Augmentation d'index par propagation sur un réseau lexical
Application aux comptes rendus de radiologie

Lionel Ramadier[1,2], Mathieu Lafourcade[2],
(2) Imaios, 34000 MONTPELLIER - France
(1) Lirmm, Université Montpellier, France
lionel.ramadier@imaios.com, mathieu.lafourcade@lirmm.fr

Résumé. Les données médicales étant de plus en plus informatisées, le traitement sémantiquement efficace des rapports médicaux est devenu une nécessité. La recherche d'images radiologiques peut être grandement facilitée grâce à l'indexation textuelle des comptes rendus associés. Nous présentons un algorithme d'augmentation d'index de comptes rendus fondé sur la propagation d'activation sur un réseau lexico-sémantique généraliste.

Abstract.
Index augmentation through propagation over a lexical network – application to radiological reports
Medical data being increasingly computerized, semantically effective treatment of medical reports has become a necessity. The search of radiological images can be greatly facilitated through textual indexing of the associated reports. We present here an index enlargement algorithm based on spreading activations over a general lexical-semantic network.

Mots-clés : réseau lexico-sémantique, propagation, indexation, recherche d'information, imagerie médicale
Keywords: lexico-semantic network, propagation, indexation, information retrieval, medical imaging

1 Introduction

L'informatisation des professions de santé et le développement du dossier médical personnalisé (DMP) entrainent une progression rapide du volume d'information numérique. Les systèmes informatiques médicaux permettent de stocker de grandes masses d'informations (dossier médical, résultats d'examens complémentaires, images et comptes rendus radiologiques, par exemple), d'y accéder en vue d'améliorer la prise en charge des patients, de collecter de nouvelles informations ou encore de fournir une aide à la décision pour l'amélioration de la qualité des soins. Un accès simple et efficace à ces documents médicaux est devenu un objectif primordial pour les établissements de santé. La recherche d'information dans le domaine médical fait à l'heure actuelle l'objet de nombreux travaux de recherche ainsi que des campagnes d'évaluation (Voorhees et al, 2011, Diaz-Galiano et al, 2009, L.Goeuriot, 2014). L'indexation efficace des divers compte rendus (opératoires, radiologiques, etc.) est une tâche nécessaire qui permet d'améliorer la recherche d'informations dans un but clinique mais aussi pédagogique. Par exemple, Dinh *et al.*, (2010) ont réalisé une indexation sémantique des dossiers médicaux des patients afin qu'elle serve de support à des processus de recherche d'information. Leur système d'indexation repose sur l'utilisation de MeSH (Medical Subject Headings), implique un traitement de la désambiguïsation, de l'extraction de valeurs cliniques et de la pondération de concepts. Pouliquen, (2002) a aussi réalisé une indexation automatique par reconnaissance et extraction des concepts médicaux. Il a exploité les mots composés et les associations de mots pour convertir une phrase en mots de référence avec l'aide d'un thésaurus médical.

Dans le domaine de l'imagerie médicale, la quantité d'images et de comptes rendus devient de plus en plus importante, ce qui fait de leur accessibilité pour le corps médical un enjeu majeur. En effet, tirer le meilleur parti d'une telle collection d'images radiologiques en identifiant rapidement l'information pertinente suppose qu'elles soient correctement indexées à partir de leurs comptes rendus. Cette tâche d'indexation, afin d'être utile aux praticiens, doit tenir compte des requêtes qu'ils effectuent. Plusieurs auteurs, notamment Hersh *et al.,* (2001) et Huang *et al.,* (2003) ont réalisé une indexation automatique des comptes rendus radiologiques en se fondant sur le métathésaurus UMLS. Pour améliorer leurs résultats (surtout en ce qui concerne la précision) ils ont utilisé une sous-section de la terminologie UMLS. Toujours dans l'optique d'augmenter la précision, Hersh *et al.,* (2001) ont délibérément écarté certaines parties des comptes rendus, en particulier la section *indications* ; ils ont ainsi obtenu un index ne contenant que les termes strictement médicaux. Or, en pratique, dans leurs requêtes, les utilisateurs d'un système de recherche dédié à la radiologie ont besoin de rechercher non seulement des termes médicaux précis (*perforation digestive, glioblastome*) mais aussi des termes composés ou des périphrases de sens général (*accident de ski*, *femmes jeunes, coups de couteau,*

