Indexation de textes médicaux par extraction de concepts, et ses utilisations

Pouliquen Bruno[1], Delamarre Denis[2], Le Beux Pierre[2]

[1]European Commission, IPSC, Joint Research Centre – 21020 ISPRA – Italie – Bruno.Pouliquen@jrc.it

[2]Laboratoire d'informatique Médicale – Faculté de Médecine – 35033 Rennes cedex – France – {Denis.Delamarre, Pierre.Lebeux}@univ-rennes1.fr

Abstract

The work presented in this paper specifically targets the accessibility to medical information. We used a French medical dictionary (specifically created for the medical domain), and built an index tool to particularly recognize a concept from a medical thesaurus that is present in a sentence written in natural language. First we indexed medical documents with a set of concepts and then demonstrated the utility of such indexing by developing a search engine and various tools which include: keyword identification, document similarity and automatic document synthesis. This indexing greatly aided in reducing the repository complexity of natural language documents. In addition, the evaluation results demonstrated that this indexing retains the main semantic information.

Résumé

Nous nous intéressons à l'accès à l'information médicale. Nous avons utilisé un dictionnaire de flexions, dérivations et synonymes de mots (spécifiquement créé pour le domaine médical), et nous avons créé un outil d'indexation permettant de reconnaître un concept d'un thésaurus médical dans une phrase en langage naturel. Nous avons ainsi pu indexer des documents médicaux par un ensemble de concepts, ensuite nous avons démontré l'utilité d'une telle indexation en développant un système de recherche d'information et divers outils : extraction de mots-clé, similarité de documents et synthèse automatique de documents. Cette indexation diminue considérablement la complexité de la représentation des connaissances contenues dans les documents en langage naturel. Les résultats des évaluations montrent que cette indexation conserve néanmoins la majeure partie de l'information sémantique.

Mots-clés : Traitement automatique des langues naturelles, Indexation, Médecine, Système de recherche d'information, Dictionnaire, Thesaurus, Web.

1. Introduction

L'informatisation des hôpitaux, des cabinets médicaux et le développement d'Internet entraînent une prolifération de l'information médicale. Du moins cette information médicale est de plus en plus souvent écrite, et disponible sous forme informatique. Le réseau Internet et les sites web médicaux offrent un corpus de textes médicaux de taille gigantesque. Or cette information est souvent mal exploitée car hétérogène et difficile d'accès.

En médecine, on dispose maintenant de nomenclatures et de thésaurus médicaux, qui sont en quelque sorte, une représentation des concepts médicaux les plus importants. Le thésaurus le plus connu étant l'UMLS [Lindberg 1993] (Unified Medical Language System) de la National Library of Medicine (N.L.M.), il s'agit en fait d'un "meta-thésaurus" contenant, notamment, le MeSH [MeSH 1986] et la CIM-10 [CIM 1977]. Au Laboratoire d'Informatique Médicale de la

faculté de médecine de Rennes, nous disposons d'un thésaurus extrait de la base de connaissance ADM [Lenoir et al. 1981] (Aide au Diagnostic Médical) qui contient près de 200000 concepts (essentiellement des symptômes, maladies et syndromes). Cette base de connaissance contient un dictionnaire qui couvre la plus grande partie du vocabulaire médical.

Le but de notre travail est d'utiliser ce dictionnaire médical, pour rechercher dans des textes médicaux (en langage naturel) des concepts appartenant à un thésaurus médical. Le résultat de l'indexation se résumant à l'ensemble des concepts détectés dans le document. S'il est clair que cette procédure entraîne une perte d'information importante, le résultat de l'indexation est un modèle plus simple à traiter automatiquement qu'un texte en langage naturel, l'évaluation des résultats nous montre que cela permet de répondre à la plupart des attentes des utilisateurs.

Cette indexation devra être entièrement automatisée de manière à pouvoir être utilisée de manière intensive. Le domaine d'application reste vaste, aussi nous avons fait le choix de nous baser sur les ressources existantes (dictionnaire ADM, thésaurus, corpus de textes), sans devoir décrire les spécificités morphologiques, grammaticales et syntaxiques du langage. L'indexation n'étant pas un objectif en soi, nous avons orienté notre travail sur les diverses applications de cette indexation : recherche de documents, similarité de documents, synthèse automatique et extraction de mots-clé. Nous n'avons pas abordé la classification des documents, le sujet est très vaste et dépasse le cadre de ce travail.

Après un aperçu des ressources existantes, nous présenterons le moteur d'indexation, et son utilisation dans différents contextes. Nous finirons cet exposé par une présentation des résultats obtenus.

