cedex
abacha@limsi.fr, pz@limsi.fr

RÉSUMÉ. *De nombreux travaux se sont attaqués à la reconnaissance des entités médicales à partir de textes. Cependant il n'y a pas eu, à notre connaissance, d'études comparant deux stratégies pour traiter cette tâche : (i) l'extraction en amont des syntagmes nominaux, suivie d'une étape de catégorisation de leur type et (ii) la détermination simultanée des frontières et des types des entités. C'est la question que nous nous posons ici. Nous testons ces deux stratégies en utilisant des méthodes à base de règles et/ou à base d'apprentissage. Nous comparons leur robustesse et aussi leur portabilité en les évaluant sur deux corpus médicaux standard de genres différents. Les résultats obtenus confirment que les méthodes statistiques sont plus robustes que celles à base de règles à condition qu'un nombre suffisant d'exemples soit disponible. À cette contrainte s'ajoute le manque de portabilité des méthodes à base d'apprentissage sur des corpus différents. Les méthodes hybrides combinant les aspects sémantiques et statistiques permettent d'améliorer davantage les performances obtenues par apprentissage.*

ABSTRACT. *Several research efforts tackled medical entity recognition from texts. However, to our knowledge, there are no comparative studies for the following approaches: (i) the extraction of noun phrases in an independent step before the final categorization step and (ii) identifying simultaneously entity boundaries and categories. In this paper, we focus on these approaches by experimenting with different methods based on rules and/or machine learning techniques. We compare their performance and evaluate their scalability on two standard medical corpora. The results confirm that machine learning methods are more robust than rule-based ones provided that a sufficient number of examples is available. They also point out the lack of scalability of such methods on corpora of different genres. Hybrid methods combining statistical and semantic techniques allow improving the performance obtained by machine learning.*

MOTS-CLÉS : *reconnaissance des entités médicales, extraction d'information, apprentissage.*

KEYWORDS: *medical entity recognition, information extraction, machine learning.*

1. Introduction

Dans cette section, nous présentons les motivations, la problématique ainsi que les objectifs de notre travail puis exposons le plan de l'article.

1.1. *Motivations*

Avec le volume important des connaissances médicales numérisées à large échelle, retrouver automatiquement une information de haute précision est devenu un défi. Le volume des connaissances médicales double tous les 5 ans (Engelbrecht, 1997), voire tous les 2 ans (Hotvedt, 1996). Avec la numérisation à large échelle, plusieurs moteurs de recherche spécialisés dans ce domaine ont vu le jour (*e.g.* PubMed[1]). À une requête donnée, ces moteurs retournent un ensemble de documents et délèguent à l'utilisateur la tâche de trouver l'information cherchée, si elle existe, dans les documents retournés. Pour faciliter la recherche d'information, des systèmes plus précis sont mis en œuvre, comme les systèmes de questions-réponses (Terol *et al.*, 2007 ; Embarek et Ferret, 2010). Ces outils procèdent à une analyse profonde des textes médicaux pour trouver l'information demandée. Cette analyse passe obligatoirement par une étape de reconnaissance des entités médicales présentes dans le texte. En effet, la reconnaissance des entités médicales est utilisée au niveau de l'analyse de la question pour déterminer les mots-clés et le type de la réponse attendue et aussi au niveau de la recherche des réponses possibles. La reconnaissance des entités médicales est une tâche importante voire nécessaire non seulement pour la recherche d'information et les systèmes de questions-réponses mais aussi pour d'autres tâches comme l'extraction de relations sémantiques entre entités médicales (Vintar *et al.*, 2003), la détermination de la factivité des problèmes médicaux (Bernhard et Ligozat, 2011) et la résolution de coréférence (*e.g.* le challenge i2b2/VA 2011[2]).

1.2. *Problématique*

La reconnaissance des entités médicales est une tâche complexe. Cette complexité réside dans les problèmes classiques rencontrés en domaine ouvert, mais aussi dans les spécificités du domaine médical. En effet, en domaine ouvert, les entités nommées désignent habituellement les noms de personnes, de lieux, d'entreprises, ainsi que les dates et les quantités monétaires. Mais d'autres catégories plus ou moins précises peuvent aussi être incluses (les événements, les fonctions, etc.), ce qui pose la question de la définition de cette tâche. Sa définition dépend-elle du domaine et/ou de l'application ? Ce problème a motivé plusieurs travaux qui se sont intéressés à la définition de la tâche de reconnaissance des entités nommées (Ehrmann, 2008). Le même problème se pose en domaine médical, à savoir quelle est la liste des catégories mé-

1. `http://www.pubmed.com`
2. `https://www.i2b2.org/NLP/Coreference`

dicales (*Traitement*, *Examen*, *Problème médical*, etc.) et quelle est la définition exacte de chaque catégorie (*e.g.* les plantes doivent-elles être rangées parmi les traitements ?).

La polysémie est un deuxième obstacle à la reconnaissance des entités médicales que l'on retrouve en domaine ouvert. Par exemple le mot *drug* a deux sens (médicament ou drogue). À ceci s'ajoute le cas des mots non spécialisés qui, hors contexte, peuvent porter un sens médical particulier. Par exemple le mot *ten* (dix) désigne aussi la maladie *Toxic Epidermal Necrolysis* (le *syndrome de Lyell*). La réciproque est aussi vraie : certains termes médicaux, hors contexte, peuvent prendre d'autres sens. Par exemple, le terme médical *Antimicrobial agents* est aussi le nom d'un journal et d'un site Web.

Les spécificités du domaine médical ajoutent aussi une couche de complexité aux défis communs soulevés en domaine ouvert. Ci-dessous, nous citons quelques points clés liés à la reconnaissance des entités médicales :

– la grande variation terminologique dans ce domaine de spécialité : chaque concept peut être désigné par plusieurs termes synonymes, des abréviations, etc. Par exemple, *Diabetes mellitus type 1*, *Type 1 diabetes*, *IDDM*, ou *juvenile diabetes* désignent le même concept. Il en est de même pour *Papanicolaou test*, *Pap smear*, *Pap test*, *cervical smear* ou encore *cervix smear* qui désignent le même examen médical ;

– certaines entités médicales peuvent avoir des noms différents selon les pays, c'est surtout le cas pour certaines maladies et certains médicaments. Par exemple, le médicament *Procarbazine* possède aussi les noms commerciaux suivants : *Matulane* (US), *Natulan* (Canada), *Indicarb* (Inde) ;

– l'évolution rapide de la terminologie médicale (*e.g.* nouvelles abréviations, noms de nouveaux médicaments ou maladies).

Ces obstacles limitent la généricité des méthodes qui se fondent sur les dictionnaires et les listes d'entités nommées (« gazetteers »). D'un autre côté, le domaine médical dispose de ressources spécialisées intéressantes (*e.g.* UMLS, voir plus bas). Cependant, ces ressources manquent parfois de précision et doivent être mises à jour d'une façon continue et rapide (ce qui n'est pas toujours le cas). Ceci conduit à l'adoption d'autres méthodes utilisant potentiellement des connaissances du domaine, mais aussi exploitant les outils du TAL et les techniques connues en domaine ouvert telles que l'apprentissage.

1.3. *Objectifs*

Dans cet article, nous étudions quatre méthodes pour la reconnaissance d'entités médicales en anglais : (i) une méthode à base de règles qui s'appuie sur l'outil de référence MetaMap (Aronson, 2001), (ii) une méthode qui extrait les syntagmes nominaux (avec un extracteur robuste, ou « *chunker* ») puis détecte les entités médicales parmi ces syntagmes par apprentissage supervisé (classifieur SVM), (iii) une méthode qui utilise l'apprentissage supervisé pour déterminer les frontières et les types des entités

médicales avec un classifieur CRF et l'encodage B-I-O[3] et (iv) une variante hybride de cette dernière qui combine une méthode statistique et une méthode à base de règles. Les expérimentations regroupées ici ont été présentées en partie dans (Ben Abacha et Zweigenbaum, 2011). Dans cet article nous les détaillons davantage et nous faisons une synthèse des différentes méthodes proposées en comparant leur performance sur les mêmes corpus standard. Avec ces méthodes nous étudions en particulier deux stratégies différentes : (i) l'extraction en amont des syntagmes nominaux avec des outils spécialisés, suivie d'une étape de catégorisation et (ii) l'exploitation des techniques d'apprentissage pour déterminer simultanément les frontières et les catégories des entités médicales.

Nous présentons aussi une étude comparative de la performance de trois outils pour l'extraction de syntagmes nominaux à partir de textes médicaux : TreeTagger-chunker, OpenNLP et MetaMap. Cette comparaison a mené au choix de TreeTagger-chunker, utilisé dans les deux premières méthodes de reconnaissance d'entités médicales. Les différentes approches ont été expérimentées sur deux corpus différents : (i) un corpus de textes cliniques : le corpus du challenge international i2b2/VA 2010 et (ii) un corpus de résumés d'articles scientifiques extrait de MEDLINE : le corpus de Berkeley (Rosario et Hearst, 2004).

1.4. *Plan de l'article*

Nous commençons cet article par un état de l'art sur les différentes techniques utilisées pour la reconnaissance d'entités nommées en domaine ouvert et en domaine médical (section 2). Nous consacrons la section 3 à la présentation de la tâche de reconnaissance des entités médicales. Nous présentons dans la section 4 les méthodes proposées. Les résultats expérimentaux obtenus sur le corpus i2b2 de textes cliniques sont présentés dans la section 5.1 et ceux obtenus sur le corpus de Berkeley dans la section 5.2. Nous dédions la section 6 à une synthèse des différentes méthodes et directions étudiées et des résultats obtenus. Enfin nous concluons et donnons quelques perspectives (section 7).