coup de sabot). Nous appelons *termes médicaux précis* les termes qui dans le réseau utilisé pour ce travail sont liés par la relation *domaine* à la *médecine*. Les termes généraux ne sont pas directement liés par cette relation à la médecine.

Une grande partie de la complexité de l'extraction automatique d'informations pertinentes à partir de corpus médicaux provient de la forme non structurée de la plupart des textes, de l'écriture informelle (beaucoup d'abréviations, de raccourcis, d'incorrections, etc.), de la quantité d'informations à analyser et de l'identification de leur pertinence. La difficulté d'analyse automatique du sens (en particulier la gestion précise des négations (Huang *et al.,* 2007) et des apocopes, d'identification de termes inconnus (non présents dans la base de connaissances), d'analyse syntaxique de phrases agrammaticales, de reconnaissance d'entités médicales figurant souvent sous une forme d'écriture très dégradée, d'extraction de relations sémantiques présentes dans le texte (Bundschus *et al.,* 2008), sont autant d'obstacles à une indexation fine de ce genre de documents. Pour réaliser une telle analyse, il est donc crucial de disposer d'un support sémantique non seulement de grande couverture, c'est-à-dire une base de connaissances non réduite aux formes normées, mais également dynamique (i.e. capable d'évoluer et de s'enrichir par apprentissage permanent).

A notre connaissance, l'indexation automatique de comptes rendus radiologiques a jusqu'alors essentiellement porté sur les termes strictement médicaux sans tenir compte des informations d'ordre général. Cependant, Xu *et al.,* (2014) ont réalisé une reconnaissance d'entités nommées de termes anatomiques avec l'aide de ressources externes générales comme Wikipedia et WordNet, en supplément des ressources médicales usuelles, à savoir UMLS, RadLex, MeSH et BodyPart3D (http://lifesciencedb.jp/bp3d/). Un autre type de ressource, qui n'avait encore jamais été utilisé dans le cadre médical ou biomédical, permet de prendre en compte non seulement les mots et les concepts du domaine de spécialité, mais également le langage commun couramment utilisé dans les comptes rendus (notamment dans la section *indications*) : il s'agit du réseau lexico-sémantique JeuxDeMots (http://www.jeuxdemots.org) que nous utilisons comme base de connaissances support pour l'indexation automatique des comptes rendus radiologiques.

Dans le projet IMAIOS (en collaboration avec des médecins radiologues de Montpellier), afin d'être en mesure d'indexer correctement des comptes rendus de radiologie, nous réalisons non seulement une description des termes et concepts du domaine, mais nous visons également à déterminer les sens (ou les usages) des termes ou des abréviations très fréquentes en médecine. McInnes et Stevenson (2014) ont souligné la difficulté de réaliser cette tâche dans le domaine biomédical, et Ramadier (*et al,* 2014) cherche à le faciliter à l'aide d'annotations et d'inférences de relations sémantiques. Nous décrivons dans cet article comment à partir de ses informations sémantiques, il est possible de définir une *augmentation des index bruts construits pour chaque compte rendu* afin d'améliorer le rappel de la recherche documentaire. En effet, les médecins radiologues peuvent exprimer leurs requêtes en utilisant des génériques (par exemple, *tumeur bénigne du cerveau*, *tumeur du cerveau, tumeur bénigne, tumeur*), des conséquences, des circonstances, etc. sans que ces termes soient pour autant explicitement présents dans les comptes rendus. L'article est organisé comme suit : nous présentons en premier lieu le support utilisé pour réaliser cette indexation, c'est-à-dire le réseau lexical JeuxDeMots, puis décrivons la forme précise que peut prendre un index augmenté ainsi que l'algorithme d'augmentation basé sur une propagation au sein du réseau lexical. Enfin nous discutons des expérimentations et analysons les résultats.

