Combinaison d'analyses sémantiques pour la recherche d'information médicale

Loïc Maisonnasse[1], **Eric Gaussier**[2], **Jean-Pierre Chevallet**[2]

[1] *Université de Lyon, INSA-Lyon, LIRIS*
loic.maisonnasse@insa-lyon.fr
[2] *LIG-UJF - 38041 Grenoble Cedex 9, France*
jean-pierre.chevaller@imag.fr,eric.gaussier@imag.fr

RÉSUMÉ. *Nous étudions dans ce papier la combinaison de différentes méthodes de détection de concepts dans une indexation sémantique. Les indexations conceptuelles fournissent de bons résultats lorsque de grandes bases de connaissances sont disponibles. Cependant, la détection des concepts n'est pas toujours fiable et des erreurs peuvent limiter les performances de l'indexation. Une solution pour résoudre ce problème est de combiner différentes méthodes de détection des concepts. Dans cet article, nous étudions plusieurs méthodes pour combiner ces détections, à la fois sur les requêtes et les documents, dans le cadre de l'approche modèle de langue de la recherche d'information. Nos expériences montrent que notre modèle de fusion améliore le modèle standard d'environ 17% en précision moyenne.*

ABSTRACT. *We study in this paper the combination of different concept detection methods for conceptual indexing. Conceptual indexing shows effective results when large knowledge bases are available. But concept detection is not always accurate and errors limit interest of concept usage. A solution to solve this problem is to combine different concept detection methods. In this paper, we investigate several ways to combine concept detection methods, both on queries and documents, within the framework of the language modeling approach to IR. Our experiments show that our model fusion improves the standard language model by up to 17% on mean average precision.*

MOTS-CLÉS : *recherche d'information, modèle de langue, combinaison d'analyses*

KEYWORDS: *information retrieval, modèle de langue, combinaison d'analyses*

1. Introduction

L'Indexation conceptuelle, c'est-à-dire l'utilisation des concepts plutôt que des mots ou des termes en recherche d'information, est souvent considérée comme un moyen de mieux représenter le contenu d'un document. Les concepts étant des entités abstraites qui unifient et résument un ensemble d'objets concrets ou mentaux par abstraction de traits communs, leur utilisation permet de mieux s'abstraire du contenu des documents. Cela permet de résoudre des problèmes linguistiques tels que la polysémie, la synonymie ou encore la variation des termes. De plus par ses caractéristiques, une indexation conceptuelle est naturellement multilingue.

Même si l'idée est assez ancienne(Schank *et al.,* 1980), il reste actuellement difficile d'indexer des textes au niveau conceptuel. De telles indexations nécessitent généralement de grandes bases de connaissances. Or dans certains domaines, comme le domaine médical, de telles bases existent, et plusieurs travaux ont profité de leur présence pour proposer des modèles de recherche de l'information. Dans ImageCLEFmed [1], une tâche de recherche d'information médicale, les systèmes les plus performants sur le texte (Maisonnasse *et al.,* 2008, Lacoste *et al.,* 2006, Chevallet *et al.,* 2007) utilisent des indexations conceptuelles. De même dans la piste génomique de TREC[2], où les travaux de Zhou (Zhou *et al.,* 2007) utilisent les informations conceptuelles pour étendre les requêtes.

Néanmoins, les concepts détectés à partir du texte ne sont pas toujours exacts et leur utilisation peut entrainer une dégradation des performances du système de recherche d'information. Les méthodes utilisées pour identifier des concepts dans les textes, à partir d'une base de connaissances, reposent habituellement sur une séquence de traitement de la langue naturelle. Or si de tels outils existent, leurs performances varient fortement d'un domaine à un autre, ainsi que d'une langue à l'autre. Au final, les performances d'un système de recherche d'information basé sur les concepts dépend en grande partie des performances de la méthode de détection des concepts utilisée. Pour surmonter ce problème, sans travailler sur le processus de détection lui-même, une stratégie consiste à combiner les résultats de plusieurs méthodes. L'objectif de ce document est précisément d'étudier différentes stratégies de fusion dans le cadre de l'approche modèle de langue de RI.

