Extraction d'information automatique en domaine médical par projection inter-langue : vers un passage à l'échelle

Asma Ben Abacha* Pierre Zweigenbaum* Aurélien Max*
(*) LIMSI-CNRS, BP 133 91403 Orsay cedex
abacha@limsi.fr, pz@limsi.fr, aurelien.max@limsi.fr

RÉSUMÉ

Cette recherche est issue de notre volonté de tester de nouvelles méthodes automatiques d'annotation ou d'extraction d'information à partir d'une langue L1 en exploitant des ressources et des outils disponibles pour une autre langue L2. Cette approche repose sur le passage par un corpus parallèle (L1-L2) aligné au niveau des phrases et des mots. Pour faire face au manque de corpus médicaux français annotés, nous nous intéressons au couple de langues (français-anglais) dans le but d'annoter automatiquement des textes médicaux en français. En particulier, nous nous intéressons dans cet article à la reconnaissance des entités médicales. Nous évaluons dans un premier temps notre méthode de reconnaissance d'entités médicales sur le corpus anglais. Dans un second temps, nous évaluons la reconnaissance des entités médicales du corpus français par projection des annotations du corpus anglais. Nous abordons également le problème de l'hétérogénéité des données en exploitant un corpus extrait du Web et nous proposons une méthode statistique pour y pallier.

ABSTRACT

Automatic Information Extraction in the Medical Domain by Cross-Lingual Projection

This research stems from our willingness to test new methods for automatic annotation or information extraction from one language L1 by exploiting resources and tools available to another language L2. This approach involves the use of a parallel corpus (L1-L2) aligned at the level of sentences and words. To address the lack of annotated medical French corpus, we focus on the French-English language pair to annotate automatically medical French texts. In particular, we focus in this article on medical entity recognition. We evaluate our medical entity recognition method on the English corpus and the projection of the annotations on the French corpus. We also discuss the problem of scalability since we use a parallel corpus extracted from the Web and propose a statistical method to handle heterogeneous corpora.

MOTS-CLÉS : Extraction d'information, projection d'annotation, reconnaissance des entités médicales, apprentissage.

KEYWORDS: Automatic Information Extraction, Annotation Projection, Medical Entity Recognition, Machine Learning.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 15–28,
Grenoble, 4 au 8 juin 2012.

1 Introduction

L'extraction d'information vise à extraire automatiquement à partir de textes des informations structurées pertinentes pour une tâche particulière (Poibeau, 2003). Il y a essentiellement deux types de méthodes utilisées en extraction d'information : les méthodes où une personne (un « expert ») fournit des connaissances (linguistiques ou sur le domaine) [1], et les méthodes dirigées par les données, où ces connaissances sont construites par apprentissage supervisé. Il existe également des méthodes hybrides combinant ces deux techniques. Ces deux types de méthodes ont certaines limitations ((Bach et Badaskar, 2007), (Nadeau et Sekine, 2007), (Ben Abacha et Zweigenbaum, 2011c)) :

– Les méthodes à base de connaissances expertes sont simples à mettre en place mais coûteuses en temps pour ce qui est de la construction des connaissances. Elles ont aussi un potentiel de couverture réduit comparé aux méthodes statistiques.
– Les méthodes par apprentissage peuvent être très robustes si (i) on dispose d'un bon nombre d'exemples d'entraînement et si (ii) le corpus de test est du même type que le corpus d'entraînement. Ces méthodes sont de fait dépendantes des données et des corpus annotés, ressources qui ne sont pas disponibles pour toutes les langues (par exemple, il n'existe pas de corpus médicaux annotés en français) ni pour toutes les tâches (par exemple, reconnaissance des entités médicales, extraction de relations sémantiques, etc.).

Cette observation s'applique aussi au domaine médical : pour l'anglais, plusieurs outils spécialisés d'extraction d'information existent (tels que MetaMap (Aronson, 2001), cTAKES (Guergana K Savova et Chute, 2010)), ainsi que des corpus annotés en entités nommées (tels que i2b2 (Uzuner *et al.*, 2011), Berkeley (Rosario et Hearst, 2004)). En revanche, peu de ressources sont disponibles en français : on ne trouve pas d'outils spécialisés pour l'extraction d'information, ni de corpus médicaux annotés.

L'annotation manuelle d'exemples pour l'entraînement peut être une solution pour les méthodes par apprentissage supervisé ou semi-supervisé. Cependant, cette tâche nécessite des experts du domaine ciblé, au moins pour la validation. D'après nos expériences précédentes portant sur l'annotation manuelle de corpus médicaux en français constitués (i) de résumés d'articles scientifiques et (ii) de textes extraits du corpus EQueR (Ayache, 2005), plusieurs obstacles ont été mis en évidence. Dans une première phase, nous avons annoté manuellement des textes médicaux avec le concours de deux médecins. L'obstacle principal était le fait que la tâche est longue et fastidieuse. Ensuite, et pour accéler la tâche d'annotation, nous avons développé une interface pour l'annotation de phrases (et non pas de textes entiers) permettant à davantage de médecins de prendre part à l'annotation. Le premier inconvénient de cette méthode est la perte du contexte des phrases. Un deuxième inconvénient réside dans le fait que, même si le guide d'annotation est très détaillé, les divergences dans les avis des médecins augmentent avec le nombre de médecins intervenant (par exemple dans l'annotation des symptômes et des relations dans des textes dans le domaine psychiatrique). Ces divergences, portant par exemple sur les types d'entités médicales et les relations à annoter, peuvent ralentir le processus d'annotation manuelle et le rendre moins fiable.

