Construction de lexiques pour l'extraction des mentions de maladies dans les forums de santé

Elise Bigeard[1, 2]
(1) laboratoire STL, Lille 3
(2) équipe ERIAS, laboratoire ISPED, Bordeaux
elise.bigeard@u-bordeaux.fr

RÉSUMÉ

Les forums de discussion et les réseaux sociaux sont des sources potentielles de différents types d'information, qui ne sont en général pas accessibles par ailleurs. Par exemple, dans les forums de santé, il est possible de trouver les informations sur les habitudes et le mode de vie des personnes. Ces informations sont rarement partagées avec les médecins. Il est donc possible de se fonder sur ces informations pour évaluer la qualité des soins ou les pratiques réelles des patients, en dehors des hôpitaux. Il s'agit cependant d'une source d'information difficile à traiter, essentiellement à cause des spécificités linguistiques qu'elle présente. Si une première étape pour l'exploration des forums consiste à indexer les termes médicaux présents dans les messages avec des concepts issus de terminologies médicales, cela s'avère extrêmement compliqué car les formulations des patients sont très différentes des terminologies officielles. Nous proposons une méthode permettant de créer et enrichir des lexiques de termes et expressions désignant une maladie ou un trouble, avec un intérêt particulier pour les troubles de l'humeur. Nous utilisons des ressources existantes ainsi que des méthodes non supervisées. Les ressources construites dans le cadre du travail nous permettent d'améliorer la détection de messages pertinents.

ABSTRACT

Building lexica for extraction of mentions of diseases from healthcare fora

Discussion fora and social networks potentially provide various types of information, which usually are not available in other sources. For instance, it is possible to find in discussion fora information on habits and lifestyle of persons, while this information is seldom shared with medical doctors. Hence, it is possible to rely on this information for evaluating the quality of healthcare or the real habits of patients, when they are outside hospitals. Yet, this information source is difficult to process, mainly because it offers some linguistic specificities. If the first step for exploring discussion fora consists of indexing the medical terms occurring in the messages with concepts from medical terminologies, this task is very difficult because patient utterances are different from official terminologies. We propose a method that permits to create and enrich lexica with terms and expressions meaning disorders or abnormalities, with a special interest to mood disorders. We exploit existing resources and non-supervised methods. Resources built during the work allow to improve the detection of relevant messages.

MOTS-CLÉS : Traitement automatique du langage, forums de discussion, extraction, terminologies, langage patient, maladies.

KEYWORDS: Natural Language Processing, Discussion Forums, extraction, patient language, disorders.

1 Introduction

Nous nous intéressons à l'usage réel qui est fait des médicaments par les patients et recherchons des méthodes pour détecter des mésusages de médicaments fréquents. Il est rare qu'un patient en parle au médecin qui lui a fait la prescription, soit parce qu'il n'est pas conscient d'utiliser le médicament dans des conditions non prévues (présence d'une contre-indication, d'une intéraction avec un autre produit, une erreur de dose ou de voie d'administration etc.), soit parce que ce mésusage est volontaire pour obtenir un effet particulier. Pour en savoir plus, il est nécessaire de se tourner vers d'autres sources d'information, telles que les forums de discussion. En effet, dans l'anonymat et sans avoir besoin de fournir d'effort particulier, les patients discutent volontiers de leurs maladies et de leurs pratiques sur les forums, nous ouvrant ainsi une fenêtre sur leur vie et leurs actions (Gauducheau, 2008). Une première étape pour aller vers l'exploration de forums consiste typiquement à indexer les termes médicaux correspondant aux maladies présentes dans les messages avec des termes de terminologies médicales de référence. La difficulté principale est que ces termes ont des formulations très différentes et diverses dans le langage patient.

Par exemple, pour désigner les troubles de l'humeur, des termes standards tels que *dépression* ou *déprimé* côtoient des expressions voisines comme dans cette phrase : *Je ne supporte rien je suis à fleur de peau c'est horrible, je suis hyper nerveuse et obligée de compléter avec une benzo pour me calmer tellement je suis dans un état de nerfs prononcé.* Nous pouvons ainsi observer plusieurs expressions de patients, qui ne peuvent pas être trouvées dans les terminologies ou ontologies.

D'autres obstacles, tels les fautes d'orthographe ou les abréviations non-standard, sont spécifiques au médium des réseaux sociaux, et même parfois à un réseau social particulier.

Pour résoudre ces difficultés, nous proposons une méthode permettant de créer et enrichir des lexiques avec des termes et expressions désignant une maladie. Nous portons un intérêt particulier aux troubles de l'humeur.

La constitution de lexiques est une tâche du TAL explorée depuis de nombreuses années. Nous nous intéressons ici aux lexiques qui associent une entrée à une liste de mots sémantiquement proches. La proximité sémantique peut référer à des mesures variées, dont notamment les relations synonymiques et paradigmatiques. (Budanitsky & Hirst, 2006) (Adam *et al.*, 2013).

La plupart des ressources sémantiques existantes se limitent aux relations synonymiques. Elles ne sont pas forcément accompagnées d'une mesure de proximité ou de confiance, rendant difficile la distinction entre synonymes, quasi-synonymes ou relations d'un autre type. Une autre limitation courante des ressources externes concerne le registre. Le vocabulaire utilisé par les usagers sur les forums de discussion ne coïncide pas avec le vocabulaire trouvé dans les terminologies de spécialité, plutôt destinées aux professionnels d'un secteur.