L'approche de modèle de langue appliquée à la RI a été proposée par Ponte et Croft (Ponte *et al.,* 1998). L'idée de base considère chaque document comme un échantillon d'une langue donnée, et l'interrogation comme un processus génératif. De nombreux travaux se sont appuyés sur cette méthode, et ont abouti à un modèle simple et adaptable (Lafferty *et al.,* 2001) qui fournit de bonnes performances en RI. Si la plupart des approches modèle de langue sont encore fondées sur les modèles unigrammes, certaines tentatives ont été faites pour utiliser des concepts au lieu des mots (Srikanth *et al.,* 2003, Maisonnasse *et al.,* 2008). Nous suivons ici le modèle de base présenté

1. ImageCLEFmed est une piste de la campagne d'évaluation CLEF (http ://www.clef-campaign.org/)
2. trec.nist.gov

dans (Maisonnasse *et al.*, 2008) et nous étudions comment y intégrer différentes analyses des requêtes et des documents.

2. Combinaison de modèles

Si extraire des concepts est une tâche difficile, l'impact d'une mauvaise détection n'est pas le même si cette détection est faite sur les documents ou sur les requêtes. En effet, les documents contiennent plusieurs phrases et il est possible qu'une erreur dans l'une des phrases soit compensée par une détection correcte dans le reste du document.

Ce n'est pas le cas pour les requêtes qui, dans la majorité des cas, ne contiennent que peu de mots. Une seule erreur au niveau des requêtes peut fortement dégrader le rappel. Nos travaux(Maisonnasse *et al.*, 2008), sur la collection CLEF médical, montrent que la combinaison de détections peut fortement améliorer les performances d'un système de recherche d'information conceptuel. Nous présentons ici les approches qui permettent de combiner différentes détections, que ce soit sur les requêtes ou sur les documents.

Dans cette étude, nous nous reposons sur un modèle de langue défini sur des concepts, nous y référerons en tant que *Model Conceptuel Unigramme*. Nous supposons qu'une requête q est composée par un ensemble $\mathcal{C}$ de concepts, chaque concept étant indépendant des autres sachant le modèle de documents, nous obtenons le calcul suivant :

$$P\left(C|M_d\right) = \prod_{c_i \in C} P(c_i|M_d)^{\#(c_i,q)} \quad [1]$$

où $\#(c_i, q)$ représente le nombre de fois où le concept c_i apparaît dans une requête q. La quantité $P(c_i|M_d)$ est calculé directement par un maximum de vraisemblance, combiné à un lissage de Jelinek-Mercer :

$$P(c_i|M_d) = (1 - \lambda_u)\frac{|c_i|_d}{|*|_d} + \lambda_u \frac{|c_i|_{\mathcal{D}}}{|*|_{\mathcal{D}}}$$

où $|c_i|_d$ (respectivement $|c_i|_{\mathcal{D}}$) est la fréquence d'un concept c_i dans le document d (respectivement dans la collection $\mathcal{D}$), et $|*|_d$ (respectivement $|*|_{\mathcal{D}}$) est la taille du document d, i.e. le nombre de concepts dans d (respectivement de la collection).

Cependant puisque nous nous intéressons à la combinaison d'analyses, nous n'avons plus un seul modèle de document associé à un document d mais plusieurs, chacun correspondant à une méthode de détection. Nous notons par M_d^* l'ensemble des modèles d'un document ($M_d^* = \{M_d^1, ..., M_d^p\}$). De même, une requête est constituée d'un ensemble d'ensemble de concepts C^*, chaque ensemble de concepts résultant de l'application d'une des méthodes de détection sur les requêtes ($C^* =$

$\{C^1, ..., C^p\}$). Le score final de pertinence d'un document pour une requête (RSV) est alors donné par :

$$RSV(q,d) \quad = \quad P(C^*|M_d^*)LMcontrib \qquad [2]$$

le problème est maintenant de décomposer la probabilité $P(C^*|M_d^*)$ en fonction de nos différents ensembles de concepts et de modèles de documents. Puisque chaque élément dans C^* et M_d^* est généré par une méthode de détection différente, nous posons l'hypothèse que ces éléments sont indépendants[3].

