La subjectivité dans le discours médical : sur les traces de l'incertitude et des émotions

Pierre Chauveau Thoumelin*, Natalia Grabar*

*STL UMR 8163 CNRS, Université Lille 3 et Lille 1
p.chauveau.thoumelin@gmail.com, natalia.grabar@univ-lille3.fr,
http://natalia.grabar.perso.sfr.fr/

Résumé. Les acteurs et usagers du domaine médical (médecins, infirmiers, patients, internes, pharmaciens, etc.) ne sont pas issus de la même catégorie socio-professionnelle et ne présentent pas le même niveau de maîtrise du domaine. Leurs écrits en témoignent et véhiculent, de plus, la subjectivité qui leur est propre. Nous nous intéressons à l'étude automatisée de la subjectivité dans le discours médical dans des textes en langue française. Nous confrontons le discours des médecins (articles scientifiques, rapports cliniques) à celui des patients (messages de forums de santé) en analysant contrastivement les différences d'emploi des descripteurs tels que les marqueurs d'incertitude et de polarité, les marques émotives non lexicales (smileys, ponctuations répétées, etc.) et lexicales, et les termes médicaux relatifs aux pathologies, traitements et procédures. Nous effectuons une annotation et catégorisation automatiques des documents afin de mieux observer les spécificités que présentent les discours médicaux ciblés.

1 Introduction

Le domaine médical, comme d'autres domaines de spécialité, est caractérisé par l'hétérogénéité de ses acteurs et utilisateurs. Mentionnons par exemple les médecins, les patients, les infirmiers, les pharmaciens, les internes, les brancardiers, les administratifs, les aides soignants, les chercheurs, les biologistes qui interagissent quotidiennement dans la pratique médicale. Il est évident que tous ces acteurs jouent des rôles différents et, de la même manière, les besoins de ces acteurs, y compris les besoins en information, sont aussi différents. Par exemple, un médecin recherche typiquement des informations précises qui lui permettent de faire un diagnostic ou des prescriptions appropriées, un chercheur est souvent à la recherche des derniers travaux dans un domaine bien ciblé, alors qu'un patient peut rechercher des informations plus ou moins générales afin de retrouver des explications sur une maladie ou un traitement. De manière plus générale, nous pouvons différencier les cas suivants (Pearson, 1998) : les informations créées par des spécialistes pour les spécialistes (le cas des médecins ou des chercheurs), les informations créées par des spécialistes pour les non spécialistes (le cas des patients), les informations créées par des non spécialistes pour les non spécialistes. Les textes correspondant à chaque cas ont des propriétés et fonctions différentes. De même, ils véhiculent des informations dont le niveau de spécialisation varie et qui peuvent nécessiter une expertise plus ou

moins importante pour une compréhension correcte. Pour un système de recherche d'information, il peut donc être important de pouvoir distinguer entre les différents types de textes et, de cette manière, de proposer une caractérisation supplémentaire de ces textes, en en distinguant par exemple le niveau de spécialisation. Cet objectif, distinction entre les textes véhiculant différents niveaux de spécialisation, correspond à la motivation principale de notre étude.

Parmi les travaux existants, nous pouvons par exemple citer ceux relatifs à la personnalisation de la recherche d'information (Pasi, 2010). Parmi les méthodes proposées, afin d'adapter les résultats de recherche à un utilisateurs, certaines exploitent le filtrage collectif (Herlocker et al., 2004), le filtrage basé sur le contenu (Kassab et Lamirel, 2006), le filtrage par la combinaison des deux (Basilico et Hofmann, 2004), le filtrage par la modélisation de l'utilisateur (Hadjouni Krir, 2012). Nous proposons d'aborder la question par l'analyse du contenu des documents. Nous proposons d'exploiter la particularité des documents, qui concerne la subjectivité des acteurs telle qu'elle transparaît à travers leurs incertitudes et émotions. Plus particulièrement, nous proposons d'exploiter les informations relatives à l'emploi de l'incertitude (*i.e.*, *possible, il semblerait, certain*), de la polarité (*i.e.*, *absence, pas de, ni*), et des émotions (*i.e.*, un lexique spécifique comme *peur* ou *colère*, la ponctuation répétée et intensifiée comme !!!, les émoticônes comme *:-)*, les mots avec des caractères répétés comme *maaaaal*). Nous utilisons également des modifieurs lexicaux (*i.e.*, *très, beaucoup*). Notre hypothèse est que les différents types de documents médicaux présentent des spécificités liées à la subjectivité, qui peuvent être utilisées pour une distinction automatique entre ces types de documents.

Dans la suite de ce travail, nous décrivons d'abord le matériel utilisé (section 2), ensuite la méthode (section 3). Nous présentons et discutons les résultats produits (section 4), et terminons avec des perspectives (section 5).

