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

Analyse automatique de la modalité et du niveau de certitude : application au domaine médical

Delphine Bernhard[1] Anne-Laure Ligozat[1, 2]
(1) LIMSI-CNRS, B.P. 133, 91403 Orsay Cedex
(2) ENSIIE, 1 square de la résistance, 91000 Évry
bernhard@limsi.fr, annlor@limsi.fr

Résumé. De nombreux phénomènes linguistiques visent à exprimer le doute ou l'incertitude de l'énonciateur, ainsi que la subjectivité potentielle du point de vue. La prise en compte de ces informations sur le niveau de certitude est primordiale pour de nombreuses applications du traitement automatique des langues, en particulier l'extraction d'information dans le domaine médical. Dans cet article, nous présentons deux systèmes qui analysent automatiquement les niveaux de certitude associés à des problèmes médicaux mentionnés dans des compte-rendus cliniques en anglais. Le premier système procède par apprentissage supervisé et obtient une f-mesure de 0,93. Le second système utilise des règles décrivant des déclencheurs linguistiques spécifiques et obtient une f-mesure de 0,90.

Abstract. Many linguistic phenomena aim at expressing the speaker's doubt or incertainty, as well as the potential subjectivity of the point of view. Most natural language processing applications, and in particular knowledge extraction in the medical domain, need to take this type of information into account. In this article, we describe two systems which automatically analyse the levels of certainty associated with medical problems mentioned in English clinical reports. The first system uses supervised machine learning and obtains an f-measure of 0.93. The second system relies on a set of rules decribing specific linguistic triggers and reaches an f-measure of 0.90.

Mots-clés : Modalité épistémique, Niveau de certitude, Domaine médical.

Keywords: Epistemic modality, Certainty level, Medical domain.

1 Introduction

En traitement automatique des langues, les informations contenues dans les textes sont souvent considérées comme affirmées et vérifiées. Or, de nombreux phénomènes linguistiques visent à exprimer le doute ou l'incertitude de l'énonciateur, ainsi que la subjectivité potentielle du point de vue. Il y a ainsi une gradation dans les niveaux de certitude associés à une information : elle peut être vraie, possible ou fausse. Une information peut également n'être vraie que dans certaines conditions, ou être hypothétique.

Dans certains domaines, il est particulièrement important de savoir si l'information donnée dans un document est certaine ou pas. Par exemple dans le domaine médical, si l'on tente d'analyser des relations entre un médicament et des symptômes décrites dans un texte, il est nécessaire de savoir si le symptôme est présent ou pas, ou encore s'il est susceptible d'être développé par le patient. En questions-réponses, notamment sur des corpus de documents issus du web, le niveau de certitude d'une information peut également être utile : une réponse comme «La tour Eiffel est une tour de 327 mètres de hauteur» à la question «Quelle est la taille de la tour Eiffel» est plus précise que la réponse «Je pense que la tour Eiffel fait environ 300 mètres» et devra être considérée comme plus fiable.

Ces divers aspects ne sont encore que très rarement pris en compte dans les applications développées actuellement en traitement automatique des langues, même s'il existe des travaux récents visant à détecter l'incertitude, la modalité épistémique, la spéculation ou encore les opinions.

Dans cet article, nous nous intéressons à l'analyse automatique de la modalité et du niveau de certitude dans le domaine médical. Plus particulièrement, nous nous attachons à étudier ces phénomènes lorsqu'ils portent sur des problèmes médicaux (maladies, syndromes, virus, bactéries, symptômes) mentionnés dans des compte-rendus

cliniques en anglais afin de savoir si le problème est présent, absent, hypothétique, soumis à certaines conditions, ou encore associé à une personne différente du patient dont il est question. Nous utilisons pour ce faire des données produites et annotées dans le cadre du challenge international i2b2/VA 2010 [1].

L'article est organisé comme suit : nous détaillons dans un premier temps l'état de l'art (section 2). Puis nous décrivons les données utilisées (section 3). Dans la section 4, nous présentons nos deux systèmes et exposons les résultats de leur évaluation dans la section 5.

2 État de l'art

Les notions de modalité et de niveau de certitude couvrent des phénomènes linguistiques variés qu'il est nécessaire de prendre en compte pour effectuer une analyse sémantique profonde des textes. Après une mise au point terminologique, nous détaillons le traitement de ces problématiques pour la fouille d'opinion et l'extraction d'informations dans le domaine médical.

2.1 Mise au point terminologique

La modalité est définie par Lapaire & Rotgé (1998) comme « une prise de position concernant la valeur de vérité d'une proposition » et « la manière dont un sujet pensant et parlant se prononce sur un contenu propositionnel ». Dans cet article, nous nous intéressons plus particulièrement à la modalité épistémique qui traduit le niveau de certitude de l'énonciateur par rapport à l'énoncé. Cette notion est à rapprocher des procédés euphémistiques (en anglais, *hedging*) qui visent à atténuer ou à moduler la force d'une assertion.

