Recherche d'information médicale pour le patient Impact de ressources terminologiques

Sébastien Le Maguer, Thierry Hamon, Natalia Grabar, Vincent Claveau

Recherche d'information médicale pour le patient

Impact de ressources terminologiques

Sébastien Le Maguer[*,] — Thierry Hamon[***] — Natalia Grabar[****] — Vincent Claveau[*]**

[*] *IRISA - CNRS, Rennes, France*
[**] *MMCI, University of Saarland, Germany*
[***] *LIMSI - CNRS, Orsay, France; Université Paris 13, Sorbonne Paris Cité, France*
[****] *STL UMR8163 CNRS, Université Lille 3, France*

RÉSUMÉ. *Le droit d'accès au dossier clinique par les patients est inscrit dans le code de Santé Publique. Cependant, ce contenu reste difficile à comprendre. Nous proposons une expérience, où les requêtes des patients sont utilisées pour retrouver les documents pertinents. Nous utilisons le moteur de recherche Indri, basé sur le modèle statistique de la langue, et des ressources sémantiques. L'accent est mis sur la variation terminologique (e.g. synonymes, abréviations) pour faire le lien entre la langue des experts et des patients. Différentes combinaisons de ressources et du paramétrage de Indri sont testées, essentiellement à travers l'expansion des requêtes. Notre système montre jusqu'à 0,7660 de P@10 et 0,6793 de NDCG@10.*

ABSTRACT. *The right of patients to access their clinical health record is granted by the code of Santé Publique. Yet, this content remain difficult to understand. We propose an experience, in which we use queries defined by patients in order to find relevant documents. We utilise the Indri search engine, based on statistical language modeling and semantic resources. We stress the point related to the terminological variation (e.g. synonyms, abbreviations) to make the link between expert and patient languages. Various combinations of resources and Indri settings are explored, mostly based on query expansion. Our system shows up to 0.7660 P@10 and up to 0.6793 NDCG@10.*

MOTS-CLÉS : *Recherche d'information orientée sur le patient, ressources sémantiques, variation terminologique, Indri, UMLS*

KEYWORDS: *Patient-Oriented Information Retrieval, Semantic Resources, Terminological Variation, Indri, UMLS*

1. Introduction

Depuis plusieurs années, les patients bénéficient d'un accès plus libre à leurs dossiers médicaux. En France, ce droit est inscrit dans les articles spécifiques du code de la Santé Publique [1]. Cependant, la mise à disposition de ce type d'informations ne garantit pas leur compréhension par les patients, ce qui peut créer des situations délicates. Il a été par exemple observé que l'utilisation généralisée des informations de santé sur l'Internet (Diaz *et al.*, 2002 ; Luciano *et al.*, 2013) cause un changement important dans la relation entre les médecins et les patients et dans leur communication (Jucks et Bromme, 2007 ; de Boer *et al.*, 2007 ; Tran *et al.*, 2009).

Il devient donc important de fournir des modèles de recherche d'information qui permettent aux patients de trouver et consulter les informations qui soient adaptées pour eux. Au moins deux objectifs doivent alors être satisfaits :

– les informations doivent provenir de sources fiables (Gaudinat *et al.*, 2007 ; Eysenbach et Thomson, 2007 ; Pletneva *et al.*, 2011), car souvent les patients ne peuvent pas faire la différence par eux-mêmes entre les informations de santé de qualité et les informations indésirables ;

– le modèle de recherche doit mettre en correspondance les termes spécifiques propres au domaine médical (e.g., *infarctus du myocarde*, *desmorrhexie*) et les expressions utilisées par les patients (*crise cardiaque* et *rupture des ligaments*, respectivement) (AMA, 1999 ; D'Alessandro *et al.*, 2001 ; McCray, 2005 ; Kickbusch *et al.*, 2013).

Nous proposons une contribution dans ce domaine de recherche. Nous travaillons sur des données indépendantes fournies par les compétitions CLEF eHealth 2013 et 2014 (Kelly *et al.*, 2013 ; Goeuriot *et al.*, 2014), ce qui garantit que la recherche est faite dans un espace de fiabilité et fournit des informations de confiance. Le modèle de recherche d'information est basé sur le moteur de recherche existant Indri (Strohman *et al.*, 2005) et des ressources sémantiques du domaine médical. L'accent principal est mis sur l'utilisation de ressources sémantiques qui pourraient permettre de prendre en considération quelques spécificités de la langue des patients et des médecins. Nous faisons d'abord le tour des travaux existants (section 2). Nous présentons ensuite le matériel traité (section 3) et la méthode (section 4). Nous présentons et discutons les expériences (section 5) et les résultats obtenus (section 6), et terminons avec les pistes pour le travail futur (section 7).

2. Travaux existants

Il existe plusieurs travaux qui cherchent à faire le lien entre le vocabulaire des patients et celui des médecins. Le résultat attendu se présente souvent sous forme de

1. http://www.legifrance.gouv.fr/affichCode.do?idSectionTA=LEGISCTA000006196866&cidTexte=LEGITEXT000006072665

lexiques alignés, où les expressions provenant de ces deux discours sont mises en correspondance {*terme spécialisé*, *expression non spécialisée*} :

– {*myocardial infarction*, *heart attack*}, {*abortion*, *termination of pregnancy*}, {*acrodynia*, *pink disease*} (Zeng et Tse, 2006) ;

– {*retard de cicatrisation*, *retarder la cicatrisation*}, {*apports caloriques*, *apport en calories*}, {*calculer les doses*, *doses sont calculées*}, {*efficacité est renforcée*, *renforcer son efficacité*} (Cartoni et Deléger, 2011) ;

– {*myocardial infarction*, *heart attack*}, {*SBP*, *systolic blood pressure*}, {*atrial fibrillation*, *arrhytmia*}, {*hypercholesterolemia*, *cholesterol*}, {*mental stress*, *stress*} (Elhadad et Sutaria, 2007).