2 Collecte et préparation du matériel

Nous utilisons deux types de matériel : corpus (section 2.1) et ressources (section 2.2).

2.1 Corpus

Les données étudiées portent sur le thème de la rhumatologie (maladies des os, des articulations ou des muscles). Elles sont constituées de trois corpus, tous collectés en mai 2013 :

- le *corpus scientifique* contient des articles scientifiques rédigés par des médecins ou des chercheurs. Ce corpus est constitué à partir du portail médical CISMeF [1] (Catalogue et Index des Sites Médicaux de langue Française) ;
- le *corpus clinique* regroupe des documents cliniques aussi rédigés par des médecins. Ces documents proviennent d'un hôpital français ;
- le *corpus forum* est composé de messages d'un forum Doctissimo consacré aux douleurs du dos [2]. Les message sont rédigés essentiellement par des patients.

Les données source (pdf, doc, html, texte...) sont normalisées au format texte et converties en utf-8, en faisaint attention aux ligatures (*e.g.* œ → oe, æ → ae, fl → fl), aux accents mal convertis (*e.g.* o^ → ô, i¨ → ï, e´ → é), à la suppression des caractères non imprimables (*e.g.* retour à la ligne, tabulation verticale). La taille des corpus est indiquée dans le tableau 1 : les documents

1. http://www.chu-rouen.fr/cismef/
2. http://forum.doctissimo.fr/sante/douleur-dos/liste_sujet-1.htm

du corpus scientifique sont les plus longs, viennent ensuite les fils de discussion du forum et les documents cliniques. Les corpus sont échantillonnés, pour assurer leur comparabilité. Le nivellement est effectué par rapport au corpus scientifique, le plus petit de l'ensemble.

	Scient.	Clinique	Forum	Total
Nombre de mots	840 228	5 806 158	3 351 951	9 998 337
Nombre de documents	265	8 162	4 388	12 815
Moyenne mots/document	3 170	711	763	780

TABLE 1 – *Taille des corpus en nombre de mots et de documents.*

2.2 Ressources

Une partie importante des ressources est dédiée à la détection de la subjectivité et des émotions (section 2.2.1). D'autres ressources sont spécifiques au domaine médical (section 2.2.2). Nous décrivons également comment les ressources sont ajustées (section 2.2.3).

2.2.1 Détection de la subjectivité et des émotions dans les corpus

Les ressources linguistiques exploitées contiennent plusieurs types de marqueurs :

- L'incertitude (n=101) peut être exprimée aussi bien avec des verbes (*i.e.*, *supposer*, *apparaître*, *suspecter*), des noms (*i.e.*, *possibilité*, *hypothèse*), des adjectifs (*i.e.*, *vraisemblable*, *douteux*), qu'avec des adverbes (*i.e.*, *sûrement*, *peut-être*). Deux degrés d'incertitude sont distingués : l'incertitude forte, qui influence fortement la fiabilité des informations (*i.e.*, *douteux, évocateur, hypothèse*), et l'incertitude faible, qui influence faiblement la fiabilité des informations (*i.e.*, *apparemment, certain, probablement*) ;
- La négation (n=20) peut être exprimée de différentes façons également : avec des adverbes (*i.e.*, *ne*, *pas*), des noms (*i.e.*, *absence*, *lacune*), des adjectifs (*i.e.*, *négatif*, *impossible*), des prépositions (*i.e.*, *sans*) ou encore avec le préfixe *non-* ;
- Les modifieurs (n=17) du degré d'incertitude comme *peu, très peu*, *fort peu*, *extrêmement*, *vraiment*. Leur interprétation dépend de la polarité du terme de base : par rapport à *probable*, *très probable* conduit à une diminution de l'incertitude, tandis que *très douteux*, par rapport à *douteux* seul, conduit à une augmentation de l'incertitude ;
- Un lexique des émotions (Augustyn et al., 2008) contient 1 144 entrées (verbes, noms et adjectifs). Les entrées du lexique sont associées à plus d'une trentaine d'emotions (*e.g. tristesse*, *dégoût*, *joie*, *honte*). Ceci correspond à une catégorisation plus fine que celles habituellement utilisées dans les méthodes semi-automatiques (Ekman, 1992). Pour arriver à un niveau de généralisation, les émotions sont catégorisées en trois catégories : émotions positives, négatives et neutres. Par exemple, *tristesse*, *dégoût* et *honte* sont des émotions négatives, *joie* positive, *anticipation*, *étonnement* et *surprise* neutres.