2. Contexte

Dictionnaire ADM

Depuis plus de vingt ans, la base de connaissance ADM est alimentée par des thésards de médecine, ils décrivent les signes apparaissant dans les pathologies d'un domaine particulier, le résultat est ensuite validé par un expert. Cette base de connaissance est la plus importante base sémiologique française, voire mondiale (du moins d'un point de vue quantitatif). Elle permet aux professionnels de la santé de connaître les symptômes apparaissant dans une pathologie donnée et les pathologies contenant un symptôme donné. Elle est maintenant accessible sur Internet [Pouliquen et al. 1995] sur le site http://www.med.univ-rennes1.fr/adm.dir/.

Afin de pouvoir interroger cette base de données, un dictionnaire de mots (avec leurs flexions, dérivations, synonymes et quasi-synonymes) avait été construit. Ce qui permettait de retrouver une maladie ou un symptôme d'après une phrase en langage naturel. Un dictionnaire de mots simples était insuffisant pour gérer toutes les synonymies, il a donc été enrichit par des mots multiples (unités complexes), ce qui a permis de pouvoir créer des familles de mots assez élaborées, comme : "angor", "angors", "angine de poitrine", …

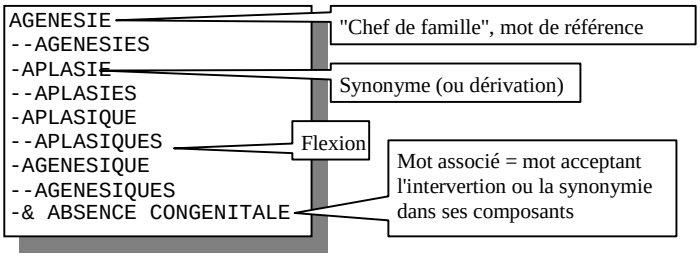

Figure 1 : Exemple de famille de mots

Ce dictionnaire comporte des mots composés, unités lexicales complexes dont la définition est différente de celle de chacun des composants (ex: "Angine de poitrine", synonyme d'"Angor", qui n'a pas du tout le sens d'"Angine"), mais comporte aussi des "mots associés" qui sont des unités complexes synonymes de mots simples ("Accident Vasculaire Cérébral" synonyme de "AVC", ou "Mal de tête" synonyme de "Céphalée").

Thésaurus

Un thésaurus médical est une représentation d'un ensemble de concepts médicaux. Chaque concept étant décrit par un ou plusieurs termes. Ces concepts sont, dans la plupart des cas, organisés de manière hiérarchique (taxonomie, méronymie…) et parfois de manière sémantique (réseau sémantique comme dans un graphe conceptuel [Sowa 1984]).

De la base de connaissance ADM, nous avons pu extraire un véritable thésaurus contenant les concepts (pathologie, symptôme), les termes (l'ADM contient les diverses terminologies utilisées pour chaque maladie ou signe), les hiérarchies de concepts, et même un réseau sémantique (la relation exprimant qu'un symptôme apparaît dans une pathologie). Par ailleurs nous avons extrait du meta-thésaurus UMLS un sous-ensemble des concepts traduits en français (il s'agit du MeSH, traduit par l'INSERM), nous avons conservé la hiérarchie de l'UMLS. L'avantage d'un thésaurus comme MeSH, est qu'il est élaboré par des professionnels du domaine, validé par un comité d'expert, et, surtout, qu'il évolue chaque année.

Nous n'avons donc pas utilisé les outils de constitution de terminologies (comme Ana, Acabit, Lexter, Termino). Cependant, il serait possible d'utiliser de tels outils pour constituer sa propre ressource terminologique, ou pour enrichir un thésaurus par de nouveaux termes, dans le cas ou les thésaurus médicaux existants se révèlent insuffisants [Bourigault et Jacquemin 2000]. Dans ce domaine, citons les travaux de [Zweigenbaum et al. 1995] ou [Bouaud et al. 1994] pour l'acquisition et la représentation du vocabulaire médical.

