Modélisation de relations dans l'approche modèle de langue en recherche d'information

L. Maisonnasse[1], **E. Gaussier**[1], **J.-P. Chevallet**[2]

[1] *Université de Grenoble - Laboratoire d'Informatique de Grenoble - 38041 Grenoble Cedex 9, France*
loic.maisonnasse@imag.fr, eric.gaussier@imag.fr
[2] *IPAL-I2R, Singapore 119613*
viscjp@i2r.a-star.edu.sg

RÉSUMÉ. *Nous abordons dans cet article le problème de la prise en compte de relations (par exemple de nature syntaxique ou sémantique) dans un modèle de langues en recherche d'information. En particulier, nous proposons, sur la base du modèle de langue, un cadre complet pour la prise en compte de relations, étiquetées ou non. Afin d'illustrer ce cadre, nous avons conduit une série d'expériences fondées sur différentes indexations structurées (grammaire de dépendances et graphes de relations entre concepts) dans le domaine médical. Nos résultats montrent que l'intégration d'information sur les relations entre termes améliore la qualité d'un système de recherche d'information sur la précision à 5 documents. Ils confirment aussi le bien-fondé du modèle que nous proposons.*

ABSTRACT. *We address in this paper the problem of modelling relations (such as syntactic or semantic relations) in the language modelling approach to information retrieval. In particular, we propose, in this framework, a new model that allows one to make use of relations (labelled or not) when matching documents and queries. In order to illustrate this model, we have conducted a series of experiments in the medical domain. Our results show that the use of relations helps improve the precision at 5 documents. They also confirm the good behavior of our model.*

MOTS-CLÉS : *Modèle de langue, Indexation par concepts et relations, Recherche d'information dans le domaine médical*

KEYWORDS: *Language model, Concept and relation indexing, Medical retrieval*

1. Introduction

Nous abordons dans cet article le problème de la prise en compte de relations, étiquetées ou non, entre termes ou concepts dans l'approche modèle de langue en recherche d'information. Dans certains domaines, comme le domaine médical que nous explorons ici, plusieurs ressources sont disponibles pour, d'une part, indexer documents et requêtes par des concepts, et, d'autre part, tenir compte de relations (de nature syntaxique ou sémantique par exemple) entre ces concepts. Nous élaborons dans cet article un modèle relationnel complet, sur la base du modèle de langue, qui permet de tenir compte de tous ces éléments en recherche d'information. Ce modèle fait l'objet de la section 2. Nous présentons ensuite, en section 3, l'application de ce modèle au domaine médical, en introduisant une nouvelle procédure de désambiguïsation des étiquettes de relation. Nous détaillons ensuite en section 4 les expériences que nous avons conduites avec ce modèle dans le domaine médical. Enfin, la section 5 tire les leçons de nos expériences.

2. Modèle de langue relationnel

L'approche modèle de langue en recherche d'information a été introduite pour la première fois dans (Ponte *et al.,* 1998). De nombreux travaux se sont depuis inscrits dans cette mouvance, et ont permis d'obtenir un cadre simple et facilement adaptable (cf. par exemple (Lafferty *et al.,* 2001)) qui fournit de bons résultats. Au-delà du modèle unigramme utilisé traditionnellement dans le modèle de langue, plusieurs travaux se sont intéressés à la prise en compte de dépendances ou de termes complexes. Ainsi, (Srikanth *et al.,* 2002) et (Song *et al.,* 1999) proposent de combiner des modèles unigrammes et bigrammes (dans l'esprit des modèles de langue utilisés par exemple en reconnaissance de la parole), les relations prises en compte ici étant celle de contiguïté. Dans une approche un peu plus générale, (Lee *et al.,* 2006) et (Gao *et al.,* 2004) ont tenté de prendre en compte, dans le modèle de langue, des dépendances syntaxiques entre termes. Le travail réalisé dans (Gao *et al.,* 2004) est à notre avis le plus complet d'un point de vue formel. Il comporte toutefois des problèmes et des limitations que nous voulons exposer.