2.2.2 Détection de notions médicales dans les corpus

Les ressources sémantiques se composent de termes appartenant à trois types sémantiques : *maladie* (maladies, problèmes médicaux ou troubles), *procédure* (actes médicaux effectués par

les médecins) et *médicament*. Les maladies et les procédures proviennent de la terminologie médicale SNOMED international (Côté, 1996), ou Systematized Nomenclature of Human and Veterinary Medicine, telle que distribuée par l'ASIP Santé [3]. La liste des médicaments est constituée à partir de : (1) la base de médicaments Thériaque [4], créée par le CNHIM (Centre National Hospitalier d'Information sur le Médicament) [5] ; (2) l'UMLS (Lindberg et al., 1993), ou Unified Medical Language System, une collection de terminologies biomédicales développées par la US National Library of Medicine ; (3) la base UCD (Unité commune de dispensation) couvrant les médicaments qui disposent d'une autorisation de mise sur le marché et sont commercialisés en France. Nous avons au total 71 449 entrées dans la catégorie *maladie*, 25 148 dans la catégorie *procédure*, et 17 571 dans la catégorie *médicament*.

2.2.3 Ajustement des ressources

Les ressources exploitées doivent être ajustées aux données traitées. Comme c'est souvent le cas avec des ressources constituées dans d'autres cadres, nous cherchons par exemple, à les rendre plus précises et/ou plus exhaustives.

Rendre les ressources plus précises. Certaines entrées peuvent avoir un sens différent dans les corpus par rapport à ce qui est prévu dans les ressources. Ceci concerne surtout le lexique des émotions. Par exemple, dans ce lexique, les mots comme *irriter* et *irritation* sont assignés à la catégorie de l'émotion *colère*, *tendu* à la catégorie *attirance*, *manque* à la catégorie *colère*. Cependant, dans un corpus médical, dans des expressions comme :

> (1) *La seule différence constatée réside dans la réponse à l'intensité de l'irritation, provoquant une extension progressive de la douleur...*
> (2) *Elever la jambe tendue jusqu'à apparition d'une douleur radiculaire*
> (3) *Les guidelines ... s'accordent sur le manque de preuve pour recommander des interventions préventives pour la lombalgie aiguë.*

les entrées *irritation*, *tendu* ou *manque* ne signifient pas les émotions prévues dans le lexique, mais reçoivent un sens propre au domaine médical. Après le nettoyage du lexique, effectué afin de réduire le bruit lors de l'annotation, nous gardons 1 032 entrées.

La situation est similaire avec la terminologie, dont l'objectif est d'assurer l'exhaustivité des notions recensées et dont certaines peuvent être ambiguës dans certains contextes. Par exemple, les entrées comme *PDF, THE, CI, base, élément, solution* s'avèrent trop ambiguës. Au total, 50 entrées de la terminologie SNOMED International ne sont pas considérées.

Rendre les ressources plus exhaustives. D'un autre côté, il existe aussi des cas où des unités linguistiques, potentiellement intéressantes pour l'annotation, sont absentes des ressources. Ceci peut être dû (1) à la spécificité des corpus, comme par exemple avec le corpus du forum, (2) aux limites de la chaîne d'annotation, ou (3) à l'incomplétude des ressources, malgré la recherche de l'exhaustivité. Pour y remedier, nous effectuons plusieurs traitements :

– Pour chaque entrée simple de la terminologie, ne se terminant pas par *s* ou *x*, susceptible de marquer le pluriel, nous générons la forme plurielle correspondante. Nous obtenons un total de 6 924 nouvelles entrées, dont le type sémantique (*maladie*, *procédure* ou *médicament*) est identique à celui de l'entrée source. Parmi les nouvelles entrées, nous

3. Agence des Systèmes d'Information Partagés de Santé : `http://esante.gouv.fr/asip-sante`
4. `http://www.theriaque.org/`
5. `http://www.cnhim.org/`

avons par exemple {*achalasies*, *achalasie*} ou {*acholuries*, *acholurie*}, appartenant à la catégorie des maladies. Ceci permet de répondre en partie aux limites de la chaîne de traitement (section 3.1.1), qui parfois n'est pas adaptée aux données médicales et qui peut être fautive dans la reconnaissance des lemmes ;