2. État de l'art

La reconnaissance des entités nommées, introduite officiellement à la campagne d'évaluation MUC-6 (Grishman et Sundheim, 1996), est étudiée depuis une vingtaine d'années en domaine ouvert. On peut distinguer trois types d'approches : (i) les approches linguistiques qui utilisent des listes d'entités nommées et des patrons de reconnaissance écrits manuellement (Poibeau, 1999 ; Elkateb-Gara, 2003), (ii) les approches statistiques qui se fondent sur des techniques d'apprentissage à partir de textes

3. Le format B-I-O permet de trouver la catégorie ainsi que les frontières des entités en indiquant pour chaque mot s'il correspond au début, à l'intérieur ou l'extérieur de l'entité.

annotés (McCallum et Li, 2003 ; Raymond et Wei, 2006) et (iii) les approches hybrides qui intègrent les deux premières méthodes (Kosseim et Poibeau, 2001 ; Fourour, 2002).

Depuis plus d'une dizaine d'années, des travaux se sont intéressés à la reconnaissance des entités nommées en domaine de spécialité. Dans le domaine biomédical Rindflesch *et al.* (2000) ont développé le système EDGAR qui extrait les informations concernant les médicaments et les gènes liés au cancer à partir de la littérature biomédicale de la base MEDLINE. Le système se fonde sur le Metathesaurus et les connaissances lexicales de l'UMLS (*Unified Medical Language System*) (Bodenreider, 2006 ; Zweigenbaum, 2004). Outre la détection de noms de gènes, la détection des noms de protéines a fait l'objet de plusieurs travaux (Liang et Shih, 2005 ; Wang, 2007). Embarek et Ferret (2008) ont utilisé une approche à base de patrons linguistiques et d'entités médicales canoniques pour la reconnaissance de termes médicaux de cinq types. Une autre famille de travaux utilise des outils comme MetaMap (Aronson, 2001) qui permettent de reconnaître et de catégoriser les termes médicaux. MetaMap est un outil développé par la NLM (U.S. National Library of Medecine) pour reconnaître les termes médicaux qui désignent des concepts de l'UMLS. MetaMap identifie la plupart des concepts présents dans les titres des articles de la base MEDLINE (Pratt et Yetisgen-Yildiz, 2003). Il a été par exemple utilisé par Shadow et MacDonald (2003) pour extraire des entités médicales à partir de rapports de pathologistes, ces entités pouvant avoir 20 types sémantiques possibles (choisis parmi ceux de l'UMLS). Meystre et Haug (2005) ont pu obtenir 89,2 % de rappel et 75,3 % de précision avec une approche qui se fonde sur MetaMap et l'algorithme NegEx de détection de négation (Chapman *et al.*, 2001) pour l'extraction de « problèmes médicaux » (signes, symptômes, diagnostics).

Le domaine médical dispose de plusieurs ressources sémantiques structurées comme le Metathesaurus et le réseau sémantique de l'UMLS. L'UMLS est organisé en trois parties (i) le Specialist Lexicon, lexique anglais incluant les termes du domaine ainsi que leurs variations syntaxiques et morphologiques, (ii) le Metathesaurus, vocabulaire de plus de deux millions de concepts (un concept « regroupe » des termes synonymes, acronymes et variantes terminologiques) et (iii) le réseau sémantique qui organise les concepts en 135 « types sémantiques » et définit 54 relations entre ces types.

À l'opposé des approches linguistiques qui requièrent plus de connaissances du domaine pour la construction des règles ou patrons, les approches statistiques sont plus robustes et sont souvent mises en avant pour leur portabilité et leur potentiel de passage à l'échelle si suffisamment d'exemples représentatifs sont disponibles. Plusieurs travaux ont utilisé des classifieurs comme les arbres de décision ou les SVM (Isozaki et Kazawa, 2002). D'autres se sont basés sur des modèles de Markov comme le modèle de Markov caché ou encore les CRF (He et Kayaalp, 2008). La performance des approches fondées sur ces algorithmes supervisés est dépendante de la présence d'un corpus d'entraînement bien annoté et de la conception d'un ensemble pertinent d'attributs.

Les approches hybrides tentent de cumuler les avantages des méthodes à base de règles et des méthodes statistiques tout en éliminant certains de leurs inconvénients (*e.g.* problème de passage à l'échelle des méthodes à base de règles, performances réduites pour les méthodes statistiques avec des corpus d'entraînement de taille réduite). Proux *et al.* (1998) ont construit un système pour la détection des symboles et noms de gènes à partir de textes biomédicaux. Le système traite les mots inconnus avec des règles lexicales afin d'obtenir des catégories candidates qui sont ensuite désambiguïsées en utilisant un modèle de Markov. Liang et Shih (2005) utilisent à la fois des règles empiriques et une approche statistique pour la reconnaissance des noms de protéines.

Il est aussi important de noter que différents genres de corpus existent dans le domaine biomédical (Zweigenbaum *et al.*, 2001). Parmi les plus récurrents nous pouvons citer les textes cliniques et les articles scientifiques (Friedman *et al.*, 2002). La première catégorie de corpus a intéressé plusieurs travaux (Sager *et al.*, 1995 ; Meystre *et al.*, 2008) mais aussi des challenges internationaux comme i2b2 2010 (Uzuner *et al.*, 2011). Les articles scientifiques du domaine biomédical ont aussi fait l'objet de différents travaux (Rindflesch *et al.*, 2000), en particulier depuis plus de dix ans en génomique (*e.g.* le challenge BioCreAtIvE (Yeh *et al.*, 2005)).

3. Reconnaissance des entités médicales : définitions et principes

Nous consacrons cette section à l'introduction de la problématique de reconnaissance des entités médicales, en répondant aux questions suivantes : (i) Qu'est-ce qu'une entité médicale ? (ii) En quoi consiste le processus de reconnaissance d'une entité médicale ?

3.1. *Entités médicales*

En domaine ouvert, les entités nommées désignent classiquement les noms propres, à savoir les noms de personnes, d'organisations et de lieux (Grishman et Sundheim, 1996), mais aussi les dates et les quantités. Certains vont plus loin et cherchent à annoter plus finement leurs corpus, par exemple (Sekine, 2004), ou encore (Ehrmann et Jacquet, 2006) qui considèrent la reconnaissance d'entités de catégories plus précises (*e.g. professeur* ou *président* plutôt que simplement *personne*).

Nous désignons par « entité médicale » une instance d'un concept médical générique comme « Maladie » (*e.g. l'Alzheimer*) ou « Examen » (*e.g. la laryngoscopie*). Dans le domaine médical, Khelif et Dieng-Kuntz (2004) ont utilisé le réseau sémantique de l'UMLS comme une ontologie du domaine biomédical et les termes du Metathesaurus comme désignant des instances possibles de concepts biomédicaux. De même, Delbecque *et al.* (2005) ont considéré les types sémantiques du réseau sémantique de l'UMLS comme types d'entités nommées spécifiques au domaine médical (*e.g. Drug*, ou encore *Therapeutic or Preventive Procedure*).

Plus récemment, Embarek et Ferret (2008) se sont intéressés à la reconnaissance de termes médicaux de cinq types (*Maladie*, *Traitement*, *Médicament*, *Examen* et *Symptôme*).

Nos choix : dans cet article nous travaillons sur les trois catégories les plus importantes dans le domaine médical, à savoir *Problème* (signes, symptômes, diagnostics, etc.), *Traitement* (y compris médicaments et matériel médical) et *Test* (examens) qui sont aussi les catégories ciblées dans la tâche de reconnaissance d'entités médicales dans le challenge international i2b2 2010 (Uzuner *et al.*, 2011). Le tableau 1 présente les types sémantiques de l'UMLS correspondant aux catégories médicales traitées.

Catégories médicales	**Types sémantiques correspondants dans l'UMLS**
Problème	Virus, Bacterium, Anatomical Abnormality, Congenital Abnormality, Acquired Abnormality, Sign or Symptom, Pathologic Function, Disease or Syndrome, Mental or Behavioral Dysfunction, Neoplastic Process, Cell or Molecular Dysfunction, Injury or Poisoning, Sign or Symptom
Traitement	Medical Device, Drug Delivery Device, Clinical Drug, Steroid, Pharmacologic Substance, Antibiotic, Biomedical or Dental Material, Therapeutic or Preventive Procedure
Test	Laboratory Procedure, Diagnostic Procedure

Tableau 1. *Catégories médicales traitées et types sémantiques correspondants dans l'UMLS*

3.2. *La reconnaissance des entités médicales : principes*

La reconnaissance des entités médicales consiste en (i) un repérage des termes médicaux dans les textes (*e.g. beta cell replacement*, *pyogenic liver abscess*, *infection of biliary system*) et (ii) l'identification de la catégorie sémantique des termes repérés. L'exemple 1 montre les résultats de la reconnaissance d'entités médicales dans une phrase extraite d'un texte clinique. Les termes médicaux y sont marqués par les étiquettes *<Treatment>* et *<Disease>*.

(1) *<Treatment> Adrenal-sparing surgery </Treatment> is safe and effective , and may become the treatment of choice in patients with <Disease> hereditary phaeochromocytoma </Disease>.*

Ces deux étapes amènent à effectuer des choix sur (i) les catégories médicales à traiter (cf. section 3.1) et (ii) les règles de délimitation des frontières des entités médicales dans le texte, telles que :

– inclure ou non les articles (*e.g.* ***The*** *West Nile Virus* ou juste *West Nile Virus*) et les possessifs (*e.g.* ***his*** *cancer therapy* ou *cancer therapy*) ;

– inclure ou non les adjectifs (*e.g.* ***recurrent*** *or* ***persistent*** *angina*), sachant qu'il est parfois important de conserver les adjectifs pour déterminer correctement l'entité médicale comme par exemple ***Severe Acute Respiratory*** *Syndrome (SARS)* ;

– inclure ou non les adverbes (*e.g.* ***all*** *drugs*) ;

– inclure ou non les pourcentages, les chiffres (*e.g.* ***30*** *cancers*) et les doses des médicaments (*e.g. clamoxyl* ***1g***) ;

– annoter les abréviations qui suivent les entités médicales (exp : *Hepatitis A Virus* ***(HAV)***) les deux ensemble ou chacune à part ;

– annoter ou non une entité médicale à l'intérieur d'une autre (*e.g. The BC* ***Cancer*** *Agency*, *Canadian Network For* ***Asthma*** *Care*) (voir par exemple les entités imbriquées de Grouin *et al.* (2011)).