2.1. *Le modèle DLM*

(Gao *et al.,* 2004) introduit un modèle de dépendances (que nous appelons *DLM* pour *Dependency Language Model* dans la suite) qui, sur la base de l'approche modèle de langue (Ponte *et al.,* 1998, Lafferty *et al.,* 2001), intègre des dépendances syntaxiques dans le calcul du score de pertinence d'un document. Le modèle *DLM* repose sur une variable L qui décrit un ensemble de relations entre termes de la re-

quête. Nous appellerons un tel ensemble de relations un *linkage*. En désignant par Q une requête et M_d le modèle d'un document, on a :

$$P\left(Q|M_d\right) = \sum_L P\left(L|M_d\right) P\left(Q|L, M_d\right) \quad [1]$$

où la somme est prise sur tous les *linkage*s possibles. Toutefois, pour des raisons d'efficacité, cette somme est approchée dans *DLM* par le *linkage* le plus probable, conduisant à :

$$P\left(Q|M_d\right) = P\left(L|M_d\right) P\left(Q|L, M_d\right)$$

avec :

$$L = \text{argmax}_L P\left(L|Q\right) \quad [2]$$

Dans le cas des analyseurs de dépendances tels que ceux considérés dans (Gao *et al.*, 2004), chaque terme a exactement un gouverneur dans chaque *linkage* possible. La quantité ci-dessus se simplifie alors sous la forme (voir (Gao *et al.*, 2004) pour les détails) :

$$\begin{aligned}\log P\left(Q|M_d\right) = \quad & \log P\left(L|M_d\right) + \textstyle\sum_{i=1..n} \log P\left(q_i|M_d\right) \\ & + \textstyle\sum_{(i,j)\in L} MI\left(q_i, q_j|L, M_d\right)\end{aligned}$$

où MI dénote l'information mutuelle :

$$MI\left(q_i, q_i|L, M_d\right) = \log \frac{P(q_i, q_j|L, M_d)}{P(q_i|L, M_d)P(q_j|L, M_d)}$$

et :

$$P(L|M_d) \propto \prod_{(i,j)\in L} \hat{P}(R|q_i, q_j) \quad [3]$$

L est toujours donné par l'équation 2 et $\hat{P}(R|q_i, q_j)$, dans l'équation ci-dessus, représente une estimation empirique de la probabilité que les termes q_i et q_j soient en relation de dépendance dans le document d.

Il y a toutefois dans *DLM* une certaine ambiguïté sur ce que représente un *linkage*. En particulier, l'équation 3 suggère d'interpréter un *linkage* comme un ensemble de liens définis sur un ensemble de termes connus (d'après le conditionnement sur q_i et q_j). En revanche, une telle interprétation n'est pas compatible avec l'équation 1 car elle conduit à ne pas prendre en compte le terme $P\left(Q|L, M_d\right)$ (qui vaut alors 1 car tous les termes de la requête sont connus dans L). L'ambiguïté dans la définition de L peut ne pas être importante en pratique. Elle est néanmoins gênante d'un point de vue théorique. Enfin, le modèle *DLM* ne s'applique, en théorie, qu'à des dépendances syntaxiques (de type grammaires de dépendance) non étiquetées.

Nous présentons maintenant un modèle relationnel qui revient sur ces problèmes. Ce modèle est le modèle relationnel le plus complet à notre connaissance proposé sur la base du modèle de langue en recherche d'information.

2.2. *Un modèle relationnel complet*

De façon générale, nous supposons que, dans le cadre d'une indexation relationnelle, requêtes et documents sont représentés sous forme d'un graphe $G = (C, E)$ où C représente l'ensemble des noeuds (concepts ou termes par exemple) du graphe et E l'ensemble des relations, que nous supposons étiquetées par une ou plusieurs étiquettes. Dans la suite, nous utiliserons le terme "concept" pour désigner un élément de C. G_q et G_d désigneront un graphe associé à une requête ou un document. Soit $\mathcal{L}$ l'ensemble des étiquettes possibles pour une relation, $\mathcal{P}(\mathcal{L})$ désignera l'ensemble des parties de $\mathcal{L}$. Avec nos hypothèses, E définit deux applications : une application, que nous noterons L, de $C \times C$ dans $\{0, 1\}$ qui permet de rendre compte s'il existe ou non une relation entre deux concepts, et une application, que nous noterons R, de $C \times C$ dans $\mathcal{P}(\mathcal{L})$ qui permet d'associer à chaque relation un ensemble d'étiquettes.