– Dans les corpus de forum, la difficulté principale est différente : elle est liée à la présence fréquente des mots mal orthographiés. Nous faisons l'adaptation suivante :
  1. constitution d'un *lexique de référence* contenant les mots simples du français. Ce lexique est généré à partir des entrées simples de deux lexiques : L*efff*[6] (Sagot, 2010) et Lexique 3 [7] (New, 2006) (125 348 dans L*efff* et 405 793 dans Lexique 3) ;
  2. constitution du *lexique du corpus* contenant les mots graphiques, qui ne font pas partie du *lexique de référence* ni de la terminologie. Il s'agit de mots *a priori* inconnus et supposés correspondre à des formes graphiques fautives ;
  3. calcul de la distance d'édition entre chaque mot du *lexique du corpus* avec ceux de la terminologie. Nous utilisons le module `Text::Levenshtein::XS`. La distance de Levenshtein (Levenshtein, 1966) est une mesure de similarité entre deux chaînes de caractères. Elle considère trois opérations : la suppression d'un caractère, l'ajout d'un caractère et la substitution d'un caractère par un autre. Chacune de ces opérations a pour valeur 1. Par exemple, la distance de Levenshtein entre *ambolie* et *embolie* est de 1, que coûte la substitution de *a* par *e*. Les mots traités doivent avoir au moins six caractères, car avec une longueur inférieure les propositions contiennent trop de bruit. Deux seuils sont testés [1 ; 2]. Lors de cette étape, lorsqu'un mot inconnu est jugé similaire à un terme de la terminologie, ce mot hérite le type sémantique du terme. Par exemple, *ambolie* hérite le type sémantique *maladie* de *embolie*. Nous enregistrons également l'information sur le fait qu'il s'agit d'une forme mal orthographiées : le type sémantique devient $maladie_{orth}$.
  4. évaluation des propositions générées. Avec le seuil 1, parmi les 1 120 générations, 789 (73 % de précision) sont considérées correctes et ajoutées à la terminologie. Avec le seuil 2, 6 679 propositions sont générées. Une analyse des 100 premières montre que seules 15 sont correctes. Nous nous sommes alors limités aux propositions générées avec le seuil 1.

– Un autre moyen pour répondre au manque d'exhaustivité de la terminologie consiste en ajouts manuels de notions qui n'y sont pas enregistrées. Les ajouts de ce type restent minoritaires et sont repérés lors des analyses des corpus. Parmi les 317 entrées ajoutées, nous avons surtout les médicaments (*i.e.*, *actiskenan, alprazolam, depakote, anti-tnf*), mais aussi des maladies (*i.e.*, *DMLA, cécité bilatérale, leuconeutropénie*) et des procédures (*i.e.*, *embolisation artérielle, contrôle endoscopique, ostéodensitométrie*).

3 Méthode

La méthode proposée et mise en œuvre comporte deux étapes : l'annotation linguistique et sémantique des documents de corpus (section 3.1), la catégorisation automatique des documents (section 3.2). Nous indiquons également les modalités d'évaluation de ces deux aspects.

6. Le L*efff* est téléchargeable à cette adresse : `http://atoll.inria.fr/~sagot/lefff.html`
7. Lexique 3 est téléchargeable à cette adresse : `http://www.lexique.org/telLexique.php`

3.1 Annotation des corpus

L'annotation est appliquée aux données textuelles brutes afin d'obtenir des documents enrichis d'annotations linguistiques (*i.e.*, catégories syntaxiques) et sémantiques (*e.g.* termes médicaux, marqueurs d'incertitude et de négation, modifieurs, marques émotives).

3.1.1 Annotations linguistique et sémantique

Les annotations linguistique et sémantique sont réalisées grâce à la plateforme Ogmios (Hamon et Nazarenko, 2008), qui permet d'articuler plusieurs outils du Traitement Automatique de Langues (TAL), plusieurs ressources et plusieurs niveaux d'annotations.

Annotation linguistique. L'annotation linguistique est effectuée avec l'étiqueteur morpho-syntaxique `TreeTagger` (Schmid, 1994), qui assure la segmentation des documents en mots, la catégorisation des mots selon leurs catégories syntaxiques (*i.e.*, *alimentations* est un nom, *saignent* un verbe), et leur lemmatisation (*i.e.*, *alimentations* est lemmatisé vers *alimentation*, *saignent* vers *saigner*).

Annotation sémantique. L'annotation sémantique consiste en repérage des termes et de divers marqueurs des ressources (incertitude, négation, modifieurs, émotions). Pour chaque entrée des ressources, détectée par la plateforme Ogmios, le type sémantique correspondant lui est associé. L'annotation sémantique est effectuée avec les formes et les lemmes. En supplément des ressources, nous détectons également des marques émotives non lexicales, particulièrement fréquentes dans le corpus du forum :

- smileys ou émoticônes : comme par exemple *=)*, *;-)*, *:-/*, *XD* ;
- marques de rire : comme *lol*, *mdr*, *haha*, *hihi* ;
- ponctuations expressives : comme *!!!??*, *!!!!!!!!!!* ;
- mots avec lettres répétées : comme *maaaaaaal*, *grrrrrr*, *nooooooon* ;

Chaque marque émotive non lexicale est typée selon qu'elle dénote une émotion positive (*i.e.*, *=)*, *mdr*, *looool*), négative (*i.e.*, *:-(*, *:-/*) ou neutre (*i.e.*, *??!!?*, *grrrrrrrrrr*, *ohhhhh*). La détection de ces marques est réalisée avant la tokenisation de façon à s'assurer que chaque marque est considérée comme un seul token par `TreeTagger`. Dans le cas contraire, `TreeTagger` fait par exemple de *!!!* une suite de trois tokens.