Dans cet article, nous exploitons un autre type de méthode, la *projection d'annotations* d'une langue à une autre (Yarowsky et Ngai, 2001), et testons son application au domaine médical. L'idée générale consiste à transférer des annotations d'une langue L1 (pour laquelle plus de

1. Méthodes souvent appelées improprement « à base de règles ».

ressources sont disponibles) à une langue L2 en utilisant des corpus parallèles et leur alignement au niveau des mots. Cette approche devrait nous permettre d'exploiter, pour l'annotation automatique de textes en français, les ressources disponibles en anglais ainsi que les méthodes d'extraction d'information développées pour cette même langue. Notre premier objectif, présenté à travers cet article, consiste à annoter automatiquement les entités médicales de textes en français par transfert d'entités détectées dans les textes anglais correspondants par des outils existants de reconnaissance d'entités médicales. La table 1 présente un exemple de ce que nous cherchons à obtenir.

Phrase en anglais	The role of carotid endarterectomy in the management of asymptomatic carotid stenosis is much less clear.
Phrase équivalente en français	Le rôle de l'endartériectomie carotidienne dans le traitement d'une sténose carotidienne asymptomatique est beaucoup moins clairement défini.
Alignement au niveau des mots	0-0 1-1 2-2 3-3 3-4 4-3 5-5 6-6 7-7 8-8 9-11 10-8 11-9 11-10 12-12 13-13 14-14 15-15 15-16
Entités médicales (en anglais)	"carotid endarterectomy" 3–4 [treatment] "asymptomatic carotid stenosis" 9–11 [problem]
Résultat final (annotations en français)	"l'endartériectomie carotidienne " 3–4 [treatment] "une sténose carotidienne asymptomatique" 8-11 [problem]

TABLE 1 – Exemple illustratif de l'approche proposée

Après un rappel des travaux similaires (section 2), nous présentons les deux étapes principales de l'approche que nous proposons ici (telle qu'illustrée sur la figure 1) :

1. L'extraction d'information à partir de la partie L1 du corpus parallèle, en utilisant des méthodes déjà développées ou des outils disponibles (section 3). Pour ce faire, nous utilisons une méthode à base de connaissances expertes, MetaMapPlus (section 3.2) que nous adaptons pour traiter des corpus hétérogènes (section 3.3).
2. L'alignement des mots des parties L1 et L2 du corpus (section 4.1) et la projection des entités repérées sur L1 vers la partie L2 en utilisant ces alignements (section 4.2). Nous mettons en place quelques heuristiques pour réparer certaines erreurs et améliorer la précision de la projection en diminuant le bruit des alignements.

Nous évaluons notre approche (section 5) sur une partie du corpus Santé Canada [2] et discutons ses résultats, puis concluons (section 6) sur des perspectives de travaux futurs [3].

2 Travaux similaires

Des travaux sur la projection d'analyses linguistiques ou d'annotations d'une langue à l'autre se sont développés essentiellement à partir des années 2000. Yarowsky et Ngai (2001) ont proposé

2. http://www.hc-sc.gc.ca
3. Ce travail a été partiellement soutenu par OSEO dans le cadre du programme Quæro.

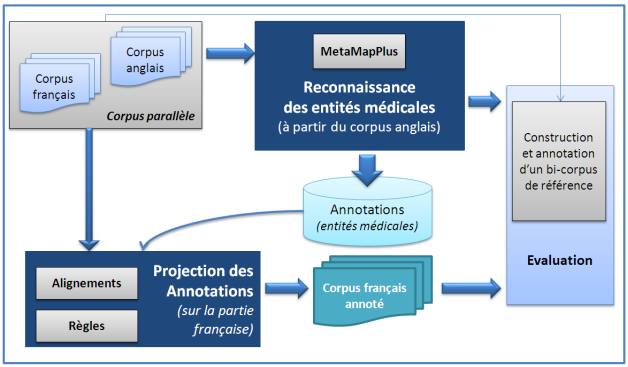

FIGURE 1 – Approche proposée pour l'annotation automatique d'un corpus médical français, utilisant un corpus parallèle et des méthodes d'extraction d'information à partir de textes anglais

d'utiliser des corpus parallèles alignés au niveau des mots pour transférer de façon robuste de l'anglais au français ou au chinois des étiquettes morphosyntaxiques et des frontières de syntagmes nominaux. Lopez *et al.* (2002) ont étudié comment transférer un arbre de dépendances de l'anglais vers le chinois. Lü *et al.* (2002) se sont également intéressés au transfert d'analyses syntaxiques de l'anglais vers le chinois Padó et Pitel (2007) ont traité le problème de l'annotation automatique de rôles sémantiques dans une langue ne disposant pas de lexique FrameNet [4], en s'intéressant au couple de langue (anglais-français).