Pour le domaine médical, sur l'exemple de l'initiative Consumer Health Vocabularies (Zeng & Tse, 2006) en anglais, des initiatives existent pour créer des vocabulaires patient en français : sur le cancer du sein (Eholié *et al.*, 2016; Messai *et al.*, 2006), ou pour le domaine médical en général (Grabar & Hamon, 2014). Mais ces travaux ne couvrent que peu de termes et sont rarement adaptés au média particulier des forums de discussion.

Le crowdsourcing peut être une autre approche, mais est limité par la motivation et l'expertise des participants (Khare *et al.*, 2016).

Outre des ressources externes, il est possible de se fonder sur des méthodes distributionnelles pour

faire émerger du corpus les relations sémantiques qui nous intéressent. Ces méthodes utilisent les relations paradigmatiques comme dimension de la proximité sémantique entre des termes. Ce sont des méthodes non supervisées, capables de tirer parti de gros corpus et de créer des ressources fournies. De nombreux travaux récents s'intéressent à ces méthodes pour traiter des données médicales (Choy *et al.*, 2016) (Liu *et al.*, 2016).

Nous proposons d'explorer les ressources existantes, créées par spécialistes ou le crowdsourcing, et de construire de nouvelles ressources issues de nos corpus en exploitant des méthodes non supervisées. Ces ressources vont nous permettre de détecter les messages relatifs à un trouble donné.

2 Ressources

2.1 Forums de discussion

Notre corpus provient du forum de santé *www.doctissimo.fr* et de sa section dédiée aux anti-dépresseurs [1]. Les utilisateurs y parlent de leurs troubles de l'humeur, mais pas nécessairement d'anti-dépresseurs ou de dépression. Nous avons collecté les messages postés de 2010 à 2015. Dans chaque sujet, nous avons conservé uniquement le premier message car nous avons estimé que des informations à propos des actions d'un utilisateur sont plus susceptibles d'y apparaître, tandis que les réponses contiennent le plus souvent des opinions et des réactions à ce premier message. Notre corpus contient 6 185 posts composés de 828 250 occurrences de mots. Chaque message contient généralement quelques phrases. Ce corpus est utilisé pour détecter des informations sur l'usage des médicaments, et pour analyser le comportement des patients.

2.2 Ensemble de référence avec les noms de maladies

Un ensemble de 23 noms de maladies est extrait des Résumés de Caractéristiques du Produit (RCP) publiés par l'Agence nationale de sécurité du médicament et des produits de santé (ANSM). Il s'agit d'une ressource indépendante de notre travail. Pour permettre sa commercialisation, chaque médicament doit recevoir une autorisation de mise sur le marché (AMM) dont le RCP est une annexe : elle constitue par la suite un document de référence qui contient toutes les informations plus particulièrement destinées aux professionnels de santé. Les termes de notre ensemble sont liés aux maladies pour lesquelles sont prescrits des antidépresseurs, anxiolitiques et autres médicaments utilisés pour traiter les troubles de l'humeur. Cette liste contient des termes tels que *dépression*, *anxiété*, *nerveux*, *phobie*, *panique* ou *angoisse*. Chaque terme est associé à un code CIM10 (OMS, 1995), comme par exemple *anxiété / F41.9* ou *dépression / F32*. Le choix de la CIM-10 plutôt qu'une autre classification a été fait par des experts. Notre objectif est de trouver les messages dans lesquels ces termes, ou leurs équivalents sémantiques, sont présents. Dans la suite de ce papier, ces 23 termes seront appelés *seeds* lorsqu'ils sont utilisés comme source pour créer ou ajuster des ressources sémantiques.

Ces 23 seeds sont associées à deux jeux de codes CIM10. Le premier, appelé *seeds1*, correspond aux codes tels qu'extraits automatiquement des RCP. Dans ce cas, chaque code correspond à un diagnostic distinct. Cependant, les patients ne font pas forcément la distinction entre certains diagnostics

1. *http ://forum.doctissimo.fr/medicaments/antidepresseurs-anxiolytiques/liste sujet-1.htm*

proches et ne marquent pas cette distinction dans leur façon de parler des maladies. En effet, si nous recherchons les 30 termes les plus proches de chaque seed en exploitant les sorties de l'algorithme Word2Vec (Mikolov *et al.*, 2013a,b) (voir section 3.2 pour sa description) nous pouvons observer que les seeds comme *phobie* et *agoraphobie* ont 38 % de mots en commun, *angoisse* et *panique* 35 % de mots en commun, *stress* et *angoisse* ainsi que *phobie* et *panique* partagent 26 % de leurs mots. Il s'agit donc de termes qui apportent potentiellement des ambiguïtés lors de leur utilisation par les patients.

Nous avons donc créé un second jeu de codes, plus généraux et appelés *seeds2*, où les maladies le plus souvent associées et confondues par les patients figurent sous un même code. Ainsi, *agoraphobie/F40.0* est rapproché du terme moins spécifique *phobie/F40*. *angoisse/F41.0*, *anxiété/F41.9* et *anxiété généralisée/F41.1* sont également convertis vers un code plus général *autres troubles de l'anxiété/F41*. Ce regroupement est effectué à la main par un expert et se fonde d'une part sur les recoupements détectés par Word2Vec et notre connaissance du domaine et du corpus, afin de déterminer quelles seeds devraient être fusionnées.