Bilan des annotations. Grâce aux différentes ressources et annotations, nous obtenons un jeu de plusieurs types sémantiques :

- incertitude, négation et modifieurs ;
- émotions (lexicales et non lexicales) : positives, négatives et neutres ;
- notions médicales : maladies, médicaments et procédures ;
- les mêmes notions médicales mais dont les unités comportent des erreurs d'orthographe.

Évaluation de l'annotation. L'évaluation des annotations concerne les termes médicaux, les marqueurs de la négation et de l'incertitude, et les marques émotives non lexicales. Pour chaque corpus, 500 annotations sont contrôlées selon trois critères :

- détection : est-ce qu'une entrée donnée est détectée ? Si oui, est-ce qu'elle est contextuellement correcte ou incorrecte ?
- typage : pour chaque entrée détectée, reçoit-elle un type sémantique ? Si oui, est-ce que ce type est correct ou incorrect ?
- lemmatisation : est-ce que la lemmatisation de chaque entrée est correcte ou incorrecte ?

La précision (proportion d'annotations correctes par rapport à toutes les annotations effectuées) est calculée de deux manières :

- précision stricte P_s : on considère les vrais positifs comme étant uniquement des tokens pour lesquels tous les paramètres sont corrects (détection, typage et lemmatisation) ;
- précision lâche P_l : la notion de vrais positifs est élargie aux tokens dont la détection est correcte, alors que le type peut être absent (dû à un défaut de format de sortie) ou le lemme incorrect (dû à la difficulté de `TreeTagger` à traiter les termes médicaux).

3.2 Catégorisation automatique

La catégorisation automatique selon le degré de spécialisation des documents est effectuée avec une approche par apprentissage supervisé. En apprentissage supervisé, le système automatique a besoin d'exemples annotés en fonction des catégories visées (corpus d'apprentissage) pour pouvoir construire un modèle de classification. Ce modèle peut être appliqué à des nouvelles données, et le système peut faire des prédictions sur leur catégorisation. Nous utilisons divers algorithmes d'apprentissage supervisé implémenté dans la plate-forme `Weka` (Witten et Frank, 2005), et dont nous gardons le paramétrage par défaut.

3.2.1 Catégories à reconnaître

Nous cherchons à distinguer trois catégories : *scientifiques*, *cliniques* et *forum*. En rapport avec notre hypothèse, les utilisateurs de forum laissent libre cours à leurs émotions et jugements subjectifs, tandis que les documents scientifiques et cliniques doivent montrer plus de détachement et d'objectivité. Nous nous attendons donc à ce que les documents du corpus forum soient beaucoup plus faciles à discriminer que les documents des deux autres corpus. Nous effectuons une catégorisation bicatégorie en testant les trois couples possibles de catégories : documents cliniques et des forum, documents cliniques et scientifiques, documents de forum et scientifiques. Nous effectuons aussi un test multicatégorie, où les trois catégories sont à discriminer en même temps. Les corpus (tableau 1) sont nivellés par rapport aux corpus scientifique, avec 265 documents dans chaque catégorie.

3.2.2 Descripteurs utilisés et leur pondération

Les descripteurs proviennent de l'annotation sémantique (section 3.1.1) :

- incertitude, négation et modifieurs ;
- émotions (lexicales et non lexicales) : positives, négatives et neutres ;
- notions médicales : maladies, médicaments et procédures ;
- les mêmes notions médicales mais dont les unités comportent des erreurs d'orthographe.

Ces descripteurs sont pondérés de trois manières :

- *freq* correspond à la fréquence brute du descripteur ;
- *norm* correspond à la normalisation de la fréquence par le nombre de mots du document ;
- *tfidf* correspond à la pondération de la fréquence brute par *tfidf* (term frequency*inverse document frequency) (Salton, 1991) : $freq * \log(\frac{tot}{nbdoc})$, où $freq$ est la fréquence du descripteur, tot le nombre de documents dans le corpus et $nbdoc$ le nombre de documents où ce descripteur apparaît. Cette mesure permet d'évaluer l'importance du descripteur

par rapport au corpus. Le poids augmente proportionnellement à la fréquence du descripteur dans le document.

Nous avons au total 46 descripteurs.