Dans le domaine médical, plusieurs travaux ont attaqué le transfert de connaissances d'une langue à une autre (Névéol *et al.*, 2005; Deléger *et al.*, 2006). En particulier, Deléger *et al.* (2009) se sont intéressés à l'acquisition de nouvelles traductions de termes issues de trois terminologies différentes («MeSH», «SNOMED CT» et «the MedlinePlus Health Topics»). Ces auteurs se sont basés sur l'alignement des mots à partir d'un corpus parallèle anglais-français.

3 Reconnaissance d'entités médicales dans des textes anglais

Dans cette section, nous présentons la tâche de Reconnaissance des Entités Médicales (REM) (section 3.1). Ensuite, nous décrivons notre méthode à base de connaissances expertes pour la REM à partir de textes anglais (section 3.2). L'application de cette méthode sur des grands volumes de données hétérogènes a révélé certains problèmes liés à l'ambiguité de certains termes. Nous proposons dans la section 3.3 une solution pour pallier cette ambiguité : un filtre statistique utilisé en amont pour améliorer la précision des entités extraites. En effet, en vue de la projection vers un autre corpus à des fins de détection d'entités correctes dans la langue cible, il est fortement souhaitable que les entités du corpus source soient correctes : c'est ce que vise à obtenir le filtrage mis en place, qui privilégie la précision par rapport au rappel.

4. `http://framenet.icsi.berkeley.edu/fndrupal/`

3.1 Description de la tâche de reconnaissance d'entités médicales

La REM est la tâche de base de l'extraction d'information à partir de textes médicaux. Nous désignons par « entité médicale » une instance d'un concept médical ou une catégorie générique (par exemple, *l'Alzheimer* est une instance de la catégorie « Maladie », *la laryngoscopie* est une instance de « Examen »). Cette définition soulève deux questions : (i) quelle est la liste des catégories médicales traitées (Problème médical, Examen, Traitement, etc.) et (ii) quelle est la définition exacte de chaque catégorie (par exemple, les plantes peuvent-elles être considérées comme des traitements ?). Dans cet article nous travaillons sur les trois grandes catégories les plus importantes dans le domaine médical, à savoir : « Problème », « Traitement » et « Examen ». Nous utilisons les types sémantiques de l'UMLS [5] pour définir chaque catégorie (cf. la table 2), en suivant le guide d'annotation i2b2/VA 2010 (Uzuner *et al.*, 2011).

Catégorie	Types sémantiques de l'UMLS correspondants
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
Traitement	Medical Device, Drug Delivery Device, Clinical Drug, Steroid, Pharmacologic Substance, Antibiotic, Biomedical or Dental Material, Therapeutic or Preventive Procedure
Examen	Laboratory Procedure, Diagnostic Procedure.

TABLE 2 – Les catégories médicales traitées

La reconnaissance des entités médicales consiste en (i) le repérage des termes médicaux dans les textes (tels que *beta cell replacement*, *pyogenic liver abscess*, *infection of biliary system*, etc.) et (ii) l'identification de la catégorie sémantique des termes repérés (telles que *Maladie*, *Médicament*, *Examen*, etc.). L'exemple suivant illustre les résultats de la REM pour une phrase extraite d'un résumé MEDLINE. Les termes médicaux sont annotés par des étiquettes *<Treatment>* et *<Disease>*.

<Treatment> Adrenal-sparing surgery </Treatment> is safe and effective , and may become the treatment of choice in patients with <Disease> hereditary phaeochromocytoma </Disease>. [PMID : 10027369]

Ces deux étapes amènent à effectuer des choix dans les catégories médicales à traiter, mais également dans les règles de délimitation des frontières des entités médicales dans le texte. Dans ce travail, nous avons effectué les choix suivants pour la délimitation des frontières : (i) inclure les possessifs, adjectifs, adverbes et chiffres dans les entités nommées (ii) annoter les abréviations séparément et (iii) ne pas annoter une entité médicale incluse dans une autre.

3.2 La méthode MetaMapPlus pour la reconnaissance d'entités médicales

Le domaine médical dispose de grandes bases terminologiques telles que l'UMLS (McCray et Nelson, 1995) ainsi que d'outils qui permettent de détecter les termes médicaux. Un des outils

5. L'UMLS (Unified Medical Language System) comporte (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.

les plus largement utilisés est MetaMap (Aronson, 2001), un système à base de connaissances qui se fonde sur l'UMLS. MetaMap permet de segmenter les textes médicaux en phrases et syntagmes nominaux qui correspondent à des termes médicaux. L'outil identifie les entités médicales et leurs catégories (concepts et types sémantiques du réseau sémantique UMLS). Cependant l'étude de l'utilisation simple de MetaMap a révélé qu'il présente certains problèmes. Afin d'améliorer la précision des résultats de MetaMap, nous avons proposé la méthode MetaMapPlus (Ben Abacha et Zweigenbaum, 2011a), qui comporte les quatre étapes suivantes :