La probabilité que le graphe de la requête q soit généré par le modèle du document d peut alors se décomposer sous la forme :

$$P\left(G_q|M_d\right) = P\left(C|M_d\right) P\left(E|C, M_d\right) \quad [4]$$

En faisant l'hypothèse, d'une part que les concepts sont indépendants les uns des autres conditionnellement au modèle du document, et d'autre part que les relations sont indépendantes les unes des autres conditionnellement aux concepts et au modèle du document (hypothèse d'indépendance standard en recherche d'information), nous obtenons :

$$P\left(C|M_d\right) = \prod_{c_i \in C} P\left(c_i|M_d\right) \quad [5]$$

$$P\left(E|C, M_d\right) = \prod_{(i,j), i \leq j} P(L(c_i, c_j)|\ C, M_d)) \times P(R(c_i, c_j) = \mathcal{L}_{ij}|L(c_i, c_j), M_d) \quad [6]$$

où $\mathcal{L}_{ij}$ est l'ensemble vide si c_i et c_j ne sont pas reliés dans G_q, et l'ensemble des étiquettes de leur relation sinon. Nous posons :

$$P(R(c_i, c_j) = \{\emptyset\}|L(c_i, c_j) = 0, M_d) = 1$$

Dans notre prise en compte de la structure du graphe (équation 6), $P(L(c_i, c_j)|\ C, M_d))$ désigne la probabilité que les concepts c_i et c_j de la requête soient reliés (conditionnellement à l'ensemble des concepts de la requête et au modèle du document), alors que $P(R(c_i, c_j) = \mathcal{L}_{ij}|L(c_i, c_j), M_d)$ désigne la probabilité que cette relation soit étiquetée par l'ensemble $\mathcal{L}_{ij}$. On voit ici que les variables aléatoires $L(c_i, c_j)$ et $\mathcal{L}_{ij}$ ont une interprétation claire et prennent des valeurs différentes suivant les concepts considérés. Les problèmes posés par le modèle *DLM* sont ici évités.

L'équation 5 correspond au modèle de langue unigramme standard. En notant λ_u le paramètre de lissage, la quantité $P(c_i|M_d)$ est traditionnellement estimée (en se reposant sur un lissage de Jelinek-Mercer) par :

$$P(c_i|M_d) = (1-\lambda_u)\frac{D(c_i)}{D(*)} + \lambda_u \frac{C(c_i)}{C(*)} \quad [7]$$

où $D(c_i)$ (respectivement $C(c_i)$) est le nombre de fois que c_i apparaît dans le document d (respectivement dans la collection), et $D(*)$ (respectivement $C(*)$) le nombre de concepts dans d (respectivement la collection).

La quantité $P(L(c_i,c_j)|\ C, M_d))$ est estimée de manière analogue par :

$$\begin{aligned} P(L(c_i,c_j)=x|C,M_d) &= (1-\lambda_r)\frac{xD(c_i,c_j,\mathcal{R})+(1-x)D(c_i,c_j,\neg\mathcal{R})}{D(c_i,c_j,\mathcal{R})+D(c_i,c_j,\neg\mathcal{R})} \\ &\quad +\lambda_r \frac{xC(c_i,c_j,\mathcal{R})+(1-x)C(c_i,c_j,\neg\mathcal{R})}{C(c_i,c_j,\mathcal{R})+C(c_i,c_j,\neg\mathcal{R})} \quad [8] \end{aligned}$$

où $D(c_i,c_j,\mathcal{R})$ (respectivement $\mathcal{C}(c_i,c_j,\mathcal{R})$) représente le nombre de fois que c_i et c_j sont reliés dans le documents d (respectivement la collection). De façon similaire, $D(c_i,c_j,\neg\mathcal{R})$ (respectivement $\mathcal{C}(c_i,c_j,\neg\mathcal{R})$) représente le nombre de fois que c_i et c_j sont dans le document (respectivement la collection) sans être reliés. x prend les valeurs 0 ou 1 suivant que la relation est observée ou non.

La quantité $P(R(c_i,c_j)=\mathcal{L}_{ij}|L(c_i,c_j)=1,M_d)$ de l'équation 6 n'intervient pas dans le calcul final lorsque $L(c_i,c_j)=0$ (car dans ce cas $R(c_i,c_j)$ est l'ensemble vide et sa probabilité vaut 1 par définition). Lorsque $L(c_i,c_j)=1$, l'ensemble d'étiquettes $\mathcal{L}_{ij}$ est pris en compte. La considération de cet ensemble d'étiquettes traduit ici le fait que, en général, plusieurs étiquettes co-existent pour qualifier une relation. Pour rendre compte de toutes ces étiquettes, nous nous reposons donc sur la décomposition suivante de $P(R(c_i,c_j)=\mathcal{L}_{ij}|L(c_i,c_j)=1,M_d)$:

$$P(R(c_i,c_j)=\mathcal{L}_{ij}|L(c_i,c_j)=1,M_d) = \prod_{l\in\mathcal{L}_{ij}} P(r(c_i,c_j)=l|L(c_i,c_j),M_d)$$