De manière générale, il s'agit de proposer des moyens pour atteindre l'interopérabilité sémantique entre les deux discours. La première initiative de ce type, Consumer Health Vocabulary (Zeng et Tse, 2006), a fourni le premier lexique en anglais qui fait actuellement partie d'UMLS (NLM, 2008). Pour détecter de telles correspondances, les travaux exploitent le plus souvent des corpus comparables et l'expertise humaine.

Système	Moteur	Modèle	Expansion	Ressources
(Shenwei *et al.*, 2014)	Indri	LM	IM	UMLS, Metamap
(Choi et Choi, 2014)	Indri	LM	TP, DS	Metamap, UMLS
(Oh et Jung, 2014)	Lucene	LM	abbr, PRF	
(Thakkar *et al.*, 2014)	Indri	Vary	QLH, BRF	-
(Yang *et al.*, 2014)	Indri	-	MRF, PRF	GeniaSS
(Ozturkmenoglu *et al.*, 2014)	Terrier	VSM	KLE	Weka

Tableau 1. *Description des systèmes de la compétition CLEF eHealth 2014.*

Un autre type de travaux est effectué dans le domaine de recherche d'information. Il semble être moins étudié, certainement parce que les jeux de données appropriées sont très rares. Les données fournies par les compétitions CLEF eHealth 2013 et 2014 (Kelly *et al.*, 2013 ; Goeuriot *et al.*, 2014) semblent être assez uniques pour ce domaine de recherche. En 2014, une tâche était plus explicitement orientée sur les utilisateurs non experts et nous présentons ici quelques systèmes participants. Dans le tableau 1, nous indiquons les moteurs de recherche utilisés (Lucene [2], Indri (Strohman *et al.*, 2005) et Terrier (Ounis *et al.*, 2006)), les modèles de recherche d'information (e.g., LM (Language Model), VSM (Vector Space Model), Hiemstra, Vary), la méthode pour l'expansion de requêtes (e.g., PRF (Pseudo Relevance Feeback), IM (Information Mutuelle), TP (Terme Préféré), QLH (Query-likelihood), BRF (Blind Relevance Feeback), MRF (Markov Random Fields), KLE (Kullback–Leibler Expansion)) et les ressources externes utilisées. Pour plus de détail sur ces systèmes, il est possible de consulter les publications référencées. Nous pouvons voir que les stratégies de recherche sont variées et sont souvent accompagnées de l'utilisation de

2. `http://lucene.apache.org/`

ressources externes pour étendre les requêtes. Les meilleurs résultats de la compétition (précision@10 = 0,7560, NDCG@10 = 0,7445) sont obtenus par le système utilisant le moteur de recherche Indri, des ressources externes médicales (UMLS et Metamap) et la pondération avec l'information mutuelle (Shenwei *et al.*, 2014). Les expansions sont effectuées en utilisant l'UMLS au niveau des SUI (essentiellement les variations morpho-syntaxiques des termes), et en les consolidant par l'information mutuelle. Ce système ne recoure pas à l'utilisation de ressources linguistiques complémentaires. Dans notre travail, nous mettons justement l'accent sur l'utilisation de telles ressources afin d'élargir la notion de la variation terminologique.

3. Matériel utilisé et traité

3.1. *Données des compétitions CLEF eHealth*

Nous utilisons les données en anglais fournies par les compétitions CLEF eHealth. Il s'agit de requêtes créées par des vrais patients suite à la consultation de leurs dossiers cliniques. Un ensemble de documents donnant réponses à ces requêtes est défini. Ces documents sont collectés dans le cadre du projet KHRESMOI [3]. Les systèmes participants doivent donc retrouver les documents pertinents. L'ensemble de requêtes comporte 5 requêtes pour l'entraînement et 50 pour le test. Chaque requête est composée de la partie question :

Can one experience pain if there is a severe brain injury ?
What are the connections between respiratory failure and CHF ?

et de la partie description :

The document should contain information about brain injury and pain. Relevant documents should contain information about respiratory failure and CHF.

Nous avons effectué une analyse manuelle des requêtes en comparant les informations présentes dans la partie question avec les informations de la partie description, et donc les informations attendues dans les documents. Cette analyse montre que la grande majorité des requêtes concerne des problèmes médicaux. Seules deux requêtes portent sur des procédures médicales (*laryngectomy* et *aortic valve replacement* dans les requêtes 13 et 27, respectivement). Nous avons également observé qu'une trentaine de requêtes visent à obtenir des informations sur un problème médical, mais aussi ses causes, ses conséquences ou son traitement. Dans les autres requêtes, nous pouvons avoir également d'autres situations, où les informations attendues :

– sont implicites : par exemple, dans la requête 24, le lien entre la fonction cardiaque (*heart function*), le système circulatoire (*cirkulatory*) et les *vaisseaux du cœurs* (*heart vessels*) ;

3. `http://www.khresmoi.eu`

– nécessitent une inférence : dans la requête 15, il faut faire l'inférence entre l'hérédité et le fait que c'est lié à la famille ;

– supposent une généralisation : dans la requête 32, il faut faire le lien entre *myocardial infarction* et *heart patient*.

Notons aussi que la description peut être plus précise que la question. Par exemple, dans la requête 1, il s'agit de maladies coronariennes, mais seule la description mentionne qu'il est nécessaire d'en trouver les traitements.

Le corpus contient plus d'un million de documents (environ 200 M occurrences), dont les réponses aux requêtes. Le corpus 2013 est utilisé pour le réglage du système et celui de 2014 pour les tests. Le corpus contient les documents de santé collectés dans le cadre du projet KHRESMOI. Le corpus couvre un large éventail de questions médicales, mais ne contient pas de documents cliniques. Les documents proviennent de plusieurs sources en ligne, y compris les sites web certifiés par Health On the Net, mais aussi des sites et bases de données de santé connus (e.g. *Genetics Home Reference, ClinicalTrial.gov, Diagnosia*). Les documents sont prétraités afin de permettre leur indexation par le moteur de recherche : conversion de HTML en texte (la structure des documents n'est donc pas exploitée), conversion en UTF-8, segmentation en phrases, racinisation avec Porter et Krovetz, suppression de mots vides.

3.2. *Ressources sémantiques*

Plusieurs ressources sémantiques sont utilisées pour l'expansion de requêtes.