– Extraire les syntagmes nominaux à l'aide d'un segmenteur (*chunker*). Nous utilisons TreeTagger-chunker qui offre une bonne segmentation et permet de diminuer le bruit de la REM (voir (Ben Abacha et Zweigenbaum, 2011c) pour une comparaison de trois segmenteurs).
– Filtrer les syntagmes candidats avec une liste de *mots vides* en amont de MetaMap.
– Rechercher les termes candidats dans des listes d'entités médicales construites à partir du Web.
– Pour le reste des termes candidats, déterminer leurs catégories avec MetaMap, après un filtrage par une liste des erreurs les plus fréquentes de MetaMap et en contraignant les types sémantiques utilisés par ce dernier.

Les résultats de MetaMapPlus, mesurés sur le corpus i2b2 (rappel de 48,68 %, précision de 56,46 % et F-mesure de 52,28 %), sont significativement meilleurs que ceux de MetaMap (F-mesure de 15,80 %) mais restent limités à cause de la performance du procédé de segmentation. L'approche a cependant permis d'identifier le type correct pour 52,28 % des entités, sachant que seules 60,76 % des entités ont été extraites correctement par le segmenteur (i.e. avec des frontières correctes).

En appliquant la méthode MetaMapPlus sur un grand corpus médical extrait du web, nous avons constaté que des ambiguïtés lexicales apparaissaient plus souvent. En effet, plusieurs termes généraux sont considérés par MetaMap comme des entités médicales. Cette ambiguïté peut être divisée en deux catégories principales : (i) les homonymes (e.g. ten, qui désigne dix en domaine ouvert et la maladie « Toxic Epidermal Necrolysis » en domaine médical) et (ii) les termes généraux ayant un sens qui se spécialise dans le domaine médical (e.g. case, form). Ces ambiguïtés causent du bruit dans la reconnaissance d'entités médicales.

3.3 Traitement des ambiguïtés entre acception générale et spécialisée

Pour résoudre le problème d'ambiguïté lexicale, nous proposons une étape supplémentaire intégrée à la méthode MetaMapPlus. Cette étape (appelée Maxent_SNG) consiste à utiliser un classifieur pour distinguer les termes médicaux et les termes généraux, avant d'appliquer MetaMap. Le but de ce module est de :

1. Réduire le bruit lié à l'ambiguïté lexicale, en éliminant les syntagmes nominaux (SN) « généraux » fréquents en domaine ouvert même lorsqu'ils sont utilisés dans le domaine médical (par exemple « table »).
2. Réduire le volume à traiter par la catégorisation via MetaMap en éliminant une bonne partie des syntagmes nominaux à classifier, ce qui devrait réduire le temps d'exécution.

Les méthodes statistiques à base d'apprentissage supervisé peuvent être très robustes. Cependant, ces méthodes présentent deux inconvénients importants :

1. La dépendance aux données annotées disponibles (cf. (Ben Abacha et Zweigenbaum,

2011c)), ce qui constitue un obstacle à l'utilisation de ce type de méthodes pour des tâches et domaines pour lesquels on ne dispose pas de corpus annotés, considérant en outre que la constitution de ces corpus est une tâche coûteuse.

2. Le problème de portabilité sur des corpus différents de ceux utilisés en entraînement (cf. (Ben Abacha et Zweigenbaum, 2011c)), la dégradation des performances une fois appliquées sur des corpus ayant des caractéristiques différentes de ceux utilisés pour l'entraînement constitue un obstacle pour le passage à l'échelle de ces méthodes.

Ces deux inconvénients constituent un important défi pour la mise en place d'une méthode statistique efficace et portable. Pour différencier les données d'entraînement (ce qui offrira une meilleur *adaptabilité*) et éviter le sur-apprentissage (en apprenant correctement et non pas « par coeur »), nous traitons deux problèmes : (i) comment choisir les exemples d'entraînement ? (ii) et quels sont les attributs à utiliser ?

3.3.1 Sélection des données d'apprentissage

À l'instar des travaux sur l'apprentissage actif (Active Learning) (Thompson *et al.*, 1999; Tomanek et Olsson, 2009) qui sélectionnent des exemples diversifiés et représentatifs à annoter manuellement, nous avons trouvé utile de sélectionner les exemples à utiliser pour « bien » apprendre. Deux questions clés se posent alors :
– le nombre des exemples *positifs* et *négatifs* à utiliser ;
– le choix de ces exemples qui doivent être *représentatifs*.
Pour choisir ces exemples, nous proposons d'utiliser :

1. la fréquence des mots/syntagmes nominaux (positifs et négatifs) dans un même corpus ;
2. la présence des mots/syntagmes nominaux (positifs et négatifs) dans des corpus textuels médicaux de genres différents ;
3. le Web pour collecter des données (des exemples positifs et négatifs).

