Présent, hypothétique, conditionnel ? Annotation du statut des problèmes médicaux dans des comptes-rendus cliniques en français

Anne Garcia-Fernandez, Anne-Laure Ligozat, Delphine Bernhard

Présent, hypothétique, conditionnel ? Annotation du statut des problèmes médicaux dans des comptes-rendus cliniques en français

Anne Garcia-Fernandez *, Anne-Laure Ligozat*,**, Delphine Bernhard*

*LIMSI-CNRS, BP133, 91403 Orsay cedex
prenom.nom@limsi.fr,
http://www.limsi.fr
**ENSIIE, 1 square de la résistance, 91000 Evry

Résumé. Dans le but d'extraire automatiquement le *statut* (présent, absent, possible...) d'un problème médical cité dans un document, l'utilisation d'une ressource annotée est nécessaire. Cet article présente une méthode d'annotation d'assertions dans un corpus français de comptes-rendus médicaux. L'annotation concerne des *concepts* (ou problèmes médicaux), des *catégories d'assertions* (valeurs de vérité associées aux concepts) et des *justifications* du choix d'assertion donné à un concept. Une annotation manuelle a été effectuée en plusieurs phases en observant l'accord inter-annotateur, sur la base d'un guide d'annotation précis. Nous présentons les choix d'annotation effectués, et les difficultés d'annotation, puis les caractéristiques du travail d'annotation et du corpus obtenu. Ce corpus permettra de développer des systèmes d'extraction d'information dans des comptes-rendus médicaux en français.

1 Introduction

Les dossiers médicaux de patients contiennent de nombreuses informations, qui pourraient être exploitées plus largement pour l'aide à la décision médicale. Cependant, l'accès à ces informations n'est pas simple car elle sont sous forme textuelle et ne se sont donc pas structurée. Dans ces documents, plusieurs types d'informations peuvent être extraits, en particulier les problèmes médicaux des patients, c'est-à-dire les maladies et symptômes évoqués dans les dossiers, et dont la détection a fait l'objet de nombreux travaux en extraction d'information. Les problèmes médicaux évoqués dans les dossiers de patients peuvent être de plusieurs natures : un symptôme que le patient ressent actuellement, un antécédent médical, une maladie soupçonnée et testée... Il est donc primordial de connaître la nature des problèmes si l'on veut pouvoir exploiter efficacement ces informations. Dans le but d'extraire automatiquement le *statut* (présent, absent, possible...) d'un problème médical cité dans un document, l'utilisation d'une ressource annotée est nécessaire.

Dans ce travail, nous nous sommes intéressées à la valeur de vérité associée à un problème médical, que nous appellerons *catégorie d'assertion* du problème, selon la terminologie

utilisée dans la campagne i2b2 2010 [1]. Nous avons étudié ce phénomène dans des dossiers médicaux d'hôpitaux français, et avons exploité pour cela le corpus du projet ANR Akenaton [2]. Dans cet article, nous présentons la première phase d'annotation de ce corpus, en présentant et justifiant la méthode d'annotation mise en œuvre, les choix d'annotation effectués, puis détaillons les caractéristiques du corpus obtenu, et évaluons notamment le coût d'annotation de ce corpus, qui pourra servir à entraîner et évaluer des systèmes de reconnaissance automatique des problèmes médicaux et de leurs assertions.

2 État de l'art

Dans le domaine biomédical, les expressions d'incertitude sont particulièrement fréquentes. En effet, elles permettent d'indiquer des impressions, des explications possibles ou des résultats négatifs (Vincze et al., 2008), comme le montrent les exemples suivants :

*These findings that **may be** from an acute pneumonia include minimal bronchiectasis as well.*

*Stable appearance the right kidney **without** hydronephrosis.*

*The treatment **seems to be** successful.*

*Right upper lobe volume loss and **probably** pneumonia.*

Différents niveaux de certitude peuvent être distingués. Le niveau le plus traité est celui de la négation, mais une gradation plus fine peut également être utilisée avec, par exemple, des niveaux signalant une possibilité ou une condition. La détection du niveau de certitude peut se faire soit au niveau global de la phrase, soit à l'intérieur d'une phrase.