Synonymes d'UMLS. L'UMLS (Unified Medical Language System) (NLM, 2008) propose la fusion d'environ 200 terminologies biomédicales. Chaque concept reçoit un identifiant unique CUI et peut regrouper plusieurs termes. Il s'agit d'un niveau de granularité moins fin que celui des SUI. Nous exploitons cette propriété : les termes avec le même identifiant ({*theophyllamine*, *ammophyllin*}) sont considérés comme synonymes. La motivation supplémentaire est que l'UMLS contient aussi le Consumer Health Vocabulary (Zeng et Tse, 2006). 227 887 synonymes entre les termes simples et tous les synonymes entre les termes complexes sont exploités.

Synonymes induits. Une ressource de synonymes est induite à partir des synonymes d'UMLS (Grabar et Hamon, 2010). Grâce au principe de compositionnalité et à l'analyse syntaxique de termes complexes, les composants de ces termes peuvent être mis en relation de synonymie. Par exemple, les termes *acetone anabolism* et *acetone biosynthesis* sont synonymes et nous pouvons induire que *anabolism* et *biosynthesis* sont aussi synonymes. Cette ressource comporte 1 314 paires de synonymes, dont la majeure partie n'apparaît pas parmi les synonymes UMLS d'origine.

Variantes morpho-syntaxiques. Les variantes morpho-syntaxiques ont une sémantique proche entre elles car les modifications subies sont liées à l'ordre de mots, l'organisation syntaxique et les modifications morphologiques. Ces variantes vont au-delà de la racinisation. Nous utilisons FASTR (Jacquemin, 1996) pour l'identification

de ces variantes avec plusieurs types de règles de transformation : insertion (*cardiac disease/cardiac valve disease*), dérivation morphologique (*artery restenosis/arterial restenosis*) et permutation (*aorta coarctation/coarctation of the aorta*). Plusieurs opérations sont effectuées sur une partie du corpus 2013 : segmentation du corpus et des requêtes en mots et phrases ; étiquetage morpho-syntaxique avec TreeTagger (Schmid, 1994) ; analyse syntaxique avec YaTeA (Aubin et Hamon, 2006) pour l'extraction de syntagmes nominaux. Finalement, FASTR est appliqué pour l'acquisition de variantes de termes : les termes extraits des requêtes sont les termes de référence pour lesquels les variantes sont recherchées parmi les termes extraits du corpus. Au total, pour les 284 termes extraits de 50 requêtes nous générons 575 variantes.

Inclusion lexicale et relations hiérarchiques. L'inclusion lexicale (Kleiber et Tamba, 1990) suppose que lorsqu'un terme est inclus lexicalement dans un autre terme il existe une relation sémantique de subsomption hiérarchique entre eux. La proximité sémantique des inclusions lexicales est plus faible que celle des synonymes. Néanmoins, nous pensons que ce type de relations peut être utile pour la recherche d'information, surtout dans le contexte de recherche par les non experts, car ces derniers peuvent faire des requêtes qui sont souvent moins précises et spécifiques que les termes utilisés dans la littérature scientifique. Pour acquérir ce type de relation, le traitement est effectué en plusieurs étapes : la segmentation en mots et phrases ; l'étiquetage avec TreeTagger (Schmid, 1994) ; l'extraction de groupes nominaux avec YaTeA (Aubin et Hamon, 2006) ; les termes extraits par YaTeA sont analysés syntaxiquement en tête et expansion. Par exemple, l'analyse syntaxique de *muscle pain* donne la tête *pain* et l'expansion *muscle*. La relation sémantique est alors établie entre le terme *muscle pain* et sa tête syntaxique *pain*. Les relations identifiées de cette manière sont hiérarchiques : le terme long *muscle pain* est le fils hiérarchique du terme court *pain*. En effet, *muscle pain* contient une information plus spécifique. Ce traitement est appliqué à un sous-ensemble du corpus, et nous obtenons 1 114 959 paires de termes.

Abréviations. Les abréviations sont très fréquentes dans la littérature biomédicale et nous nous attendons à ce que le lien entre la forme abrégée et étendue d'un terme soit utile pour la recherche d'information. La ressource utilisée comporte 1 897 abréviations.

Liste de mots vides. Nous utilisons une liste de 627 mots vides. Il s'agit de mots grammaticaux et de mots très fréquents dans les documents (e.g., *accordance, amongst, indicate...*).

4. Modèle de recherche d'information

Le système de recherche d'information, qui se trouve au coeur de notre travail, est basé sur le modèle de langue (LM) statistique tel qu'il est implémenté par Indri (Strohman *et al.*, 2005). Ce système a montré de bonnes performances dans de nombreuses tâches de recherche d'information. Nous pensons que Indri fournit des fonctionnalités intéressantes pour le traitement des données biomédicales.

4.1. *Champ aléatoire de Markov*

En considérant la collection de N variables aléatoires $x = (x_1, \ldots, x_N)$, le champ aléatoire de Markov (MRF) est un modèle probabiliste graphique qui peut être représenté par un graphe non orienté $\mathcal{G}$. Ce graphe est défini par $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ où les noeuds sont les variables aléatoires ($\mathcal{V} = x$) et les arêtes représentent les dépendances conditionnelles entre ces variables. Ainsi, une clique $c \in C(\mathcal{G})$ représente un ensemble de variables dépendantes. Nous pouvons donc définir $P(x)$ comme suit :

$$P(x) = \frac{1}{Z_\Lambda} \prod_{c \in C(\mathcal{G})} \psi(c; \Lambda) \quad [1]$$

où $Z_\Lambda = \sum_x \prod_{c \in C(\mathcal{G})} \psi(c; \Lambda)$ est le coefficient de normalisation, $\psi_c(c; \Lambda)$ les fonctions potentielles et Λ le modèle.