De nombreux marqueurs textuels servent à exprimer le degré de certitude de l'énonciateur : adjectifs épistémiques (« possible », « impossible »), adverbes épistémiques (« probablement »), verbes (« savoir », « croire ») et locutions verbales (« il est possible que »), propositions conditionnelles, etc (Rubin *et al.*, 2006; Rubin, 2007; Saurí & Pustejovsky, 2009).

Rubin (2010) distingue cinq niveaux de certitude : absolue, élevée, modérée, faible et incertaine. À ces niveaux de certitude sont associées diverses dimensions contextuelles : *perspective* (point de vue de l'auteur / discours rapporté), *focus* (opinion / faits) et *temporalité* (passé, présent, futur, hors de propos ou ambigu). L'annotation manuelle de 2 243 phrases issues d'articles de journaux en anglais selon cette typologie a montré que plus de la moitié des phrases contiennent des marqueurs de modalité épistémique, avec 1 727 occurrences de marqueurs, classés dans 47 classes syntaxico-sémantiques (Rubin, 2010). Les marqueurs les plus fréquents sont les auxiliaires modaux (« must », « could »), les adjectifs superlatifs et les adverbes intensifieurs (« much », « so »).

Saurí & Pustejovsky (2009) définissent quant à eux les notions de fait avéré (*fact*) et fait non avéré (*counterfact*), modulés par un niveau de certitude (certain, probable, possible). Ces catégories ont été utilisées pour l'annotation du corpus anglais FactBank.

2.2 Fouille d'opinion

La fouille d'opinion vise à détecter de manière automatique les unités textuelles (mots, syntagmes, phrases ou textes) qui portent une marque de subjectivité et à déterminer leur polarité ou orientation (positive, négative ou mixte) (Pang & Lee, 2008). Les systèmes de fouille d'opinion utilisent souvent des lexiques affectifs regroupant des mots et expressions subjectifs, associés à leur polarité et éventuellement l'intensité de leur orientation. L'utilisation en contexte de tels lexiques nécessite la prise en compte des phénomènes linguistiques pouvant modifier l'orientation *a priori* des expressions subjectives. Ces phénomènes incluent notamment la modification adverbiale, conduisant à une modulation de l'intensité d'une opinion (« bon », « très bon »), voire son inversion par des adverbes de négation (« bien », « pas bien »). En pratique, les phénomènes de négation sont identifiés par diverses heuristiques. Ainsi, Pang *et al.* (2002) marquent les mots modifiés par une négation s'ils apparaissent entre un négateur et une marque de ponctuation. Na *et al.* (2005) utilisent quant à eux des patrons syntaxiques pour identifier de manière plus précise les expressions incluant un négateur.

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

L'analyse automatique de la subjectivité est également influencée par d'autres phénomènes, tels que la présence d'une proposition conditionnelle (Narayanan *et al.*, 2009). Une phrase conditionnelle peut contenir des expressions subjectives sans pour autant formuler une opinion (ex : « Si cet auteur écrit un bon livre, je l'achèterai »). D'autre part, l'opinion exprimée est souvent déterminée conjointement par la proposition conditionnelle et le conséquent (ex : « Si tu veux lire un bon livre, je te conseille celui-ci »). L'analyse de phrases conditionnelles en fouille d'opinion repose en général sur des critères spécifiques, tels que le temps des verbes ou la présence de connecteurs conditionnels.

2.3 Extraction d'information en domaine biomédical

En domaine biomédical, il est particulièrement important d'analyser la modalité et le niveau de certitude car les formes modalisatrices sont très largement utilisées dans les documents biomédicaux pour 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.

Concernant la négation, sa détection dans les documents biomédicaux a été largement étudiée. Chapman *et al.* (2001) ont proposé un algorithme appelé NegEx détectant les négations et identifiant les termes médicaux dans la portée d'une négation [2]. Le principe est le suivant : les déclencheurs de négation sont annotés dans la phrase (comme par exemple «no» ou «denies») et répartis en deux classes selon qu'ils apparaissent avant ou après les termes dans leur portée (les deux déclencheurs précédents apparaissent par exemple avant les termes qu'ils qualifient) ; des pseudo-déclencheurs sont également annotés, c'est-à-dire des termes contenant des déclencheurs de négation, mais qui n'en sont pas, comme «no increase». Puis les portées de ces déclencheurs sont définies : par défaut, la portée d'un déclencheur va de ce déclencheur à la fin (ou au début) de la phrase, mais peut être interrompue par la présence d'un autre déclencheur ou d'un déclencheur de fin de clause (comme par exemple «presenting», qui marque la fin de portée de «History» dans la phrase «History of COPD, presenting with shortness of breath»).

Mutalik *et al.* (2001) ont également proposé un outil de détection des négations dans des documents médicaux. Des marqueurs de négation sont repérés puis une grammaire associe ces marqueurs avec des concepts. Enfin, Huang & Lowe (2007) ont utilisé une approche mixte combinant des expressions régulières et une analyse syntaxique. Le traitement de la négation est également inclus dans certains systèmes plus généraux comme le système Medical Language Extraction and Encoding (MedLEE) (Friedman *et al.*, 1994).