La détection du niveau de certitude est importante afin de déterminer le statut des problèmes étudiés. Pour faciliter la construction de systèmes de reconnaissance de ces niveaux et permettre de les évaluer, des corpus ont été construits à partir de textes biomédicaux.

Le corpus BioScope (Vincze et al., 2008) constitue ainsi une ressource pour l'étude de la négation et de l'incertitude dans des textes biomédicaux et permet le développement de systèmes statistiques de détection de ces phénomènes. Ce corpus est constitué de comptes-rendus cliniques, d'articles scientifiques et de résumés. Il contient environ 3 000 documents pour plus de 20 000 phrases. L'annotation a consisté à identifier des marqueurs de négation et d'incertitude et leur portée. Ainsi, dans l'exemple suivant "raises the question of" est annoté comme un marqueur d'incertitude, dont la portée inclut le problème médical "cystitis" :

Mild bladder wall thickening (<raises the question of> cystitis).

Le processus d'annotation a été guidé par une stratégie *min-max* : pour l'annotation des marqueurs d'incertitude, l'unité minimale qui exprime l'incertitude a été privilégiée ; en revanche, pour la portée de ces marqueurs, l'unité syntaxique la plus grande possible a été annotée. L'annotation a été faite indépendamment par deux annotateurs linguistes puis un troisième annotateur linguiste avait pour rôle de résoudre les conflits. Environ 10% des phrases contiennent des négations et environ 15% contiennent des incertitudes.

1. Fourth i2b2/VA Shared-Task and Workshop, Challenges in Natural Language Processing for Clinical Data, https ://www.i2b2.org/NLP/Relations/Main.php

2. Extraction d'information à partir de comptes-rendus cliniques, http ://resmed.univ-rennes1.fr/AKENATON/

Deux autres corpus relativement similaires existent également : le corpus Genia Event (Kim et al., 2008) qui comprend des annotations d'entités biologiques avec leur niveau d'incertitude dans 1 000 résumés d'articles, et le corpus BioInfer (Pyysalo et al., 2007), dans lequel la négation est annotée pour des relations biologiques dans environ 1 100 phrases. Dans ces deux corpus, les catégories d'incertitude ont été annotées, mais pas les marqueurs.

La campagne d'évaluation i2b2 2010 [3] a proposé une tâche de catégorisation des assertions portant sur des concepts médicaux. Un corpus spécifique a été constitué pour cette tâche et a été diffusé aux participants à la campagne. Ce corpus est constitué de comptes-rendus cliniques provenant de plusieurs hôpitaux américains, et comprend environ 800 documents. L'annotation a consisté à identifier les problèmes médicaux et leur catégorie d'assertion. Six catégories ont été définies : *présent*, *absent*, *possible*, *conditionnel*, *hypothétique* et *associé à quelqu'un d'autre*. Ainsi, dans l'exemple suivant, les concepts "pleural effusion" et "pneumothorax" ont été annotés, avec comme catégorie d'assertion "absent" :

No pleural effusion or pneumothorax.

Un guide d'annotation a été défini. Il a servi de base de travail pour l'annotation de notre travail et est détaillé dans la section 3. Lors de la campagne i2b2 2010, l'annotation a été faite par 12 annotateurs, dont des experts médicaux et des non-experts. Le corpus final (entraînement plus évaluation) comprend environ 30 000 concepts avec leur catégorie d'assertion.