Pour utiliser le MRF en recherche d'information, on considère que le graphe $\mathcal{G}$ contient le noeud requête Q et le noeud document D (Metzler et Croft, 2005). Ainsi, les noeuds faisant partie de la clique c, qui contient le noeud document, sont les requêtes auxquelles correspondent les documents. Comme l'objectif est de classer les documents, le score RSV (Retrieval Status Value) est utilisé :

$$RSV(Q, D) = log(P_\Lambda(D|Q)) = log(\frac{P_\Lambda(Q, D)}{P_\Lambda(Q)}) = \sum_{c \in C(\mathcal{G})} log(\psi(c; \Lambda)) \quad [2]$$

$\psi(c; \Lambda)$ est défini comme $\psi(c; \Lambda) = exp[\lambda_c * f(c)]$ où $f(c)$ est une fonction caractéristique à valeur réelle et λ_c correspond à son poids (Metzler et Croft, 2005). Nous utilisons l'implémentation proposée dans (Metzler et Croft, 2005) qui considère trois types de voisinage : unigrammes, bigrammes et bigrammes non ordonnés dans la fenêtre définie par l'utilisateur. La fonction des unigrammes $f_T(t, D)$ est définie comme suit :

$$f_T(t, D) = \frac{nb(t, D) + \mu \cdot P(t|C)}{|D| + \mu} \quad [3]$$

où t est le terme courant de l'index de la requête q_i et μ le coefficient de lissage de Dirichlet.

Les bigrammes $(t, t+1)$ et les bigrammes non ordonnés $\{t, t+1\}$ dans la fenêtre w définie par l'utilisateur sont donc considérés comme $f_{Bi}((t, t+1), D)$ et $f_W(\{t_i, t_{i+1}\}, w, D)$, respectivement. Le score RSV peut alors être calculé :

$$\begin{aligned} RSV(Q, D) \quad = \quad & \lambda_T * \prod_{t \in Q} f_T(t, D) + \\ & \lambda_{Bi} * \prod_{i=1}^{|Q|-1} f_{Bi}((t_i, t_{i+1}), D) + \\ & \lambda_W * \prod_{i=1}^{|Q|-1} f_W(\{t_i, t_{i+1}\}, w, D) \end{aligned} \quad [4]$$

Ces paramètres sont d'habitude fixés aux valeurs par défaut : $w = 8$, $\mu = 2500$, $\lambda_T = 0.85$, $\lambda_{Bi} = 0.1$ et $\lambda_W = 0.05$. Nous proposons de les ajuster à notre tâche.

4.2. *Ajustement de paramètres*

Pour optimiser le paramètre de lissage μ et les paramètres de combinaison λ, nous utilisons les données de 2013. Nous choisissons d'optimiser la MAP (Mean Average Precision, cf. infra). Pour μ, nous testons systématiquement les valeurs de 0 à 30 000 avec des seuils variant d'un pas de 500 ; pour λ, nous testons les valeurs de 0 à 1 avec des seuils variant d'un pas de 0,05. Dans la table 2, nous indiquons les performances obtenues avec les paramètres par défaut et ceux qui maximisent la MAP sur l'ensemble 2013. La différence significative des résultats montre l'importance de l'optimisation de ces paramètres.

Paramètres	MAP	P@5	P@10	P@100	NDCG@5	NDCG@10	NDCG@100
défaut	0,2627	0,3920	0,4040	0,1414	0,3967	0,4075	0,4420
meilleur	0,3053	0,4895	0,4653	0,1558	0,4895	0,4763	0,4943

Tableau 2. *Performances des paramètres définis sur les données de CLEF eHealth 2013 (valeurs par défaut et valeurs optimisées).*

4.3. *Évaluation*

Le test de la méthode est effectué sur le corpus 2014. Seuls les 1 000 premiers résultats (RSV le plus élevé) de chaque requête sont retenus et évalués. Plusieurs mesures sont utilisées pour l'évaluation :

– P@N, R@N : évalue le taux de précision ou de rappel sur les N premiers documents retournés ;

– MAP : avec Σ l'ensemble des réponses attendues (vérité terrain), Δ l'ensemble des réponses retournées par le système, $P@rank(d)$ la précision au rang du document d : $\text{MAP} = \frac{\sum_{d \in \Sigma \cap \Delta} P@rank(d)}{|\Sigma|}$;

– NDCG@N (Normalized Discounted Cumulative Gain) évalue le gain pour les N premières réponses, par exemple cinq (NDCG@5) et dix (NDCG@10) premières réponses.

5. Expériences

La méthode et les ressources dont nous disposons offrent plusieurs possibilités pour effectuer les expériences. Nous présentons ici quelques expériences testées. Dans chaque expérience, la question et sa description sont traitées.

Baseline. L'objectif est de mesurer la performance du système sans modification et sans l'utilisation de ressources supplémentaires. Cette expérience consiste donc à utiliser Indri avec les paramètres réglés sur le corpus 2013. Il s'agit de la `baseline`.

exp_UMLS. Pour cette expérience, les requêtes sont étendues avec les synonymes d'UMLS : les termes UMLS de longueur maximale sont recherchés dans les requêtes ; les synonymes de ces termes sont recherchés dans l'UMLS (termes avec le même CUI) ; ces synonymes sont ajoutés à la requête initiale. Il est important de noter que certains termes sont ambigus et se retrouvent dans différents CUI. Cet aspect n'est pas traité et il est possible qu'un terme soit étendu avec les synonymes qui n'ont pas la sémantique attendue. Pour limiter l'effet éventuellement négatif des termes plus éloignés sémantiquement, les termes de l'expansion reçoivent un poids plus bas que les termes de la requête initiale. Ici encore, nous nous basons sur le corpus 2013 pour déterminer le poids optimal : il apparaît que ce poids ne doit pas dépasser 0.1, alors que le poids maximal est de 1.

D'autres expériences. Ces expériences exploitent les ressources sémantiques autres que l'UMLS et bénéficient également des traitements suivants effectués grâce à la plate-forme Ogmios (Hamon et Nazarenko, 2008) :

– la segmentation de requêtes en mots et phrases, si nécessaire ;

– l'étiquetage morpho-syntaxique avec TreeTagger (Schmid, 1994) ;

– l'analyse syntaxique avec YaTeA (Aubin et Hamon, 2006) pour l'extraction de groupes nominaux.