3.2.3 Évaluation de la catégorisation

Nous effectuons une validation croisée (Sebastiani, 2002), qui permet aux algorithmes d'utiliser deux ensembles distincts de données pour les étapes d'entraînement et de validation. La validation croisée à n plis est effectuée n fois sur des partitions de données différentes et le résultat global correspond à la moyenne des performances. Nous effectuons une validation croisée à 10 plis. Les mesures d'évaluation correspondent aux moyennes de toutes les itérations. Nous calculons les mesures d'évaluation standard : précision (pourcentage de documents correctement catégorisés parmi tous les documents assignés à une catégorie), rappel (pourcentage de documents correctement catégorisés par rapport aux documents qui doivent être assignés à une catégorie) et f-mesure (leur moyenne harmonique).

La *baseline* correspond à l'assignation des documents dans la catégorie par défaut. Typiquement, pour un test avec deux catégories (*e.g.* forum et clinique) et un nombre de documents égal dans chaque catégorie, une telle *baseline* produirait une précision de 50 % : tous les documents sont donc assignés à une seule catégorie. Par rapport à une telle *baseline*, nous calculons aussi le gain obtenu, qui correspond à l'amélioration effective de la performance P par rapport à la *baseline* BL (Rittman, 2008) : $\frac{P-BL}{1-BL}$.

4 Présentation et discussion des résultats

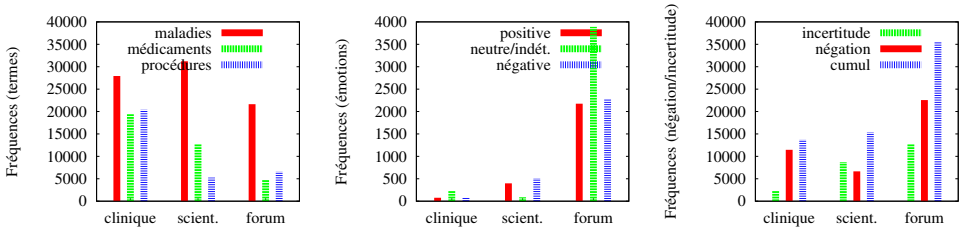
4.1 Annotations et leur évaluation

La méthode et les ressources sémantiques ont été exploitées pour effectuer une annotation sémantique des documents. Dans la figure 1, nous présentons les fréquences de différents types sémantiques dans les trois corpus. Nous pouvons par exemple voir que les notions médicales sont plus fréquentes dans les corpus scientifique et clinique (figure 1(a)), tandis que les émotions, incertitudes et négations sont plus fréquentes dans le corpus forum (figures 1(b) et 1(c)). Parmi les émotions les plus fréquentes, nous observons par exemple *peur, joie, tristesse, attirance* et *colère* dans le corpus forum, *joie* (*i.e.*, *plaisir, rassuré, satisfaisant*), *peur* (*i.e.*, *inquiet, anxieux, crainte, souci*) et *tristesse* (*i.e.*, *désolé, effondrement, destabilisé*) dans le corpus clinique, et *doute, étonnement, peur* (*i.e.*, *appréhender, menacer, craindre*) et *joie* (*i.e.*, *plaisir, heureux, rassuré*) dans le corpus scientifique.

Les résultats d'évaluation des annotations sont présentés dans le tableau 2, en termes des précisions stricte et lâche. Nous pouvons voir que la précision stricte, tout type sémantique confondu, est supérieure à 80 %, tandis qu'avec la précision lâche nous gagnons 10 %. Cela veut dire que dans 90 % de cas, les entités sont correctement reconnues, bien que leurs lemmes ou leurs types sémantiques peuvent être mal détectés.

4.2 Catégorisation automatique

Dans le tableau 3, nous présentons les performances de la catégorisation automatique des documents. Les résultats indiqués sont obtenus avec l'algorithme *RandomForest* (Breiman,



(a) Répartition des types de termes. (b) Répartition des types d'émotions. (c) Répartition de marqueurs d'incertitude et de négation.

FIGURE 1 – *Répartition des types de termes, d'émotions, et de l'incertitude et de la négation.*

Corpus	P_s	P_l
Clinique	0,87	0,91
Scientifique	0,81	0,89
Forum	0,88	0,90
Moyenne	0,85	0,90

TABLE 2 – *Évaluation de l'annotation sémantique en termes de précision stricte P_s et lâche P_l, dans les trois corpus traités.*