D'autres travaux concernent de façon plus générale le traitement de la modalité. La première étude analysant ce problème d'un point de vue linguistique et informatique est celle de Light *et al.* (2004) ; les auteurs ont analysé les marqueurs de modalité dans des résumés d'articles scientifiques médicaux issus de PubMed [3], annoté manuellement des phrases selon la présence de modalisation et entraîné un classifieur à détecter de telles phrases dans des résumés d'articles. Plusieurs travaux se sont également intéressés à la classification de phrases en fonction de la présence de modalité. Medlock & Briscoe (2007) se fondent sur une approche faiblement supervisée pour cette classification, et utilisent uniquement les mots des phrases comme attributs. Leur classification obtient un point d'équilibre (*break-even point* ou BEP) entre rappel et précision de 0,76. Ils ont par ailleurs rendu public le corpus d'articles scientifiques biologiques qu'ils utilisent dans leurs travaux. Szarvas (2008) a étendu cette approche en ajoutant les bigrammes et trigrammes aux attributs du classifieur, en sélectionnant les attributs pour réduire le nombre de mots-clés candidats et en ajoutant des listes de mots-clés externes, pour obtenir un BEP de 0,85. Kilicoglu & Bergler (2008) soulignent le rôle de la syntaxe dans la détection des marqueurs de modalité, et ajoutent aux attributs des informations syntaxiques sur les relations de dépendance afin de tenir compte du contexte syntaxique des marqueurs.

2. Cet outil nous a servi de système de référence lors de notre évaluation.
3. `http://www.ncbi.nlm.nih.gov/pubmed`

Une limitation des ces travaux est qu'ils ne considèrent la modalité qu'au niveau des phrases, alors que plusieurs modalités peuvent en réalité être présentes dans une même phrase comme le notent par exemple Wilbur *et al.* (2006). Un niveau de détection plus précis consiste à identifier des marqueurs de modalité et leur portée. Ainsi, dans l'exemple «Right upper lobe volume loss and probably pneumonia.», «probably» est un marqueur de modalité dont la portée est le seul terme «pneumonia». Le corpus BioScope (Vincze *et al.*, 2008) a été utilisé pour évaluer la reconnaissance de telles informations, notamment dans le cadre de la tâche d'évaluation «Learning to detect hedges and their scope in natural language text» à CoNLL 2010 [4]. Morante & Daelemans (2009) par exemple ont considéré la détection des marqueurs et de leur portée comme un problème de classification supervisée, en entraînant et combinant plusieurs classifieurs. Ils obtiennent une f-mesure de 0,85 pour la détection des marqueurs, et de 0,66 pour la détection des portées de ces marqueurs.

Comme nous venons de le montrer, l'analyse de la modalité et du niveau de certitude a fait l'objet de divers travaux dans le domaine biomédical. Les méthodes utilisées sont soit à base de règles et de lexiques, soit à base d'apprentissage, lorsqu'il existe des données annotées en quantité suffisante. Dans cet article, nous décrivons et comparons de manière détaillée deux méthodes différentes pour une tâche précise, à savoir la classification d'assertions portant sur des problèmes médicaux en anglais.

3 Description des données

Nous nous sommes intéressées à la détection de modalité et du niveau de certitude intra-phrases dans le cadre de la campagne d'évaluation i2b2 2010 [5]. L'objectif d'une des tâches de cette campagne était la classification d'assertions relatives à des problèmes médicaux. Il s'agissait donc de donner une valeur de certitude ou de modalité à des concepts médicaux qui avaient été préalablement annotés dans des compte-rendus de patients.

Plutôt que de calculer la portée de chaque marqueur possible de modalité, l'accent est mis ici sur les concepts étudiés, pour lequels on cherche à connaître une modalité. Nous détaillerons dans cette section les données que nous avons utilisées pour nos expériences, puis présenterons dans la section suivante les deux systèmes d'analyse développés.

3.1 Corpus

Les corpus sont composés de 826 compte-rendus médicaux en langue anglaise provenant de trois hôpitaux américains. Ils ont préalablement été annotés en concepts [6], qui sont :
– les **problèmes médicaux**, définis comme les observations concernant le corps ou l'esprit du patient et considérées comme anormales ou provoquées par une maladie, ce qui englobe les maladies, syndromes, virus, bactéries, symptômes...
– les **traitements**, regroupant les procédures, interventions ou médicaments donnés au patient pour résoudre un problème médical ;
– les **tests**, c'est-à-dire les procédures ou mesures faites sur le patient pour trouver des informations sur un problème médical.

Ces compte-rendus ont également été anonymisés, afin de supprimer les noms de personne, de lieu, les adresses, et autres informations géographiques précises, les dates relatives à un individu comme une date de naissance, les numéros de téléphone et fax, les courriels, les numéros de sécurité sociale etc. Dans les documents, toutes ces informations sont remplacées par des balises génériques comme «NAME», «ADDRESS», ou encore «AGE».

3.2 Catégories d'assertions

Une classe d'assertion a été attribuée à chaque problème médical du corpus. Six catégories d'assertions ont été définies :

4. http://www.inf.u-szeged.hu/rgai/conll2010st/
5. https://www.i2b2.org/NLP/Relations/Main.php
6. Pour une méthode d'annotation automatique des concepts, voir (Minard *et al.*, 2011).