où r est une application $C\times C$ dans $\mathcal{L}$ qui définit une relation étiquetée élémentaire. La quantité $P(r(c_i,c_j)=l|L(c_i,c_j)=1,M_d)$, pour toute étiquette l de $\mathcal{L}$, peut alors être estimée de manière analogue aux autres quantités qui entrent en jeu dans le modèle :

$$P(r(c_i,c_j)=l|L(c_i,c_j)=1,M_d) = (1-\lambda_e)\frac{D(c_i,c_j,l)}{D(c_i,c_j)} + \lambda_e\frac{C(c_i,c_j,l)}{C(c_i,c_j)}$$

où $D(c_i,c_j,l)$ (respectivement $C(c_i,c_j,l)$) représente le nombre de fois que c_i et c_j sont reliés, avec l'étiquette l, dans le document d (respectivement dans la collection), et $D(c_i,c_j)$ (respectivement $C(c_i,c_j)$) le nombre de fois où ils sont reliés dans le document (respectivement la requête).

2.3. *Cas des relations non étiquetées*

Dans le cas où les relations considérées ne portent pas d'étiquettes (c'est par exemple le cas des relations syntaxiques considérées dans (Gao *et al.,* 2004)), l'ensemble $\mathcal{L}$ est réduit à un seul élément l, et nous posons :

$$P(r(c_i, c_j) = l | L(c_i, c_j) = 1, M_d) = 1$$

Nous avons toujours, comme précédemment :

$$P(R(c_i, c_j) = \{\emptyset\} | L(c_i, c_j) = 0, M_d) = 1$$

Le modèle obtenu est alors équivalent à un modèle sans étiquette, dont la forme générale est donnée par l'équation 4 avec :

$$\begin{aligned} P\left(C|M_d\right) &= \prod_{c_i \in C} P\left(c_i|M_d\right) \\ P\left(E|C, M_d\right) &= \prod_{(i,j), i \leq j} P(L(c_i, c_j)|\ C, M_d)) \end{aligned}$$

L'estimation des paramètres de ce modèle passe directement par les équations 7 et 8, ce qui permet de retrouver le modèle proposé dans (Maisonnasse *et al.,* 2007). A la différence du modèle *DLM,* ce dernier modèle est bien fondé théoriquement et peut être appliqué à toute représentation graphique (avec des relations non étiquetées). De plus, il ne repose que sur deux termes, relativement faciles à estimer, là où le modèle *DLM* en comporte trois, avec une estimation plus complexe pour l'un d'entre eux. Enfin, tout comme le modèle *DLM,* il généralise le modèle bigramme introduit dans (Srikanth *et al.,* 2002). Nous présentons en section 4 une comparaison expérimentale entre ce modèle et le modèle de dépendance, sur la base d'une analyse syntaxique des phrases des documents et requêtes.

3. Application au domaine médical

Le domaine médical se prête bien à l'indexation par graphes 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[1]) a été étudiée dans plusieurs articles, comme par exemple (Lacoste *et al.,* 2006) dans le cadre des campagnes ImageCLEFmed de CLEF[2] ou (Zhou *et al.,* 2007) dans la piste génomique de TREC[3] (dans ce

1. Unified Medical Language System - umlsinfo.nlm.nih.gov
2. www.clef-campaign.org
3. trec.nist.gov

dernier cas, ce sont toutes les variantes terminologiques d'un concept qui sont utilisés 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 de UMLS. Une relation entre deux concepts (d'un document ou d'une requête) est postulée dès lors que les deux concepts apparaissent dans la même phrase et sont reliés dans le thésaurus. Des restrictions sur les concepts mis en jeu, ainsi que des verbes "marqueurs" de relations, sont également utilisés. Nous nous inscrivons dans la lignée de ces travaux, et détaillons maintenant notre processus d'indexation.

3.1. *Extraction de concepts et relations 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 d'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. En plus de ces concepts, UMLS propose un réseau sémantique, qui constitue une catégorisation de tous les concepts à un niveau relativement général et définit des relations entre ces catégories, chacune de ces relation étant spécifiée par une ou plusieurs étiquettes. La procédure d'extraction de concepts et relations que nous avons suivie exploite ces différentes ressources, tout d'abord pour la détection des concepts, puis pour la détection des relations entre les concepts retenus.

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 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 denière étape permet de retrouver toutes les variantes, 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

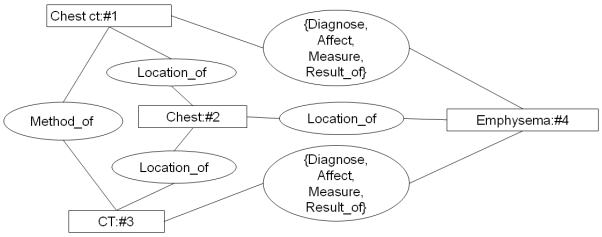

Figure 1. *Graphe extrait pour la phrase "Show me chest CT images with emphysema"*

portant sur un domaine différent de celui couvert par la collection[4]. 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.