2 Augmentation d'index et propagation

La base de connaissances sur laquelle s'appuie notre stratégie d'indexation des comptes rendus médicaux est le réseau lexical JeuxDeMots (Lafourcade 2007). Bien que généraliste, le réseau JDM contient un grand nombre de données de spécialités, notamment des données de médecine/radiologie introduites dans le cadre du projet IMAIOS. Ce réseau sert de base à un algorithme de propagation visant à augmenter un index brut obtenu par des moyens classiques en recherche d'informations.

2.1 Le réseau JeuxDeMots

Le réseau JDM est un graphe lexico-sémantique pour le français obtenu via des jeux - des GWAP, voir (Lafourcade *et al*. 2015) - et un outil contributif nommé Diko. Au moment de l'écriture de cet article, le réseau JDM contient près de 20 millions de relations entre 490 000 termes. Les propriétés de ce réseau qui sont d'intérêt ici sont les suivantes :

- il existe environ 80 types de relations différentes. Celles qui nous intéressent sont les relations essentiellement sémantiques, comme l'hyperonymie, les caractéristiques typiques, les lieux typiques, les constituants typiques, le domaine (de spécialité), etc. ;
- les termes polysémiques sont associés (via la relation raffinement) à leurs différents usages. Environ 9 000 termes sont ainsi raffinés en plus de 25 000 usages, comme par exemple, fracture → fracture (lésion), fracture (rupture), fracture (sociologie). Le terme entre parenthèse est une *glose* qui permet de savoir (ou de deviner) de quel sens (*raffinement*) il s'agit ;
- les relations sont pondérées, le poids traduisant la force d'association entre les termes. Environ 70 000 relations ont des poids négatifs, qui indiquent une relation fausse (mais intéressante à conserver, car pouvant aider à la désambiguïsation lexicale) comme par exemple : *fracture du tibia *hyperonyme (< 0)* fracture (sociologie) ;

- il existe une relation d'inhibition qui à un terme *t*, associe un usage d'un autre terme dont un sens frère est en rapport avec *t*. Par exemple : fracture → fracture talus (pente), talus (imprimerie), talus (remblai), astragale (architecture), astragale (botanique), … Au moins un des autres sens des termes associés est en rapport avec *fracture* : ou talus (os) ou astragale (os).

Figure 1. Capture d'écran de la page Diko pour l'entrée « fracture du tibia ». Diko est un outil de visualisation et de contribution en ligne pour le réseau lexical JeuxDeMots. On remarquera dans cet exemple que l'entrée contient à la fois des informations médicales précises (voir symptômes, diagnostiques etc.) et des associations d'ordre plus général (voir causes, conséquences, etc.).

L'indexation de mots clés dans le domaine médical concerne souvent certains aspects d'une maladie (Andrade, 2000) ou une partie de l'anatomie. Comme le but de cette indexation est la recherche de documents dans un objectif de pratique quotidienne, nous indexons non seulement les termes anatomiques au sens large (*genou, paroi antérieure du côlon, genou du corps calleux*,..), les signes cliniques et les maladies, mais aussi des termes du langage commun susceptibles de faire l'objet de requêtes par le radiologue.

Concernant les domaines (de spécialité) pouvant nous intéresser spécifiquement pour le projet IMAIOS, le tableau suivant donne une idée de l'ordre de grandeur relatif à la quantité d'informations dont nous disposons :

terme	nb de liens sortants	nb de liens entrants
médecine	21408	22666
anatomie	10477	11453
radiologie	382	502
accident	741	956
imagerie médicale	541	556

Tableau 1 : Nombres de liens entre termes dans JeuxDeMots pour certains termes clés.