– **présent** : le problème associé au patient est présent. C'est également la catégorie par défaut pour les problèmes médicaux qui n'ont pu être classés dans une autre catégorie.
  Exemples : «*the wound* was noted to be clean with *mild serious drainage*», «history of *chest pain*»
– **absent** : le patient ne présente pas ce problème. Cette catégorie inclut les problèmes résolus.
  Exemples : «patient denies *pain*», «history inconsistent with *stroke*»
– **possible** : le patient peut avoir le problème, mais ce n'est pas certain.
  Exemples : «this is very likely to be *an asthma exacerbation*», «We suspect this is not *pneumonia*»
– **conditionnel** : le patient ne rencontre le problème que dans certaines conditions, comprenant en particulier les allergies.
  Exemples : «Penicillin causes *a rash*», «patient reports *shortness of breath* upon climbing stairs»
– **hypothétique** : le patient est susceptible de développer le problème dans le futur.
  Exemples : «if you experience *wheezing* or *shortness of breath*», «ativan 0.25 to 0.5 mg IV q 4 to 6 hours prn *anxiety*»
– **associé à quelqu'un d'autre** : le problème concerne quelqu'un d'autre que le patient.
  Exemples : «Family history of *prostate cancer*», «brother had *asthma*»

Catégorie d'assertion	Corpus d'entraînement	Corpus d'évaluation
Présent	8 052	13 025
Absent	2 536	3 609
Possible	535	883
Conditionnel	103	171
Hypothétique	651	717
Associé à qqun d'autre	92	145
Total	11 969	18 550

TABLE 1 – Nombre d'assertions par catégorie dans les corpus d'entraînement et d'évaluation

La table 1 détaille le nombre d'assertions annotées pour chaque catégorie dans les corpus d'entraînement et d'évaluation. La majorité des problèmes médicaux évoqués dans les documents sont *présents* : 67% des assertions appartiennent à cette catégorie. La répartition des diverses catégories est donc très déséquilibrée.

Les organisateurs d'i2b2 ont calculé l'accord inter-annotateur sur le corpus d'entraînement, annoté par 12 annotateurs. Cette information, présentée dans la table 2, est intéressante car elle montre que pour certaines catégories, les annotateurs ont fait des choix d'annotation relativement différents. L'accord inter-annotateurs sur l'ensemble des catégories est élevé, mais pour certaines catégories notamment «conditionnel», il est très faible. On peut donc supposer que les résultats obtenus par des systèmes automatiques seront également plus faibles pour ces catégories.

Catégorie d'assertion	Accord
Présent	0,89
Absent	0,94
Possible	0,60
Conditionnel	0,44
Hypothétique	0,79
Associé à qqun d'autre	0,89
Total	0,91

TABLE 2 – Accords inter-annotateurs pour chaque catégorie d'assertion

4 Description des systèmes

Dans cette section, nous décrivons tout d'abord les systèmes de référence (ou *baseline*) et les résultats qu'ils obtiennent. Puis nous détaillons les systèmes que nous avons développés afin d'identifier de manière automatique la catégorie d'assertion associée à un problème médical.

4.1 Références

Nous avons considéré plusieurs références pour évaluer l'apport de nos systèmes :
– la première méthode de référence consiste à tout étiqueter comme « présent », ce qui donne une f-mesure totale de 0,67 (proportion de présents dans le corpus) ;
– la deuxième référence consiste à utiliser le système NegEx [7] sans aucune modification, qui n'annote donc que les catégories «présent» et «absent», ce qui donne une f-mesure de 0,83 ;
– la troisième référence consiste à utiliser le système ConText (extension de NegEx pour tenir compte de la catégorie associé à quelqu'un d'autre) sans aucune modification, donc sans les catégories «possible», «conditionnel» et «hypothétique», ce qui donne une f-mesure de 0,86.

4.2 Prétraitements

4.2.1 Annotation des concepts

Nous avions à notre disposition l'annotation en concepts faite manuellement par les organisateurs d'i2b2. Nous avons donc annoté les documents en balisant les termes qui correspondaient à un concept médical reconnu, c'est-à-dire à un problème, un traitement ou un test.

4.2.2 Traitement des coordinations

L'étude des données de développement a montré que de nombreux concepts sont coordonnés à l'aide de virgules ou de conjonctions de coordination, comme par exemple « pleural effusion or pneumothorax ». Nos deux systèmes utilisent les contextes directs des concepts, sous forme de fenêtre de mots, afin d'identifier la catégorie de l'assertion. La présence de séquences de concepts coordonnés peut donc conduire à la prise en compte d'un contexte gauche ou droit réduit, comportant principalement d'autres problèmes coordonnés. Dans ce cas, des indices essentiels pour l'identification de la catégorie de l'assertion peuvent se trouver en dehors de la fenêtre contextuelle considérée. Le rôle important de la coordination a également été démontré pour une autre tâche d'extraction d'information, l'extraction d'événements (Kilicoglu & Bergler, 2009). Nous avons donc pré-traîté les données pour identifier les séquences de concepts coordonnés et ainsi redéfinir les fenêtres contextuelles utilisées : les fenêtres gauches se terminent au début d'une séquence de concepts coordonnés tandis que les fenêtres droites débutent à la fin d'une séquence de concepts. Ces fenêtres contextuelles sont partagées par tous les concepts qui apparaissent dans la même séquence. Plus de la moitié des concepts étaient inclus dans une coordination de concepts dans le corpus d'entraînement.