L'étape de détection des concepts est suivie d'une étape de détection des relations entre concepts. Les relations utilisées sont celles du réseau sémantique. Nous émettons l'hypothèse qu'une relation sémantique existe entre deux concepts si ces concepts sont détectés dans la même phrase et si le réseau sémantique définit une relation entre ces deux concepts. Pour détecter l'existence de ces relations, nous effectuons une étape où nous attribuons à chaque concept ses catégories sémantiques, et ajoutons ensuite les relations sémantiques qui relient les catégories des concepts. La figure 1 illustre le résultat de ces étapes, où chaque relation porte l'ensemble des étiquettes figurant dans chaque relation trouvée dans le réseau sémantique. Cet ensemble d'étiquettes est parfois important et la question se pose de savoir s'il ne serait pas intéressant de le filtrer. Nous présentons ci-dessous une procédure de désambiguïsation qui réalise un tel filtrage.

3.2. *Désambiguïsation des étiquettes*

Notre procédure de désambiguïsation vise à sélectionner une ou plusieurs étiquettes par relation suivant le contexte graphique de chaque étiquette. Par contexte graphique nous entendons ici les autres étiquettes (et donc les autres relations) qui mettent en jeu les concepts reliés par une étiquette donnée. Sur chaque

4. 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.

graphe d'indexation, nous extrayons ainsi les n-uplets ($n = 1, 2, 3$) d'étiquettes reliées (dans le cas où $n = 1$, l'étiquette courante est directement extraite). Sur l'exemple de la figure 1, les triplets *(Method_of,Diagnose,Location_of)*, *(Method_of,Affect,Location_of)*, entre autres, sont ainsi extraits. Nous comptons de plus, pour chaque n-uplet, le nombre de fois où il apparaît dans la collection.

Une fois cette étape réalisée, nous désambiguïsons ensuite l'ensemble $\mathcal{N}$ des n-uplets ($n = 1, 2, 3$) de chaque graphe d'indexation de la façon suivante :

1) Tant que l'ensemble des triplets du graphe n'est pas vide, sélectionner le triplet le plus fréquent, et supprimer tous les n-uplets de $\mathcal{N}$ qui mettent en jeu une des relations étiquetées du triplet sélectionné.

Dans l'exemple de la figure 1, la sélection du triplet *(Method_of,Diagnose,Location_of)* conduirait ainsi à ne retenir, pour la relation entre *Chest ct* et *Emphysema* que l'étiquette *Diagnose* ;

2) Répéter l'étape précédente sur les paires ;

3) Traiter enfin les 1-uplets de la même façon.

L'ensemble des procédures ci-dessus permet de réaliser des indexations par graphes des documents et requêtes dans le domaine médical. Nous voulons maintenant montrer comment se comporte le modèle que nous avons introduit en section 2 avec ces différents types d'indexation.

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). Voici un extrait de document anglais issu de la collection 2005 :

```
<MetadataID>6068</MetadataID>
<GlobalID>1112</GlobalID>
<Title>RESPIRATORY</Title>
<Description>RESPIRATORY: Lung: Arteriosclerosis Grade 3: Micro
low mag H&E same as in slide  grade 3 lesion with dilated appearing
arteriole distal to the small artery lesion
</Description>
```

Les expériences que nous avons menées ont trois buts :

1) Valider le comportement du modèle avec relations non étiquetées (présenté en 2.3), modèle que nous appellerons *MRSE*[5], en le comparant au modèle *DLM*, qui a les mêmes visées ;

2) Evaluer l'utilité d'une désambiguïsation des étiquettes dans le schéma d'indexation retenu ;

3) Evaluer l'impact de la prise en compte de relations en recherche d'information, dans le domaine médical.

Pour chaque expérience, les meilleurs résultats obtenus sont notés en gras dans les tableaux.

4.1. *Comparaison entre MRSE et DLM*

Pour cette comparaison, nous nous sommes concentrés sur la partie anglaise des années 2005 et 2006, et avons considéré deux types de relations non étiquetées : les dépendances syntaxiques produites par MiniPar (cf. (Lin, 1998)), et les relations sémantiques produites selon la stratégie détaillée en section 3.1. Dans ce dernier cas, seuls les parties du discours et les lemmes produits par MiniPar sont pris en compte.