2.2 Indexation standard de comptes rendus

Notre corpus de travail est constitué d'environ 40 000 comptes rendus de radiologie (Exemple 1) englobant les différentes modalités d'imagerie médicale (imagerie par résonance magnétique, scanner, échographie, radiologie conventionnelle, radiologie vasculaire). Ces comptes rendus sont écrits de manière semi-structurée, c'est-à-dire qu'ils sont généralement divisés en quatre parties distinctes (*indications, technique, résultats*, et une *conclusion* optionnelle). Chaque partie est rédigée par le médecin radiologue sous une forme très libre, avec souvent une profusion d'acronymes (*ATCD* pour *antécédent*, *ACR* pour *american college of radiologie*, *tt* pour *traitement* etc.) d'élisions (par exemple, la *communicante antérieure* au lieu de *artère communicante antérieure*), et toutes sortes d'incorrections diverses. Les comptes rendus contiennent une grande quantité d'informations implicites, devant être explicitées si nous souhaitons obtenir une indexation répondant aux besoins des praticiens.

La création de l'index à partir des comptes rendus reste relativement simple. Nous utilisons les méthodes classiques de la recherche documentaire, soit la fréquence des termes (TF) et la fréquence documentaire (DF) pour calculer l'IDF (Inverse Document Frequency). La reconnaissance des termes composés est effectuée en amont par comparaison au contenu de JDM. Un tiret bas remplacera l'espace entre chaque élément d'un terme composé afin de les conserver comme unité autonome lors de l'extraction (*fracture_du_tibia*).

Malgré le filtrage fréquentiel, nous conservons les termes situés au voisinage de la médecine même pour de faibles valeurs du TFR-IDF. Si un mot simple ou composé du texte est présent dans le réseau JDM, et qu'il est lié par la relation domaine au terme *médecine* (voisinage à distance 1) alors il est ajouté à l'index. De même, des termes non médicaux (*accident de moto*, *prise de drogue, boule de pétanque*) sont également capturés et ajoutés à l'index dès lors qu'ils possèdent une relation avec un terme lui-même lié à *médecine* (voisinage à distance 2) : ainsi *accident de moto* est ajouté car lié à *polytraumatisé* par la relation conséquence et *polytraumatisé* est lui-même lié à *médecine* par la relation domaine.

indications : fracture du tibia droit, chute de ski **technique** : une série de coupes axiales transverses sur l'ensemble de la cheville droite sans injection de produit de contraste **étude** : en fenêtres parties molles et osseuses. **résultats** : fractures diaphysaires spiroïdes à trois fragments principaux du 1/3 distal du tibia et de la fibula avec discret déplacement vers l'avant, sans retrait de refend articulaire. Fractures de la base de M2 et de M3 non articulaire et non déplacée. Fracture articulaire de la partie interne de la base de M1 non déplacée. Atrophie avec dégénérescence marquée des corps musculaires de l'ensemble des loges.	atrophie • cheville • chute • corps musculaire • coupe axiale transverse • dégénérescence • déplacement • fibula • fracture • fracture du tibia • loge • non articulaire • non déplacée • ski • spiroïde • tibia

Exemple 1 : compte rendu de radiologie typique (à gauche) et index brut (à droite) : liste de termes présents extraits, ordonnée par ordre alphabétique (les pondérations ne sont pas représentées et la liste est simplifiée). Les termes composés sont identifiés pour peu qu'ils soient présents dans le réseau JDM sous une forme connexe.