Les mots vides sont supprimés et les requêtes sont enrichies avec les ressources et leurs différentes combinaisons :

– `exp_ogmios` (`ogmios`) : expansion avec les synonymes induits,

– `exp_ogmios+ab` (`ab`) : expansion avec les synonymes induits et les abréviations,

– `exp_ogmios+ab+terms` (`terms`) : expansion avec les synonymes induits, les abréviations, et les termes complexes extraits,

– `exp_ogmios+ab+FASTER` (`FASTER`) : expansion avec les synonymes induits, les abréviations, et les variantes morpho-syntaxiques de FASTR,

– `exp_ogmios+ab+IL_gen` (`IL_gen`) : expansion avec les synonymes induits, les abréviations, et les inclusions lexicales dans le sens de généralisation ({*crohn disease*, *disease*}, où *crohn disease* est réduit à *disease*),

– `exp_ogmios+ab+IL_spec` (`IL_spec`) : expansion avec les synonymes induits, les abréviations, les inclusions lexicales dans le sens de spécialisation ({*disease*, *crohn disease*}, où *disease* est spécifié vers *crohn disease*).

6. Résultats

Résultats globaux. La figure 1 et le tableau 3 présentent les résultats globaux obtenus sur le corpus de test (corpus 2014) avec les différentes expériences. Globalement, nous pouvons discerner trois groupes : la `baseline` et `exp_UMLS` qui obtiennent les meilleurs résultats ; les expériences basées sur l'utilisation de FASTER (`exp_ogmios+ab+FASTER_gen` et `exp_ogmios+ab+FASTER_spec`) aboutissent à

Expe.	MAP	R-prec	P@5	P@10	P@100	tps exec.	Nb. expansions
Baseline	0,4409	0,4389	0,7680	0,7600	0,3686	6m29s	/
UMLS	0,4450	0,4414	0,7760	0,7660	0,3678	16m24s	255,68
ogmios	0,3692	0,3839	0,7360	0,7120	0,3198	4m2s	15,38
ogmios+UMLS	0,4070	0,4210	0,7560	0,7540	0,3440	20m34	271,06
ab	0,3723	0,3834	0,7560	0,7020	0,3194	4m2s	15,48
ab+terms	0,3697	0,3833	0,7240	0,7060	0,3194	4m25s	22,32
FASTER_gen	0,2597	0,2769	0,5720	0,5580	0,2372	7m42s	26,34
FASTER_spec	0,2399	0,2599	0,5120	0,4880	0,2218	8m42s	37,50
IL_gen	0,3875	0,3992	0,7520	0,7260	0,3298	4m50s	25.88

Tableau 3. *Performances obtenues sur le corpus 2014 pour les différentes mesures. La dernière colonne correspond au nombre moyen d'expansions par requête.*

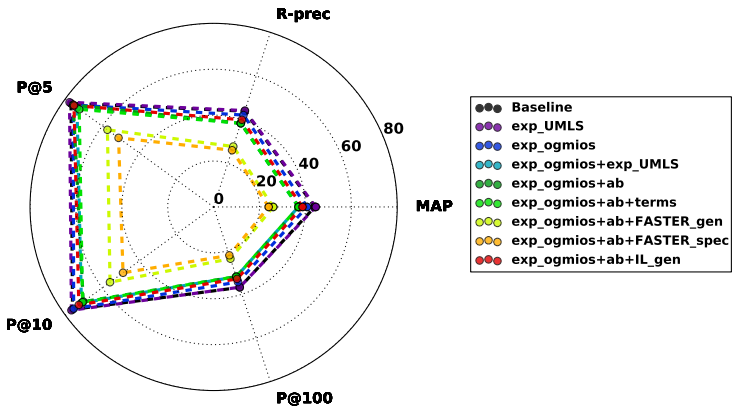

Figure 1. *Performances comparatives de différentes expériences.*

des résultats nettement inférieurs. L'ensemble des autres expériences se situent à un niveau intermédiaire. La figure montre les expériences effectuées avec une autre représentation. Nous pouvons par exemple voir que deux des expériences effectuées (`exp_ogmios+ab+FASTER_gen` et `exp_ogmios+ab+FASTER_spec`) sont nettement inférieures aux autres expériences quelles que soient les mesures d'évaluation. Par contre, `baseline`, `exp_UMLS` et dans la moindre mesure `exp_ogmios+exp_UMLS` sont compétitifs entre eux. Notons aussi que ce sont les mesures P@5 et P@10 qui sont privilégiées par nos expériences. Le temps d'exécution reste raisonnable, sauf pour les expériences utilisant l'UMLS, qui comporte plus de 10 M de concepts. La dernière colonne du tableau 3 indique le nombre moyen d'expansions par requête : lorsque l'UMLS est utilisé ce nombre est largement supérieur par rapport à d'autres expériences.

Différence de performances entre 2013 et 2014. La figure 2 montre la différence de performances obtenues sur les deux corpus traités : 2013 et 2014. Le paramétrage

du système est le même dans les deux cas, en particulier avec les mêmes paramètres RSV et le même index. Nous pouvons voir que les résultats obtenus sur le corpus de test (2014) sont bien meilleurs que ceux obtenus sur le corpus d'entraînement (2013). Cette différence est difficile à expliquer. Il est possible que les requêtes de 2014 soient plus faciles que celles de 2013, ou bien que la référence comporte des données plus homogènes.