Notre objectif est de constituer un corpus similaire à celui utilisé dans i2b2, mais pour le français. Nous nous sommes concentrées sur les catégories d'assertion afin de pouvoir développer, à terme, un système de catégorisation pour le français. Nous sommes donc parties d'un corpus de comptes-rendus cliniques constitué dans le cadre du projet Akenaton, dans lequel les documents ont été anonymisés. Nous avons annoté une partie de ces documents en nous attachant principalement à la définition des catégories d'assertion. Notre objectif est d'obtenir un très bon accord inter-annotateur sur ces catégories avant d'entamer une phase d'annotation plus massive.

3 Notions annotées

Trois notions ont été définies ou redéfinies pour cette tâche : les notions de concept, d'assertion et de justification.

Tout d'abord, nous avons repris la notion de *concept* d'i2b2, restreinte aux problèmes médicaux, c'est-à-dire les "expressions qui contiennent des observations faites par les patients ou le personnel médical à propos du corps ou de l'esprit du patient et qui sont considérées comme anormales ou causées par une maladie". Ces expressions comprennent notamment les maladies et les symptômes, et correspondent approximativement aux types sémantiques UMLS de "pathologic functions", "disease or syndrome", "mental or behavioral dysfunction", "cell or molecular dysfunction", "congenital abnormality", "acquired abnormality", "injury or poisoning", "anatomic abnormality", "neoplastic process", "virus/bacterium", "sign or symptom", mais ne sont pas limitées à la seule couverture de l'UMLS.

Nous avons également réutilisé la notion d'*assertion* qui est définie par i2b2 comme la valeur de vérité donnée à un concept et qui constitue le cœur de notre étude. Sept catégories d'assertion ont été définies :

3. https ://www.i2b2.org/NLP/

Présent Catégorie par défaut ; en particulier, si un antécédent médical est évoqué et est toujours présent ou sans indication de guérison ;

il existe une discrète hypocinésie ventriculaire gauche

Absent Le problème n'existe pas chez le patient. Cette catégorie est un peu différente de celle d'i2b2 car elle n'inclut pas les cas où le problème a existé, mais a disparu, cas dans lesquels la catégorie est "historique" ;

*Il n'y a pas d'*hépatomégalie.

Historique Le patient ne présente plus le problème ;

Dans les antécédents de cette patiente, (...) une chirurgie de cataracte de l'oeil gauche.

Possible Le patient est susceptible de présenter le problème ;

De toutes façons, si il s'agissait de tachycardies paroxystiques*, (...)*

Conditionnel Le patient présente le problème sous certaines conditions, après la prise d'un médicament ou lors d'un effort par exemple ;

une dyspnée *d'effort*

Hypothétique Le problème pourrait être développé par le patient ;

l'hypothèse d'une nouvelle ischémie myocardique *semble peu vraisemblable*

Associé à quelqu'un d'autre Le problème concerne quelqu'un d'autre, ce qui comprend notamment les antécédents familiaux.

un frère et une sœur auraient des problèmes cardiaques

Enfin, nous avons étudié les *justifications*, que nous avons définies comme les portions de texte permettant de déterminer quelle catégorie d'assertion doit être attribuée à un concept.

4 Corpus et méthode d'annotation

Le corpus que nous avons annoté est extrait du corpus Akenaton. Ce corpus est constitué de 20 000 comptes rendus cliniques en cardiologie et a été distribué au LIMSI par l'Équipe d'Accueil 3888 [4] (EA3888). Ce corpus a fait l'objet d'un premier niveau d'anonymisation à la source par l'EA3888 puis d'un second niveau d'anonymisation effectué par le LIMSI (Grouin et al., 2009). Parmi ce corpus, ont été annotés manuellement 50 documents correspondant plus de 1300 phrases et près de 22 000 mots.