Par ailleurs, pour chaque terme de l'index brut, il est intéressant d'essayer de déterminer le ou les bons raffinements sémantiques (s'il en possède). Par exemple, dans le compte rendu ci-dessus, les termes *fracture, cheville, chute* et *loge* sont polysémiques. Les résultats montreront que la détermination des bons raffinements a de l'importance. Ainsi, *l'augmentation est un processus visant à rajouter dans l'index des termes pertinents, mais non présents dans le texte*.

accident de ski • **accident de sports d'hiver** • atrophie • cheville • **cheville>anatomie** • chute • **chute>tomber** • corps musculaire • coupe axiale transverse • dégénérescence • **dégénérescence musculaire** • déplacement • fibula • fracture • **fracture articulaire** • **fracture des membres inférieurs** • **fracture multiple** • **fracture diphysaire** • **fracture du tibia** • **fracture non articulaire** • **fracture non déplacée** • **fracture spiroïde** • **fracture avec déplacement** • **fracture>lésion** • **imagerie médicale** • **jambe** • **lésion** • **lésion osseuse** • loge • **loge>anatomie** • **médecine** • non articulaire • non déplacée • **péroné** • **radiologie** • ski • **spiroïde** • **sports d'hiver** • tibia • **traumatisme des membres inférieurs** • …

Exemple 2 : Index augmenté correspondant à l'exemple présenté ci-dessus (termes triés par ordre alphabétique avec en gras les termes ajoutés). Les thématiques générales du texte sont bien identifiées (médecine, imagerie médicale, radiologie). Les termes polysémiques ont été raffinés avec leur usage correct en contexte.

2.3 Algorithme d'augmentation par propagation

Pour constituer l'index augmenté à partir de l'index brut, nous adoptons une stratégie consistant à propager des signaux sur le réseau JDM à partir des termes de l'index brut. L'idée principale est *d'allumer* les termes de l'index brut et de récupérer à leur suite les termes du réseau qui s'allument également.

A chaque cycle, les termes déchargent en parallèle leur activation courante vers leurs voisins. L'activation totale n'est que la mémoire des décharges reçues par un terme sur l'ensemble du processus. Pour les relations à poids négatif ou inhibitrices, l'activation n'est pas ajoutée mais soustraite. Un terme avec une AC < 0 ne décharge pas. La séquence itérée est réalisée en parallèle pour tous les termes. On remarquera que la distribution du signal se fait proportionnellement aux logarithmes des poids (et non pas proportionnellement aux poids eux-mêmes).

A l'issue de l'itération (lignes 5 à 7), nous obtenons une liste de termes pondérés que nous ordonnons par poids décroissants. Nous retenons (filtrage) les N termes de poids les plus forts tels que la somme de leur poids représente S% du poids total des termes de cette liste.

Plus précisément, l'algorithme que nous avons mis au point s'énonce informellement comme suit :

1 Init : les termes T du réseau sont associés à un couple de valeurs (AC, AT), *activation courante* et *activation totale*.
2 pour les termes **T** appartenant à l'index brut, nous fixons AC = AT = 1. // les **T** sont les sources d'activation
3 pour tous les autres termes, AC= AT = 0.
4 nous fixons un nombre d'itérations NBI
5 nous répétons NBI fois l'opération suivante :
6 pour chaque terme T du réseau ayant des voisins $\{t_1,\dots,t_n\}$
via une relation de type *r* de T vers t_i de pondération positive w_i, nous modifions les AC et AT des t_i :

$$AC(t_i) = AC(t_i) + AC(T) \times \frac{\log(w_i)}{\sum_{k=0}^{n} \log(w_k)}$$

$AT(t_i) = AT(t_i) + AC(t_i)$ // on mémorise ce que reçoit t_i dans $AT(t_i)$
7 $AT(\mathbf{T}) = 1$ // tous les T ont déchargé leur activation, on recharge les **T**
8 filtrage des termes activés avec pourcentage de surface S ; nous retournons les termes activés restants.

Algorithme 1 : calcul d'un index augmenté à partir d'un index brut, à l'aide d'une propagation sur le réseau lexical JDM. Les deux principaux paramètres sont NBI (nombre d'itérations) et S (% de surface retenu pour le filtrage).