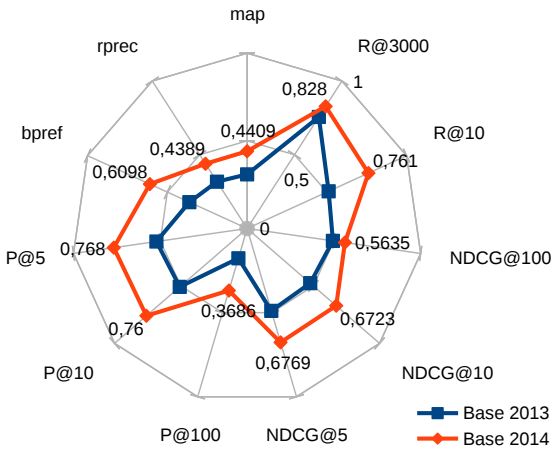

Figure 2. *Performance de la baseline sur les corpus 2013 et 2014.*

Utilisation de synonymes d'UMLS. La particularité de `exp_UMLS` est qu'il effectue l'expansion avec les termes d'UMLS. Il peut s'agir de termes simples et complexes. Pour cette expérience, nous notons un nombre important d'expansions. Sur la figure 3, nous présentons la différence entre `baseline` et `exp_UMLS`. Nous pouvons voir que ces deux expériences sont assez équivalentes, avec des améliorations obtenues pour certaines requêtes (e.g. 9, 11, 26, 30, 33, 36) avec le `exp_UMLS`. Nous pensons que l'avantage principal d'UMLS est qu'il offre des termes contrôlés pour un concept donné, qui peuvent provenir de terminologies orientées sur le médecin ou sur le patients (CHV), et qu'il établit le lien d'équivalence entre ces termes.

Utilisation de synonymes induits. À la figure 4, nous présentons une comparaison entre plusieurs expériences qui recourent à l'expansion avec les synonymes induits. La comparaison est effectuée avec le `exp_UMLS`. Contrairement au `exp_UMLS`, étendu avec des termes simples et complexes, les synonymes induits consistent majoritairement en termes simples : l'extension de requêtes repose donc sur l'hypothèse de compositionnalité des termes et la possibilité de retrouver la sémantique des termes complexes après leur segmentation en mots et l'extension sémantique de ces mots. Le risque induit avec cette hypothèse cause une détérioration de résultats pour plusieurs requêtes. Seules quelques requêtes bénéficient de ces traitements (1, 12, 14, 23, 27, 31, 33, 43, 44). Nous notons aussi un nombre d'expansions plus réduit qu'avec `exp_UMLS`.

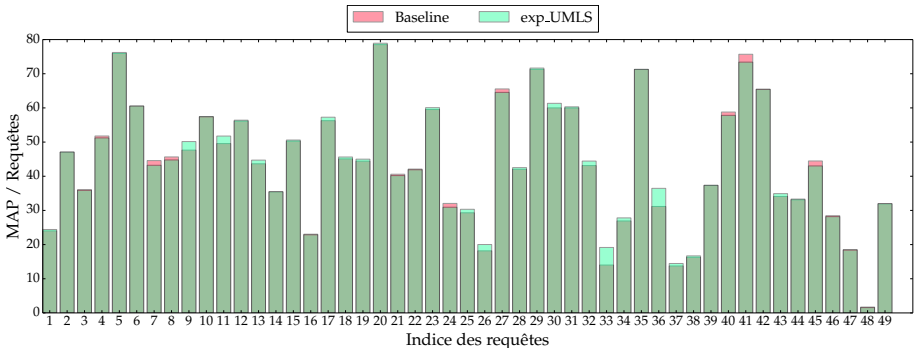

Figure 3. *Comparaison entre les runs* `baseline` *et* `exp_UMLS`.

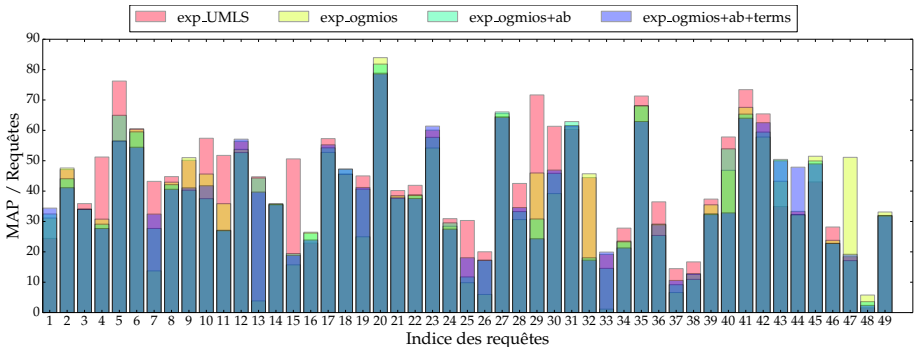

Figure 4. *Comparaison entre les runs* `exp_UMLS` *et l'utilisation de synonymes induits.*

Utilisation des variantes morpho-syntaxiques, des inclusions lexicales et des abréviations. La figure 5 présente les résultats obtenus suite à l'utilisation des variantes morpho-syntaxiques générées avec FASTER, des inclusions lexicales et des abréviations. La comparaison est effectuée avec le `exp_UMLS`. Ici encore nous pouvons voir que, par rapport au `exp_UMLS`, l'expansion avec ces ressources induites est bénéfique pour plusieurs requêtes (1, 13, 14, 16, 18, 23, 31, 32, 36, 43, 44, 45, 48, 49), pour lesquelles des éléments d'expansion intéressants sont alors ajoutés. Si on essaie de positionner les différentes expériences de la série entre elles, nous pouvons observer que le plus souvent l'expérience `exp_ogmios+ab+IL_gen` (utilisation des abréviations et des inclusions lexicales) s'avère la plus performante. Notons aussi que les abréviations concernent en réalité très peu de requêtes, ce qui peut expliquer leur faible impact. À la figure 6, nous effectuons une comparaison entre la `baseline`, `exp_UMLS`, `exp_ogmios` et la combinaison des deux dernières `exp_ogmios+exp_UMLS`. Chacune

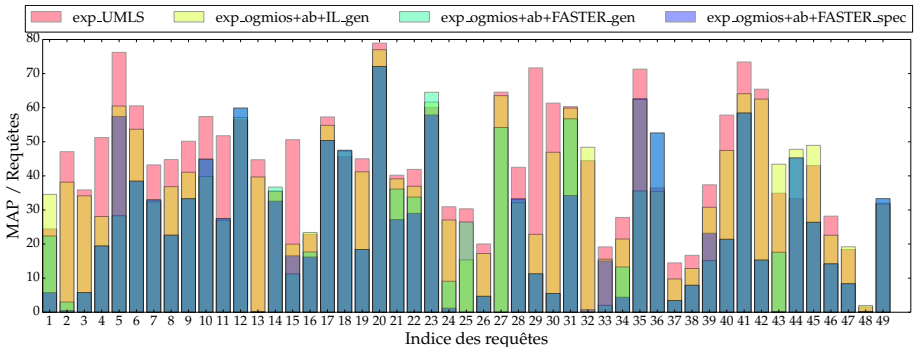

Figure 5. *Comparaison entre le run* `exp_UMLS` *et les expériences utilisant les variantes FASTER, les inclusions lexicales et les abréviations.*