2.3 Messages annotés manuellement

Pour constituer le corpus d'évaluation, 100 messages ont été sélectionnés aléatoirement. Ont été exclus certains messages très longs et peu cohérents car difficiles à annoter et peu représentatifs. Ces 100 messages ont été annotés manuellement par un expert, dont la tâche était d'y associer une expression d'un trouble de l'humeur en relation avec la liste de seeds. Ces annotations ont ensuite été converties en codes *seeds1* et *seeds2*. L'annotation a été effectuée au niveau des phrases. Ces 100 messages constituent les données de référence par rapport auxquelles nous évaluons les résultats produits dans ce travail. L'annotation s'est révélée difficile pour certains codes seeds1 proches. C'est pourquoi seulement 100 posts ont été annotés. Pour des travaux futurs, nous nous baserons uniquement sur seeds2 pour faciliter l'annotation et produire un corpus annoté plus fourni.// Ci-dessous des exemples de phrases et de leurs annotations.

	annotation	phrase
1.	*angoisse, dépression*	Bonjour, J'ai besoin de conseils car je ne suis pas bien en ce moment : angoissée en permanence et idées noires très présentes.
2.	*dépression*	Salut Je repose la question mais pour un nouvelle ad que je prends.
3.	*dépression*	J'ai vraiment du mal à me tirer du lit le matin, même après trois semaines.
4.	*phobie*	ça part ces obsession de peur d'avoir peur ?
5.	*angoisse*	pcq mon medecin a dit que mon mal de centre c de l'angoisse qu'en penser vous ? ?
6.	*boulimie*	Bonjour à tous, Je viens de suivre un traitement au Séroplex pendant un an, qui m'a été moralement bénéfique certes, mais dont je paie le prix fort : Plus de 15 kilos en plus, sachant que je faisais 65 kilos il y'a un peu moins d'un an ! ! !

Le but de ces annotations n'est pas de poser un diagnostic, mais d'annoter les termes faisant référence à un trouble. Ainsi, dans la dernière phrase présentée ci-dessus, la mention de prise de poids a été annotée en *boulimie*, bien que le reste de la phrase ne semble pas pointer vers un diagnostic de boulimie. Dans les messages non standards, il est rare que le nom de la maladie soit mal orthographié (phrase 5), en revanche les abréviations y sont fréquentes (phrase 2). Parfois, un terme n'est pas assez

précis pour être associé à un trouble donné. Dans ce cas, un choix arbitraire mais consistant a été fait par l'annotateur. C'est par exemple le cas d'idées noires qui est toujours annoté comme dépression. La phrase 3 donne un exemple de phrase où il est difficile d'isoler un terme univoquement associé à une maladie.

3 Méthodes

3.1 Prétraitement du corpus

Les messages sont segmentés en phrases et mots, étiquetés morpho-syntaxiquement et lemmatisés par Treetagger (Schmid, 1994). Les chiffres sont remplacés par une marque de substitution. Les diacritiques et la casse sont neutralisés pour diminuer la variation orthographique. Aucune correction orthographique n'est effectuée. Les mots vides peuvent avoir une importance pour les méthodes exploitées et n'ont donc pas été supprimés.

3.2 Construction des ressources sémantiques

La construction des ressources sémantiques est l'étape principale de notre méthode. La tâche consiste à créer des lexiques constitués d'expressions proches des troubles étudiés. Pour réaliser cette tâche, nous explorons deux approches : (1) utilisation de ressources existantes (*Lexique.org, Wikidata* et *JeuxdeMots*) desquelles les expressions pertinentes pour notre travail seront extraites ; (2) exploitation d'approches basées sur le corpus, en l'occurrence de méthodes distributionnelles (Brown clustering et Word2Vec). Ces algorithmes distributionnels sont non supervisés : ils se fondent sur la cooccurrence des mots et de leur voisinage pour regrouper les mots partageant une distribution similaire et créant ainsi des classes sémantiques. Chaque lexique existe en deux versions : création à partir des ensembles *seeds1* et *seeds2*. Dans le cas de Word2Vec, le processus est différent pour les deux séries de codes.

Lexique.org est un lexique existant (*http ://lexique.org*) construit par des psycholinguistes. Il contient des liens entre mots morphologiquement liés. Nous utilisons ce lexique pour enrichir l'ensemble initial de seeds avec leurs familles morphologiques. Ainsi *nerveux* est enrichi avec *nerveusement* et *nervosité*. Ce lexique apporte 36 termes.

Wikidata (*http ://wikidata.org*) est une base de connaissances sémantiques d'ordre général, dont le rôle est d'héberger et structurer le contenu sémantique de Wikipédia et d'autres projets Wikimédia. Cette base rend ces connaissances sémantiques accessibles à d'autres applications du web sémantique. Cette ressource contient plus de 24 millions d'objets, dont au moins 8 000 objets dans le domaine médical. Plusieurs étapes sont nécessaires pour l'exploitation de cette ressource : (1) Extraction des objets représentant des maladies. Wikidata contient des propriétés telles que *subclass of disease* mais celles-ci ne sont pas utilisées systématiquement. Nous proposons donc d'extraire les objets possédant un code CIM10. Nous nous intéressons aux objets dont le code CIM10 est identique au code d'une seed dans l'ensemble *seeds1*. Les codes *seeds2* correspondant à des classes de maladies et non pas à des maladies, nous nous basons toujours sur les codes *seeds1* pour trouver les objets pertinents. (2) Nous utilisons la propriété *alternative label* pour récolter différentes appellations. Ainsi *agoraphobia* (Q174589) est associée aux labels français *agoraphobe* et *peur sociale*. Nous excluons les labels variant uniquement par l'utilisation de diacritiques {*schizophrénie*, *schizophrènie*, *schizophrenie*},

celles-ci étant neutralisées durant l'analyse des posts, mais nous conservons les variations de flexion {*phobie*, *phobies*} afin de pallier un éventuel défaut de lemmatisation. Nous obtenons 61 termes.