Nos choix : dans ce travail, nous avons effectué les choix suivants : (i) inclure dans les entités médicales : les possessifs, les adjectifs, les adverbes ainsi que les chiffres, (ii) annoter les abréviations séparément et (iii) ne pas annoter une entité médicale qui fait partie d'une autre entité. Ces principes rejoignent les règles qui ont été fixées pour l'annotation du corpus i2b2 (décrit dans la section 5.1.1).

4. Méthodes proposées pour reconnaître les entités médicales

Dans cette section nous présentons deux stratégies différentes pour la reconnaissance des entités médicales. Nous présentons aussi les méthodes proposées au niveau de chaque stratégie.

4.1. *Présentation générale*

Comme annoncé plus haut, ce travail a principalement deux objectifs. Le premier est d'étudier et de comparer deux stratégies différentes pour traiter le problème de reconnaissance des entités : (i) une première reposant sur l'extraction en amont des syntagmes nominaux avec des outils spécialisés, suivie d'une étape de catégorisation et (ii) une deuxième exploitant des techniques d'apprentissage pour déterminer simultanément les frontières et les catégories des entités médicales. La figure 1 illustre avec un exemple ces deux stratégies.

Notre second objectif consiste à tester différentes méthodes fondées d'une part sur des règles utilisant les connaissances du domaine (fournies par l'UMLS) et d'autre part sur des techniques d'apprentissage reposant sur deux classifieurs différents. La figure 2 présente les quatre méthodes proposées : une méthode à base de règles, deux méthodes différentes à base d'apprentissage supervisé et une méthode hybride. Nous détaillons ces méthodes dans le reste de cette section.

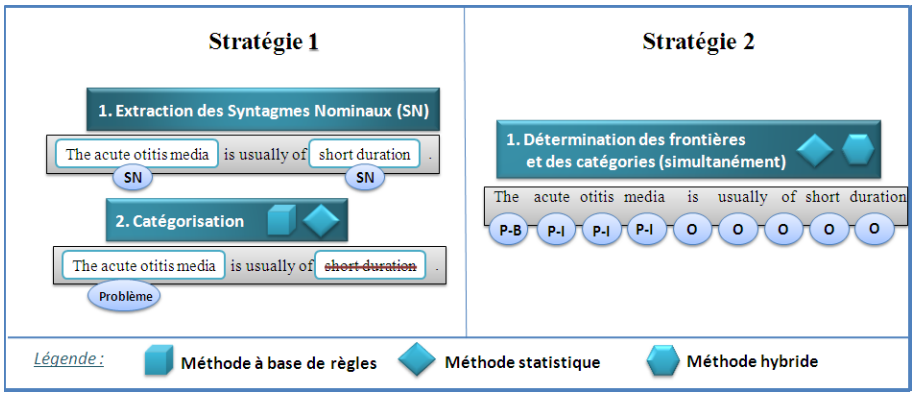

Figure 1. *Reconnaissance des entités médicales : stratégies, étapes et méthodes*

Etapes	Stratégie 1		Stratégie 2	
	Méthode 1 (MetaMapPlus)	**Méthode 2 (TT-SVM)**	**Méthode 3 (CRF-BIO)**	**Méthode 4 (CRF-BIO-H)**
(1) Identification des frontières	Extraction des syntagmes nominaux avec un chunker	Extraction des syntagmes nominaux avec un chunker	▪ Apprentissage supervisé avec un classifieur *CRF* ▪ Utilisation de l'encodage *BIO* ▪ Classification des *mots* ▪ Nbr de classes = *2N+1*	▪ Combinaison des 2 méthodes statistique (CRF-BIO) et à base de règles ▪ Ajout des résultats de MetaMapPlus aux attributs du classifieur CRF
***(2)* Catégorisation** (en N catégories)	▪ Méthode à base de *règles* ▪ Classification des *syntagmes nominaux* ▪ Nbr de classes = *N+1*	▪ Apprentissage supervisé avec un classifieur *SVM* ▪ Classification des *syntagmes nominaux* ▪ Nbr de classes = *N+1*		

Figure 2. *Méthodes proposées pour la reconnaissance des entités médicales*

4.2. *Stratégie 1 : deux étapes, frontières puis catégories*

Les deux premières méthodes que nous proposons utilisent un *chunker* pour l'extraction des syntagmes nominaux. Ces syntagmes seront les cibles pour une catégorisation médicale effectuée dans une seconde étape : si un syntagme est associé à une catégorie médicale il sera considéré comme étant une entité médicale. Les performances des méthodes qui utilisent cette stratégie pour l'identification des entités médicales dépendront de la qualité de l'extraction des syntagmes nominaux et donc de la performance du *chunker* choisi.

4.2.1. *Extraction des syntagmes nominaux*

Malgré l'importance de cette tâche pour l'extraction d'information, il n'y a pas eu beaucoup d'études comparatives des outils disponibles dans le domaine médical. Une étude comparative récente (Kang *et al.*, 2010) a comparé des outils de segmentation en phrases et en syntagmes nominaux ainsi que l'étiquetage morphosyntaxique pour le domaine biomédical. Les auteurs ont comparé 6 outils fréquemment utilisés en domaine ouvert ou médical : GATE Chunker, Genia Tagger, Lingpipe, MetaMap, OpenNLP et Yamcha. Les résultats de cette étude montrent que OpenNLP a la meilleure performance : F-mesure de 89,9 % et 95,5 % respectivement pour l'extraction des syntagmes nominaux et verbaux. Il est important de noter ici que MetaMap est un outil spécialisé pour le domaine médical qui n'est pas dédié à l'extraction des syntagmes nominaux mais à la détection des entités médicales, bien que cette détection passe par une étape d'extraction de syntagmes nominaux. Comparer MetaMap à des *chunkers* consiste donc uniquement à comparer sa composante de « *chunking* » avec ces outils.

Afin de choisir le *chunker* à utiliser au sein de nos méthodes d'extraction d'entités médicales nous comparons le meilleur outil évalué par Kang *et al.* (2010) (OpenNLP[4]) à TreeTagger-chunker[5] (un outil performant en domaine ouvert) et au module de *chunking* de MetaMap. Nous dissocions aussi un syntagme nominal en plusieurs groupes nominaux s'il contient des conjonctions ou des disjonctions.

Nous évaluons ces outils sur nos deux corpus (décrits dans les sections 5.1.1 et 5.2.1). Nous considérons qu'un groupe nominal est extrait correctement s'il correspond exactement à une entité médicale annotée (bien que ces groupes ne soient pas encore catégorisés à ce niveau). Nous calculons uniquement la valeur du rappel étant donné que les groupes nominaux retrouvés comportent aussi beaucoup de groupes nominaux non médicaux (non pertinents pour connaître la performance des différents outils pour le domaine médical).

	Corpus de textes cliniques (i2b2)			**Corpus d'articles scientifiques (Berkeley)**		
	MetaMap	**TreeTagger**	**OpenNLP**	**MetaMap**	**TreeTagger**	**OpenNLP**
E1	58 115	58 115	58 115	3 371	3 371	3 371
E2	6 532	35 314	26 862	151	2 106	1 874
E3	212 227	129 912	122 131	22 334	19 796	18 850
R	11,24 %	**60,76 %**	46,22 %	4,48 %	**62,47 %**	55,59 %

Tableau 2. *Évaluation du rappel de trois chunkers (E1 = entités de référence, E2 = entités correctes, E3 = entités trouvées, R = rappel = E2/E1)*

Suivant les résultats obtenus et présentés dans le tableau 2, TreeTagger-chunker a obtenu les meilleurs résultats et a été choisi pour l'extraction des syntagmes nominaux dans les deux premières méthodes (cf. sections 4.2.2 et 4.2.3).

4. `http://incubator.apache.org/opennlp`

5. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`

4.2.2. *Catégorisation des syntagmes nominaux : méthode à base de règles (MetaMapPlus)*

En domaine ouvert, ce type de méthode consiste à utiliser des règles écrites manuellement qui exploitent potentiellement des listes de noms (personnes, organisations, etc). Pour le domaine médical, ce genre d'informations est disponible grâce à des bases de connaissances telles que l'UMLS.

Plusieurs outils se sont intéressés à l'extraction des entités médicales. Un des outils les plus largement utilisés pour cette tâche est MetaMap.

L'outil MetaMap (Aronson, 2001) permet de segmenter les textes médicaux en phrases et syntagmes nominaux qui correspondent à des termes médicaux. Il identifie les entités médicales et leurs catégories en utilisant le Metathesaurus et le réseau sémantique de l'UMLS et fournit potentiellement plusieurs catégories candidates aux entités qu'il retrouve avec des scores de confiance. Plus précisément, ces catégories sont les types du réseau sémantique UMLS jugés comme étant pertinents pour l'entité retrouvée. Le tableau 3 montre un exemple de sortie de MetaMap sur une phrase.

Cependant, l'étude de l'utilisation simple de MetaMap a révélé qu'il présente certains problèmes résiduels, principalement à trois niveaux : (i) la segmentation en syntagmes nominaux n'est pas toujours pertinente et n'est pas du même niveau de performance que d'autres outils connus en TAL, (ii) la détection des entités médicales n'est pas toujours performante car MetaMap considère certains mots généraux et certains verbes comme des termes du domaine et (iii) la catégorisation des entités médicales peut rester ambiguë, car MetaMap peut proposer plusieurs concepts pour un même terme ainsi que plusieurs types sémantiques pour un même concept ; donc plusieurs combinaisons « terme-concept-type » sont possibles (cf. tableau 3).

Afin de pallier ces problèmes, nous proposons une méthode, nommée MetaMapPlus, qui comprend quatre étapes :

1) extraire les syntagmes nominaux avec un *chunker*. Nous utilisons TreeTagger-chunker qui offre une meilleure segmentation en syntagmes nominaux et permet de diminuer le bruit de la reconnaissance d'entités médicales (cf. tableau 2) ;

2) filtrer les syntagmes candidats avec une liste de mots vides en amont de MetaMap ;

3) rechercher des termes candidats dans des listes de problèmes médicaux, traitements et tests médicaux obtenus du corpus d'entraînement, de Wikipedia, Health on the Net, et Biomedical Entity Network ;

4) pour les entités candidates qui n'ont pas été détectées dans les listes, déterminer leurs concepts et types sémantiques correspondant dans l'UMLS avec MetaMap, après un filtrage avec (i) une liste des erreurs les plus fréquentes et (ii) la limitation des types sémantiques utilisés par MetaMap afin d'éviter certains concepts trop généraux (*e.g. Quantitative Concept, Functional Concept, Qualitative Concept*).