2001), qui est apparaît parmi les plus efficaces pour cette tâche. Les résultats correspondent aux valeurs de f-mesure telles que calculées par `Weka`. Nous indiquons les performances selon les pondérations des descripteurs (fréquences brutes, normalisées par la taille de documents ou par $tfidf$). Avec les tests à deux catégories, les performances sont systématiquement supérieures à 90 %. Avec le test à trois catégories, les performances sont proches ou supérieures à 90 %. La normalisation de la fréquence par la taille des documents produits les meilleurs résultats. Le gain par rapport à la *baseline* varie entre 0,818 et 0,896 pour les tests à deux catégories, et entre 0,824 et 0,861 pour le test à trois catégories : il s'agit d'une bonne amélioration par rapport à une *baseline* de base. Globalement, il s'agit de bons résultats, qui montrent qu'il existe une très forte corrélation des trois types de documents du domaine médical avec la notion de subjectivité et les notions médicales. Par ailleurs, il apparaît que l'emploi de ces descripteurs reste spécifique aux types de documents. Nous pouvons en effet discriminer assez facilement entre les types de documents dans les tests à deux ou trois catégories. Si les documents du corpus forum sont les plus aisés à distinguer, les documents scientifiques et cliniques montrent aussi des différences discriminantes. Nous ne nous attendions pas à ce que ce dernier test soit aussi performant. Nous nous sommes alors intéressés de voir de plus près le rôle de différents descripteurs de la subjectivité dans les corpus traités :

– dans le discours clinique des médecins, la subjectivité correspond le plus souvent à un marqueur de précaution par rapport à un diagnostic. Les médecins peuvent ainsi faire des hypothèses quant à un diagnostic pas certain, qui requiert des analyses biologiques ou d'autres analyses supplémentaires. Cependant, c'est la négation qui occupe une place prépondérante dans les documents cliniques (Chapman et al., 2001) ;

Catégories	freq	norm	tfidf
Clin./Forum	0,937	0,948	0,946
Clin./Scient.	0,909	0,946	0,911
Forum/Scient.	0,936	0,928	0,940
Clin./Scient./Forum	0,891	0,903	0,877

TABLE 3 – *Catégorisation automatique des discours médicaux : valeurs de F-mesure.*

– dans le discours scientifique, la subjectivité a été bien étudiée (Hyland, 1995; Light et al., 2004). Il apparaît qu'elle peut avoir plusieurs rôles : entre autres, elle devient incontournable car elle permet à l'auteur de se positionner par rapport aux travaux d'autres chercheurs ou par rapport à ses propres résultats expérimentaux obtenus lors des expériences scientifiques. Là aussi, la subjectivité sert de moyen de précaution, de distance et de protection ;
– dans les messages de forums, la subjectivité est exprimée de façon beaucoup plus générale (très souvent sans relation avec les notions médicales par exemple). De plus, elle montre souvent un lien fort avec les émotions des patients.

5 Conclusion et Perspectives

Dans le domaine médical, où il existe plusieurs types d'acteurs et d'utilisateurs avec des besoins informationnels souvent différents, nous proposons d'effectuer les expériences afin d'observer si les traces de subjectivité permettent de différencier entre les documents scientifiques, cliniques et les messages de forum. Nous utilisons pour ceci les descripteurs relatifs à l'incertitude, la négation, les modifieurs, les émotions (lexicales et non lexicales) et les notions médicales. Nos expériences montrent qu'il existe en effet une corrélation forte de la subjectivité et des notions médicales avec ces différents types de documents, avec les performances de catégorisation automatique étant souvent supérieures à 0,90. Le travail effectué est réalisé avec les documents médicaux relevant du thème *rhumatologie*, mais nous pensons que ces résultats sont généralisables à d'autres thèmes de la médecine. Cela voudrait dire que les descripteurs proposés peuvent être utilisés dans les moteurs de recherche pour apporter une caractérisation supplémentaire aux documents (ici, un niveau de spécialisation et les destinataires attendus). Dans cette optique, les résultats de recherche ne sont pas filtrés *a priori*, mais enrichis avec des annotations supplémentaires : il revient à l'utilisateur de décider s'il veut consulter un document avec un niveau de spécialisation donné. De la même manière, nous pensons que les descripteurs proposés et testés peuvent être appliqués dans la tâche de recherche d'information appliquée à d'autres domaines de spécialité.

Parmi les perspectives de ce travail, nous prévoyons d'effectuer la catégorisation automatique de phrases, afin de détecter les catégories d'émotions et de subjectivité des acteurs. Nous voulons aussi tester l'impact individuel de différents types de descripteurs. Les classes de marqueurs portant sur l'incertitude, la négation et leur interaction avec les modifieurs peuvent être affinées (Zadeh, 1972; Akdag et al., 1992; Cornelis et al., 2004; Akdag et al., 2001). Les descripteurs proposés peuvent être combinés avec d'autres descripteurs exploités dans la littérature (*e.g.* le lexique de manière générale, les informations syntaxiques et stylistiques, l'analyse

morphologique des termes). De même, une autre *baseline*, plus évoluée, peut être utilisée. Finalement, la méthode peut être adaptée à d'autres domaines (*e.g.* juridique, financier), où il existe aussi des utilisateurs avec de différents types d'expertise.