Nous n'exploitons pas tous les types de relations disponibles dans JDM, certaines, très lexicales, auraient tendance, dans le cadre de notre application, à dégrader la précision. Les relations que nous utilisons sont les suivantes (avec leur pondération éventuelle, sinon le poids par défaut est 1) : idées associées (poids 1/2), hyperonymes (poids 2), synonymes, caractéristiques typiques, symptômes, diagnostiques, parties/tout, lieux typiques, causes, conséquences, domaine, et fréquemment associé à. Dans l'algorithme 1 ci-dessus, par souci de simplification, tous les types de relations ont un poids identique (le poids serait rajouté de part et d'autre de la fraction).

3 Evaluation des index augmentés

Nous avons évalué l'algorithme de propagation de façon statistique par une sélection aléatoire de 200 index augmentés (sur 30 000 calculés). Chaque terme de l'index augmenté a été manuellement évalué comme pertinent ou non. L'évaluation manuelle a été réalisée par des experts en radiologie. Un terme (ou multi termes) est considéré pertinent par un spécialiste lorsqu'il est susceptible de faire l'objet de requêtes. Les couples de valeurs du Tableau 2 sont donc (a) le nombre moyen de termes de l'index augmenté qui ne sont pas dans l'index brut (valeur *nouv*) et (b) le pourcentage moyen de termes pertinents de l'index augmenté (valeur *pert*).

NBI \ S	10 %	20 %	30 %	40 %	50 %
1	22 / 82 %	45 / 80 %	67 / 78 %	93 / 53 %	127 / 38 %
2	31 / 95 %	55 / 92 %	83 / 89 %	211 / 57 %	439 / 41 %
3	48 / 99 %	90 / 97 %	139 / 95 %	356 / 53 %	755 / 34 %
4	111 / 97 %	223 / 92 %	335 / 87 %	747 / 45 %	1259 / 23 %
5	387 / 96 %	774 / 87 %	1161 / 76 %	1671 / 26 %	2089 / 15 %

Tableau 2 : Présentation des valeurs *nouv* (à gauche de chaque colonne) et *pert* (à droite) en fonction des paramètres NBI et S. NBI est le nombre d'itérations effectuées dans le réseau lexical. S est la part retenue de la surface sous la courbe des poids cumulés des termes atteints par l'algorithme de propagation.

En pratique, l'évaluation manuelle de la valeur *pert* n'a besoin d'être réalisée qu'une fois indépendamment des paramètres NBI et S. En effet, il suffit pour un compte rendu de considérer l'ensemble des termes obtenus pour toutes les valeurs possibles des paramètres, puis d'évaluer la pertinence de chaque terme dans l'ordre de leurs poids décroissants. Au bout de 5 termes consécutifs non pertinents, on considère que tout ce qui suit est également non pertinent. La valeur *nouv*, elle, peut être calculée automatiquement. Pour un même nombre d'itérations, plus la surface retenue est grande plus le nombre de termes atteints est important (filtrage faible). C'est-à-dire que le rappel est d'autant plus important, mais en contrepartie la précision à tendance à diminuer (voire s'écroule au-delà de 30%), les termes ajoutés à l'index brut ayant tendance être de moins en moins pertinents. A l'inverse, plus le nombre d'itérations augmente, plus les termes pertinents sont renforcés (ce sont les voisins mutuels des termes de l'index brut). Le réseau

lexical contient des boucles (directes et indirectes) qui agissent comme autant d'auto-renforcements. Le temps de calcul croît considérablement à chaque nouvelle itération, le nombre de termes déchargeant leur activation augmentant très fortement. Pour NBI = 5, la quasi-totalité du réseau est atteinte (si on exclut le filtrage par S), le diamètre étant d'environ 6 (le réseau JDM est de type petit monde). Globalement, la zone qui semble la plus intéressante pour un temps de calcul raisonnable (quelques secondes) correspond à 3 à 4 itérations pour une surface inférieure à 30%.