L'annotation consiste à repérer les concepts dans les documents, à déterminer pour chacun sa catégorie d'assertion et la justification correspondante. Un concept correspond à un unique

4. http ://www.ea3888.univ-rennes1.fr/

segment textuel continu. Pour chaque concept, une catégorie d'assertion exactement doit être attribuée. Une justification consiste en un segment textuel continu. À chaque concept peut être attribué zéro, une ou plusieurs justifications. Les cas où aucune justification n'est associée à un concept sont le plus souvent dus au style même d'un compte-rendu médical qui peut facilement contenir des phrases nominales (sans verbe principal) comme on peut le voir dans l'exemple suivant [5] :

Régurgitation mitrale grade II. *Péricarde libre*

Les quelques cas où un concept est associé à plusieurs justifications s'expliquent par la présence d'une coordination comme on peut le voir dans l'exemple suivant :

L'interrogatoire ne retrouve pas (...) de *notion de mort subite* *ou de* *trouble du rythme paroxystique* *ni de* *pathologie cardiaque*

D'autre part, une justification peut être attribuée à plusieurs concepts :

Les antécédents de cette patiente sont principalement marqués par *une artérite des membres inférieurs* , *une phlébite* *(...)*

L'outil d'annotation utilisé est Knowtator [6] (Ogren, 2006). Il s'agit d'un plug-in du logiciel de gestion d'ontologie Protégé [7]. Les notions annotées (concepts, assertions et justifications) constituent l'ontologie utilisée pour annoter le corpus. Cet outil permet aussi de calculer l'accord inter-annotateur (IAA) concernant les différents éléments annotés. La mesure utilisée est un calcul du pourcentage d'accord entre annotateur ou l'accord observé. Cette mesure correspond à la proportion d'éléments sur lesquels les annotateurs sont d'accord, c'est-à-dire au nombre total d'éléments pour lesquels les annotateurs sont en accord divisé par le nombre total d'éléments annoté :

$$IAA = \frac{nombre\ d'annotations\ en\ accord}{nombre\ d'annotations\ au\ total} \quad (1)$$

Le travail d'annotation a été effectué par trois annotateurs. L'identification des problèmes médicaux (ou concept) n'est pas le point crucial de notre travail. En effet, nous nous intéressons plus précisément à la catégorie d'assertion de ces concepts. De plus, l'annotation porte aussi sur l'identification de segments de texte permettant de justifier le choix d'une catégorie d'assertion pour un problème donné. Nous n'avons donc pas fait appel à des experts du domaine médical mais à des expert linguistes. L'annotation des documents doit permettre de disposer d'un corpus de référence pour l'étude des assertions dans le domaine médical. Si notre souci n'était pas nécessairement l'exhaustivité au niveau de l'annotation des concepts (ou problèmes médicaux), il était très important d'obtenir une annotation précise concernant les assertions, c'est-à-dire la valeur de vérité associée à ces concepts. Nous avons ainsi travaillé en plusieurs phases : une phase d'apprentissage au cours de laquelle les annotateurs ont annoté les mêmes documents et une phase au cours de laquelle ils ont pu annoter des documents différents. Lors de la première phase, cinq documents a été annotés par les trois annotateurs. L'accord inter-annotateur a été calculé et la première version d'un guide d'annotation a été élaborée. Cette

5. Dans cet exemple et les exemples suivants, les concepts sont encadrés et les justifications sont soulignées.
6. http ://knowtator.sourceforge.net/
7. http ://protege.stanford.edu/

opération a été répétée deux fois : tant que l'accord inter-annotateur concernant l'annotation des assertions n'était pas satisfaisant. Nous avons considéré un accord de 88% comme suffisant. Le tableau 1 résume les accords calculés lors de cette phase d'apprentissage.

Afin de faciliter leur travail d'annotation, les annotateurs avaient accès à un wiki sur lequel ils pouvaient échanger leurs remarques. La page en question a été modifiée plus de 200 fois. Les sous-sections suivantes traitent de ces remarques. Bien que Knowtator offre la possibilité d'ajouter des remarques aux annotations, on constate que cette fonctionnalité, qui ne permet qu'un partage en différé des remarques, n'a été que peu utilisée.