2.1. *Fusion de la requête*

Du côté des requêtes nous proposons une méthode pour décomposer l'ensemble d'ensemble de concepts C^*. Cette méthode considère qu'un document est pertinent s'il génère toutes les analyses de la requête q, ce qui conduit à calculer :

$$P(C^*|M_d^*) \propto \prod_{C \in C^*} P(C|M_d^*) \qquad [3]$$

En utilisant cette décomposition, nous pouvons maintenant nous intéresser à la décomposition de M_d^*.

2.2. *Fusion des modèles de documents*

Dans l'approche modèle de langue de RI, un modèle de langue est calculé en fonction du contenu du document. Comme nous utilisons différentes méthodes de détection, un document possède différentes représentations conceptuelles. Il existe donc plusieurs méthodes pour regrouper les différentes représentations fournies en sortie de chaque détection conceptuelle d'un document. Nous explorons maintenant ces différentes possibilités.

En utilisant le théorème de Bayes, nous obtenons la réécriture suivante de la probabilité $P(C|M_d^*)$:

$$P(C|M_d^*) = \frac{P(C)}{P(M_d^*)} P(M_d^*|C)$$

Dans le contexte da la recherche d'information, nous souhaitons obtenir une liste triée de documents, par l'utilisation de leur rsv (retrieval status value). Le terme $P(C)$, commun à tous les documents, n'influence pas ce classement. N'ayant pas de connaissance *à priori* sur les performances de chaque méthode de détection des concepts, nous

3. cette hypothèse est évidement une simplification du fait que chacune des méthodes de détection des concepts utilise la même base de connaissance.

considérons que les modèles de document M_d^* sont équiprobables, et par conséquent que les probabilités $P(M_d^*)$ sont les mêmes pour chaque document. Nous pouvons alors écrire :

$$P(C|M_d^*) \propto P(M_d^*|C) \quad [4]$$

Comme précédemment, nous avons plusieurs possibilités pour décomposer la probabilité $P(M_d^*|C)$. D'une part, nous pouvons considérer que chaque document doit être associé à l'ensemble des concepts C, ou qu'au moins un document doit être associé à C. La première considération correspond à regrouper les contributions des différents modèles de document à l'aide d'un produit. La seconde, quant à elle, consiste à sommer les contributions de chaque modèle de document. De plus dans la seconde considération, nous pouvons également choisir de ne prendre en compte que le meilleur modèle de document, ce qui correspond à prendre le maximum sur toutes les contributions des modèles de document. Ces deux dernières décompositions, utilisant la somme et le maximum sont en fait similaire quand la probabilité $P(M_d|C)$ est prise sur un seul modèle, et le maximum est souvent utilisé comme approximation de la somme quand celle-ci est difficile à calculer. Nous la considérons ici pour la complétude, mais nous ne l'utiliserons pas dans la décomposition de la requête où la somme peut facilement être calculée. Les décompositions produites par ces différentes approches sont résumées ci-dessous :

$$P(M_d^*|C) \quad \propto \quad \begin{cases} \prod_{M_d \in M_d^*} P(M_d|C) \\ \sum_{M_d \in M_d^*} P(M_d|C) \\ \max_{M_d \in M_d^*} P(M_d|C) \end{cases} \quad [5]$$

En appliquant le théorème de Bayes au terme $P(M_d|C)$ nous obtenons :

$$P(M_d|C) = \frac{P(M_d)}{P(C)} P(C|M_d)$$

Avec les mêmes hypothèses que précédemment, cette quantité peut se simplifier en :

$$P(M_d|C) \propto P(C|M_d)$$

En substituant cette expression dans les équations 5 et 4 nous obtenons :

$$P(C|M_d^*) \quad \propto \quad \begin{cases} \prod_{M_d \in M_d^*} P(C|M_d) \\ \sum_{M_d \in M_d^*} P(C|M_d) \\ \max_{M_d \in M_d^*} P(C|M_d) \end{cases} \quad [6]$$

cet ensemble de décompositions peut être directement combiné avec la décomposition obtenue précédemment pour la requête. Il existe cependant une deuxième méthode pour décomposer la probabilité $P(C^*|M_d^*)$, cette méthode est présentée dans la partie suivante.