Plus précisément, pour la construction des données d'apprentissage pour le module qui permet de classifier les syntagmes nominaux (SN) en entités médicales (EM) ou termes généraux (SNG), nous utilisons les exemples positifs et négatifs suivants :

1. Exemples positifs : entités médicales
   – les EM les plus fréquentes dans le corpus i2b2 de textes cliniques ;
   – les EM les plus fréquentes dans le corpus Berkeley d'articles scientifiques (Rosario et Hearst, 2004) ;
   – les EM communes aux deux corpus ;
   – des EM extraites du Web (notamment de Wikipedia [6], HON [7]) ;
2. Exemples négatifs : SN « généraux » (SNG) qui ne correspondent pas à des entités médicales :
   – les SNG les plus fréquents dans le corpus i2b2 ;
   – les SNG les plus fréquents dans le corpus de Berkeley ;
   – les SNG les plus fréquents qui existent dans ces deux corpus ;

6. Différentes listes d'entités médicales ont été extraites à partir de Wikipedia : medical tests, diseases, disorders, treatments, procedures (diagnostiques, therapeutiques, surgicales,..)
7. HON (Health On the Net) : `http://www.hon.ch/`

– des SNG extraits du Web, à partir de sites thématiquement distant du domaine médical. Nous avons choisi des sites d'histoires pour enfants [8] [9]). Notre motivation est d'utiliser des corpus ne contenant pas ou peu d'entités médicales.

La table 3 décrit les types d'exemples positifs et négatifs que nous avons utilisés, selon trois critères : corpus, nombre d'exemples et nombre d'occurrences de chaque exemple.

	Corpus	Nb d'exemples	Fréquence des exemples
Exemples positifs	extraits du Web (Wikipedia, HON, etc.)	1 114	entre 1 et 3
	corpus médical 1 (i2b2 : textes cliniques)	3 974 (sur 26 187 EM)	>= 3 (allant jusqu'à 347 pour « hypertension »)
	corpus médical 2 (Berkeley : articles scientifiques)	391 (sur 2 463 EM)	>=2 (allant jusqu'à 28 pour « chemotherapy »)
	Total	**5 479** entités médicales	
Exemples négatifs	extraits du Web (sites d'histoires pour enfants)	2 127	1 et 2
	corpus médical 1 (i2b2 : textes cliniques)	2 031 (sur 15 882 SN)	>= 3 (allant jusqu'à 855 pour « the patient »)
	corpus médical 2 (Berkeley : articles scientifiques)	1 639 (sur 10 464 SN)	>= 2 (allant jusqu'à 278 pour « patients »)
	Total	**5 797** syntagmes nominaux (généraux)	

TABLE 3 – Classification des syntagmes nominaux en termes médicaux et termes généraux, et sélection des exemples positifs et négatifs selon trois critères : corpus, nombre d'exemples et nombre d'occurrences de chaque exemple.

3.3.2 Attributs utilisés par le classifieur

Pour cette tâche, nous utilisons un classifieur à maximum d'entropie [10]. Pour chaque syntagme nominal (médical ou général), les attributs utilisés par le classifieur sont :
– la longueur du SN, son nombre de tokens ;
– le SN est un mot en majuscules / le SN est en majuscules / le SN contient un mot en majuscules ;
– les mots / lemmes / catégories syntaxiques du SN ;
– la présence et la fréquence des mots du SN dans la liste des mots du corpus général BNC [11] ;
– la présence des mots du SN dans un dictionnaire général (nous avons utilisé le dictionnaire standard du système d'exploitation Linux).

8. `http://www.goodnightstories.com/read/pnkbook1.htm`
9. `http://www.vtaide.com/png/stories.htm`
10. Nous avons utilisé l'implémentation disponible à : `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`.
11. British National Corpus, `http://www.natcorp.ox.ac.uk/`.

4 Projection des annotations sur des textes en français par alignement

4.1 Alignement au niveau des mots

La projection que nous réalisons se fonde sur des alignements calculés au niveau des mots. Pour les obtenir, nous avons utilisé les programmes d'alignement de corpus parallèles du système de traduction statistique MOSES (Koehn *et al.*, 2007), en utilisant le paramétrage par défaut. Celui-ci utilise l'outil GIZA++ (Och et Ney, 2004), qui calcule des modèles statistiques d'alignement de mots de complexité croissante. L'alignement est réalisé dans les deux directions puis ses résultats sont *symétrisés*. Finalement, des *tables de traduction*, qui regroupent l'ensemble des bi-segments pouvant être extraits du corpus, sont construites par application d'heuristiques d'extraction de bi-segments *cohérents*, qui imposent que tout mot d'un segment dans un langue doit être aligné avec au moins un mot du segment dans l'autre langue, mais avec aucun mot en dehors de celui-ci.

4.2 Projection

La figure 2 présente quelques exemples de bi-phrases alignées au niveau des mots.