De plus, à partir des annotations des cinq premiers textes et de la terminologie UMLS, nous avons effectué une pré-annotation des textes suivants. La quantification et l'évaluation de cette pré-annotation sont présentés dans la section 5.

Phase	#documents	# moy. de problèmes/doc	#annotateurs	IAAs
1	5	13	3	37%
2	5	19	3	47%
3	5	5	3	89%

TAB. 1 – *Accord inter-annotateur (IAA) calculé lors de la phase d'apprentissage.*

Un guide d'aide à l'annotation a été élaboré par collaboration entre les annotateurs et a permis d'affiner la qualité des annotations. Ce guide concerne la détection des problèmes dans les comptes-rendus médicaux, l'attribution d'une modalité à chaque problème détecté et la sélection d'une portion de texte justifiant ce dernier choix. Il ne constitue pas un guide d'annotation complet mais doit être considéré comme un complément à celui utilisé lors de l'évaluation i2b2 [8].

4.1 Annotation des concepts

Nous avons suivi le guide d'annotation d'i2b2 concernant l'annotation des concepts mais en se limitant à annoter les problèmes médicaux. Ainsi l'annotation d'un concept doit être le groupe nominal ou adjectival maximal désignant le problème médical. Il inclut déterminants et modifieurs (dont les termes tels que "épisode", "séquelle"...) mais pas les adjectifs indiquant la possibilité (comme "probable"), ni les expressions comme "signe de", "tableau de", "séropositif à". D'autre part, un concept ne doit contenir qu'un unique problème. Ainsi on annote séparément les concepts même s'ils font partie du même groupe nominal.

probable [*séquelle postérieure*]

[*des épisodes de lipothymies*], [*quelques extrasystoles auriculaires*], [*une dyspnée*], *séropositif à* [*HIV*]...

Il existe [*des calcifications aortique et mitrale*] *sans* [*rétrécissement significative*] *ni* [*fuite aortique*].

8. https ://www.i2b2.org/NLP/Relations/Documentation.php

4.2 Annotation de la catégorie d'assertion

L'annotation de la catégorie d'assertion d'un problème suit les définitions présentées dans la section 3. Les précisions suivantes viennent compléter ces définitions.

- **Historique** : seuls les problèmes médicaux qui apparaissent comme clairement résolus sont annotés dans cette catégorie. Dans le cas contraire, ils sont considérés comme présents. Ceci est valide aussi pour les antécédents d'un patient. Du point de vue temporel, il faut se placer au moment où le compte-rendu a été rédigé.
- **Conditionnel** : les termes tels que "dyspnée d'effort" ou "orthopnée" qui, par définition, indiquent un problème médical qui a lieu sous certaines conditions doivent être annotés comme conditionnel.
- **Hypothétique** : concerne les problèmes médicaux qui pourraient survenir dans le futur.
- **Possible** : concerne les problèmes médicaux que le patient a peut-être au moment où le compte-rendu a été rédigé.
- **Associé à quelqu'un d'autre** : cette valeur est prévalente sur les autres valeurs. Si un problème, qu'il soit absent ou présent, n'est pas associé au patient (mais à un membre de sa famille par exemple), l'annotation doit prendre la valeur "associé à quelqu'un d'autre".

4.3 Annotation des justifications

L'annotation des justifications n'est pas basée sur le guide d'annotation i2b2 puisqu'elles n'ont pas donné lieu à annotation lors de cette campagne. L'objectif de cette annotation est d'identifier le segment permettant de justifier le choix d'une catégorie d'assertion pour un problème donné. Nous avons soumis l'annotation des justifications aux règles suivantes. Une justification consiste en un groupe syntaxique complet. Elle doit être suffisante pour permettre de justifier le choix de la catégorie d'assertion. Cependant, il peut ne pas y avoir de justification pour un concept donné :