JeuxdeMots (Lafourcade, 2007) est une ressource existante issue du crowdsourcing qui propose des relations sémantiques entre deux mots. Chaque relation possède un poids, accordé en fonction de la fréquence à laquelle ces mots sont entrés en association par les participants. Le type de relation peut être indiqué, mais ce n'est pas le cas le plus fréquent. Nous utilisons donc les relations de manière non-typée, ce qui correspond généralement à une relation sémantique faible. Nous utilisons également les relations synonymiques et hyponymiques. Nous conservons les premiers 80 % des termes associés à chaque seed, à l'exception de 5 seeds pour lesquelles les mots associés sont particulièrement ambigus (*anxiété, tristesse, angoisse, panique et fatigue*). Pour ces 5 mots, nous conservons uniquement les premiers 30 %. La ressource finale apporte 1 566 termes.

Brown clustering (Brown *et al.*, 1992; Liang, 2005) est un algorithme distributionnel qui permet de générer de nouvelles ressources à partir de notre corpus. Nous avons empiriquement choisi de générer 500 clusters. Chaque terme du corpus est placé de façon ordonnée dans un cluster. Nous avons d'abord soumis à l'algorithme l'ensemble de discussions du forum médicament de Doctissimo, contenant 61% de messages dans la catégorie contraception, et seulement 10% dans la catégorie antidépresseurs et anxiolitiques. En conséquence, les termes liés aux antidépresseurs se retrouvaient regroupés dans un petit nombre de clusters, sans distinction entre les différentes maladies. Afin d'obtenir des distinctions plus fines, nous avons choisi d'utiliser uniquement le corpus des antidépresseurs et anxiolitiques. Nous générons ensuite deux ressources distinctes : *Brown1* est construit en conservant les 30 premiers termes de chaque cluster contenant une seed. *Brown2* est construit de la même façon, mais en ajoutant également les clusters où se trouve un mot de la même famille morphologique qu'une seed, tel que fourni par *Lexique.org*. Le corpus d'évaluation est exclu du corpus d'entraînement de Brown. Les deux ressources résultantes contiennent respectivement 397 et 725 termes.

Word2Vec (Mikolov *et al.*, 2013a,b) est également un algorithme distributionnel qui permet de créer des clusters autour d'une seed donnée. Afin de pouvoir comparer les résultats avec ceux de Brown, nous avons utilisé le même corpus constitué uniquement de la catégorie antidépresseurs et anxiolitiques de Doctissimo. Cependant, ce corpus étant composé d'environ 800 000 mots seulement, cela peut limiter les performances de word2vec. Le corpus d'évaluation est exclu de ce corpus d'entraînement. Nous utilisons l'algorithme *cbow* avec une fenêtre de 10 mots et l'utilisation de bigrammes. Pour créer un cluster, nous soumettons à l'algorithme une requête pouvant être constituée d'un ou de plusieurs mots. Pour l'ensemble *seeds1*, chaque requête est constituée d'une seule seed. Pour l'ensemble *seeds2*, nous avons combiné toutes les seeds correspondant à un même code en une seule requête. Pour chaque cluster, nous gardons les 30 premiers mots retournés. Nous créons ainsi quatre lexiques :

1. *W2Vseeds*, basé uniquement sur la liste de seeds.
2. *W2Vmorph*, basé sur les seeds ainsi que sur leurs familles morphologiques.
3. *W2Vortho*, où chaque seed est associée à un mot mal orthographié rencontré souvent dans le corpus (nous avons choisi *medcin*) pour former une requête de deux mots.
4. *W2Vcomb* est l'association de *W2Vmorph* et *W2Vortho* : les seeds et leurs familles morphologiques sont utilisées, et chaque terme est également associé à un mot mal orthographié.

Ces lexiques comportent respectivement 313, 616, 242 et 395 termes pour les codes *seeds1* ; 213, 242, 200 et 219 termes pour les codes *seeds2*. La variation de taille entre chaque lexique s'explique par le recoupement entre chaque cluster généré, qui peut être plus ou moins important. Les lexiques générés pour les codes *seeds2* sont plus petits, car moins de requêtes ont été réalisées pour leur création.

La **Combinaison** de toutes ces ressources génère deux lexiques supplémentaires : *Total* correspond à l'addition de tous les lexiques, et *Vote* contient chaque terme apparaissant dans au moins deux ressources différentes. Dans *Vote*, les seeds sont toujours conservées. Avec l'ensemble *seeds1*, ces lexiques contiennent 2 969 et 135 entrées, alors qu'avec l'ensemble *seeds2*, ces lexiques contiennent 2 508 et 130 termes, respectivement.