Phrase	
Adrenal-sparing surgery is safe and effective, and may become the treatment of choice in patients with hereditary phaeochromocytoma.	
Extrait des résultats	

Phrase : "Adrenal-sparing surgery"
`Meta Mapping (825):`
660 Adrenal [Finding]
589 Sparingly [Intellectual Product,Qualitative Concept]
827 Surgery (Operative Surgical Procedures) [Therapeutic or Preventive Procedure]
`Meta Mapping (825):`
660 Adrenal [Finding]
589 Sparingly [Intellectual Product,Qualitative Concept]
827 Surgery (Surgery specialty) [Biomedical Occupation or Discipline]
`Meta Mapping (825):`
660 Adrenal [Finding]
589 Sparingly [Intellectual Product,Qualitative Concept]
827 surgery (Surgical aspects) [Functional Concept]
`Meta Mapping (825):`
660 Adrenal (Adrenal Glands) [Body Part, Organ, or Organ Component]
589 Sparingly [Intellectual Product,Qualitative Concept]
827 Surgery (Operative Surgical Procedures) [Therapeutic or Preventive Procedure]
`Meta Mapping (825):`
660 Adrenal (Adrenal Glands) [Body Part, Organ, or Organ Component]
589 Sparingly [Intellectual Product,Qualitative Concept]
827 surgery (Surgical aspects) [Functional Concept]
(...)
	Phrase : "effective,"
	`Meta Mapping (1000):`
	1000 Effective [Qualitative Concept]
	Phrase : "the treatment"
	`Meta Mapping (1000):`
	1000 Treatment (Administration procedure) [Therapeutic or Preventive Procedure]
	`Meta Mapping (1000):`
	1000 Treatment (Biomaterial Treatment) [Conceptual Entity]
	`Meta Mapping (1000):`
	1000 Treatment (Therapeutic procedure) [Therapeutic or Preventive Procedure]
	`Meta Mapping (1000):`
	1000 Treatment (Treating) [Functional Concept]
	`Meta Mapping (1000):`
	1000 treatment (therapeutic aspects) [Functional Concept]
	Phrase : "of choice"
	`Meta Mapping (1000):`
	1000 choice (Choice Behavior) [Individual Behavior]
	Phrase : "in patients"
	`Meta Mapping (1000):`
	1000 Patients [Patient or Disabled Group]
	Phrase : "with hereditary phaeochromocytoma."
	`Meta Mapping (888):`
	694 Hereditary [Functional Concept]
	861 pheochromocytoma (Benign pheochromocytoma of adrenal gland) [Neoplastic Process]
	`Meta Mapping (888):`
	694 Hereditary [Functional Concept]
	861 PHAEOCHROMOCYTOMA (Pheochromocytoma) [Neoplastic Process]

Tableau 3. *Extrait des résultats de MetaMap pour une phrase donnée. Chaque concept est précédé de son score et suivi [entre crochets] de son type sémantique.*

Ces améliorations visent à augmenter la précision des résultats de MetaMap. Afin d'avoir une première idée sur les modifications proposées et une première évaluation de la méthode MetaMapPlus, nous avons construit un corpus d'évaluation de 20 articles scientifiques anglais variés extraits de PubMedCentral (d'autres évaluations seront présentées dans les sections 5.1 et 5.2). Nous avons ensuite annoté manuellement les entités médicales correspondant à 16 types sémantiques donnés. Comme il est difficile d'annoter manuellement toutes les entités médicales présentes dans notre corpus, nous avons mesuré uniquement la précision de la reconnaissance d'entités médicales de 16 types sémantiques. La précision dépend de l'exactitude de leurs catégories (types sémantiques) mais aussi de la précision de localisation de ces entités (correcte, avec du bruit, partielle ou fausse). Dans cette évaluation, une erreur liée à la localisation partielle (respectivement avec du bruit) d'un terme médical coûte un demi point, et la précision est calculée selon la formule suivante :

$$Precision = \frac{C + 0,5 \times B}{Ref}$$

– C : le nombre d'entités correctes ;

– B (boundary) : le nombre d'entités avec une catégorie correcte mais une localisation imprécise (partielle ou avec bruit) ;

– Ref : le nombre total des entités de référence.

Le tableau 4 compare la précision obtenue avec notre méthode MetaMapPlus et celle obtenue avec l'utilisation simple de MetaMap sur un sous-ensemble de types sémantiques. Les erreurs liées aux types sémantiques sont notées par T, celles liées aux frontières des entités sont notées par B et la précision est notée par P. Notre méthode conduit à une augmentation significative de la précision par rapport à MetaMap (le total indiqué a été calculé sur toutes les occurrences des 16 types sémantiques traités). La méthode MetaMapPlus est aussi basée sur le *chunker* TreeTagger qui a obtenu un meilleur rappel que MetaMap (cf. section 4.2.1) et sur des listes de termes médicaux supplémentaires au thésaurus de MetaMap qui permettent de retrouver plus d'entités médicales.

	MetaMap			MetaMapPlus		
	T	B	P	T	B	P
Disease or Syndrome	9,09 %	52,27 %	64,77 %	9,81 %	26,48 %	76,94 %
Injury or poisoning	33,33 %	34,84 %	49,24 %	26,19 %	35,71 %	55,95 %
Total	30,24 %	34,56 %	**54,62 %**	12,23 %	27,10 %	**74,21** %

Tableau 4. *Précision de la reconnaissance des entités médicales de 16 types sémantiques sur un premier corpus de test*

4.2.3. *Catégorisation des syntagmes nominaux : méthode statistique (TT-SVM)*

La deuxième méthode que nous avons mise en place et testée consiste à extraire les syntagmes nominaux à partir du texte, puis à utiliser un classifieur pour déterminer

les syntagmes qui correspondent à des entités médicales et leurs catégories (*e.g. Traitement, Test, Maladie*). Il s'agit d'une classification des syntagmes nominaux en $n+1$ classes (où n est le nombre des catégories d'entités médicales).

Nous avons choisi d'utiliser un classifieur de type SVM (machine à vecteurs de support). Ekbal et Bandyopadhyay (2010) affirment que les classifieurs SVM ont un certain avantage sur les algorithmes d'apprentissage statistique conventionnels, tels que les arbres de décision, les modèles de Markov cachés, les modèles à entropie maximale et cela sur deux aspects : (i) les SVM ont une forte capacité de généralisation indépendante de la dimension des vecteurs d'attributs et (ii) les SVM peuvent effectuer leur entraînement avec toutes les combinaisons des attributs choisis sans augmenter la complexité algorithmique, par l'introduction d'une fonction noyau. La bibliothèque LIBSVM (Chang et Lin, 2001) a été utilisée pour la mise en place du classifieur. La sélection du modèle, et plus particulièrement la recherche du paramètre relatif au noyau (gamma) et du paramètre de régularisation (C), a été effectuée automatiquement avec un script disponible en complément de cette bibliothèque. Keerthi et Sundararajan (2007) ont effectué une expérimentation qui montre que les classifieurs SVM structurés et les CRF sont assez proches en termes de performance si les mêmes attributs sont utilisés pour la classification.

Nous avons sélectionné des attributs lexicaux, orthographiques et morphosyntaxiques. Le tableau 5 décrit ces attributs.

Attributs lexicaux	– Mots du syntagme lui-même – Nombre de mots du syntagme – Lemmes des mots du syntagme (obtenus avec TreeTagger) – Trois mots avant le syntagme et leurs lemmes – Trois mots après le syntagme et leurs lemmes
Exemples d'attributs orthographiques	- Le premier mot, un mot du syntagme ou tous les mots sont capitalisés – Le premier mot, un mot ou tous les mots sont en majuscules – Le premier mot, un mot ou tous les mots sont en minuscules – Le syntagme contient une abréviation – Le syntagme contient un seul caractère en majuscule, un chiffre, ou un signe spécial (*e.g.* –, +, &, /).
Attributs morphosyntaxiques	Les catégories morphosyntaxiques des mots du syntagme, des trois mots avant le syntagme et des trois mots après le syntagme (obtenues avec TreeTagger).

Tableau 5. *Les attributs lexicaux, orthographiques et morphosyntaxiques utilisés avec le classifieur SVM pour la classification des syntagmes nominaux*

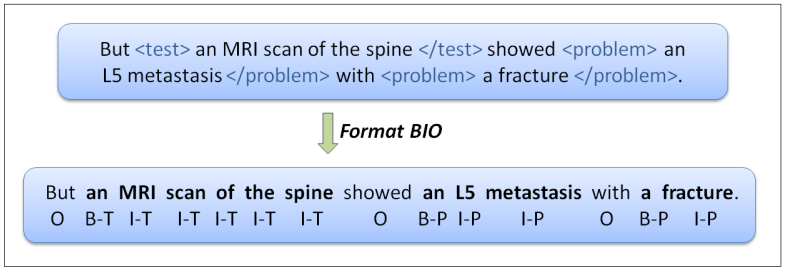

Figure 3. *Exemple de phrase au format BIO (T = Test, P = Problème)*

4.3. *Stratégie 2 : une seule étape, frontières et catégories conjointement*

4.3.1. *Méthode statistique pour la détermination des frontières et des catégories (CRF-BIO)*

Cette méthode comporte une seule étape : déterminer les frontières et les types des entités médicales. Le passage par un *chunker* n'est plus indispensable. Pour ce faire, nous utilisons le format BIO : B (*beginning*), I (*inside*) et O (*outside*) qui permet de représenter un marquage de segments de texte (les entités) par un étiquetage individuel des mots. Une entité de type *Problème* formée de plusieurs mots (par exemple, *an L5 metastatis*) voit son premier mot (*an*) étiqueté B-P (début de *Problème*) et ses autres mots (*L5* et *metastatis*) étiquetés I-P (dans un *Problème*). Une entité de type *Problème* comprenant un seul mot est étiquetée B-P. Les mots hors entité sont étiquetés O (voir la figure 3). Si nous avons n catégories (*e.g. Problème, Traitement, Test*), nous avons alors n classes de type B-, n classes de type I- (*e.g.* les classes P-B et P-I associées à la catégorie Problème) et une classe de type *O*. Il s'agit alors d'un problème de classification de chaque mot (et non plus de chaque syntagme nominal) en $2n + 1$ classes possibles (où n est le nombre de catégories médicales).