FIGURE 2 – Exemples : trois bi-phrases alignées au niveau des mots

Pour projeter les annotations, nous utilisons le principe suivant :

Soit $E_1 = \{m_{11}, \ldots, m_{1n}\}$ l'ensemble des mots constituant une entité médicale dans le corpus anglais et $E_2 = \{m_{21}, \ldots, m_{2p}\}$ l'ensemble des mots constituant la projection de E_1 (i.e. l'union des projections de chaque mot dans E_1). En notant $position(m_i)$ la fonction retournant la position d'un mot dans sa phrase, nous considérons que la projection de l'entité médicale anglaise est la séquence ordonnée de mots $E_3 = \{m_{31}, \ldots, m_{3k}\}$ telle que :
– $position(m_{31}) = Min_{m_{2i} \in E_2}(position(m_{2i}))$
– $position(m_{3k}) = Max_{m_{2i} \in E_2}(position(m_{2i}))$

Le bruit produit par l'alignement et les annotations affecte la qualité des annotations projetées. Pour diminuer ce bruit et améliorer la phase de projection, (i) nous définissons des heuristiques (telles que la longueur de l'entité trouvée en français par rapport à l'entité originale en anglais) et (ii) nous utilisons un *antidictionnaire* [12] pour filtrer les entités médicales obtenues et supprimer les « mots vides ».

12. `http://members.unine.ch/jacques.savoy/clef/index.html`

5 Expérimentations et évaluation

Le corpus utilisé pour les expérimentations a été construit à partir du site bilingue « Santé Canada [13] » aligné au niveau des phrases (Deléger *et al.*, 2009). La table 4 présente le corpus parallèle (anglais-français) Santé Canada.

	Corpus anglais	**Corpus français**
Nombre de lignes	395600	395600
Nombre de mots	4 465 672	5 052 543
Nombre de caractères	29 845 733	33 901 471
Nombre de mots par ligne (en moyenne)	11	13
Nombre de caractères par mot (en moyenne)	7	7

TABLE 4 – Le corpus parallèle Santé Canada

5.1 Construction et annotation manuelle d'un bi-corpus de référence

Pour évaluer notre approche, nous avons besoin d'un bi-corpus de référence annoté. Deux éléments sont à déterminer : (i) la taille du corpus de référence et (ii) la manière de choisir ce corpus à partir du corpus initial Santé Canada. Nous nous sommes pour cela basés sur des travaux en statistiques.

Taille du corpus. Pour déterminer la taille (acceptable) du corpus de référence à sélectionner, nous utilisons la formule utilisée en statistiques (Sim et Wright, 2005) pour déterminer la taille d'un échantillon :

$$N = \frac{T^2 P(1-P)}{E^2}$$

$$\begin{cases} N & = & \text{La taille de l'échantillon attendu.} \\ T & = & \text{Niveau de confiance déduit du taux de confiance} \\ & & \text{(traditionnellement 1,96 pour un taux de confiance de 95 \%).} \\ P & = & \text{Proportion estimée de la « population » présentant la caractéristique} \\ & & \text{étudiée dans l'étude. Lorsque cette proportion est ignorée, une pré-} \\ & & \text{étude peut être réalisée ou sinon p = 0,5 sera retenue.} \\ E & = & \text{Marge d'erreur (traditionnellement fixée à 5 \%).} \end{cases}$$

Nous fixons les valeurs suivantes : $P = 0{,}5$, $T = 1{,}96$ et $E = 0{,}05$, ce qui nous donne la valeur : $\mathbf{N = 385}$.

Sélection du corpus. Différentes méthodes sont possibles, telles que l'échantillonnage aléatoire simple (*simple random sampling*) ou l'échantillonnage stratifié (*stratified sampling*). Nous avons choisi d'utiliser l'échantillonnage aléatoire simple et sélectionné aléatoirement 385 phrases, contenant 4 613 mots.

Annotation manuelle du corpus de référence. Nous avons annoté manuellement les 385 phrases sélectionnées avec trois types de catégories médicales : Traitement, Problème et Maladie. Nous avons annoté les deux parties françaises et anglaises du corpus de référence en utilisant le guide d'annotation de i2b2 2010 (tâche 1), en respectant les règles de délimitation des frontières décrites dans la section 3.1.

13. `http://www.hc-sc.gc.ca`

5.2 Évaluation de l'annotation du corpus anglais

Dans cette section, nous évaluons la REM à partir du corpus anglais. Nous différencions le cas où les entités médicales ont été reconnues avec des frontières précises ou exactes et le cas où les frontières ne sont pas précises (par exemple, «as antimicrobial resistance» au lieu de «antimicrobial resistance», «Pap smear» au lieu de «a Pap smear» ou «the Pap smear» dans le texte). Nous utilisons les mesures standard de rappel, de précision et de F-mesure.

Nous avons entraîné le module Maxent_SNG de classification des syntagmes nominaux en entités médicales et entités générales sur le corpus d'entraînement i2b2 et nous l'avons testé sur le corpus de test i2b2. Nous avons obtenu une correction (proportion d'exemple de test correctement classés) de 90,99 % (16 169/17 769). La table 5 présente la contribution à la méthode MetaMapPlus de ce module (Maxent_SNG), entraîné sur le corpus d'entraînement décrit dans la section 3.3.