2.3. *Décomposition commune*

Au lieu de décomposer d'abord l'ensemble des requêtes, puis les modèles de document, il est possible de décomposer les modèles de document en premier. Cela revient en utilisant le théorème de Bayes à :

$$P(C^*|M_d^*) = \frac{P(C^*)}{P(M_d^*)} P(M_d^*|C^*)$$

et avec les hypothèses faites dans le contexte de la recherche d'information :

$$P(C^*|M_d^*) \propto P(M_d^*|C^*)$$

nous pouvons à nouveau décomposer M_d^* soit par un produit, une somme ou un maximum sur les modèles de document, ce qui revient à :

$$P(M_d^*|C^*) \propto \begin{cases} \prod_{M_d \in M_d^*} P(M_d|C^*) \\ \sum_{M_d \in M_d^*} P(M_d|C^*) \\ \max_{M_d \in M_d^*} P(M_d|C^*) \end{cases}$$

en utilisant à nouveau le théorème de Bayes et en nous appuyant sur nos hypothèses, nous obtenons le même développement que celui utilisé pour l'équation 6, ce qui nous mène à :

$$P(C^*|M_d^*) \quad \propto \quad \begin{cases} \prod_{M_d \in M_d^*} P(C^*|M_d) \\ \sum_{M_d \in M_d^*} P(C^*|M_d) \\ \max_{M_d \in M_d^*} P(C^*|M_d) \end{cases} \qquad [7]$$

2.4. *Résumé*

En combinant les décompositions produites par les équations 3,6 et 7, nous obtenons finalement cinq décompositions (certaines décompositions étant identiques). Nous les nommons comme il suit entre parenthèses avec (*Fus* qui signifie *fusion*) :

$$P(C^*|M_d^*) \propto \begin{cases} \prod_{C \in C^*} & \prod_{M_d \in M_d^*} & P(C|M_d) & (Fus1) \\ \prod_{C \in C^*} & \sum_{M_d \in M_d^*} & P(C|M_d)) & (Fus2) \\ \prod_{C \in C^*} & \max_{M_d \in M_d^*} & P(C|M_d)) & (Fus3) \\ \sum_{M_d \in M_d^*} & \prod_{C \in C^*} & P(C|M_d)) & (Fus4) \\ \max_{M_d \in M_d^*} & \prod_{C \in C^*} & P(C|M_d)) & (Fus5) \end{cases} \qquad [8]$$

équations dans lesquelles la quantité $P(C|M_d)$ est calculée par l'équation 1.

3. Application au domaine médical

Le domaine médical se prête bien à l'indexation par concepts dans la mesure où de nombreuses ressources ont été développées de façon à indexer plus finement le contenu de textes médicaux. L'utilisation de thésaurus permettant une indexation conceptuelle (par exemple à partir de UMLS[4]) a été étudiée dans plusieurs articles, par exemple(Lacoste *et al.*, 2006) dans le cadre des campagnes ImageCLEFmed de CLEF[5] ou (Zhou *et al.*, 2007) dans la piste génomique de TREC[6] (dans ce dernier cas, ce sont toutes les variantes terminologiques d'un concept qui sont utilisées en indexation). Au-delà d'une indexation conceptuelle, plusieurs chercheurs se sont intéressés à la prise en compte de relations entre concepts pour la recherche d'information. En particulier (Vintar *et al.*, 2003) indexe documents et requêtes d'un corpus médical sur la base d'UMLS. Une relation entre deux concepts (d'un document ou d'une requête) est détectée dès lors que les deux concepts apparaissent dans la même phrase et qu'ils sont reliés dans le thésaurus. Nous nous inscrivons dans la lignée de ces travaux, et détaillons maintenant notre processus d'indexation.

3.1. *Extraction de concepts dans le domaine médical*

UMLS est un méta-thésaurus qui résulte de la fusion de différentes sources (thésaurus, listes d'autorité). Même s'il n'est ni complet ni consistant, il contient plus de 1 million de concepts reliés à plus de 5,5 millions de termes dans 17 langues. UMLS ne constitue cependant pas une ontologie au sens strict du terme, car aucune description formelle des concepts n'est fournie. UMLS définit plutôt des groupes de termes, chaque groupe, identifié à un concept, étant constitué d'un ou plusieurs termes et de leurs variantes. La procédure d'extraction des concepts que nous avons suivie exploite ces ressources. La détection de concepts dans un document à partir d'un thésaurus est une procédure relativement bien établie. Elle consiste en quatre grandes étapes :

1) Analyse morpho-syntaxique (*POS tagging*) du document avec lemmatisation des formes fléchies ;

2) Filtrage des mots vides sur la base de leur catégorie grammaticale ;

3) Repérage dans le document des mots ou groupes de mots apparaissant dans le méta-thésaurus ;

4) Filtrage éventuel des concepts ainsi identifiés.