3.3 Rechercher les posts pertinents

En utilisant les seeds décrites à la section 2.2 et les ressources sémantiques construites précédemment, nous recherchons les messages de forum contenant une mention des maladies étudiées. Nous effectuons cette recherche en enrichissant la liste de seeds avec les termes collectés dans les ressources sémantiques. Les termes récoltés sont rassemblés dans des lexiques. Les termes de ces lexiques sont associés avec le code CIM10 de la seed qui a permis de les collecter. Ce code peut être précisément le code de la maladie (*seeds1*) ou un code plus générique, désignant plusieurs maladies proches (*seeds2*). Nous évaluons ensuite la qualité de l'étiquetage selon deux niveaux de granularité : au niveau des phrases et au niveau des posts, le niveau des phrases étant supposé être plus précis, mais aussi plus difficile à traiter.

4 Résultats et discussion

4.1 Acquisition des ressources sémantiques

TABLE 1 – Desciption des ressources sémantiques obtenues

Lexique	*Taille seeds1*	*Taille seeds2*	*Exemples*
seeds	23	23	*crise d'angoisse*
Lexique.org	36	36	*angoissant* , *angoissé*
Wikidata	61	61	*attaque de panique*
JdM	1 566	1 566	*convulsion* , *crampe* , *médicament*
Brown1	397	397	*dépersonalisation* , *hystérie* , *alzheimer*
Brown2	725	725	*hystérique* , *stresser* , *suicidaire* , *bailler*
W2V seeds	313	213	*spasmophilie* , *violent* , *gros*
W2V morph	616	242	*cercle vicieux* , *trembler* , *devenir fou*
W2V ortho	258	200	*spasmo* , *dangoisse* , *fesais*
W2V comb	395	219	combinaison de *W2Vmorph* et *W2Vortho*
Total	2 969	2 508	fusion de tous les lexiques
Vote	135	130	vote de tous les lexiques

Dans le tableau 1, nous décrivons les lexiques générés à partir des ressources existantes et du corpus. Nous indiquons leur taille et donnons quelques exemples de termes dans la dernière colonne. Les exemples sont donnés pour la seed *panique*, *crise d'angoisse*, dont le code CIM10 est F41.0. Nous avons tenté de sélectionner des exemples représentatifs du contenu de chaque ressource. Nous constatons que les termes proposés par les différentes méthodes sont plus ou moins sémantiquement

proches de la seed. *Lexique.org*, qui propose des mots de la même famille morphologique, contient des mots très proches. *Wikidata* contient également des mots très proches. Les autres ressources peuvent contenir des mots plus distants (*convulsion, crampe, médicament, alzheimer, bailler, spasmophilie, violent, gros, trembler, fesais...*) aux côtés de mots plus pertinents (*dépersonalisation, hystérie, hystérique, stresser, suicidaire, cercle vicieux, devenir fou, spasmo, dangoisse...*).

Le lexique issu de *Wikidata* contient 61 termes, qui peuvent être des expressions composées de plusieurs mots : *dépression, dépression clinique, dépression caractérisée, dépression majeure*. Nous trouvons également des variantes morphologiques et syntaxiques intéressantes (*intolérance au glucose, intolérance aux glucides, intolérance glucidique*) ainsi que des fautes d'orthographe. Toutes ces variantes ne sont pas forcément présentes dans notre corpus.

Le lexique *JeuxdeMots* contient 1 566 termes. Ces termes peuvent être plus ou moins sémantiquement proches des seeds. Par exemple pour la seed *dépression* nous obtenons :

— *anxiété, tristesse, angoisse, panique, fatigue*
— *pauvreté, rater, malheureusement, sanglot, résigner, souffrira*

Si la première série contient des termes pertinents et intéressants, la deuxième série risque d'apporter surtout du bruit.

Les lexiques *Brown* tendent à regrouper les termes d'une même classe sémantique : les maladies, les médicaments, etc. Par exemple, *anxiété* se trouve dans le même cluster que *hypocondrie, timidité, hyperphagie, isolement, urticaire...* . Nous présumons que ces clusters assez larges nous seront moins utiles pour notre tâche.

L'algorithme *Word2Vec* donne généralement des candidats assez proches. Pour l'exemple de *dépression*, les premiers termes obtenus sont *dépressif* et *anxieux*. Une autre piste qui sera explorée dans des travaux futurs est de réutiliser les premiers termes pour obtenir davantage de candidats pertinents, en procédant de proche en proche. Ainsi, *anxieux* nous donne à son tour *nerveux, anxiété, stresser, panique, déprime*. En ajoutant aux seeds leur famille morphologique, nous avons construit les lexiques *W2V morph* et *W2V comb*. Construire une requête constituée de mots de classes morphologiques différentes nous permet de neutraliser la tendance de l'algorithme à sélectionner des mots de même classe morphologique que la requête. Nous avons utilisé des mots mal orthographiés pour générer *W2V ortho* et *W2V comb* afin d'obtenir des mots également mal orthographiés dans le cluster, comme exemplifié dans le tableau 1.

Le lexique *Total* contient bien sûr le plus grand nombre d'entrées, indiquant que nos lexiques sont complémentaires et peuvent chacun apporter de nouveaux termes. Mais nos lexiques se chevauchent également. Voici le nombre d'entrées exclusives à chaque lexique : 47 dans *Wikidata*, 1 472 dans *JdM*, 625 dans les deux *Brown* regroupés et 639 dans les quatre *W2V* regroupés.