4.3 Système par apprentissage supervisé

Le système par apprentissage supervisé considère l'identification d'assertions comme une tâche de classification. Nous avons entraîné une machine à vecteurs de support (SVM) avec libsvm (Chang & Lin, 2001) sur la base d'attributs binaires et d'un noyau RBF [8]. Les paramètres optimaux ont été sélectionnés automatiquement par validation croisée en 5 sous-ensembles [9]. Pour le développement du système, nous avons utilisé un corpus d'entraînement de 241 fichiers et un corpus de test de 54 fichiers.

Nous avons utilisé quatre types d'attributs :
– **attributs contextuels lexicalisés** : mots et mots désuffixés dans une fenêtre de cinq mots à gauche et à droite du concept cible. Nous avons également effectué des expériences complémentaires avec les étiquettes morpho-syntaxiques mais celles-ci n'apportent pas d'améliorations significatives ;

7. `http://code.google.com/p/negex/`
8. Il aurait également été possible d'utiliser les CRF (Conditional Random Fields) afin d'étiqueter de manière plus précise les déclencheurs et leur portée. Toutefois, ces annotations n'étaient pas fournies dans le corpus d'apprentissage et il aurait donc été nécessaire d'utiliser un corpus externe, comme Bioscope par exemple (Vincze *et al.*, 2008). Or, seules les catégories de négation et de spéculation sont annotées dans Bioscope. Nous avons donc renoncé à cette possibilité.
9. Cette étape a été réalisée de manière automatique en utilisant le script `easy.py` fourni avec libsvm.

– **déclencheurs** : nous avons utilisé les déclencheurs définis pour le système à base de règles (voir section suivante), avec quelques déclencheurs supplémentaires. Ces déclencheurs sont identifiés dans une fenêtre de cinq mots à gauche et à droite du concept cible. Nous avons également identifié quelques déclencheurs internes au concept, tels que « on exertion » (« à l'effort ») qui indique la catégorie conditionnel ;
– **attributs internes au concept cible** : mots et mots désuffixés formant le concept et présence du préfixe « non » dans un des mots ;
– **attributs spécifiques à une séquence de concepts coordonnés.** Dans le cas où le concept cible apparaît dans une séquence de concepts coordonnés, nous utilisons des attributs spécifiques qui correpondent aux mots et aux mots désuffixés de la séquence. Par exemple, pour la séquence « pleural effusion **or** pneumothorax », les mots « pneumothorax », « pleural » et « effusion » ainsi que la racine « effus » sont utilisés comme attributs.

4.4 Système à base de règles

L'algorithme utilisé est très proche de celui de NegEx (voir section 2.3) mais est étendu aux six catégories d'assertions à détecter. Quatre types de déclencheurs ont été définis (les exemples suivants sont donnés pour la catégorie «absent») :
– ceux qui précèdent le problème comme «denies», «never had» ou «negative for» ;
– ceux qui suivent le problème comme «was ruled out», «is absent» ou «was stopped» ;
– ceux qui sont inclus dans le problème comme «afebrile» (dans le cas, le déclencheur est le problème lui-même) ou «allergy» ;
– ceux qui limitent la portée d'un des déclencheurs précédents, comme «but».
Un même terme peut être de plusieurs types : ainsi, «ruled out» peut précéder ou suivre le problème associé. Les termes de «pseudo-négation» de NegEx (qui ressemblent à des termes de négation mais n'en sont pas comme la double négation «not ruled out») ne constituent pas un type distinct dans notre système, mais sont éliminés en vérifiant le contexte des termes de négation (par exemple on vérifiera que le mot «not» ne précède pas «ruled out»).

Les déclencheurs ont d'abord été définis manuellement par une étude de corpus, puis les listes ont été complétées grâce aux résultats du système par apprentissage : les attributs les plus utiles à la classification ont été étudiés, et éventuellement ajoutés aux listes. Le nombre moyen de déclencheurs par catégorie d'assertion varie d'une cinquantaine pour les déclencheurs précédant le problème à une quizaine pour les déclencheurs inclus dans le problème.

Trois expressions régulières sont utilisées pour rechercher des déclencheurs avant, après ou à l'intérieur des problèmes, sachant que des mots non déclencheurs peuvent être autorisés entre le déclencheur et le problème (le nombre de ces mots a été fixé par une étude du corpus d'entraînement pour chaque catégorie d'assertion) :
– <déclencheur> <mots non déclencheurs> {0,n} <problème> ;
– <problème> <mots non déclencheurs> {0,n} <déclencheur> ;
– <problème> <déclencheur> >.
Ainsi, dans la phrase «Also with *multiple masses consistent with* metastasis», «consistent with» est un déclencheur de la catégorie «possible» suivi directement du problème «metastasis». La première expression régulière sera donc déclenchée, et donnera la catégorie «possible» à ce problème.