Pour la première étape, différents outils peuvent être utilisés suivant les langues considérées. Une fois les documents analysés, les deuxième et troisième étapes sont mises en œuvre directement, d'une part par un filtrage des mots grammaticaux (prépositions, déterminants, pronoms, conjonctions), d'autre part par un *look-up* des séquences de mots pleins dans UMLS. Cette dernière étape permet de retrouver toutes les variantes,

4. Unified Medical Language System -umlsinfo.nlm.nih.gov
5. www.clef-campaign.org
6. trec.nist.gov

attestées dans UMLS, d'un concept donné. On peut toutefois essayer de l'améliorer par des techniques permettant de regrouper les variantes terminologiques (cf. (Jacquemin, 1999)). Il est à noter ici que nous n'avons pas utilisé la totalité de UMLS pour la troisième étape : les thésaurus NCI et PDQ n'ont pas été pris en compte, car portant sur un domaine différent de celui couvert par la collection[7]. Une telle restriction est également utilisée dans (Huang *et al.*, 2003). La quatrième étape du processus d'indexation conceptuelle vise essentiellement à éliminer un certain nombre d'erreurs générées par les étapes précédentes. Toutefois, les travaux présentés dans (Radhouani *et al.*, 2006) montrent qu'il est préférable de garder un plus grand nombre de concepts pour la recherche d'information. Nous n'avons donc appliqué aucun filtrage ici. Notons enfin que l'outil MetaMap (cf. (Aronson, 2001)), associé à UMLS, permet de réaliser l'ensemble de ces étapes, mais pour l'Anglais seulement. Nous reviendrons sur son utilisation dans la section 4.

4. Validation expérimentale

Nos expériences portent sur la collection CLEF Medical (cf. (Müller *et al.*, 2007)), composée de comptes-rendus médicaux multilingues associés à des images et fournis dans le cadre des campagnes CLEF. Ces comptes-rendus peuvent être rédigés en anglais, en français ou en allemand. Nous nous sommes servis des collections des années 2005, 2006 et 2007. Le corpus utilisé en 2005 et 2006 comporte 50412 documents, et celui utilisé en 2007 (qui contient celui des années précédentes) 55485 documents. Sur ces trois années, 85 requêtes avec jugements de pertinence sont disponibles (chaque année comporte respectivement 25, 30 et 30 requêtes). Les jugements de pertinence sur cette collection sont faits au niveau des images ; nous considérons dans la suite qu'un diagnostic est pertinent s'il correspond à au moins une image pertinente. Cela permet d'évaluer directement notre modèle au niveau textuel.

Pour estimer les paramètres de lissage de notre modèle, nous avons divisé les 85 requêtes en deux sous-groupes : 43 pour la première partie et 42 pour la deuxième. Cela nous permet d'effectuer une validation croisée entre les deux parties. Nous évaluons deux mesures : la précision moyenne *MAP* et la précision à 5 documents*P@5*.

En correspondance avec les différentes méthodes d'extraction de concepts proposées dans la section **??**, nous utilisons les stratégies d'indexation suivantes, (a) pour les documents :

– (MM) analyse avec (mm) pour l'anglais, et avec (tt) pour le français et l'allemand,

– (MP) analyse avec (mp) pour l'anglais, et avec (tt) pour le français et l'allemand,

– (TT) analyse avec (tt) pour l'anglais, le français et l'allemand,

7. Ce filtrage se justifie ici par le fait que ces thésaurus portent sur des points précis de cancérologie alors que la collection considérée est plus générale, et concerne l'ensemble des pathologies.