Remerciements

Ce travail travail est en partie soutenu par l'Agence Nationale de la Recherche (ANR) et la DGA, sous le numéro Tecsan ANR-11-TECS-012, et par le programme de recherche *Parlons de nous* de la Maison des Sciences de l'Homme de Montpellier (MSH-M). Les auteurs remercient également les relecteurs anonymes pour les remarques pertinentes qui ont permis d'améliorer la qualité du travail.

Références

Akdag, H., M. DeGlas, et D. Pacholczyk (1992). A qualitative theory of uncertainty. *Fundamenta Informaticae 17*(4), 333–362.

Akdag, H., I. Truck, A. Borgi, et N. Mellouli (2001). Linguistic modifiers in a symbolic framework. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9*(SI), 49–62.

Augustyn, M., S. Ben Hamou, G. Bloquet, V. Goossens, M. Loiseau, et F. Rynck (2008). *Constitution de ressources pédagogiques numériques : le lexique des affects*, pp. 407–414. Presses Universitaires de Grenoble.

Basilico, J. et T. Hofmann (2004). Unifying collaborative and content-based filtering. In *International Conference on Machine learning*, pp. 65–72.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Chapman, W., W. Bridewell, P. Hanbury, G. Cooper, et B. Buchanan (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform. 2001 Oct ;34(5) : 34*(5), 301–10.

Cornelis, C., M. DeCock, et E. Kerre (2004). *Efficient Approximate Reasoning with Positive and Negative Information*, pp. 779–85.

Côté, R. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Sherbrooke, Québec : Université de Sherbrooke.

Ekman, P. (1992). An argument for basic emotions. *Cognition and emotion 6*(3-4), 169–200.

Hadjouni Krir, M. (2012). *Un système de recherche d'information personnalisée basé sur la modélisation multidimensionnelle de l'utilisateur*. Thèse de doctorat, Université de Paris-Sud, Paris, France.

Hamon, T. et A. Nazarenko (2008). Le développement d'une plate-forme pour l'annotation spécialisée de documents Web : retour d'expérience. *TAL 49*(2), 127–154.

Herlocker, J., J. Konstan, L. Terveen, et J. Riedl (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems 22*(1), 5–53.

Hyland, K. (1995). The author in the text : Hedging in scientific writing. *Hong Kong papers in linguistics and language teaching 18*, 33–42.

Kassab, R. et J. Lamirel (2006). A new approach to intelligent text filtering based on novelty detection. In *Australasian Database Conference*, pp. 149–156.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady 707*(10).

Light, M., X. Y. Qiu, et P. Srinivasan (2004). The language of bioscience: facts, speculations and statements in between. In *ACL WS on Linking biological literature, ontologies and databases*, pp. 17–24.

Lindberg, D., B. Humphreys, et A. McCray (1993). The unified medical language system. *Methods Inf Med 32*(4), 281–291.

New, B. (2006). Lexique 3 : une nouvelle base de données lexicales. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006)*, Louvain, Belgique.

Pasi, G. (2010). Issues in personalizing information retrieval. *IEEE Intelligent Informatics Bulletin 11*(1), 3–7.

Pearson, J. (1998). *Terms in Context*, Volume 1 of *Studies in Corpus Linguistics*. Amsterdam/Philadelphia : John Benjamins.

Rittman, R. (2008). *Automatic discrimination of genres*. Saarbrucken, Germany : VDM.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malte.

Salton, G. (1991). Developments in automatic text retrieval. *Science 253*, 974–979.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, pp. 44–49.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys 34*(1), 1–47.

Witten, I. et E. Frank (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.

Zadeh, L. (1972). A fuzzy-set-theoretic interpretation of linguistic hedges. *Journal of Cybernetics 2*(3), 4–34.

Summary

Actors and users of the medical field (doctors, nurses, patients, interns, pharmacists, etc.) are neither from the same socio-professional category nor they have the same expertise level of the field. Their writings testify about this fact, as they show their subjectivity. We address the automatic study of the subjectivity in the medical discourse in texts written in French. We compare the doctors discourse (scientific literature, clinical reports) with the patients discourse (discussions from health fora) through a contrastive analysis of differences observed in the use of descriptors like the uncertainty and polarity markers, non-lexical (smileys, repeated punctuations, etc.) and lexical emotional markers, some personal deictics, and medical terms related to pathologies, treatments and procedures. We perform automatic annotation and categorization in order to better observe the specificities of the studied medical discourses.