Les mots d'une phrase forment une séquence, et la décision sur la catégorie d'un mot peut être influencée par la décision portant sur la catégorie du mot précédent. Cette dépendance est prise en compte dans les modèles séquentiels comme les modèles de Markov cachés (HMM) ou les champs aléatoires conditionnels (*Conditional Random Fields*, ou CRF). Contrairement aux HMM, l'apprentissage dans les CRF maximise la probabilité conditionnelle des classes relativement aux observations plutôt que leur probabilité conjointe. Cela leur permet d'utiliser un nombre quelconque d'attributs concernant des aspects quelconques de la séquence de mots d'entrée. Ces propriétés expliquent l'intérêt des CRF pour un certain nombre de tâches de traitement automatique des langues, comme l'étiquetage morphosyntaxique, la détection de syntagmes non récursifs (*chunks*) ou la reconnaissance d'entités nommées (Tellier et Tommasi, 2010).

Nous avons donc testé la détection d'entités médicales avec un CRF. Nous avons utilisé pour cela l'outil CRF++[6], qui permet de décrire facilement les attributs à utiliser à travers des patrons d'attributs (*feature templates*). Ces attributs sont listés ci-dessous. Nous avons aussi réglé CRF++ pour qu'il utilise la dépendance entre catégories successives (instruction `B` du fichier d'attributs).

Pour chaque mot, nous utilisons un ensemble d'attributs lexicaux, orthographiques et morphosyntaxiques. Le tableau 6 décrit ces attributs.

Attributs lexicaux	– Le mot (M) lui-même, deux mots avant et trois mots après – Les lemmes de ces mots.
Exemples d'attributs orthographiques	- Le mot contient un signe spécial (*e.g.* +, –, &, /) – Le mot est un chiffre, alphabétique, un signe de ponctuation ou un symbole – Le mot est en majuscules, capitalisé, en minuscules (AA, Aa, aa) – Préfixes de différentes longueurs (de 1 à 4) – Suffixes de différentes longueurs (de 1 à 4)
Attributs morphosyntaxiques	Les catégories morphosyntaxiques du mot M lui-même, des deux mots avant et des trois mots après M.
Autres attributs	Premier verbe suivant, premier nom suivant, longueur du mot par rapport à un seuil, etc.

Tableau 6. *Les attributs lexicaux, orthographiques et morphosyntaxiques utilisés avec le classifieur CRF pour la classification des mots*

4.3.2. *Méthode hybride pour la détermination des frontières et des catégories (CRF-BIO-H)*

Cette méthode consiste à utiliser les résultats de la méthode à base de règles (MetaMapPlus) comme attributs pour la méthode statistique (i.e. CRF-BIO). Plusieurs choix sont possibles : (i) l'utilisation de la catégorie sémantique associée au mot en appliquant MetaMapPlus, (ii) l'utilisation du type sémantique associé de l'UMLS (en appliquant MetaMapPlus sur le mot) et (iii) la transformation des résultats de MetaMapPlus au format BIO et en les considérant comme attributs pour le classifieur CRF. À chaque mot est associée une classe de type B-problem, I-problem, B-treatment, I-treatment, B-test et I-test (obtenues grâce aux résultats de MetaMapPlus sur le corpus i2b2). Nous testons l'apport de chacun de ces choix de combinaisons dans nos expérimentations.

6. `http://crfpp.sourceforge.net/`

5. Expérimentations

Nous avons mené des expériences de reconnaissance d'entités médicales dans des textes cliniques en anglais (section 5.1). Pour tester la robustesse des systèmes créés, nous les avons également appliqués à des résumés d'articles scientifiques extraits de MEDLINE (section 5.2). Tous les résultats présentés dans cette section ont été calculés sur la base de la correspondance stricte des frontières.

5.1. *Expériences sur des textes cliniques*

Dans cette section, nous présentons le corpus i2b2 et les résultats des différentes méthodes.

5.1.1. *Corpus i2b2 de textes cliniques*

Le corpus i2b2 a été construit dans le cadre du challenge i2b2 2010[7]. Ce corpus comporte des entités médicales de trois catégories (*Problème*, *Traitement* et *Test*) annotées dans 76 165 phrases (663 476 mots au total), pour une moyenne de 8,7 mots par phrase.

L'exemple 2 montre une phrase annotée du corpus i2b2.

(2) *<problem> CAD </problem> s/p <treatment> 3v-CABG </treatment> 2003 and subsequent <treatment> stenting </treatment> of <treatment> SVG </treatment> and LIMA.*

L'annotation manuelle du corpus a été conduite suivant un guide d'annotation. Le tableau 7 présente l'accord interannotateurs pour chaque catégorie médicale.

i2b2	**Accord interannotateurs (frontières strictes)**	**Accord interannotateurs (frontières non strictes)**
Problème	0,84	0,91
Traitement	0,83	0,90
Test	0,83	0,88
Total	**0,83**	**0,90**

Tableau 7. *Corpus i2b2 : Accord interannotateurs pour chaque catégorie médicale*

Le tableau 8 présente le nombre de phrases d'entraînement et de test (nous avons conservé la répartition officielle du challenge).

7. `http://www.i2b2.org`

i2b2	Phrases	Mots
Corpus d'entraînement	31 238	267 304
Corpus de test	44 927	396 172

Tableau 8. *Nombre de phrases et de mots d'entraînement et de test*

5.1.2. *Configurations et résultats*

Nous avons expérimenté les cinq configurations suivantes :

1) MM : MetaMap

2) MM+ : MetaMapPlus

3) TT-SVM : Catégorisation des syntagmes nominaux avec SVM

4) CRF-BIO : Apprentissage format BIO avec CRF

5) CRF-BIO-H : Méthode hybride (CRF-BIO utilisant des attributs sémantiques construits à partir des résultats de MM+)

Le tableau 9 présente les résultats obtenus sur le corpus i2b2 avec les mesures classiques de rappel, précision et F-mesure[8]. Le tableau 10 détaille les résultats pour chaque catégorie médicale (*Problème*, *Traitement* et *Test*). Nous avons mis en gras la meilleure performance totale (tableau 9) et pour chaque catégorie (tableau 10).

Configuration	Rappel	Précision	F-mesure
MM	15,52	16,10	15,80
MM+	48,68	56,46	52,28
TT-SVM	43,65	47,16	45,33
CRF-BIO	70,15	83,31	76,17
CRF-BIO-H	**71,92**	**83,83**	**77,42**

Tableau 9. *Résultats de chaque configuration sur le corpus i2b2 (frontières strictes)*

5.2. *Expériences sur un corpus d'articles scientifiques*

Cette section présente les résultats obtenus avec des expériences supplémentaires effectuées sur un corpus de résumés scientifiques extraits de MEDLINE.

5.2.1. *Corpus de Berkeley*

Le corpus de Berkeley (Rosario et Hearst, 2004) est construit à partir de titres et résumés d'articles scientifiques de MEDLINE. L'objectif du corpus était l'étude des relations sémantiques entre les entités médicales de type *Maladie* et *Traitement*, qui

8. Lors du challenge i2b2 2010, la meilleure F-mesure (85,23 %) a été obtenue par de Bruijn *et al.* (2011)

Configurations	Catégorie	**Précision**	**Rappel**	**F-mesure**
MM+	Problème	60,84	53,04	56,67
	Traitement	51,99	61,93	56,53
	Test	56,67	28,48	37,91
TT-SVM	Problème	48,25	43,16	45,56
	Traitement	42,45	50,86	46,28
	Test	57,37	35,76	44,06
CRF-BIO-H	Problème	**82,05**	**73,14**	**77,34**
	Traitement	**83,18**	**73,33**	**77,95**
	Test	**87,50**	**68,69**	**77,00**

Tableau 10. *Résultats pour chaque catégorie sémantique sur le corpus i2b2*

sont : *cures*, *prevents* et *side effect*. Dans ce travail, nous exploitons les annotations des entités médicales uniquement.

Le corpus contient deux catégories d'entités médicales : *Maladies* (1 660 entités) et *Traitements* (1 179 entités) dans 3 654 phrases (74 754 mots), donc en moyenne 20,45 mots par phrase. Cette moyenne est nettement supérieure à la moyenne de mots par phrase du corpus i2b2. Cela est dû au fait que le corpus i2b2, constitué de textes cliniques, comporte des phrases factuelles courtes avec beaucoup d'abréviations. L'exemple 3 montre une phrase annotée du corpus de Berkeley.

(3) *We investigated the hypothesis that <TREAT PREV> an antichlamydial macrolide antibiotic , roxithromycin </TREAT PREV> , can prevent or reduce recurrent major ischaemic events in patients with <DIS PREV> unstable angina </DIS PREV>.*

Plusieurs étiquettes sont utilisées pour chaque catégorie suivant la relation qui les lie. Par exemple, il y a 7 étiquettes différentes pour les maladies (ou problèmes médicaux) : <DIS>, <DIS_NO>, <DIS_VAG>, <DISONLY>, <DIS_PREV>, <DIS_SIDE_EFF>, <DIS_EFF>. Nous avons uniformisé ces étiquettes dans une étape de prétraitement avant les expérimentations finales.

Le tableau 11 présente le nombre de phrases d'entraînement et de test.