4.2 Formes recherchées

Le trouble le plus représenté dans le corpus d'évaluation est la dépression, qui constitue un bon exemple de la variété des formes recherchées. 103 extractions dans 96 phrases ont été identifiées dans le corpus d'évaluation. Les termes *dépression, anti-dépresseur* et *dépressif*, ainsi que leurs variations orthographiques et abréviations, constituent 76% des termes à identifier. Avec une simple neutralisation de la casse, sans autre prétraitement, ces trois termes totalisent 40 formes distinctes. Le terme *anti-dépresseur* en particulier totalise 13 formes différentes, à cause de variations orthographiques pouvant se combiner les unes avec les autres : présence ou non de l'accent sur *dépression*, forme au

TABLE 2 – Résultats pour l'ensemble *seeds1*

	Message				Phrase			
Lexique	*TP*	*P*	*R*	*F*	*TP*	*P*	*R*	*F*
seeds	81	0.920	0.529	0.672	102	0.918	0.430	0.586
Lexique.org	94	**0.921**	0.614	**0.737**	123	**0.924**	0.518	**0.664**
Wikidata	86	0.914	0.562	0.696	108	0.915	0.455	0.608
Jdm	115	0.392	**0.751**	0.515	135	0.242	**0.569**	0.340
Brown1	91	0.590	0.594	0.592	108	0.486	0.455	0.470
Brown2	91	0.590	0.594	0.592	108	0.486	0.455	0.470
w2v seeds	82	0.577	0.535	0.555	102	0.528	0.430	0.474
w2v morph	91	0.476	0.594	0.529	108	0.421	0.455	0.438
w2v ortho	85	0.658	0.555	0.602	105	0.668	0.443	0.532
w2v comb	84	0.677	0.549	0.606	104	0.688	0.438	0.536
vote	109	0.698	0.712	0.705	135	0.613	0.569	0.590
total	135	0.289	0.882	0.436	165	0.170	0.696	0.273

pluriel ou au singulier, séparateur entre *anti* et *dépresseur*, qui peut être un espace, un tiret ou être absent. Il existe aussi l'abréviation *AD*, qui peut être mise au pluriel *ADs*, ainsi que les abréviations *anti dep* et *anti dd*, chacune trouvée une seule fois.

Neutraliser les variations de la liste suivante permet de diminuer le nombre de formes distinctes pour *dépression, antidépresseur* et *dépressif* de 40 formes à 26 formes.

- diacritique (dépression / depression)
- flexion (dépressif / dépressive)
- séparateur dans un mot composé (antidépresseur / anti dépresseur / anti-dépresseur)
- séparateur entre les mots : à la suite d'une erreur typographique, l'espace entre le terme recherché et le mot voisin est manquant

15% des annotations restantes sont des phrases pour lesquelles il est difficile d'identifier un terme isolé comme synonyme du trouble recherché. On trouve par exemple *"moins de joie de vivre qu'avant"* et *"je n'arrive plus a reflechir ni imaginer"*.

Les 10% restants sont des termes plus variés, qui sont présents une à trois fois dans le corpus d'évaluation. On trouve par exemple *déprime* (3 occurences), *idées noires* (3 occurences) et *mal-être* (2 occurences).

Sur l'exemple de *dépression*, il apparaît que la majorité des termes à retrouver sont des variations orthographiques de seulement 3 termes différents. Sur ces trois termes, *dépression* est la seed, *dépressif* fait partie de la même famille morpho-syntaxique, et *anti-dépresseur* est formé sur la même base, mais n'est pas un synonyme de la seed.

4.3 Recherche des messages pertinents et évaluation

Dans les tables 3 et 2, nous présentons les résultats de la recherche de messages pertinents pour les maladies étudiées. Pour chaque lexique et chaque unité (phrase ou message) nous indiquons les mesures suivantes : nombre de vrais positifs TP, précision P, rappel R et F-mesure F (Sebastiani,

TABLE 3 – Résultats pour l'ensemble *seeds2*

	Message				Phrase			
Lexique	*TP*	*P*	*R*	*F*	*TP*	*P*	*R*	*F*
seeds	82	0.911	0.522	0.663	98	0.859	0.388	0.535
Lexique.org	93	0.902	0.592	**0.715**	120	**0.869**	0.476	**0.615**
Wikidata	83	**0.912**	0.528	0.669	100	0.862	0.396	0.543
JDM	129	0.452	**0.821**	0.583	159	0.296	**0.630**	0.403
Brown1	93	0.654	0.592	0.622	105	0.541	0.416	0.470
Brown2	103	0.490	0.656	0.561	125	0.378	0.496	0.429
w2v seeds	94	0.556	0.598	0.576	109	0.480	0.432	0.455
w2v morph	100	0.628	0.636	0.632	109	0.542	0.432	0.481
w2v ortho	85	0.720	0.541	0.618	102	0.698	0.404	0.512
w2v comb	86	0.722	0.547	0.623	100	0.689	0.396	0.503
vote.txt	102	0.75	0.649	0.696	130	0.702	0.515	0.594
total.txt	136	0.349	0.866	0.498	171	0.222	0.678	0.335

2002). Le meilleur résultat pour chaque mesure est marqué en gras. Comme seulement 100 messages ont pu être annotés pour l'évaluation, la représentativité de ces résultats est limitée. Aucune validation croisée n'a été effectuée.