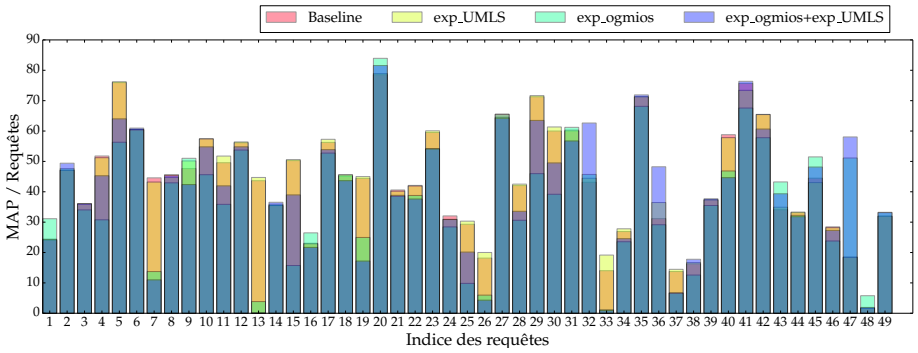

Figure 6. *Combinaison d'UMLS avec les synonymes induits.*

de ces expériences peut devancer les autres : `baseline` (e.g., 4, 7, 24, 40), `exp_UMLS` (e.g., 11, 17, 26, 30, 33, 37), `exp_ogmios` (e.g., 1, 9, 16, 20, 31, 43, 45, 48), et la combinaison des deux dernières `exp_ogmios+exp_UMLS` (e.g., 2, 32, 36, 38, 47). L'objectif serait de combiner ces expériences de manière plus efficace pour améliorer les résultats globaux.

Comparaison avec d'autres systèmes de la compétition CLEF eHealth. Par rapport aux résultats de la compétition CLEF eHealth 2014, les systèmes présentés dans le tableau 1 ont des résultats d'expériences meilleurs que ceux obtenus par notre système. Certains des paramètres sont comparables (e.g., moteur de recherche, ressources externes) à ceux que nous avons utilisés. D'autres systèmes participants montrent des résultats moins performants, pouvant descendre jusqu'à 0,0640 de P@5, 0,0560 de NDCG@10 et 0,0625 de MAP. Un des défauts de notre système est qu'il

fournit des résultats instables : l'expansion de requêtes peut être très bénéfique pour certaines requêtes mais apporter des résultats négatifs pour d'autres.

7. Conclusion et travaux futurs

Nous avons proposé un travail en recherche d'information médicale, où la tâche consiste à trouver des réponses aux questions des patients. Les données de référence ont été définies par des utilisateurs réels et ont été utilisées lors de la compétition CLEF eHealth. Notre travail est basé sur un moteur de recherche de l'état de l'art Indri et des ressources sémantiques qui cherchent à faire le lien entre le langage des patients et celui des médecins. Notre système a été entraîné sur le corpus 2013 et testé sur le corpus 2014. Comparé à d'autres tâches de recherche d'information et à d'autres systèmes participant à CLEF eHealth en 2014, les résultats globaux de notre système sont bons avec P@10 allant jusqu'à 0.65. Cependant, contrairement à nos attentes, plusieurs des stratégies testées avec l'utilisation de ressources sémantiques supplémentaires (synonymes induits, abréviations, inclusions lexicales et variantes morpho-syntaxiques) n'ont pas permis d'améliorer les résultats de manière systématique. Ces ressources sont bénéfiques pour certaines requêtes mais détériorent les résultats d'autres. Seuls les synonymes d'UMLS apportent une amélioration assez systématique par rapport à la `baseline`. Des résultats similaires quant à l'expansion de requêtes ont été également observés dans d'autres études effectuées sur des jeux de données de la langue générale (Voorhees, 1994).

Dans les travaux futurs, nous voudrions améliorer notre expertise dans l'utilisation de terminologies et de ressources sémantiques pour la recherche d'information. Une analyse détaillée de nos résultats peut nous permettre de mieux maîtriser les pistes à explorer. Par ailleurs, l'exploitation de connaissances terminologiques dès l'indexation est aussi une piste prometteuse mais peut devenir vite coûteuse pour le système. Le contexte lié à la recherche d'information par les patients introduit d'autres défis à relever. Typiquement, les documents corrects retrouvés sont des documents fiables créés par les experts et scientifiques. Ces documents, tout en proposant des réponses plus étendues aux questions des patients, comportent des termes et notions difficiles. Il est nécessaire, ici aussi, de proposer des méthodes et ressources qui aideraient les patients à mieux comprendre le contenu des documents retrouvés. Il s'agit donc d'une piste de recherche prometteuse, qui a le potentiel de fournir un système de recherche d'information à destination de patients assez complet.

Remerciements

Ce travail a été partiellement financé dans le cadre du Labex Comin'Labs.

8. Bibliographie

AMA, « Health literacy : report of the Council on Scientific Affairs. Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, American Medical Association », *JAMA*, vol. 281, n° 6, p. 552-7, 1999.

Aubin S., Hamon T., « Improving Term Extraction with Terminological Resources », *FinTAL 2006*, n° 4139 in *LNAI*, Springer, p. 380-387, 2006.

Cartoni B., Deléger L., « Découverte de patrons paraphrastiques en corpus comparable : une approche basée sur les n-grammes », *TALN*, 2011.

Choi S., Choi J., « Exploring Effective Information Retrieval Technique for the Medical Web Documents : SNUMedinfo at CLEFeHealth2014 Task 3 », *Proceedings of CLEF 2014*, Lecture Notes in Computer Science (LNCS), Springer, p. 167-175, 2014.