Berkeley	**Phrases**	**Mots**
Corpus d'entraînement	1 462	36 642
Corpus de test	2 193	38 112

Tableau 11. *Nombre de phrases et de mots d'entraînement et de test*

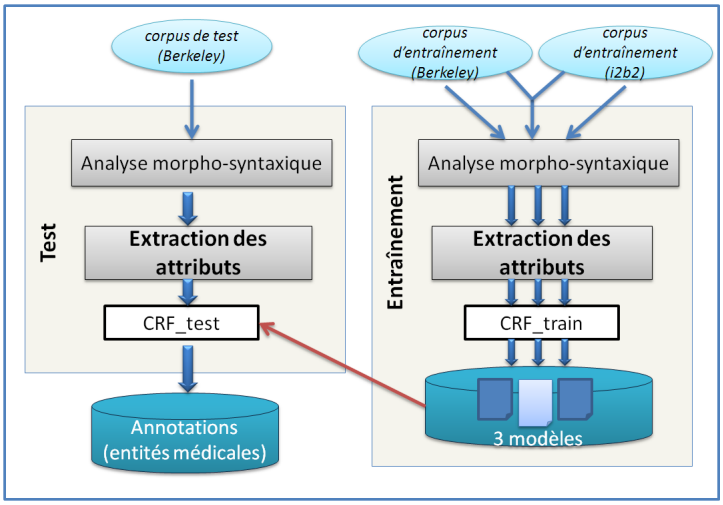

Figure 4. *Reconnaissance des entités médicales à partir du corpus de Berkeley : la méthode CRF-BIO et les trois modèles testés*

5.2.2. *Résultats*

Nous avons testé la méthode fondée sur MetaMap (MM+) sur le corpus de Berkeley. Le but de ce test n'est pas d'obtenir les meilleures performances en termes d'extraction des entités médicales mais de pouvoir comparer les deux stratégies étudiées dans cet article sur un deuxième corpus. La méthode TT-SVM (deuxième méthode de la première startégie) n'est pas présentée ici car elle a abouti à de moins bon résultats que MM+. Le tableau 12 présente les résultats obtenus avec les mesures classiques de précision, rappel et F-mesure.

		Précision	**Rappel**	**F-mesure**
MM	Maladie	5,32	7,63	6,27
	Traitement	6,37	18,84	9,52
	Total	5,35	12,34	7,46
MM+	Maladie	**34,47**	**44,97**	**39,02**
	Traitement	**18,11**	**39,36**	**24,81**
	Total	**23,43**	**42,47**	**30,20**

Tableau 12. *Résultats de la méthode à base de règles (MM+) sur le corpus de Berkeley*

Nous avons aussi testé une méthode statistique sur ce corpus (CRF-BIO). Nous avons construit trois modèles différents pour le classifieur CRF. Un premier modèle est construit à partir du corpus d'entraînement de Berkeley, un deuxième à partir du corpus de i2b2 et un troisième à partir de ces deux corpus d'entraînement (cf. figure 4).

Nous avons obtenu les meilleurs résultats avec le troisième modèle construit à partir des deux corpus d'entraînement. Le tableau 13 décrit les résultats obtenus avec les trois modèles en utilisant les valeurs classiques de rappel, précision et F-mesure. Nous avons obtenu 34,37 % de F-mesure (40,3 % pour *Problème* et 26,5 % pour *Traitement*). Ces résultats sont obtenus avec un sous-ensemble d'attributs avec lequel nous avons obtenu 76,17 % sur le corpus i2b2. Ces attributs sont les mots, les lemmes, les catégories morphosyntaxiques, des attributs orthographiques, les suffixes et préfixes (cf. ensemble A4, tableau 14).

Modèles testés	**Rappel**	**Précision**	**F-mesure**
Modèle 1 (Berkeley)	14,64	**53,29**	22,97
Modèle 2 (i2b2)	**36,09**	26,38	30,48
Modèle 3 (Berkeley+i2b2)	32,13	36,94	**34,37**

Tableau 13. *Résultats de la méthode CRF-BIO sur le corpus de Berkeley*

6. Discussion

Nous avons présenté quatre méthodes différentes pour la reconnaissance des entités médicales. Dans cette section nous analysons les différences en termes de résultats par rapport aux corpus et aux méthodes.

6.1. *Expérimentations sur le corpus i2b2*

La figure 5 compare les résultats des méthodes MetaMapPlus, TT-SVM, CRF-BIO et CRF-BIO-H expérimentées sur le corpus i2b2.

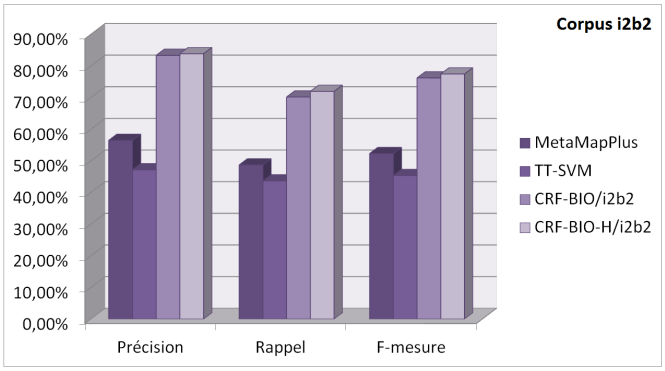

Figure 5. *Résultats sur le corpus i2b2*

L'application directe de l'outil de référence MetaMap n'a pas permis d'obtenir de bons résultats (15,80 % de F-mesure). Cela est essentiellement dû à deux points :

– les erreurs de délimitation des frontières des entités : *e.g.* extraire « ***no*** *pericardial effusion* **.** » au lieu de « *pericardial effusion* » et « ***(*** *Warfarin* » au lieu de « *Warfarin* » (cf. section 4.2.1) ;

– les termes généraux considérés par MetaMap comme entités médicales à cause de leur polysémie (*e.g. ten*) ou de leur utilisation fréquente en domaine médical (*e.g. case*, *form*).

Nous avons pu améliorer les résultats de MetaMap à travers la méthode MetaMapPlus en utilisant entre autres (i) un *chunker* externe qui a été choisi après une étude comparative, (ii) un antidictionnaire contenant les erreurs les plus fréquentes de MetaMap observées lors de tests précédents et (iii) des listes de termes médicaux supplémentaires.

Les résultats finaux atteignent 52,28 % de F-mesure. Ils restent cependant limités par comparaison à CRF-BIO et CRF-BIO-H à cause principalement de la performance du procédé de *chunking* (comme c'est le cas pour la méthode TT-SVM). La méthode MetaMapPlus a cependant permis de typer correctement 92,16 % des entités si on considère uniquement les entités candidates extraites avec de bonnes frontières avant la catégorisation. Ce point prouve l'efficacité, en termes de précision, de l'outil de référence MetaMap. Cette performance peut se justifier par le fait que MetaMap utilise des connaissances de domaine riches à travers le Metathesaurus et le réseau sémantique de l'UMLS.

La méthode TT-SVM a obtenu de moins bons résultats que MetaMapPlus. Étant donné que la méthode TT-SVM et MetaMapPlus essaient de catégoriser le même ensemble d'entités candidates (préalablement délimitées par le même *chunker*), la différence se justifie par le fait que les connaissances de domaine dont dispose MetaMap et les listes supplémentaires exploitées par MetaMapPlus ont permis une meilleure précision que le classifieur SVM qui a été entraîné sur le corpus i2b2. Le choix des attributs d'apprentissage peut aussi être amélioré mais cette étude permet cependant de confirmer que les méthodes fondées sur les connaissances du domaine (indépendantes des corpus) peuvent concurrencer les méthodes statistiques qui sont elles dépendantes des corpus d'entraînement.

La méthode statistique CRF-BIO détermine les frontières et les catégories des entités médicales simultanément par apprentissage supervisé sans avoir recours à un *chunker*. Les résultats de cette méthode sur le corpus i2b2 sont encore meilleurs que ceux obtenus avec MetaMapPlus. En effet, les méthodes statistiques dépendent fortement du nombre de données annotées disponibles (ce nombre est important pour le corpus i2b2) et aussi de l'ensemble d'attributs utilisé. Nous présentons l'apport de chaque catégorie d'attributs avec la méthode CRF-BIO dans le tableau 14.

Nous avons essayé de combiner les résultats des deux méthodes à base de règles (MetaMapPlus) et statistique. Plusieurs tests d'attributs sémantiques ont été effectués avec la méthode CRF-BIO. Par exemple, l'utilisation de la catégorie sémantique associée au mot en appliquant MetaMapPlus diminue les résultats de 76,17 % à 76,01 %. Le type sémantique associé de l'UMLS (en appliquant MetaMapPlus sur le mot) di-

Attributs	Rappel	Précision	F-mesure
A1 : mots/lemmes/POS	62,81	82,25	71,23
A2 : A1 + attributs orthographiques	63,72	82,19	71,78
A3 : A2 + suffixes	67,91	82,89	74,65
A4 : A3 + préfixes	70,15	83,31	76,17
A5 : A4 + autres attributs	70,22	83,28	76,19

Tableau 14. *Apport de chaque classe d'attributs : méthode CRF-BIO sur le corpus i2b2*

minue la F-mesure de 76,17 % à 73,55 %. Ceci peut s'expliquer par le nombre de classes (associées aux types sémantiques) qui devient plus important mais aussi par la performance réduite de MetaMap s'il est appliqué au niveau du mot et non au niveau du syntagme ou de la phrase. La meilleure solution a été obtenue en transformant les résultats de MetaMapPlus au format BIO et en les considérant comme attributs pour le classifieur CRF. À chaque mot est associée une classe de type B-problem, I-problem, B-treatment, I-treatment, B-test et I-test (obtenues grâce aux résultats de MetaMapPlus sur le corpus i2b2). Avec ces attributs sémantiques, nous avons pu passer d'une F-mesure de 76,19 % à 77,42 %. Cela peut être justifié du fait que ce dernier choix de combinaison ajoute deux niveaux d'information : la position précise de l'entité médicale et sa catégorie alors que les deux premiers ne fournissent que des indices sur la catégorie de l'entité candidate.

Le tableau 15 présente l'apport des attributs sémantiques.