De façon à estimer les paramètres (coefficients de lissage) des différents modèles, nous avons divisé l'ensemble des 55 requêtes en deux sous-ensembles, et avons retenu 25 requêtes (sélectionnées aléatoirement) pour l'apprentissage et 30 pour le test. Nous avons de plus retenu deux mesures d'évaluation : la MAP (*mean average precision*) et la précision à 5 documents. Le choix de cette dernière mesure se justifie ici par le fait que l'on peut s'attendre à ce que les relations, permettant une représentation plus fine des documents, améliorent essentiellement la précision des résultats en tête de liste. Enfin, nous avons utilisé le modèle *DLM* (en appliquant directement la formule finale) sur le graphe des relations sémantiques. Notons néanmoins que cette application n'est pas valide car la forme de ce graphe (certains concepts ont plusieurs "gouverneurs") viole les hypothèses nécessaires pour dériver le modèle *DLM*.

Les résultats obtenus sont présentés dans le tableau 1. Comme on peut le voir, les deux modèles ont des performances tout à fait similaires, avec un léger avantage pour le modèle *MRSE* pour la précision à 5 documents. Toutefois, l'utilisation d'un test de Wilcoxon signé à 0.05 ne montre aucune différence significative entre les deux modèles. Le modèle *MRSE* se comporte donc, sur cette collection, de façon similaire au modèle *DLM*, tout en étant plus simple, mieux fondé d'un point de vue théorique, et plus général car applicable à tout type de graphes.

5. Abréviation de "modèle relationnel sans étiquette".

	Dépendances syntaxiques				Relations sémantiques			
	λ_u	λ_r	training	test	λ_u	λ_r	training	test
MAP								
DLM	0.8	0.9	0.19	**0.21**	0.1	0.9	0.26	0.33
MRSE	0.4	0.9	0.22	**0.21**	0.1	0.1	0.24	**0.34**
P@5								
DLM	0.2	0.9	0.29	0.29	0.1	0.9	0.45	0.44
MRSE	0.2	0.9	0.34	**0.30**	0.1	0.1	0.43	**0.48**

Tableau 1. *Précision moyenne et précision à 5 documents pour les modèles MRSE et DLM. Les modèles sont entraînés sur 25 requêtes et évalués sur 30 autres, toutes tirées des collections 2005 et 2006 d'ImageCLEFmed.* λ_u *et* λ_r *correspondent aux paramètres de lissage pour les concepts (ou termes) et les relations*

	λ_u	λ_r	λ_e	training	test
MAP					
1 étiq.	0.1	1	0.9	0.26	**0.34**
2 étiq.	0.1	1	0.5	0.26	**0.34**
unigramme	0.1	NA	NA	0.25	**0.34**
P@5					
1 étiq.	0.1	1	0.1	0.46	**0.49**
2 étiq.	0.1	1	0.5	0.44	0.47
unigramme	0.1	NA	NA	0.41	0.46

Tableau 2. *Précision moyenne et précision à 5 documents avec désambiguïsation des étiquettes*

4.2. *Impact de la désambiguïsation des étiquettes*

Pour tester l'impact de la désambiguïsation des étiquettes, nous avons repris les relations sémantiques ci-dessus, et utilisé la procédure de désambiguïsation décrite en 3.2, en retenant soit la meilleure étiquette, soit les deux meilleures étiquettes. Nous avons ensuite utilisé le modèle relationnel complet. Les résultats obtenus sont décrits dans le tableau 2. Comme on peut le voir, la procédure de désambiguïsation améliore les résultats, mais de façon non significative. Comme les meilleurs résultats (surtout sur la précision à 5 documents) sont obtenus avec la meilleure étiquette (les résultats décroissent jusqu'à un niveau proche de celui des unigrammes sur ce jeu de test lorsque l'on augmente le nombre d'étiquettes prises en compte), nous pensons que notre procédure de désambiguïsation arrive à sélectionner une bonne étiquette dans la majorité des cas. Toutefois, la différence obtenue ne justifie pas vraiment l'utilisation de cette procédure, et nous ne l'avons pas retenue pour la suite de nos expériences.

4.3. *Impact des concepts et relations*

Pour cette dernière série d'expériences, nous avons utilisé les collections des années 2005, 2006 et 2007. Comme nous l'avons déjà mentionné, plusieurs outils peuvent être utilisés pour analyser requêtes et documents. Nous avons retenu l'outil MetaMap pour détecter les concepts dans les documents anglais, et la stratégie détaillée en section 3.1 avec l'outil TreeTagger[6] pour les documents français et allemands. Une fois les concepts identifiés, nous avons utilisé la stratégie d'extraction de relations également présentée dans la section 3.1.