La totalité des termes ambigus ont été correctement désambiguïsés. Cela signifie que l'index augmenté a systématiquement inclus le bon raffinement quand un raffinement était proposé (ce n'est pas forcément le cas pour des valeurs faibles de NBI et de S). Nous avons recalculé les index augmentés en interdisant les accès aux termes raffinés, et avons constaté une baisse globale d'environ 10% de la valeur de *pert* quelle que soit la configuration (NBI et S). Chercher à sélectionner les sens corrects des termes polysémiques peut donc être réalisé conjointement à la sélection de termes pertinents et aurait même tendance à la favoriser. Enfin, tous les domaines identifiés ont été pertinents. Rajouter les domaines pertinents dans l'index brut avant l'augmentation n'améliore pas significativement les résultats (ni ne les dégrade). Enfin, nous avons également recalculé les index (bruts et augmentés) en n'autorisant l'accès qu'aux termes directement liés au terme médecine (quelle que soit la relation) durant la propagation. Cela s'est traduit par une diminution moyenne de 12% de la valeur *pert*. Il semblerait donc que l'utilisation d'une base de connaissances non limitée au domaine de spécialité améliore grandement la pertinence de l'index produit.

On remarquera que l'ensemble du processus présenté ci-dessus fonctionne de façon thématique sur le texte et sémantique sur le réseau lexical. Il n'y a pas d'analyse sémantique fine, qui impliquerait selon toute vraisemblance une analyse en constituants et en dépendance des comptes rendus. Les cas d'erreurs manifestes (23 termes pour les 200 index soit pour environ 10 000 termes) que nous avons relevés peuvent avoir plusieurs causes :

- défaut d'information dans la base de connaissances (20% des cas d'erreur) ;
- défaut de rôle sémantique, impliquant la nécessité d'une analyse fine (55%) ;
- chimérisme – deux parties distinctes du compte rendu ont fait émerger un terme non pertinent (25%).

Perspectives et conclusion

Notre objectif est d'indexer automatiquement des comptes rendus radiologiques**,** non seulement avec les termes médicaux mais aussi avec des termes du langage courant susceptibles d'être utilisés dans des requêtes d'utilisateurs**,** notamment de praticiens hospitaliers. Pour augmenter le rappel sans notablement dégrader la précision, nous ajoutons à l'index brut des termes implicites des comptes rendus, en utilisant comme support la base de connaissances qu'offre le réseau lexico-sémantique JeuxDeMots. A notre connaissance, très peu de travaux prennent en compte des éléments non médicaux présents dans le compte rendu, ou encore effectuent de l'inférence implicite afin de trouver des termes pertinents non présents. Les approches classiques d'augmentation du rappel consistent essentiellement à inclure des termes plus généraux (hyperonymes ou synonymes) à partir d'une ontologie médicale. La présence d'informations de sens commun bonifie les résultats : l'hypothèse selon laquelle la *non séparation des connaissances* (spécialisées *et* générales) est plus intéressante que l'usage exclusif de celles de spécialité semble se confirmer, au moins dans nos travaux.

Le travail présenté ici reste préliminaire et il manque une évaluation de fond des index sur l'ensemble du corpus. Nos premiers résultats semblent prometteurs, mais pour être réalisée à grande échelle, l'évaluation doit pouvoir être mécanisée. Il serait intéressant de pouvoir comparer nos résultats avec une indexation obtenus à partir d'un métathésaurus comme l'UMLS. Nous pourrons ensuite aller plus loin dans l'analyse de comptes rendus via l'extraction des relations entre les termes du texte à l'aide de celles présentes dans le réseau JDM. L'indexation portera alors non seulement sur les termes mais également sur les relations entre ces termes. A ce moment là, on pourra réaliser une évaluation plus détaillée concernant la recherche d'information.