Ces listes de déclencheurs et expressions régulières ont été implémentés avec l'outil WMatch (Rosset *et al.*, 2008; Galibert, 2009) [10] qui présente une vitesse d'exécution bien supérieure à celle de NegEx.

Une priorité a été attribuée à chaque catégorie d'assertion, en fonction du guide d'annotation i2b2 : associé à quelqu'un d'autre > absent > possible > conditionnel > hypothétique.

5 Évaluation et discussion

Le corpus d'évaluation comporte 18 550 assertions (voir table 1). Les résultats ont été évalués en termes de rappel, précision et f-mesure pour chaque catégorie d'assertion, ainsi que pour toutes les catégories :

10. Moteur d'expressions régulières développé au LIMSI, disponible sur demande.

$$\text{rappel} = \frac{\#\text{problèmes correctement attribués à la catégorie d'assertion i}}{\#\text{problèmes de la catégorie d'assertion i}}$$

$$\text{précision} = \frac{\#\text{problèmes correctement attribués à la catégorie d'assertion i}}{\#\text{problèmes attribués à la catégorie d'assertion i}}$$

$$\text{f-mesure} = \frac{2*\text{rappel}*\text{précision}}{\text{rappel}+\text{précision}}$$

5.1 Résultats de l'évaluation

Les résultats obtenus par les deux systèmes sur le corpus d'évaluation [11] sont détaillés dans la table 3.

	Apprentissage			Règles		
Catégorie	**Rappel**	**Précision**	**F-mesure**	**Rappel**	**Précision**	**F-mesure**
Toutes	0,93	0,93	0,93	0,90	0,90	0,90
Présent	0,97	0,94	0,96	0,95	0,92	0,93
Absent	0,95	0,93	0,94	0,85	0,93	0,89
Possible	0,54	0,74	0,62	0,57	0,61	0,59
Hypothétique	0,83	0,93	0,88	0,74	0,86	0,80
Conditionnel	0,24	0,74	0,36	0,27	0,29	0,28
Associé à quelqu'un d'autre	0,78	0,86	0,82	0,95	0,76	0,84

TABLE 3 – Résultats pour le corpus d'évaluation

Les deux systèmes développés obtiennent des résultats significativement [12] meilleurs que les trois systèmes de référence. Le système par apprentissage supervisé obtient de très bons résultats, et s'est classé 5e sur 21 participants à i2b2 2010.

Si l'on compare plus en détails les deux systèmes, on constate que le système par apprentissage a tendance à privilégier la précision sur le rappel. La précision qu'il obtient est généralement supérieure ou égale à celle du système à base de règles. En outre, le système par apprentissage se caractérise par une f-mesure supérieure à celle du système à base de règles, sauf pour la catégorie «associé à quelqu'un d'autre». Dans ce dernier cas, les déclencheurs utilisés par le système à base de règles permettent d'obtenir un très bon rappel de 0,95. Cette comparaison démontre la complémentarité des deux systèmes, qui gagneraient à être combinés.

Lors du challenge i2b2, le meilleur score à cette tâche a été obtenu par l'équipe du National Research Council Canada, avec une f-mesure de 0,9362 (contre 0,9311 pour notre système par apprentissage). Les méthodes des meilleurs systèmes sont assez proches de celle utilisée par notre système par apprentissage. Les SVM sont généralement utilisés pour la classification, avec des attributs portant sur les mots, l'annotation sémantique des mots, l'étiquetage morpho-syntaxique ou encore des déclencheurs spécifiques. On trouve également des attributs différents, qu'ils serait intéressant d'intégrer aux futures versions de notre système : n-grammes de caractères, attributs spécifiques à la phrase tels que le temps du verbe principal ou encore attributs spécifiques au document considéré (longueur).

5.2 Analyse des erreurs

La table 4 détaille la matrice de confusion pour le système par apprentissage. Le système présente une tendance à sur-annoter les catégories «présent» et «absent». Ce résultat était prévisible dans la mesure où il s'agit également des catégories les plus représentées dans le corpus d'apprentissage. La catégorie «présent» est également celle qui compte le plus grand nombre de faux positifs. Pour les classes plus rares dans le corpus d'apprentissage, et notamment la classe «conditionnel», le système se caractérise par un rappel réduit, corrélé à une bonne précision, de 0,74 à 0,93 en fonction de la classe considérée.

11. Ces résultats sont très légèrement différents de ceux obtenus à l'évaluation i2b2 pour le système par règles, du fait d'un changement de système d'annotation : lors de l'évaluation, nous avions utilisé GenConText, alors que les résultats présentés dans cet article sont obtenus avec WMatch.