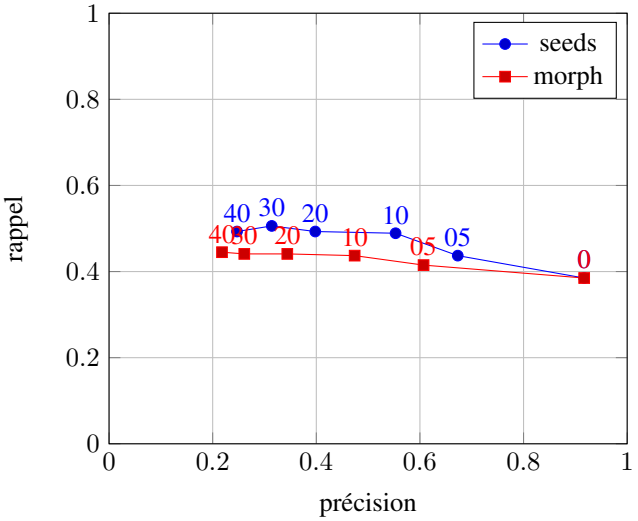
Comme attendu, identifier les messages est plus facile qu'identifier les phrases : la granularité est alors moins fine et le nombre de mots potentiellement pertinents plus important. De même, la granularité moins fine au niveau des seeds rend les résultats pour *seeds2* légèrement meilleurs que pour *seeds1*.

Nous notons une nette différence entre *seeds*, *Lexique.org* et *Wikidata* d'une part, et *JDM* et les méthodes distributionnelles d'autre part, en ce qui concerne la précision. Ce n'est pas surprenant : comme les premières ressources restent confinées aux synonymes et quasi-synonymes, la précision est donc excellente.

Pour *JDM* et les ressources distributionnelles, les termes obtenus étant classés par poids, il est possible d'ajuster les résultats en modifiant le nombre de termes sélectionnés pour chaque seed. La figure 1 montre l'évolution du rappel et de la précision selon le nombre de termes par seed sélectionné. Les paramètres suivant ont été utilisés : le lexique *W2V morph*, l'ensemble *seeds1*, les résultats calculés au niveau des phrases.

Dans le graphe de la figure 1, prendre "0 terme" correspond à ne prendre que les seeds. Nous observons que la précision diminue rapidement pour une faible amélioration du rappel. Au regard de ces résultats, nous avons choisi de conserver les 30 premiers termes pour nos résultats, car cette configuration est la plus représentative de la ressource, malgré sa faible précision. En effet, nos ressources fournissent des termes pertinents, améliorant le rappel, mais ils sont mêlés à des termes bien trop génériques, ce qui déteriore la précision assez rapidement. Par exemple, parmi les termes associés à *dépression* par Word2Vec, nous trouvons *trouble* en troisième position, alors que les premiers termes associés à *angoisse* sont, dans l'ordre, *sentiment, crise, crainte*. A cause de ces termes génériques haut placés, même si nous en sélectionnons un très petit nombre, cela ne permettrait pas de conserver une précision satisfaisante. Il est donc nécessaire d'utiliser une méthode complémentaire pour filtrer les termes fournis par ces ressources. Plusieurs sont proposées par (Claveau *et al.*, 2014) et seront explorées dans de futurs travaux.

FIGURE 1 – Résultats obtenus avec *W2V* selon le nombre de termes retenus par seed



Le meilleur rappel est obtenu avec le lexique *JDM*, qui est le plus volumineux. *Brown2* et *W2V morph*, tous deux basés sur l'exploitation de seeds et de leurs familles morphologiques, sont le plus souvent second et troisième.

Si nous observons les termes qui apportent le plus de faux négatifs dans le corpus d'évaluation, nous remarquons que le trouble le plus fréquent est la dépression, avec 46% des faux négatifs pour Lexique.org (seed1, phrase). C'est aussi le trouble que l'annotateur a généralement associé aux expressions les plus vagues, à cause de sa généricité. On trouve notamment des expressions telles que *"un peu moins de joie de vivre qu'avant","baisse de moral"* ou *"je ne me reconnais plus, plus rien de m'intéresse, je n'arrive plus à réfléchir ni à imaginer"*.

Parmi les faux négatifs de dépression, 35%, soit 19 annotations, correspondent au terme *anti-dépresseur* ou à une variante de celui-ci (*AD*, *anti dep*...). En effet, dans de nombreux messages, une personne demande conseil sur un médicament sans citer la maladie pour laquelle le médicament serait pris. On a par exemple un message dont le titre est le nom d'un médicament, et le corps de texte uniquement *"Qui a eu une amélioration avec cet AD ?"*. Dans ce cas, il existe un lien implicite entre le médicament et la maladie, et l'annotation est donc correcte. Mais *antidépresseur* n'étant pas un synonyme de *dépression* et n'étant pas utilisé dans le même contexte morpho-syntaxique, ce terme n'est retrouvé que dans deux de nos lexiques : Jeux de Mots et w2v morph. Le cas de w2v morph est le plus intéressant et celui sur lequel nous nous concentrons : parmi les variantes morphologiques utilisées comme seed, on trouve *anxiolitique*, dérivé de *anxiété*. Or, *anxiolitique* est le nom d'une classe de médicaments, auquel word2vec associe facilement *antidépresseur*. *Antidépresseur* est donc présent dans ce lexique, mais incorrectement associé à *anxiété*. Les variantes *AD* et *antidep* sont également présentes, mais incorrectement classifiées. Il apparaît donc que word2vec peut trouver les variantes orthographiques d'*anti-dépresseur* à partir de ce terme, mais pas à partir du terme *dépression*.

Le regroupement de termes au sein de *seeds2* ne produit pas d'amélioration remarquable excepté

pour les lexiques *W2V*, qui ont été générés avec des requêtes multi-mots. Ces lexiques sont alors moins fournis que leurs équivalents *seeds1*. Nous obtenons donc un rappel plus faible, mais largement contrebalancé par l'amélioration de la précision.