MetaMapPlus						**Maxent_SNG + MetaMapPlus**					
Frontières strictes			Frontières larges			Frontières strictes			Frontières larges		
R	P	F	R	P	F	R	P	F	R	P	F
61,36	22,37	32,79	82,26	28,18	41,97	50,00	40,23	**44,58**	59,26	45,98	**51,78**

TABLE 5 – Résultats de la méthode MetaMapPlus sans et avec le module Maxent_SNG sur le corpus Santé Canada (partie anglaise).

Comme attendu, ce filtrage améliore sensiblement la précision des entités médicales détectées, et en dépit d'une baisse importante de la valeur de rappel, la F-mesure connaît une nette augmentation.

5.3 Évaluation de l'annotation du corpus français par projection

Dans cette section, nous évaluons la qualité de la projection des entités médicales extraites du corpus anglais par notre méthode (i.e. Maxent_SNG+MetaMapPlus). Dans un premier temps, nous évaluons uniquement la qualité de la projection (indépendamment des erreurs d'extraction). Pour ce faire, nous étudions la qualité de la projection des entités de référence (i.e. annotées manuellement). Dans un second temps, nous évaluons l'ensemble du processus pour la REM en français (comprenant l'extraction automatique des entités médicales en anglais et leur projection). Le tableau 6 présente les résultats de la projection des entités de référence et les résultats de la projection des entités extraites avec la méthode Maxent_SNG+MetaMapPlus.

5.4 Discussion

Nous avons pu améliorer les résultats de la méthode MetaMapPlus en intégrant le module MaxEnt entraîné sur trois types différents de corpus. Notons que les résultats obtenus en exploitant ces trois types de corpus sont meilleurs que ceux obtenus en entraînant le classifieur sur un ou deux corpus uniquement (ce que nous avons testé mais ne pouvons pas détailler dans cet

Annotation du corpus français : projection des entités médicales extraites						**Annotation du corpus français : projection des entités médicales de référence**					
Frontières strictes			Frontières larges			Frontières strictes			Frontières larges		
R	P	F	R	P	F	R	P	F	R	P	F
22,39	22,90	22,64	43,08	42,75	42,91	44,78	57,69	50,42	67,91	87,50	76,47

TABLE 6 – Évaluation de la projection des entités médicales extraites avec la méthode Maxent_SNG+MetaMapPlus et les entités de référence sur le corpus Santé Canada (partie française).

article). Les résultats sont relativement acceptables (51,78 % de F-mesure) étant donné la complexité de la tâche sur un corpus hétérogène extrait du Web.

Pour la projection, nous avons utilisé les alignements au niveau des mots avec une approche simple qui consiste à prendre l'entité correspondante (projetée) la plus large. Nous avons essayé d'améliorer cette projection en utilisant un antidictionnaire pour filtrer les entités obtenues et quelques heuristiques telles qu'une différence maximale entre la longueur de l'entité initiale et celle de l'entité projetée (cf. table 7).

	Frontières larges	**Frontières strictes**
Projection sans filtrage	79,09 %	**47,15 %**
Projection + antidictionnaire	75,52 %	**49,79 %**
Projection + antidictionnaire + heuristiques	76,47 %	**50,42 %**

TABLE 7 – F-mesure de la projection des entités de référence sans et avec filtrage

Les résultats de la projection des entités médicales extraites sont relativement faibles, mais ceci dépend directement de la performance de la méthode d'extraction d'information (51,78 % de F-mesure, avec frontières larges), qui fixe le plafond de performance atteignable en projetant ses résultats sur le corpus français. Par projection, nous perdons tout de même près de 50 % des extractions correctes dans le corpus anglais. Ceci résulte principalement de la qualité des alignements au niveau des phrases puis des mots. L'alignement au niveau des mots est influencé par la qualité de l'alignement des phrases (cf. les exemples 1 et 2 ci-dessous). En effet, dans certains cas, la phrase correspondante en français n'est pas équivalente à celle en anglais (soit elle est beaucoup plus courte et contient moins d'information, soit elle est beaucoup plus longue), dans d'autres cas elle a un contenu complètement différent ou reste formulée en anglais : cela reflète un problème d'alignement de phrases qu'il nous faudra corriger dans la suite de nos travaux.

Exemple 1 :
- Statement on Immunization for <PB>Lyme Disease</PB>, 2000 (*)
- 0-0 1-1 5-2 4-3 5-3 5-4 5-5 2-7 5-9 5-10 5-11 5-12 6-13 7-14
- Déclaration sur <PB>un schéma révisé pour la vaccination des adolescents contre l'hépatite B</PB>, 2000 (*)

Exemple 2 :
- <PB>Lung Cancer</PB> : Guidelines for processing Specimens and Reporting Tumor Stage (2000)
- 0-0 1-0 2-0 1-1 1-2 3-3 4-6 9-13 8-18 6-23 7-24 10-27
- <PB>Utilisation, aux fins </PB> de la surveillance, des renseignements sur les patients atteints de cancer : Examen systématique des lois, des réglements, des politiques et des lignes directrices (2000)

Il semble que ces deux phrases ne soient pas en relation de traduction, qui peut résulter d'un mauvais appariemment de documents ou entre phrases.