12. p<0,05 selon le test de Student

		Apprentissage						
		Présent	**Absent**	**Possible**	**Cond.**	**Hypo.**	**Assoc. autre**	**Total**
Référence	**Présent**	12629	199	134	13	34	16	13025
	Absent	168	3418	20	0	2	1	3609
	Possible	381	16	475	1	10	0	883
	Cond.	117	12	1	41	0	0	171
	Hypo.	94	12	14	0	595	2	717
	Assoc. autre	16	16	0	0	0	113	145
	Total	13405	3673	644	55	641	132	18550

TABLE 4 – Matrice de confusion pour le système à base d'apprentissage

La matrice de confusion de la table 5 présente les erreurs de catégorisation détaillées du système à base de règles. Les confusions les plus fréquentes concernent principalement les catégories «présent» et «absent» ; le système a également tendance à surannoter en «possible». Enfin, le très bon rappel sur la catégorie «associé à quelqu'un d'autre» se traduit par très peu de confusion dans cette catégorie.

		Règles						
		Présent	**Absent**	**Possible**	**Cond.**	**Hypo.**	**Assoc. autre**	**Total**
Référence	**Présent**	12352	167	285	110	73	38	13025
	Absent	514	3077	9	2	1	6	3609
	Possible	331	34	505	3	10	0	883
	Cond.	117	6	1	47	0	0	171
	Hypo.	152	9	23	2	531	0	717
	Assoc. autre	5	2	0	0	0	138	145
	Total	13471	3295	823	164	615	182	18550

TABLE 5 – Matrice de confusion pour le système à base de règles

Les erreurs du système à base de règles ont de multiples causes (dans les exemples suivants, les problèmes médicaux à analyser sont remplacés par le mot PROBLEM) :

– les listes de déclencheurs sont incomplètes : ainsi, pour la catégorie «associé à quelqu'un d'autre», les cas de dons d'organes anonymes n'avaient pas été pris en compte («The liver was from a gentleman who had died from PROBLEM» [13]) et donc les déclencheurs associés («gentleman» ici) n'étaient pas dans les listes ;
– le déclencheur peut être trop loin du problème pour que les deux soient associés, notamment dans le cas d'anaphores ; ainsi pour la phrase «The patient declined the procedure stating that her mother had mitral valve regurgitation and she lived for many years without PROBLEM», le déclencheur «her mother» de la catégorie «associé à quelqu'un d'autre» est trop loin du problème à analyser ; l'utilisation d'un outil de résolution d'anaphore pourrait ici être utile ;
– un déclencheur est détecté, mais il n'est en réalité pas relié au problème, notamment lorsque l'avis de la famille est évoqué («According to the family») ; un contexte gauche au déclencheur «family» pourrait ici être précisé pour éliminer ces cas, mais les formulations plus complexes comme «The family is confident in any decision the doctor would make concerning her PROBLEM» sont plus difficiles à traiter ;
– la portée du marqueur est trop importante ou trop faible : dans la phrase «The patient is on no specific medications for his PROBLEM», le marqueur «no» est repéré, mais ne s'applique en réalité pas au problème ;
– deux marqueurs sont détectés, et l'ambigüité est mal résolue, comme par exemple dans la phrase «PROBLEM was less likely given her negative angiogram.» où «likely» est bien déclencheur de la catégorie «possible», mais le déclencheur «negative» est favorisé du fait des priorités entre catégories ;
– quelques erreurs sont liées à des problèmes d'anonymisation des données, qui perturbent l'application des règles ;
– quelques incohérences d'annotation ont également été notées, ce que prévoyait l'accord inter-annotateur.

Enfin, la matrice de confusion de la table 6 permet de comparer les annotations réalisées par les deux systèmes.

13. Les exemples présentés sont issus d'exemples réels, mais sont généralement simplifiés pour être plus lisibles.

		Règles						
		Présent	**Absent**	**Possible**	**Cond.**	**Hypo.**	**Assoc. autre**	**Total**
Apprentissage	**Présent**	12711	125	312	129	87	41	13405
	Absent	497	3142	11	2	2	19	3673
	Possible	129	22	491	0	2	0	644
	Cond.	22	1	1	31	0	0	55
	Hypo.	99	5	8	2	524	3	641
	Assoc. autre	13	0	0	0	0	119	132
	Total	13471	3295	823	164	615	182	18550

TABLE 6 – Matrice de confusion pour les deux systèmes

Les systèmes sont fortement en désaccord pour la catégorie «conditionnel» : seules 31 annotations sont partagées. C'est également la catégorie qui obtient les moins bons résultats globalement. La deuxième catégorie la plus problématique correspond aux assertions du type «possible». Le système par apprentissage tend à sous-annoter cette catégorie et par conséquent de nombreuses assertions annotées comme «possible» par le système à base de règles sont associées à la catégorie «présent» par le système par apprentissage.

6 Conclusion et perspectives

Nous nous sommes intéressées à l'analyse de la modalité et du niveau de certitude dans des textes médicaux, et avons développé deux systèmes dans le cadre de la campagne d'évaluation i2b2 2010. Le premier procède par apprentissage supervisé et s'appuie sur des attributs lexicaux, morphologiques et sémantiques. Le second est un système à base de listes et de règles. Ces deux systèmes obtiennent des résultats significativement meilleurs que les trois références considérées, et au niveau de l'état de l'art : environ 0,93 et 0,90 de f-mesure.