D'Alessandro D., Kingsley P., Johnson-West J., « The readability of pediatric patient education materials on the World Wide Web », *Arch Pediatr Adolesc Med.*, vol. 155, n° 7, p. 807-12, 2001.

de Boer M., Versteegen G., van Wijhe M., « Patients' use of the internet for pain-related medical information. », *Patient Education and Counseling*, vol. 68, n° 1, p. 86-97, 2007.

Diaz J., Griffith R., Ng J., Reinert S., Friedmann P., Moulton A., « Patients' use of the internet for medical information », *J Gen Intern Med*, vol. 17, n° 3, p. 180-185, 2002.

Elhadad N., Sutaria K., « Mining a lexicon of technical terms and lay equivalents », *BioNLP*, p. 49-56, 2007.

Eysenbach G., Thomson M., « The FA4CT algorithm : a new model and tool for consumers to assess and filter health information on the Internet », *in* K. Kuhn, A. McGray (eds), *MEDINFO*, IOS Press, Brisbane, Australia, p. 142-146, 2007.

Gaudinat A., Grabar N., Boyer C., « Automatic retrieval of webpages with standards of ethics and trustworthiness within a medical portal : What a page name tells us », *AIME 2007*, 2007.

Goeuriot L., Kelly L., Li W., Palotti J., Pecina P., Zuccon G., Hanbury A., Jones G., Müller H., « ShARe/CLEF eHealth Evaluation Lab 2014, Task 3 : User-centred Health Information Retrieval », *Proceedings of CLEF 2014*, Lecture Notes in Computer Science (LNCS), Springer, p. 43-61, 2014.

Grabar N., Hamon T., « Exploitation of linguistic indicators for automatic weighting of synonyms induced within three biomedical terminologies. », *MEDINFO 2010*, p. 1015-9, 2010.

Hamon T., Nazarenko A., « Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience », *TAL*, vol. 49, n° 2, p. 127-154, 2008.

Jacquemin C., « A Symbolic and Surgical Acquisition of Terms Through Variation », *in* S. Wermter, E. Riloff, G. Scheler (eds), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Springer, p. 425-438, 1996.

Jucks R., Bromme R., « Choice of words in doctor-patient communication : an analysis of health-related internet sites », *Health Commun*, vol. 21, n° 3, p. 267-77, 2007.

Kelly L., Goeuriot L., Suominen H., Mowery D. L., Velupillai S., Chapman W. W., Zuccon G., Palotti J., « Overview of the ShARe/CLEF eHealth Evaluation Lab 2013 », *Proceedings of CLEF 2013*, Lecture Notes in Computer Science (LNCS), Springer, 2013.

Kickbusch I., Pelikan J. M., Apfel F., Tsouros A. D., Health literacy. The solid facts, Technical report, WHO, 2013.

Kleiber G., Tamba I., « L'hyperonymie revisitée : inclusion et hiérarchie », *Langages*, vol. 98, p. 7-32, juin, 1990.

Luciano J., Cumming G., Wilkinson M., Kahana E., « The emergent discipline of health web science », *J Med Internet Res*, vol. 18, n° 8, p. e166, 2013.

McCray A., « Promoting Health Literacy », *J of Am Med Infor Ass*, vol. 12, p. 152-163, 2005.

Metzler D., Croft W., « ACM SIGIR conference on Research and development in information retrieval », *in* ACM (ed.), *A markov random field model for term dependencies*, p. 472-479, 2005.

NLM, *UMLS Knowledge Sources Manual*, National Library of Medicine, Bethesda, Maryland. 2008, www.nlm.nih.gov/research/umls/.

Oh H.-S., Jung Y., « A Multiple-stage Approach to Re-ranking Clinical Documents », *Proceedings of CLEF 2014*, Lecture Notes in Computer Science (LNCS), Springer, p. 210-219, 2014.

Ounis I., Amati G., Plachouras V., He B., Macdonald C., Lioma C., « Terrier : A High Performance and Scalable Information Retrieval Platform », *ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

Ozturkmenoglu O., Alpkoçak A., Kilinc D., « DEMIR at CLEF eHealth : The Effects of Selective Query Expansion to Information Retrieval », *Proceedings of CLEF 2014*, Lecture Notes in Computer Science (LNCS), Springer, p. 220-225, 2014.

Pletneva N., Vargas A., Boyer C., Requirements for the general public health search, Technical report, KHRESMOI project, 2011. D8.1.1.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *ICNMLP*, Manchester, UK, p. 44-49, 1994.

Shenwei W., Nie J.-Y., Liu X., Liu X., « An Investigation of the Effectiveness of Concept-based Approach in Medical Information Retrieval », *Proceedings of CLEF 2014*, Lecture Notes in Computer Science (LNCS), Springer, p. 236-247, 2014.

Strohman T., Metzler D., Turtle H., Croft W., « Indri : a language-model based search engine for complex queries », *International Conference on Intelligent Analysis*, 2005.

Thakkar H., Iyer G., Shah K., Majumder P., « Team IRLabDAIICT at ShARe/CLEF eHealth 2014 Task 3 : User-centered Information Retrieval System for Clinical Documents », *Proceedings of CLEF 2014*, Lecture Notes in Computer Science (LNCS), Springer, p. 248-259, 2014.

Tran T., Chekroud H., Thiery P., Julienne A., « Internet et soins : un tiers invisible dans la relation médecine/patient ? », *Ethica Clinica*, vol. 53, p. 34-43, 2009.

Voorhees E., « Query expansion using lexical-semantic relations », *in* I. Springer-Verlag New York (ed.), *International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 61-69, 1994.

Yang C., Bhattacharya S., Srinivasan P., « The University of Iowa at CLEF 2014 : eHealth Task 3 », *Proceedings of CLEF 2014*, Lecture Notes in Computer Science (LNCS), Springer, p. 283-295, 2014.

Zeng Q., Tse T., « Exploring and developing Consumer Health Vocabularies », *JAMIA*, vol. 13, p. 24-29, 2006.