Dans la mesure où chaque requête est donnée dans les trois langues de la collection, nous avons construit plusieurs graphes pour chaque requête. Pour les versions françaises et allemandes, nous avons également utilisé TreeTagger. Pour la version anglaise, nous avons produit un ensemble de graphes à partir de différents outils : MetaMap, MiniPar et TreeTagger. La prise en compte de l'ensemble de ces graphes s'explique par le fait que différents outils pourront produire des analyses sensiblement différentes, dont la prise en compte devrait permettre d'enrichir au mieux la requête et donc d'améliorer le rappel du système de recherche d'information. Dans la suite, nous désignerons par *E* le graphe de la version anglaise d'une requête produit par MetaMap, et par *E_Mix* les graphes produits par MetaMap, Minipar et TreeTagger. *F* et *G* désigneront les graphes des versions françaises et allemandes. La notation *EFG* désignera donc une requête indexée par 3 graphes, un graphe anglais obtenu à l'aide de MetaMap, et des graphes français et allemand obtenus à l'aide de TreeTagger. De même, *E_MixFG* désigne une requête indexée par 5 graphes : 3 graphes anglais correspondant à *E_Mix* et des graphes français et allemand obtenus à l'aide de TreeTagger.

Pour rendre compte de requêtes représentées par un ensemble de graphes, nous utilisons la décomposition suivante :

$$P\left(Q = \{G_q\} \mid M_d\right) = \prod_{G_q} P\left(G_q | M_d\right)$$

où $P\left(G_q | M_d\right)$ est donné par notre modèle relationnel complet. Cette décomposition traduit le fait que nous voulons être capable de générer l'ensemble des graphes de la requête. Si cette volonté se justifie pour les graphes des différentes versions de la requête (anglaise, française et allemande), elle se justifie moins lorsque nous utilisons plusieurs graphes pour la version anglaise. Nous comptons revenir sur ce problème dans la suite de notre travail.

Les résultats que nous avons obtenus sont présentés dans le tableau 3 (les années 2005 et 2006 ont été utilisées pour l'apprentissage des coefficients de lissage, l'année 2007 servant de test pour les modèles). Comme nous pouvons le constater, les meilleurs résultats obtenus en MAP sont fournis par l'indexation *E_MixFG*, ce qui correspond à l'hypothèse, émise plus haut, d'un plus grand enrichissement de la re-

6. www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

	Modèle unigramme			Modèle relationnel				
	λ_u	2005-2006	2007	λ_u	λ_r	λ_e	2005-2006	2007
MAP								
E	0.2	0.23	0.31	0.2	1	0.9	0.23	0.33
E_Mix	0.1	0.24	0.34	0.1	1	0.9	0.24	0.34
EFG	0.1	0.23	0.33	0.1	1	0.9	0.23	0.33
E_MixFG	0.1	0.24	**0.35**	0.1	1	0.9	0.24	**0.35**
P@5								
E	0.2	0.44	0.37	0.2	1	0.9	0.44	0.41
E_Mix	0.1	0.46	0.37	0.1	1	0.8	0.46	0.37
EFG	0.2	0.43	**0.45**	0.1	1	0.8	0.45	**0.49**
E_MixFG	0.1	0.47	0.42	0.1	1	0.8	0.47	0.42

Tableau 3. *Résultats (MAP et précision à 5 documents - P@5) pour différentes stratégies d'indexation*

quête par cette méthode. En revanche, cet enrichissement se fait au détriment du rappel, et la meilleure méthode d'indexation pour la précision à 5 documents repose sur la stratégie *EFG*. Ici encore, on peut constater une amélioration de la précision à 5 documents par la prise en compte des relations. Les résultats obtenus avec la méthode *E_MixFG* sont les meilleurs résultats obtenus lors de la campagne CLEF 2007 pour la piste CLEF Medical qui regroupait 13 groupes pour un total de 147 runs (l'évaluation ne portait que sur la MAP).