Syndrome inflammatoire *et CRP à 47.*

La justification n'inclut pas le sujet du verbe sauf s'il s'agit d'un pronom impersonnel ou est porteur d'information c'est-à-dire s'il désigne une autre personne que le patient (et permet donc de justifier la catégorie "Associé à quelqu'un d'autre"). En effet, par défaut, on considère qu'un problème est attribué au patient en question, ce n'est donc pas nécessaire de le justifier.

et il n'existe pas de *valvulopathie*

Cette jeune patiente présentait déjà de multiples antécédents : - *artériopathie oblitérante des membres inférieurs sévère*

Sa sœur présente *une insuffisance musculaire des membres inférieurs*

5 Pré-annotation des documents

Nous nous intéressons plus particulièrement dans ce travail à la catégorie d'assertion des problèmes médicaux. L'annotation des problèmes médicaux en elle-même est une tâche supplémentaire pour les annotateurs et elle pourrait par exemple être exécutée automatiquement

en utilisant un système de détection des problèmes médicaux (Grouin et al., 2011).
Afin de faciliter la tâche d'annotation, nous avons pré-annoté les problèmes médicaux dans les documents. Ne disposant pas d'un système de détection des problèmes médicaux, nous avons utilisé des listes de termes désignant des problèmes médicaux et les avons détectés et pré-annotés dans les documents. Ces listes sont issues d'une part du UMLS, d'autre part des annotations déjà effectuées par les annotateurs. Parmi l'ensemble des termes du UMLS, nous avons sélectionné uniquement ceux qui correspondent à des problèmes médicaux c'est-à-dire ceux définis lors de la campagne i2b2 [9]. Nous avons ainsi listé 114 512 termes.
Le tableau 2 montre le nombre de pré-annotations faites sur 45 des textes annotés en se basant soit sur les annotations manuelles des 5 premiers textes traités par les annotateurs, soit sur les listes du UMLS.

	1res annotations	UMLS
# termes dans la liste	55	114 512
# problèmes pré-annotés	180	630
# moy. problèmes par documents	4	14

TAB. 2 – *Quantification des pré-annotations sur 45 documents.*

On observe qu'en se basant sur la liste issue du UMLS, le nombre de pré-annotations est assez important (14 par document en moyenne). Ceci est dû au fait que le UMLS est une base extrêmement complète. Un avantage majeur est qu'elle contient des termes comportant des abréviations tout comme les comptes-rendus médicaux que nous traitons :

"fract ferm du crâne"

En revanche, cette liste comporte des termes très spécifiques comme par exemple :

"rétraction ou déficit d'abaissement de la paupière supérieure dans le regard en bas"

Ces termes ont peu de chance d'apparaître au sein de nos documents, d'autant plus que notre corpus est constitué de comptes-rendus cliniques d'un sous-domaine spécifique du domaine médical (la cardiologie) alors que le UMLS est général. Il serait donc intéressant de filtrer préalablement cette liste afin de l'adapter à notre corpus et d'accélérer le processus de pré-annotation. D'autre part, certains termes sont ambigus et peuvent faire référence à un problème médical dans certains contextes uniquement. Par exemple "faible" est un problème dans "un état général faible" mais pas dans "une faible quantité de diantalvic". Les pré-annotations à partir du UMLS ne sont donc pas adaptées au cadre de notre travail. En effet, nous ne souhaitions pas créer un détecteur de problèmes médicaux et même si une analyse linguistique aurait pu affiner la détection, le but de notre travail était autre.

Les pré-annotations issues des premières annotations sont quant à elles plus limitées en nombre (4 en moyenne par document) mais leur qualité est bien meilleure. Cette constatation n'est pas une surprise puisque les premières annotations portaient sur cinq documents seulement, mais du même corpus.