Attributs	Rappel	Précision	F-mesure
A5 : mots, lemmes, POS, attributs orthographiques, suffixes, préfixes et autres	70,22	83,28	76,19
A6 : A5 + attributs sémantiques	**71,92**	**83,83**	**77,42**

Tableau 15. *Apport des attributs sémantiques : méthode CRF-BIO-H sur le corpus i2b2*

6.2. *Expérimentations sur le corpus de Berkeley*

La figure 6 compare les résultats obtenus par les différentes méthodes sur le corpus de Berkeley.

Une comparaison avec la figure 5 montre que les résultats obtenus pour les deux corpus ne sont pas du même niveau de performance. Deux caractéristiques entrent en jeu :

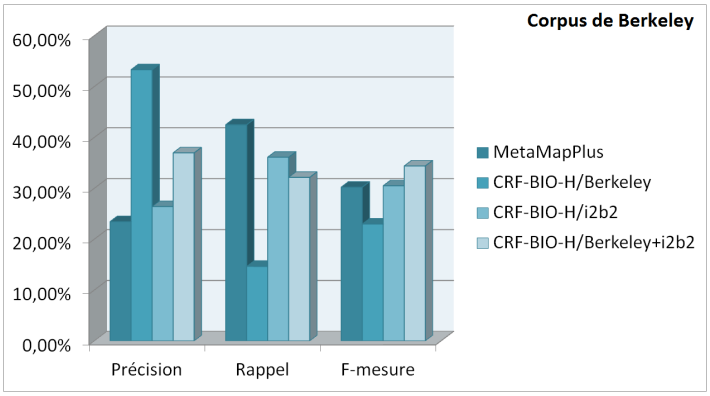

Figure 6. *Résultats sur le corpus de Berkeley*

1) *les genres différents des deux corpus :* le corpus de i2b2 2010 a un nombre moyen de mots par phrase de 8,7 alors que le corpus de Berkeley a un nombre moyen de mots par phrase de 20,45. Aussi, le corpus de i2b2 utilise un vocabulaire assez spécifique, comme les abréviations conventionnelles de termes médicaux (comme *k/p* pour *kidney pancreas* et *d&c* pour *dilation and curettage*), mais aussi des abréviations de mots généraux (*e.g. w/o* pour *without* et *y/o* pour *year old*). Ceci explique pourquoi nous n'avons pas obtenu de bons résultats en appliquant la méthode CRF-BIO-H entraînée sur le corpus i2b2 : 30,48 % de F-mesure, avec 26,38 % de précision et 36,09 % de rappel (la meilleure valeur obtenue parmi les trois modèles, grâce à la taille importante du corpus d'entraînement de i2b2) ;

2) *la qualité d'annotation des deux corpus :* le corpus de i2b2 a été annoté avec rigueur étant donné son contexte de challenge et l'utilisation, par les annotateurs, d'un guide qui explicite les principes d'annotation avec des exemples. Le corpus de Berkeley a été annoté d'une manière moins rigoureuse. En effet, aucune règle explicite n'a été énoncée. Par exemple les articles étaient parfois considérés comme faisant partie de l'entité médicale et d'autres fois exclus de l'entité. Notre évaluation sur un échantillon aléatoire de 200 entités médicales montre un taux d'erreur à l'annotation de 20 %. La qualité d'annotation du corpus de Berkeley et la taille moyenne, voire petite, du corpus d'entraînement expliquent les résultats obtenus par la méthode CRF-BIO-H entraînée sur le corpus de Berkeley.

6.3. *Comparaison des deux stratégies*

Sur les deux corpus utilisés, nous avons pu observer que la stratégie de détection conjointe des frontières et des catégories des entités médicales était plus efficace que

la stratégie de détection en deux étapes indépendantes (gain de 24 % de F-mesure sur le corpus i2b2 et de 4 % sur le corpus de Berkeley). Ces résultats confirment que (i) l'hypothèse consistant à dire qu'une entité médicale correspond à un groupe nominal extrait par un *chunker* peut être utilisée comme une heuristique (plus ou moins fiable suivant les corpus) et que (ii) la détection simultanée des frontières et des catégories par apprentissage est plus efficace si suffisamment d'exemples d'entraînement sont disponibles.

La deuxième question que l'on peut se poser est : pourquoi est-ce que le gain en performance est moins visible sur le corpus de Berkeley ? Cela est principalement dû au fait que ce corpus contient peu d'exemples d'entraînement comparé au corpus i2b2, ce qui a diminué les performances des méthodes statistiques. D'un autre côté, la méthode MetaMapPlus qui a été utilisée pour évaluer la première stratégie sur ce corpus utilise des connaissances du domaine indépendantes du corpus, ce qui lui a permis de se rapprocher des performances de la méthode statistique CRF-BIO utilisée pour évaluer la deuxième stratégie.

6.4. *Robustesse vs portabilité*

6.4.1. *Les méthodes à base de règles*

Elles ont l'avantage d'être reproductibles sur tous types de corpus sans étape de préparation ou d'apprentissage. Cependant, leur dépendance aux connaissances utilisées fait que leurs performances ne sont pas du même niveau que les processus d'apprentissage, car ces connaissances restent relativement figées par rapport à la vitesse avec laquelle les approches d'apprentissage permettent d'acquérir de nouveaux critères d'extraction et de catégorisation. Ce type de méthode a aussi l'inconvénient d'être coûteux à mettre en place si on veut obtenir une couverture satisfaisante (un bon rappel). Leur avantage est cependant d'avoir un consensus sémantique sur les informations extraites (*e.g.* réseau sémantique de l'UMLS) qui permet potentiellement un traitement plus sophistiqué de celles-ci. Il est aussi important de noter que des ressources sémantiques (listes de termes médicaux, lexiques,...) sont souvent utilisées dans les processus d'apprentissage ce qui rend le développement de telles ressources avantageux à la fois pour les méthodes à base de règles et pour les méthodes statistiques. Inversement, des méthodes statistiques peuvent aussi servir à améliorer les méthodes à base de règles en permettant d'apprendre automatiquement de nouvelles règles d'extraction ou des patrons linguistiques pertinents (Hobbs et Riloff, 2010).

6.4.2. *Les méthodes statistiques*

Avec des méthodes à base d'apprentissage supervisé nous avons obtenu de bons résultats sur le corpus i2b2. Mais ce n'est pas le cas pour le corpus de Berkeley. Ces méthodes, bien qu'elles puissent être très robustes, présentent deux inconvénients non négligeables :

1) la dépendance aux données annotées disponibles (cf. résultats de BIO-CRF-H entraîné et testé sur le corpus de Berkeley), ce qui constitue un obstacle à l'utilisation de ce type de méthodes pour des tâches et des domaines où l'on ne dispose pas de corpus annotés, d'autant plus que la constitution de ces corpus est coûteuse ;

2) le problème de portabilité sur des corpus différents de ceux utilisés en entraînement (cf. résultats de BIO-CRF-H entraîné sur i2b2 et testé sur Berkeley). La dégradation des performances de ces méthodes, appliquées sur des corpus ayant des caractéristiques différentes de ceux utilisés pour l'entraînement, constitue un grand inconvénient pour leur passage à l'échelle.

Cependant, pour un contexte ou un corpus précis, si un bon nombre de données annotées est disponible, les méthodes statistiques peuvent être très robustes et offrir la meilleure solution (77,42 % de F-mesure obtenus avec la méthode CRF-BIO-H sur le corpus i2b2 *vs* 52,28 % obtenus par la méthode MetaMapPlus). Elles offrent en plus la possibilité de découvrir des règles de classification potentiellement indétectables pour l'expert humain. Cela a été démontré dans notre étude où une méthode statistique entraînée sur un corpus médical a permis de délimiter les entités médicales dans les textes plus efficacement qu'un outil spécialisé reposant sur des connaissances et des règles du domaine (MetaMap) et plus efficacement qu'un *chunker* généraliste exploitant des règles d'extraction linguistiques.

7. Conclusions et perspectives

Nous avons étudié quatre approches pour la reconnaissance d'entités médicales dans le cadre de deux stratégies différentes pour cette reconnaissance. Les résultats obtenus montrent que l'utilisation d'un *chunker* pour l'extraction des entités limite la performance finale même si la catégorisation des entités médicales se fait de façon efficace. L'utilisation de procédés d'apprentissage pour l'extraction et la catégorisation simultanées des entités a permis de contourner cette limite. La meilleure performance a ainsi été obtenue par le classifieur CRF en exploitant le format BIO et des attributs lexicaux, orthographiques et morphosyntaxiques. Nous avons aussi présenté une méthode hybride qui a permis d'améliorer encore ces performances en exploitant des connaissances sémantiques du domaine à travers la méthode à base de règles MetaMapPlus. Les résultats des méthodes statistiques sur le premier corpus de test confirment la robustesse de ce type de méthode. Cependant les résultats obtenus sur le deuxième corpus médical, de type différent, mettent en évidence deux inconvénients des méthodes statistiques, à savoir la dépendance aux données annotées et leur portabilité limitée lorsqu'elles sont appliquées à des corpus ayant des caractéristiques différentes de ceux utilisés pour l'entraînement.

Comme perspectives, nous envisageons (i) d'améliorer la précision de la méthode MetaMapPlus en incluant un module d'apprentissage en amont de l'outil MetaMap, pour classifier les syntagmes nominaux en entités médicales ou en entités non médicales et (ii) d'élargir la liste des catégories médicales traitées. Aussi, notre objectif actuel est la reconnaissance des entités médicales dans des corpus français. L'inexis-

tence de corpus médicaux annotés en français constitue un obstacle pour l'apprentissage statistique. D'un autre côté, développer des méthodes à base de règles ou annoter manuellement des corpus pour l'apprentissage sont deux approches coûteuses en temps. Nous envisageons d'exploiter les méthodes MetaMapPlus et CRF-BIO-H pour l'extraction automatique d'entités médicales à partir de textes en français. L'idée étant d'utiliser d'une part un corpus médical parallèle (anglais/français) et d'autre part les alignements au niveau des mots pour projeter les annotations de la partie anglaise dans le corpus français. Des expérimentations dans ce cadre sont en cours (Ben Abacha *et al.*, 2012). À moyen terme nous envisageons d'exploiter ces méthodes de reconnaissance d'entités médicales dans le cadre d'un système de questions-réponses translingue.