Les requêtes de CLEF Medical ne sont en fait pas toutes du même type. Certaines portent sur des termes généraux, et mettent peu de concepts en jeu (c'est en général le cas des requêtes de type *Visual* comme *Radio d'une fracture de la hanche*), d'autres portent sur des concepts, mais souvent un seul (c'est en général le cas des requêtes de type *Textual* comme *Sclérose tubaire*), enfin d'autres mettent en jeu plusieurs concepts reliés (c'est le cas en général des requêtes de type *Mixed*, comme *Endoscopie gastroentestinale avec polype*). Nous montrons dans le tableau 4 comment nos méthodes d'indexation et modèles se comportent sur ces différentes types de requêtes. Comme nous pouvions le prévoir, les meilleurs résultats sont obtenus sur les requêtes de type *Textual*, qui concernent la recherche d'un concept précis. Il n'est pas surprenant de voir que les relations n'apportent rien ici. En revanche, les relations améliorent les résultats pour les requêtes de type *Mixed*, qui elles recherchent des concepts reliés.

5. Conclusion

Nous avons dans cet article introduit un nouveau modèle, le modèle de langue relationnel, pour la prise en compte d'une indexation par relations (étiquetées ou non) dans l'approche modèle de langue en recherche d'information. Nous avons ensuite présenté une méthode d'indexation par concepts et relations dans le domaine médical,

	concepts		relations		modèle unigramme		modèle relationnel	
	détecté	distinct	détecté	distinct	MAP	P@5	MAP	P@5
Visual								
E	19	46	213	210	**0.25**	**0.36**	**0.25**	**0.36**
E_Mix	104	48	327	210	0.22	0.24	0.22	0.24
EFG	62	50	221	218	0.23	0.34	0.23	0.34
E_MixFG	117	51	335	218	0.22	0.24	0.22	0.24
Mixed								
E	43	39	282	276	0.26	0.16	0.30	0.24
E_Mix	98	39	421	276	0.31	0.22	0.31	0.22
EFG	60	43	290	282	0.32	0.38	0.34	**0.48**
E_MixFG	115	43	429	282	**0.35**	0.38	**0.35**	0.38
Textual								
E	55	46	528	512	0.43	0.60	0.44	0.62
E_Mix	116	48	785	513	**0.49**	**0.64**	**0.49**	**0.64**
EFG	61	47	529	513	0.43	0.60	0.44	**0.64**
E_MixFG	122	49	786	514	**0.49**	**0.64**	**0.49**	**0.64**

Tableau 4. *Comportement des concepts et relations suivant les types de requêtes*

fondée sur l'utilisation d'outils et ressources existants (analyseurs mophologiques, méta-thésaurus et réseau sémantique). Au passage, nous avons introduit une méthode simple de désambiguïsation des étiquettes de relations. Enfin, nous avons montré comment se comportait notre modèle à travers un jeu d'expériences visant à mettre en évidence d'une part le bien-fondé de notre modèle, et d'autre part l'intérêt des relations (sur certains types de requêtes et pour certaines mesures d'évaluation). Certaines des expériences que nous avons présentées ont été conduites dans le cadre de la campagne CLEF 2007, pour la piste CLEF Medical. Les résultats que nous présentons dans le tableau 3 sont les meilleurs résultats obtenus au cours de cette campagne (parmi un ensemble de 13 groupes ayant soumis 147 runs).

6. Bibliographie

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

Gao J., Nie J.-Y., Wu G., Cao G., « Dependence language model for information retrieval », *Research and Development in Information Retrieval*, 2004.

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*, 2003.

Jacquemin C., « Syntagmatic and paradigmatic representations of term variation », *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, 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 W., « Document language models, query models, and risk minimization for information retrieval », *Proceedings of the 24th ACM SIGIR Conference, September 2001*, 2001.

Lee C., Lee G. G., Jang M. G., « Dependency Structure Language Model for Information Retrieval. », *ETRI journal*, 2006.

Lin D., « Dependency-based Evaluation of MiniPar », *Workshop on the Evaluation of Parsing Systems, Granada, Spain, May*, ACM, 1998.

Maisonnasse L., Gaussier E., Chevallet J.-P., « Revisiting the Dependence Language Model for Information Retrieval », *Research and Development in Information Retrieval*, 2007.

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 », *Research and Development in Information Retrieval*, 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.

Song F., Croft W. B., « A general language model for information retrieval », *CIKM '99 : Proceedings of the eighth international conference on Information and knowledge management*, ACM Press, New York, NY, USA, p. 316-321, 1999.

Srikanth M., Srihari R., « Biterm language models for document retrieval », *Research and Development in Information Retrieval*, 2002.

Vintar S., Buitelaar P., Volk M., « Relations in Concept-Based Cross-Language Medical Information Retrieval », 2003.

Zhou W., Yu C., Smalheiser N., Torvik V., Hong J., « Knowledge-intensive Conceptual Retrieval and Passage Extraction of Biomedical Literature », *Research and Development in Information Retrieval*, 2007.