9. La liste des types sémantiques concernés est donnée section 3

Nous avons évalué l'apport de la pré-annotation dans le travail des annotateurs d'un point de vue quantitatif. Nous avons demandé à deux annotateurs de traiter des comptes-rendus de tailles comparables et contenant un nombre équivalent de pré-annotations. L'un des deux était présenté pré-annoté et l'autre pas. Les annotateurs devaient se chronométrer. Nous avons ainsi observé que sur quatre couples de documents (pour un total de 121 problèmes annotés), le temps moyen d'annotation par problème est supérieur lorsque le document est pré-annoté (35 secondes) que lorsqu'il ne l'est pas (28 secondes). Les annotateurs expliquent le non gain de temps par le fait que, dans les deux situations, l'intégralité du compte-rendu doit être lu et que l'annotation de la catégorie d'assertion du problème ainsi que de la justification doivent être faits. Ils justifient la perte de temps par la présence d'erreurs de pré-annotation (segments pré-annotés alors qu'ils ne le devraient pas ou bien qui n'incluent pas l'intégralité segment désignant le problème) qui impliquent une manipulation supplémentaire pour les supprimer ou les modifier.

6 Analyse du coût des annotations

Nous avons cherché à estimer le coût d'une telle tâche d'annotation. Le tableau 3 montre les estimations de temps calculées. Les durées sont estimées en incluant la phase d'apprentissage. Les temps moyens par annotation et par document sont en revanche calculés sur l'annotation des 35 documents annotés au cours de la seconde phase de travail. Le temps moyen par problème annoté permet d'estimer le coût de la tâche d'annotation dans le cas où l'on dispose d'un corpus pré-annoté à l'aide d'un outil de détection automatique de problèmes médicaux.

#documents	50
Durée totale	44 heures
Durée moyenne par document	9 minutes
Durée min. pour un document	30 secondes
Durée max. pour un document	30 minutes
# annotation	2274
Temps moyen par annotation	12 secondes
Temps moyen par problème annoté	35 secondes

TAB. 3 – *Estimation du temps d'annotation.*

Nous avons vu dans la section 5 que la pré-annotation ne fait pas gagner de temps. Cependant, notre technique de pré-annotation ne vaut pas le travail d'un vrai détecteur de problèmes médicaux.

7 Analyse des annotations

Le tableau 4 indique le nombre d'annotations pour chaque élément annoté (problème, assertion et justification) et détaille le nombre d'annotations pour chaque catégorie d'assertion. Nous y proposons une comparaison des taux de présence des différentes catégories d'assertion pour notre corpus et le corpus d'évaluation utilisé lors de la campagne i2b2 2010.

		Notre corpus			Corpus i2b2
		# moy par document	Total		
Problème		15	775		18 550
Catégorie d'assertion	Présent	10	500	64,5%	70%
	Absent	3	165	21%	19,5%
	Possible	0,7	38	5%	4,5%
	Conditionnel	0,5	26	3,5%	1%
	Historique	0,5	26	3,5%	-
	Associé à quelqu'un d'autre	0,25	12	1,5%	1%
	Hypothétique	0,2	8	1%	4%
Justification		14,5	724		-

TAB. 4 – *Quantification des annotations sur 50 documents en comparaison avec les annotations du corpus d'évaluation utilisé lors de la campagne i2b2 2010.*

Nous observons que les taux d'annotation concernant les catégories d'assertion sont à peu près équivalents dans les deux corpus. La catégorie d'assertion la plus représentée est "présent" : la majorité des problèmes médicaux du corpus sont évoqués parce que le problème est présent chez le patient. La catégorie "historique" non utilisée lors de la campagne i2b2 représente 3,5% des annotations d'assertion. L'ajout de cette catégorie nous semble donc justifié.

On note qu'il y a moins de justifications (724) que de problèmes (775). Ceci reflète les cas où un problème apparaît seul au sein d'une phrase nominale. L'ensemble des justifications annotées, associées à la catégorie d'assertion qu'elles justifient, constitue une base d'exemples (on dénombre plus de 400 segments de texte différents) de justifications pour chaque catégorie d'assertion.