Tableau 1. *Résultats en précision moyenne (MAP) et en précision à 5 documents (P@5) avec différentes méthodes de détection des concepts. Les paramètres du modèle sont appris sur la partie d'apprentissage et sont testés sur la partie d'évaluation. La méthode considérée comme référence est mise en italique*

			apprentissage	évaluation		apprentissage	évaluation
Documents	Query	λ_u	part 1	part 2	λ_u	part 2	part 1
MAP							
MM	*Q.MM*	*0.1*	*0.260*	*0.246*	*0.4*	*0.251*	*0.259*
MP	Q.MP	0.3	0.285	0.246	0.4	0.246	**0.284**
TT	Q.TT	0.1	0.264	**0.258**	0.2	0.258	0.263
P@5							
MM	*Q.MM*	*0.8*	*0.428*	*0.357*	*0.4*	*0.433*	*0.419*
MP	Q.MP	0.2	0.493	0.424	0.1	0.433	**0.488**
TT	Q.TT	0.1	0.451	**0.462**	0.1	0.462	0.451

– (Mix) analyse avec (mm), (mp) et (tt) pour l'anglais, et avec (tt) pour le français et l'allemand,

et (b) pour les requêtes :

– (Q.MM) un ensemble de concept détecté de l'anglais par (mm)

– (Q.MP) un ensemble de concept détecté de l'anglais par (mp)

– (Q.TT) un ensemble de concept détecté de l'anglais par (tt)

– (Q.Mix) le regroupement des trois ensembles Q.MM, Q.MP, Q.TT

Nous présentons d'abord les résultats des différentes méthodes de détection utilisées séparément (tableau 1). Comme MetaMap est souvent utilisé pour indexer les documents dans le domaine médical, nous le considérons ici comme notre référence (ligne *MM* dans le tableau 1). Nous remarquons que cette approche fournit les moins bons résultats, que ce soit pour la MAP ou la P@5.

Nous présentons ensuite les résultats obtenus par les méthodes de fusion présentées dans ce papier (tableau 2). Le meilleur résultat en précision moyenne est obtenu avec la somme sur les modèles de document associée à un produit (*Fus4*) sur les requêtes E_mix pour la partie 2, mais avec le produit sur les modèles de document et sur les requêtes (*Fus1*) sur la partie 1. Cela étant dit, sur les concepts, tous les résultats qui utilisent nos méthodes de fusion sur les requêtes et sur les documents, donnent des résultats proches et significativement meilleurs que les résultats de référence.

Les différents résultats obtenus montrent que combiner différentes analyses améliore les résultats. Combiner seulement les modèles de document fournit des améliorations significatives, mais nécessite de trouver la bonne analyse de la requête à utiliser. Combiner les documents et les requêtes fournit des résultats équivalents voire supérieurs et ne nécessite pas de sélectionner la bonne analyse.

Tableau 2. *Résultats obtenus par la combinaison des méthodes de détection des concepts sur les requêtes et sur les documents. Les meilleurs résultats sont présentés en gras. Une * indique que la différence avec les résultats de référence est significative (Wilcoxon test, p=0.05).*

		part2				
Documents	Query	*fus1*	*fus2*	*fus3*	*fus4*	*fus5*
MAP						
E_mix	E.MM	0.272*	0.261	0.260	0.261	0.260
E_mix	E.MP	0.265	0.257	0.254	0.257	0.254
E_mix	E.TT	0.283	0.280	0.277	0.280	0.277
E_mix	E_mix	0.285*	0.292*	0.289*	**0.301***	0.299*
MM MP	MM MP	0.285	0.275	0.277	0.287	0.289
P@5						
E_mix	E.MM	0.448*	0.414	0.409	0.414	0.409
E_mix	E.MP	0.452	0.443	0.433	0.443	0.433
E_mix	E.TT	0.476*	0.462	0.462	0.462	0.462
E_mix	E_mix	0.500*	0.500*	0.495*	0.524*	**0.529***
MM MP	MM MP	0.500*	0.462	0.462	0.481	0.481
		part1				
Documents	Query	*fus1*	*fus2*	*fus3*	*fus4*	*fus5*
MAP						
E_mix	E.MM	0.283	0.276	0.259	0.276	0.259
E_mix	E.MP	0.298	0.275	0.273	0.275	0.273
E_mix	E.TT	**0.310***	0.294	0.293	0.294	0.293
E_mix	E_mix	0.299*	0.301	0.289	0.299	0.300
MM MP	MM MP	0.295	0.298	0.299	0.313	0.312
P@5						
E_mix	E.MM	0.460	0.442	0.442	0.442	0.442
E_mix	E.MP	0.479	0.446	0.446	0.446	0.446
E_mix	E.TT	0.474	0.484	0.484	0.484	0.484
E_mix	E_mix	0.437	0.493	0.516	**0.521**	0.497
MM MP	MM MP	0.475	0.456	0.460	0.488	0.488