Un des buts poursuivis dans le projet IMAIOS est aussi de découvrir dans les comptes rendus de nouvelles connaissances permettant d'alimenter le réseau lexical. Nous envisageons également de déduire à partir du corpus des règles d'inférence et de faire ainsi un raisonnement authentique, c'est-à-dire de proposer par déduction et induction de nouvelles informations médicales, voire des diagnostics.

Références

ANDRADE M. A. & BORK, P. (2000). *Automated extraction of information in molecular biology.* FEBS letters, Elsevier, 476/1, pp. 12–17.

BUNDSCHUS M., DEJORI M., STETTER M., TRESP V. & KRIEGEL H.-P. (2008). *Extraction of semantic biomedical relations from text using conditional random fields.* BMC bioinformatics, **9**:207, 14 p.

DINH D., TAMINE L. *et al.* (2010). *Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients.* In Conférence francophone en Recherche d'Information et Applications, CORIA 2010, pp. 325–336.

DÍAZ-GALIANO, Manuel Carlos, MARTÍN-VALDIVIA, Maite Teresa, et UREÑA-LÓPEZ, L.A. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in biology and medicine*, 2009, vol. 39, no 4, p. 396-403.

GOEURIOT, Lorraine, KELLY, Liadh, et LEVELING, Johannes. An analysis of query difficulty for information retrieval in the medical domain. In : *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014. p. 1007-1010.

HERSH W., MAILHOT M., ARNOTT-SMITH C. & LOWE H. (2001). *Selective automated indexing of findings and diagnoses in radiology reports.* Journal of biomedical informatics, 34(4), pp. 262–273.

HUANG Y. & LOWE H. J. (2007). A *novel hybrid approach to automated negation detection in clinical radiology reports.* Journal of the American Medical Informatics Association, 14(3), pp. 304–311.

HUANG Y., LOWE H. J. & HERSH W. R. (2003). *A pilot study of contextual UMLS indexing to improve the precision of concept-based representation in xml-structured clinical radiology reports.* Journal of the American Medical Informatics Association, 10(6), pp. 580–587.

LAFOURCADE M. (2007). *Making people play for lexical acquisition with the JeuxDeMots prototype.* In SNLP'07 : 7th international symposium on natural language processing.

LANGLOTZ C. P. (2006). *Radlex : A new method for indexing online educational material.* Radiographics, 26(6), pp. 1595–1597.

MCINNES B. T. & STEVENSON M. (2014). *Determining the difficulty of word sense disambiguation.* Journal of biomedical informatics, 47, pp. 83–90.

POULIQUEN B. (2002). *Indexation de textes médicaux par extraction de concepts, et ses utilisations.* Thèse de doctorat, Faculté de Médecine, Université Rennes 1, juin 2002, 163 p.

RAMADIER L., ZARROUK M., LAFOURCADE M. & MICHEAU A. (2014). *Annotations et inférences de relations dans un réseau lexico-sémantique : application à la radiologie.* TALN 2014, Marseille, juillet 2014, pp. 103-112.

RAMOS J. (2003*). Using TF-IDF to Determine Word Relevance in Document Queries.* In Proceedings of the first instructional conference on machine learning., 4 p.

ROBERTSON S. E. & JONES K. S. (1976). *Relevance weighting of search terms.* Journal of the American Society for Information science, 27(3), pp. 129–146.

ROBERTSON, S. (2004). "Understanding inverse document frequency: On theoretical arguments for IDF". *Journal of Documentation* **60** (5): pp. 503–520.

VOORHEES, E. et TONG, R. Overview of the TREC 2011 medical records track. In : *Proc. of TREC*. 2011.

XU Y., HUA J., NI Z., CHEN Q., FAN Y., ANANIADOU S., ERIC I., CHANG C. & TSUJII J. (2014). *Anatomical entity recognition with a hierarchical framework augmented by external resources.* PloS one, 9(10), e108396.

ZARROUK M., LAFOURCADE M. & JOUBERT A. (2013). *Inference and reconciliation in a crowdsourced lexical semantic network.* Computación y Sistemas, 17(2), pp. 147–159.