6 Conclusion et perspectives

Nous avons proposé dans cet article une approche pour l'annotation automatique de textes médicaux en français par projection depuis l'anglais, et présenté nos premières expérimentations en REM. L'approche présentée utilise un corpus parallèle aligné au niveau des mots pour projeter les annotations obtenues sur la partie anglaise vers la partie française. L'application de notre méthode de REM sur un grand corpus de données hétérogènes extrait du Web a posé une problématique de passage à l'échelle pour laquelle nous avons proposé une solution qui consiste à intégrer un module de filtrage statistique en amont des entités candidates pour améliorer la précision des entités extraites.

Nous envisageons principalement quatre perspectives à ce travail :
– L'annotation automatique des relations sémantiques dans des textes français en reprenant la méthode présentée. Nous avons déjà développé des méthodes à base de patrons et des méthodes statistiques pour l'extraction de relations sémantiques à partir de textes médicaux en anglais (Ben Abacha et Zweigenbaum, 2011b).
– L'utilisation ou la construction d'autres corpus médicaux parallèles de meilleur qualité.
– L'exploitation de corpus français annotés pour la mise en place de méthodes statistiques pour l'extraction d'information à partir de textes en français.
– L'intégration de ces méthodes d'extraction d'information dans un système de questions-réponses translingue.

Références

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus : the MetaMap program. *In AMIA Annu Symp Proc*, pages 17–21.

Ayache, C. (2005). Campagne EVALDA/EQueR – Évaluation en question-réponse, rapport final. Rapport technique, ELDA, Paris. Available at`http://www.technolangue.net/IMG/pdf/rapport_EQUER_1.2.pdf`.

Bach, N. et Badaskar, S. (2007). A Review of Relation Extraction.

Ben Abacha, A. et Zweigenbaum, P. (2011a). Automatic extraction of semantic relations between medical entities : a rule based approach. *Journal of Biomedical Semantics*, 2(Suppl 5):S4.

Ben Abacha, A. et Zweigenbaum, P. (2011b). A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. *In Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011*, volume 6608 de *Lecture Notes in Computer Science*, pages 139–150, Tokyo, Japan.

Ben Abacha, A. et Zweigenbaum, P. (2011c). Medical entity recognition : A comparison of semantic and statistical methods. *In Actes BioNLP 2011 Workshop*, pages 56–64, Portland, Oregon, USA. Association for Computational Linguistics.

Deléger, L., Merkel, M. et Zweigenbaum, P. (2006). Contribution to terminology internationalization by word alignment in parallel corpora. *In AMIA Annu Symp Proc.*, pages 185–189, Washington, DC.

Deléger, L., Merkel, M. et Zweigenbaum, P. (2009). Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4):692–701. Epub 2009 Mar 9.

GUERGANA K SAVOVA, James J Masanz, P. V. O. J. Z. S. S. K. C. K.-S. et CHUTE, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes) : architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17:507–513.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proceedings of ACL*, Czech Republic.

LOPEZ, A., NOSSAL, M., HWA, R. et RESNIK, P. (2002). Word-level alignment for multilingual resource acquisition. *In Actes LREC Workshop on Linguistic Knowledge Acquisition and Representation : Bootstrapping Annotated Data*, Las Palmas, Spain. ELRA.

LÜ, Y., LI, S., ZHAO, T. et YANG, M. (2002). Learning Chinese bracketing knowledge based on a bilingual language model. *In Proceedings of COLING-2002*, pages 591–598.

MCCRAY, A. T. et NELSON, S. J. (1995). The semantics of the UMLS knowledge sources. *Methods of Information in Medicine*, 34(1/2).

NADEAU, D. et SEKINE, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26. Publisher : John Benjamins Publishing Company.

NÉVÉOL, A., MORK, J., ARONSON, A. et DARMONI, S. (2005). Evaluation of French and English MeSH indexing systems with a parallel corpus. *In AMIA Annu Symp Proc.*, pages 565–9, Washington, DC.

OCH, F. J. et NEY, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4).

PADÓ, S. et PITEL, G. (2007). Annotation précise du français en sémantique de rôles par projection cross-linguistique. *In Actes TALN 2007, Toulouse, France*.

POIBEAU, T. (2003). *Extraction automatique d'information : du texte brut au web sémantique*. Hermès science publications.

ROSARIO, B. et HEARST, M. A. (2004). Classifying semantic relations in bioscience text. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 430–437, Barcelona.

SIM, J. et WRIGHT, C. C. (2005). The kappa statistic in reliability studies : Use, interpretation, and sample size requirements. *Physical Therapy*.

THOMPSON, C. A., CALIFF, M. E. et MOONEY, R. J. (1999). Active learning for natural language parsing and information extraction. *In Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, pages 406–414, Bled, Slovenia.

TOMANEK, K. et OLSSON, F. (2009). A web survey on the use of active learning to support annotation of text data. *In Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 45–48, Boulder, Colorado. Association for Computational Linguistics.

UZUNER, O., SOUTH, B. R., SHEN, S. et DUVALL, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*. [Epub ahead of print].

YAROWSKY, D. et NGAI, G. (2001). Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. *In Actes NAACL 2001*.