Lexique.org obtient la meilleure F-mesure dans toutes les configurations, principalement due à son excellente précision. Si nous souhaitons accorder davantage d'importance au rappel, *Vote* fournit un meilleur équilibre entre les deux mesures.

5 Conclusion et travaux futurs

Nous avons proposé de construire et enrichir automatiquement des lexiques pour la détection de messages de forums parlant de troubles de l'humeur. La particularité de ce travail est la nécessité d'analyser et traiter des messages écrits par des patients sur un forum de discussion, qui contiennent un vocabulaire et des expressions spécifiques. Nous avons proposé d'utiliser des ressources existantes et de créer de nouvelles ressources grâce aux méthodes non supervisées exploitant notre corpus. La ressource *Lexique.org* fournit la meilleure F-mesure, alors que la ressource *Vote*, qui contient les termes proposés par au moins deux ressources, permet d'obtenir le meilleur équilibre entre précision et rappel tout en gardant une F-mesure élevée.

Les ressources existantes que nous avons exploitées contiennent des quasi-synonymes ou bien de nombreux termes faiblement liés sémantiquement ; tandis que les méthodes distributionnelles fournissent des termes pertinents mais aussi des termes plus généraux qui déteriorent la précision. Pour améliorer les ressources distributionnelles, il est possible d'utiliser des corpus plus grands afin d'améliorer la granularité des clusters. Un filtrage supplémentaire de clusters obtenus est également souhaitable. Cela permettrait d'utiliser davantage de critères de sélection, notamment en ce qui concerne la spécificité des termes. Nous comptons également étudier davantage le fonctionnement de Word2Vec et différentes méthodologies pour la combinaison de termes dans les requêtes adressées à Word2Vec afin d'obtenir des clusters plus précis et plus diversifiés. Enfin, nous étendrons notre méthode à un éventail plus large de maladies afin d'évaluer son adaptabilité dans un contexte plus varié.

Remerciements

La présente publication s'inscrit dans le programme *Drugs Systematized Assessment in real-liFe Environnement (DRUGS-SAFE)* financé par l'Agence Nationale de Sécurité du Médicament et des Produits de Santé. Cette publication ne représente pas nécessairement l'opinion de l'ANSM.

Je remercie vivement Natalia Grabar et Frantz Thiessard pour leur supervision et leur aide à la rédaction, Pierre Simonetti pour l'annotation et toute l'équipe ERIAS.

Références

ADAM C., FABRE C. & MULLER P. (2013). Évaluer et améliorer une ressource distributionnelle : protocole d'annotation de liens sémantiques en contexte. *Traitement Automatique des Langues*, **vol.**

54(n 1), pp. 71–97.

BROWN P., DESOUZA P., MERCER R., DELLA PIETRA V. & LAI J. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, **18**(4), 467–479.

BUDANITSKY A. & HIRST G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, **32**(1), 13–47.

CHOY Y., CHIU C. & SONTAG D. (2016). Learning low-dimentional representations of medical concepts. In *AMIA Jt Summits Transl Sci Proc*.

CLAVEAU V., KIJAK E. & FERRET O. (2014). Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels. In *21ème conférence sur le Traitement Automatique des Langues Naturelles, TALN 2014*, p. 12 p., Marseille, France.

EHOLIÉ S., TAPI NZALI M. D., BRINGAY S. & JONQUET C. (2016). MuEVo, a breast cancer Consumer Health Vocabulary built out of web forums. In A. PASCHKE, A. BURGER, A. SPLENDIANI, M. MARSHALL & P. ROMANO, Eds., *SWAT4LS : Semantic Web Applications and Tools for Life Sciences*, Amsterdam, Netherlands.

GAUDUCHEAU N. (2008). La communication des émotions dans les échanges médiatisés par ordinateur : bilan et perspectives. *Bulletin de psychologie*, p. 389–404.

GRABAR N. & HAMON T. (2014). Automatic extraction of layman names for technical medical terms. In *ICHI 2014*, Pavia, Italy.

KHARE R., GOOD B. M., LEAMAN R., SU A. I. & LU Z. (2016). Crowdsourcing in biomedicine : challenges and opportunities. *Briefings in Bioinformatics*, **17**(1), 23.

LAFOURCADE M. (2007). Making people play for lexical acquisition. *7th Symposium on Natural Language Processing*. jeuxdemots.

LIANG P. (2005). *Semi-Supervised Learning for Natural Language*. Master, Massachusetts Institute of Technology, Boston, USA.

LIU F., CHEN J., JAGANNATHA A. & YU H. (2016). Learning for biomedical information extraction : Methodological review of recent advances. *CoRR*, **abs/1606.07993**.

MESSAI R., ZENG Q., MOUSSEAU M. & SIMONET M. (2006). Building a bilingual french-english patient-oriented terminology for breast cancer. In *MedNet*.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *Workshop at ICLR*.

MIKOLOV T., SUSTKEVER I., CHEN K., CORRADO G. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *NIPS*.

OMS (1995). *Classification statistique internationale des maladies et des problèmes de santé connexes — Dixième révision*. Organisation mondiale de la Santé, Genève.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, p. 44–49, Manchester, UK. treetagger.

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

ZENG Q. & TSE T. (2006). Exploring and developing consumer health vocabularies. *JAMIA*, **13**, 24–29.