5. Conclusion

Dans l'indexation conceptuelle, le principal problème est de gérer les erreurs de détection. Nous avons étudié dans ce papier comment résoudre ce problème en combinant différentes méthodes de détection des concepts. Nous avons basé notre proposition sur l'approche modèle de langue de la recherche d'information. Cette approche est flexible et permet de proposer facilement différentes stratégies de fusion. Nous testons ces fusions sur le domaine médical, où nous utilisons différentes méthodes de détection des concepts. Nos résultats obtenus sur une collection standard de recherche d'information montrent que combiner les analyses des documents et des requêtes permet d'améliorer significativement les résultats.

6. Bibliographie

Aronson A., « Effective Mapping of Biomedical Text to the UMLS Metathesaurus : The MetaMap Program », *Proc AMIA 2001*, p. 17-21, 2001.

Chevallet J.-P., Lim J. H., Le T. H. D., « Domain Knowledge Conceptual Inter-Media Indexing, Application to Multilingual Multimedia Medical Reports », *ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007), Lisboa, Portugal*, November 6–9, 2007.

Huang Y., Lowe H., Hersh W., « A pilot study of contextual UMLS indexing to improve the precision of concept-based representation in XML-structured clinical radiology reports. », *Conference of the American Medical Informatics Association*, p. 580 - 587, 2003.

Jacquemin C., « Syntagmatic and paradigmatic representations of term variation », *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 341-348, 1999.

Lacoste C., Chevallet J.-P., Lim J.-H., Wei X., Raccoceanu D., Le T.-H.-D., Teodorescu R., Vuillenemot N., « IPAL Knowledge-based Medical Image Retrieval in ImageCLEFmed 2006 », *Working Notes for the CLEF 2006 Workshop, 20-22 September, Alicante, Spain*, 2006.

Lafferty J., Zhai C., « Document language models, query models, and risk minimization for information retrieval », *SIGIR '01 : Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, p. 111-119, 2001.

Maisonnasse L., Gaussier E., Chevallet J., « Multiplying Concept Sources for Graph Modeling », *In C. Peters, V. Jijkoun, T. Mandl, H. Muller, D.W. Oard, A. Peñas, V. Petras, D. Santos, (Eds.) : Advances in Multilingual and Multimodal Information Retrieval. LNCS #5152. Springer-Verlag.*, p. 585-592, 2008.

Müller H., Deselaers T., Kim E., Kalpathy-Cramer J., Deserno T. M., Clough P., Hersh W., « Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks », *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September, 2007.

Ponte J. M., Croft W. B., « A language modeling approach to information retrieval », *In SIGIR '98 : Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, p. 275-281, 1998.

Radhouani S., Maisonnasse L., Lim J.-H., Le T.-H.-D., Chevallet J.-P., « Une Indexation Conceptuelle pour un Filtrage par Dimensions, Experimentation sur la base medicale ImageCLEFmed avec le meta thesaurus UMLS », *COnference en Recherche Information et Applications CORIA'2006*, p. 257-271, mars, 2006.

Schank R. C., Kolodner J. L., DeJong G., « Conceptual information retrieval », *Proceedings of the 3rd annual conference ACM SIGIR*, Kent, UK, p. 94-116, 1980.

Srikanth M., Srihari R., « Exploiting syntactic structure of queries in a language modeling approach to IR », *CIKM '03 : Proceedings of the twelfth international conference on Information and knowledge management*, ACM, New York, NY, USA, p. 476-483, 2003.

Vintar S., Buitelaar P., Volk M., « Relations in Concept-Based Cross-Language Medical Information Retrieval », *Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM)*, 2003.

Zhou W., Yu C., Smalheiser N., Torvik V., Hong J., « Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature », *SIGIR '07 : Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, p. 655-662, 2007.