8 Conclusions

Nous avons présenté dans cet article la méthode d'annotation d'un corpus de documents médicaux en français en catégories d'assertion, c'est-à-dire les valeurs de vérités associées à des problèmes médicaux. Le résultat de ce travail est un corpus de 50 documents annotés avec plus de 700 problèmes médicaux, leur catégorie d'assertion associée, et les justifications permettant d'attribuer la catégorie. Nous avons produit un guide d'annotation précis pour cette tâche, et étudié le coût d'annotation de ces documents.

La ressource annotée constituée, unique pour le français, est utilisable comme corpus d'entraînement par un système d'apprentissage automatique permettant ainsi d'avoir un classifieur des problèmes médicaux.

Une étape nécessaire sera de faire valider ces annotations par un expert médical, au moins pour l'annotation des problèmes médicaux, qui nous semble la plus délicate pour des non-experts, même si, notre travail ne se focalisant pas sur ces problèmes, nous avons plutôt choisi de mettre de côté les cas qui nous semblaient ambigus.

Concernant les concepts, nous souhaiterions tester l'utilisation d'un système de reconnaissance de ce type d'entités pour le français, sachant néanmoins qu'un tel outil ne serait pas forcément adapté à notre corpus.

Enfin, notre objectif final est d'adapter au français le système de reconnaissance de catégories d'assertions développé pour l'anglais (Bernhard et Ligozat, 2011), donc nous pensons utiliser ce corpus comme base d'apprentissage d'un tel système. La taille du corpus ne permettant pas *a priori* un apprentissage efficace pour toutes les catégories d'assertion (et le coût d'annotation d'un corpus d'une taille suffisante étant trop élevé), nous envisageons plutôt des techniques d'apprentissage semi-supervisé, comme de l'active learning, pour construire notre système.

9 Remerciements

Ce travail a été partiellement financé par le projet DOXA et le programme QUAERO (financé par OSEO, agence française pour l'innovation). Nous remercions le projet Akenaton pour avoir mis à notre disposition le corpus de comptes-rendus médicaux.

Références

Bernhard, D. et A.-L. Ligozat (2011). Analyse automatique de la modalité et du niveau de certitude : application au domaine médical. In *Actes de TALN 2011*. à paraître.

Grouin, C., L. Deléger, B. Cartoni, S. Rosset, et P. Zweigenbaum (2011). Accès au contenu sémantique en language de spécialité : extraction des prescriptions et concepts médicaux. In *Actes de TALN 2011*. à paraître.

Grouin, C., A. Rosier, O. Dameron, et P. Zweigenbaum (2009). Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers. In M. Fieschi, P. Staccini, O. Bouhaddou, et C. Lovis (Eds.), *Risques, technologies de l'information pour les pratiques médicales*, Volume XVII of *Informatique et santé*. France : Springler-Verlag.

Kim, J., T. Ohta, et J. Tsujii (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics 9*, 10.

Ogren, P. V. (2006). Knowtator : a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Morristown, NJ, USA, pp. 273–275. Association for Computational Linguistics.

Pyysalo, S., F. Ginter, J. Heimonen, J. Bjorne, J. Boberg, J. Jarvinen, et T. Salakoski (2007). Bioinfer : a corpus for information extraction in the biomedical domain. *BMC Bioinformatics 8*, 50. doi :10.1186/1471-2105-8-50.

Vincze, V., G. Szarvas, R. Farkas, G. Mora, et J. Csirik (2008). The bioscope corpus : biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics 9*(Suppl 11), S9.

Summary

In this paper we present a method to annotate assertions in a French corpus of medical reports. The annotation concerns *concepts* (or medical problems), *assertion categories* (truth values associated to a problem), and *justifications* (spans of text which justificate the assertion choice of a problem). A multi-phase manual annotation, based on a annotation guide, has been done observing the inter-annotator aggreement. We present our annotation choices and process. We also evaluate the cost of such an annotation. Finally we present the obtained corpus which is a base to developp information extraction system in French medical reports.