Les perspectives de poursuite de ces travaux sont nombreuses. Ces systèmes pourraient notamment être étendus afin de travailler éventuellement sur plusieurs phrases, et de tenir compte des phénomènes discursifs. Il serait également souhaitable de combiner les deux systèmes, en particulier pour profiter du très bon niveau de rappel du système à base de règles pour la catégorie «associé à quelqu'un d'autre».

Nous envisageons également d'adapter les systèmes présentés au français. Pour le système à base de règles, il sera nécessaire d'identifier les déclencheurs correspondant ou aditionnels pour le français. Le système à base d'apprentissage requiert quant à lui de larges quantités de données annotées. Dans la mesure où ceci constitue une tâche nécessitant des annotateurs experts humains, nous nous appuierons sur la méthode de l'apprentissage actif (*active learning*).

Nous souhaitons également généraliser ces travaux à d'autres domaines, afin d'analyser la modalité et les niveaux de certitude dans des contextes plus larges. En particulier, nous envisageons l'étude des phénomènes similaires dans les textes journalistiques et, plus globalement, les textes subjectifs visant à exprimer des opinions.

Remerciements

Ce travail a été partiellement financé par le projet Quæro (financement Oseo, agence française pour l'innovation et la recherche). Les données médicales utilisées proviennent du consortium Informatics for Integrating Biology to the Bedside (i2b2) grâce aux financements numéros U54LM008748 de la National Library of Medicine, VA HSR HIR 08-374 du Consortium for Healthcare Informatics Research (CHIR), et VA HSR HIR 08-204 du VA Informatics and Computing Infrastructure (VINCI).

Références

CHANG C.-C. & LIN C.-J. (2001). *LIBSVM : a library for support vector machines*. Outil disponible à l'adresse

suivante : `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

CHAPMAN W., BRIDEWELL W., HANBURY P., COOPER G. & BUCHANAN B. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, **34**(5), 301–310.

FRIEDMAN C., ALDERSON P. O., AUSTIN J. H., CIMINO J. J. & JOHNSON S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, **1**(2), 161.

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

HUANG Y. & LOWE H. (2007). A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *Journal of the American Medical Informatics Association*, **14**, 304–311.

KILICOGLU H. & BERGLER S. (2008). Recognizing speculative language in biomedical research articles : a linguistically motivated perspective. *BMC Bioinformatics*, **9**(Suppl 11).

KILICOGLU H. & BERGLER S. (2009). Syntactic dependency based heuristics for biological event extraction. In *BioNLP '09 : Proceedings of the Workshop on BioNLP*, p. 119–127.

LAPAIRE J.-R. & ROTGÉ W. (1998). *Linguistique et grammaire de l'anglais, 3e édition*. Amphi 7. Presses Universitaires du Mirail.

LIGHT M., QIU X. & SRINIVASAN P. (2004). The language of bioscience : Facts, speculations, and statements in between. In *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases : tools for users*, p. 17–24.

MEDLOCK B. & BRISCOE T. (2007). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 992–999, Prague, Czech Republic.

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

MORANTE R. & DAELEMANS W. (2009). Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on BioNLP*, p. 28–36.

MUTALIK P., DESHPANDE A. & NADKARNI P. (2001). Use of general-purpose negation detection to augment concept indexing of medical documents. *Journal of the American Medical Informatics Association*, **8**(6), 598.

NA J.-C., KHOO C. & WU P. H. J. (2005). Use of negation phrases in automatic sentiment classification of product reviews. *Library Collections, Acquisitions, and Technical Services*, **29**(2), 180 – 191.

NARAYANAN R., LIU B. & CHOUDHARY A. (2009). Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 1 of *EMNLP '09*, p. 180–189, Stroudsburg, PA, USA.

PANG B. & LEE L. (2008). *Opinion mining and sentiment analysis*, volume 2 of *Foundations and Trends in Information Retrieval*. Now Publishers Inc.

PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 79–86.

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

RUBIN V., LIDDY E. & KANDO N. (2006). Certainty Identification in Texts : Categorization Model and Manual Tagging Results. In W. B. CROFT, J. SHANAHAN, Y. QU & J. WIEBE, Eds., *Computing Attitude and Affect in Text : Theory and Applications*, volume 20 of *The Information Retrieval Series*, p. 61–76. Springer Netherlands.

RUBIN V. L. (2007). Stating with Certainty or Stating with Doubt : Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In *Proceedings of the Human Language Technologies Conference : The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007) ; Companion Volume, Short Papers*, p. 141–144.

RUBIN V. L. (2010). Epistemic modality : From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management*, **46**(5), 533 – 540.

SAURÍ R. & PUSTEJOVSKY J. (2009). FactBank : a corpus annotated with event factuality. *Language Resources and Evaluation*, **43**, 227–268.

SZARVAS G. (2008). Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL-08 : HLT*, p. 281 – 289, Colombus, Ohio, USA.

VINCZE V., SZARVAS G., FARKAS R., MORA G. & CSIRIK J. (2008). The bioscope corpus : biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, **9**(Suppl 11), S9.

WILBUR W., RZHETSKY A. & SHATKAY H. (2006). New directions in biomedical text annotation : definitions, guidelines and corpus construction. *BMC bioinformatics*, **7**(1), 356.