Remerciements

Nous voudrions remercier les organisateurs du challenge i2b2 2010 et aussi Barbara Rosario et Marti Hearst pour les ressources qu'ils ont mises à disposition.

8. Bibliographie

Aronson A. R., « Effective mapping of biomedical text to the UMLS Metathesaurus : the MetaMap program », *AMIA Annu Symp Proc*, p. 17-21, 2001.

Ben Abacha A., Zweigenbaum P., « Medical Entity Recognition : A Comparison of Semantic and Statistical Methods », *Actes BioNLP 2011 Workshop*, Association for Computational Linguistics, Portland, Oregon, USA, p. 56-64, 2011.

Ben Abacha A., Zweigenbaum P., Max A., « Extraction d'information automatique en domaine médical par projection inter-langue : vers un passage à l'échelle », *Proceedings of TALN 2012 (Traitement automatique des langues naturelles)*, Grenoble, 2012.

Bernhard D., Ligozat A.-L., « Analyse automatique de la modalité et du niveau de certitude : application au domaine médical », *Proceedings of TALN'11*, Montpellier, 2011.

Bodenreider O., « Lexical, terminological and ontological resources for biological text mining », *in* S. Ananiadou, J. McNaught (eds), *Text mining for biology and biomedicine*, Artech House, Boston, Massachusetts, p. 43-66, 2006.

Chang C.-C., Lin C.-J., *LIBSVM : a library for support vector machines*. 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chapman W. W., Bridewell W., Hanbury P., Cooper G. F., Buchanan B. G., « A simple algorithm for identifying negated findings and diseases in discharge summaries », *Journal of Biomedical Informatics*, vol. 34, n° 5, p. 301-310, 2001.

de Bruijn B., Cherry C., Kiritchenko S., Martin J. D., Zhu X., « Machine-learned solutions for three stages of clinical information extraction : the state of the art at i2b2 2010 », *JAMIA*, vol. 18, n° 5, p. 557-562, 2011.

Delbecque T., Jacquemart P., Zweigenbaum P., « Utilisation du réseau sémantique de l'UMLS pour la définition de types d'entités nommées médicales dans un système de Questions-

Réponses : impact de la source des documents explorés », *CORIA*, CLIPS, Grenoble, p. 101-115, 2005.

Ehrmann M., Les Entités Nommées, de la linguistique au Tal : Statut théorique et méthodes de désambiguïsation, PhD thesis, Université Paris 7, JUN, 2008.

Ehrmann M., Jacquet G., « Vers une double annotation des Entités Nommées », *Traitement Automatique des Langues*, vol. 47, n° 3, p. 63-88, 2006.

Ekbal A., Bandyopadhyay S., « Named Entity Recognition using Support Vector Machine : A Language Independent Approach », *International Journal of Electrical and Electronics Engineering*, vol. 4, n° 2, p. 155-170, 2010.

Elkateb-Gara F., « Extraction d'entités nommées pour la recherche d'informations précises », 4^e *Congrès ISKO France, Grenoble*, 2003.

Embarek M., Ferret O., « Learning Patterns for Building Resources about Semantic Relations in the Medical Domain », *LREC'08*, May, 2008.

Embarek M., Ferret O., « Can Esculape cure the complex of Oedipe in the medical domain ? », *Proceedings of the 9th International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information RIAO'10, Paris, France, April 28-30, 2010*, 2010.

Engelbrecht R., « Expert systems for medicine–functions and developments », *Zentralbl Gynakol*, vol. 119, n° 9, p. 428-434, 1997.

Fourour N., « Nemesis : un système de reconnaissance incrémentielle des entités nommées pour le français », *in* J.-M. Pierrel (ed.), *Actes TALN 2002 (Traitement automatique des langues naturelles)*, ATALA, ATILF, Nancy, p. 255-264, June, 2002.

Friedman C., Kra P., Rzhetsky A., « Two biomedical sublanguages : a description based on the theories of Zellig Harris », *Journal of Biomedical Informatics*, vol. 35, p. 222-235, 2002.

Grishman R., Sundheim B., « Message Understanding Conference - 6 : A Brief History », *Proc. of COLING*, Copenhagen, Denmark, p. 466-471, August, 1996.

Grouin C., Rosset S., Zweigenbaum P., Fort K., Galibert O., Quintard L., « Proposal for an Extension of Traditional Named Entities : From Guidelines to Evaluation, an Overview », *Proc. of the Fifth Linguistic Annotation Workshop (LAW-V)*, Association for Computational Linguistics, Portland, OR, June, 2011.

He Y., Kayaalp M., « Biological entity recognition with conditional random fields », *AMIA Annu Symp Proc*, p. 293-297, 2008.

Hobbs J. R., Riloff E., « Information Extraction », *in* N. Indurkhya, F. J. Damerau (eds), *Handbook of Natural Language Processing, Second Edition*, CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921.

Hotvedt M., « Continuing medical education : actually learning rather than simply listening », *JAMA, 275 :1638*, vol. 275, n° 21, p. 1637-1638, 1996.

Isozaki H., Kazawa H., « Efficient Support Vector Classifiers for Named Entity Recognition », *Proceedings of COLING-2002*, p. 390-396, 2002.

Kang N., van Mulligen E., Kors J., « Comparing and combining chunkers of biomedical text », *Journal of Biomedical Informatics*, vol. 44, n° 2, p. 354-360, nov, 2010.

Keerthi S., Sundararajan S., CRF versus SVM-struct for sequence labeling, Technical report, Yahoo Research, 2007.

Khelif K., Dieng-Kuntz R., « Web sémantique et mémoire d'expériences sur les biopuces », *Web Sémantique Médical (WSM'2004)*, Rouen, 2004.

Kosseim L., Poibeau T., « Extraction de noms propres à partir de textes variés : problématique et enjeux », *in* D. Maurel (ed.), *Actes TALN 2001 (Traitement automatique des langues naturelles)*, ATALA, Université de Tours, Tours, p. 365-371, July, 2001.

Liang T., Shih P.-K., « Empirical textual mining to protein entities recognition from PubMed corpus », *NLDB'05*, p. 56-66, 2005.

McCallum A., Li W., « Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons », *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, 2003.

Meystre S. M., Haug P. J., « Comparing natural language processing tools to extract medical problems from narrative text », *AMIA Annu Symp Proc*, p. 525-529, 2005.

Meystre S. M., Savova G. K., Kipper-Schuler K. C., Hurdle J. F., « Extracting information from textual documents in the electronic health record : a review of recent research », *Yearb Med Inform*, vol. 35, p. 128-144, 2008.

Poibeau T., « Le repérage des entités nommées, un enjeu pour les systèmes de veille », *Terminologies Nouvelles (actes du colloque Terminologie et Intelligence Artificielle, TIA'99, Nantes)*, n° 19, p. 43-51, 1999.

Pratt W., Yetisgen-Yildiz M., « Concept Identification : MetaMap vs. People », *AMIA Annu Symp Proc*, 2003.

Proux D., Rechenmann F., Julliard L., Pillet V., Jacq B., Miyano S., Takagi T., « Detecting Gene Symbols and Names in Biological Texts : A First Step toward Pertinent Information Extraction », *Proceding of Genome Informatics*, Tokyo, Japan : Universal Academy Press, p. 72-80, 1998.

Raymond C., Wei W., « Named entity recognition using hybrid machine learning approach », *IEEE ICCI*, p. 578-583, 2006.

Rindflesch T. C., Tanabe L., Weinstein J. N., Hunter L., « EDGAR : Extraction of Drugs, Genes And Relations from the Biomedical Literature », *Proceedings of Pacific Symposium on Biocomputing*, p. 517-528, 2000.

Rosario B., Hearst M. A., « Classifying Semantic Relations in Bioscience Text », *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, July, 2004.

Sager N., Lyman M., Nhàn N. T., Tick L. J., « Medical language processing : applications to patient data representation and automatic encoding », *Meth Inform Med*, vol. 34, n° 1–2, p. 140-146, 1995.

Sekine S., « Definition, dictionaries and tagger of Extended Named Entity hierarchy », *Proc. of LREC*, Lisbon, Portugal, 2004.

Shadow G., MacDonald C., « Extracting Structured information from free text pathology reports », *AMIA Annu Symp Proc*, Washington, DC, 2003.

Tellier I., Tommasi M., « Champs Markoviens Conditionnels pour l'extraction d'information », *in* E. Gaussier, F. Yvon (eds), *Modèles probabilistes pour l'accès à l'information textuelle*, Hermés, Paris, 2010.

Terol R. M., Martínez-Barco P., Palomar M., « A knowledge based method for the medical question answering problem », *Computers in Biology and Medicine*, vol. 37, n° 10, p. 1511-1521, 2007.

Uzuner O., South B. R., Shen S., Duvall S. L., « 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text », *Journal of the American Medical Informatics Association*, vol. 18, n° 5, p. 552-556, Sep-Oct, 2011. Epub 2011 Jun 16.

Vintar S., Buitelaar P., Volk M., « Semantic relations in concept-based cross-language medical information retrieval », *ECML/PKDD Workshop on adaptive text extraction and mining (ATEM)*, Cavtat-Dubrovnik, 2003.

Wang X., « Rule-based protein term identification with help from automatic species tagging », *Proceedings of CICLING 2007*, p. 288-298, 2007.

Yeh A., Morgan A., Colosimo M., Hirschman L., « BioCreAtIvE task 1A : gene mention finding evaluation. », *BMC Bioinformatics*, 2005.

Zweigenbaum P., « L'UMLS entre langue et ontologie : une approche pragmatique dans le domaine médical », *Revue d'Intelligence Artificielle*, vol. 18, p. 111-137, 2004.

Zweigenbaum P., Jacquemart P., Grabar N., Habert B., « Building a Text Corpus for Representing the Variety of Medical Language », *in* V. L. Patel, R. Rogers, R. Haux (eds), *Actes Medinfo 2001*, Londres, p. 290-294, 2001.