Indexation

Il existe trois types de systèmes d'indexation : les systèmes manuels (une personne a préalablement désigné les termes d'indexation : les descripteurs associés à chaque texte), les systèmes semi-automatiques (ou supervisés, un processus automatique propose des descripteurs, qui sont ensuite validés manuellement), enfin les systèmes d'indexation automatique (non supervisés, un programme indexe les documents sans intervention humaine). Les descripteurs (l'unité d'indexation choisie) peuvent être les mots du texte, les lemmes, les concepts (mots-clé ou termes) et, plus rarement, les *N*-grammes (séquences de *n* caractères consécutifs), ou encore les contextes (cas du "Latent Semantic Indexing" ou des méthodes basées sur l'Analyse Factorielle des Correspondances). Les modèles utilisant les mots peuvent fonctionner avec la langue anglaise (peu de flexions, peu d'homographies), mais se révèlent nettement insuffisants pour les autres langues (particulièrement pour les langues agglutinantes), on peut alors utiliser les lemmes, mais, pour avoir de bonnes performances, il faut recourir à une analyse linguistique pour lever certaines ambiguïtés [Fluhr 2000]. Dans le cas d'attribution de mots-clé, des méthodes statistiques permettent d'extraire des concepts par apprentissage automatique, depuis un corpus indexé manuellement, en étudiant les probabilités d'apparition d'un mot-clé selon les mots du texte, ces méthodes donnent de relativement bons résultats si le corpus est suffisamment important (qualitativement et quantitativement), ils n'ont pas de barrière linguistique (par exemple : [Steinberger et al. 2002]). Ils fonctionnent très mal sur des phrases (trop peu de mots). Si l'on veut pouvoir

indexer de petites unités textuelles (comme une requête de l'utilisateur), il semble indispensable de recourir aux outils linguistiques. Encore peu utilisés car ils nécessitent de décrire de manière formelle le langage utilisé, ils sont cependant susceptibles d'apporter d'intéressantes améliorations [Berrut 1988]. L'opposition entre statistique et linguistique (qui, malheureusement, a longtemps été de mise) tend maintenant à s'estomper [Fluhr 2000].

À tous les niveaux de l'indexation (qu'elle soit statistique ou linguistique) se pose le problème des mots composés (unité lexicale composée de plusieurs mots), il semble maintenant indispensable de les prendre en compte et de les considérer comme un seul mot [Gross 1986].

Systèmes de recherche d'information

Le résultat de l'indexation est utilisé le plus souvent pour un Système de Recherche d'Information, il s'agit d'un système qui permettra de retrouver certains documents du corpus correspondant à une question de l'utilisateur. Citons les modèles booléens, vectoriels, probabilistes, ou encore les modèles spécifiques au langage naturel, ou les modèles statistiques basés sur la réduction de dimensions (comme le "Latent Semantic Indexing" ou l'analyse factorielle des correspondances). Cf. [Baeza & Ribeiro 1999].

Extraction de mots-clé

Ici encore, deux méthodes existent : par apprentissage ou par extraction. La première méthode (statistique) calcule, à partir d'un corpus indexé manuellement, une matrice de co-occurrence de chaque mot avec chaque mot-clé. L'attribution de mots-clé à un nouveau texte consistera à calculer la probabilité d'apparition de chaque mot-clé en fonction des mots du texte. La seconde méthode (linguistique) consiste à décrire chaque mot-clé par des termes (les différentes variantes linguistiques pouvant le représenter), et à essayer de reconnaître ces termes dans le texte. Cette méthode a l'énorme avantage de fonctionner sur de petites unités textuelles. La principale difficulté étant de définir la liste des termes associés à chaque mot-clé, il faut recourir aux outils linguistiques pour reconnaître un terme quelle que soit sa forme syntaxique [Jacquemin et Tzoukermann 1999].

Synthèse automatique

On distingue trois méthodes [Desclès et Minel 2000] : Celles qui sont fondées sur la compréhension automatique de textes, utilisant les méthodes linguistiques et d'intelligence artificielle, le principe est de "représenter" le texte sous forme de graphe, de réduire ce graphe en ne gardant que les nœuds les plus importants, et de régénérer un texte correspondant au graphe réduit. Les méthodes par extraction utilisent des ressources linguistiques légères, elles consistent à sélectionner les unités textuelles par calcul d'un score de similarité par rapport au document, ou par rapport aux autres unités, et n'extraire que les unités les plus importantes. Une troisième méthode, par filtrage sémantique, repère les unités textuelles importantes par des marqueurs linguistiques, et les restitue dans leur contexte.

Corpus de textes

Notre corpus de textes (à indexer) se doit d'être le plus étendu possible, nous avons tout naturellement pris le parti d'indexer n'importe quel document HTML. Nous avons intégré un "aspirateur web" ("crawler", "Web agent") au moteur d'indexation. Ceci nous a permis d'indexer les cours de médecine disponibles sur le réseau pédagogique de Rennes [Fresnel et al. 1997], mais aussi des documents trouvés sur le CISMEF [Darmoni et al. 2000]. Cet outil pourra être utilisé pour l'Université Médicale Virtuelle Francophone [Le Beux et al. 2000].

3. Choix et réalisations

Notre objectif est de créer un moteur d'indexation entièrement automatique (pour ses performances quantitatives), nous utiliserons certains des outils linguistiques (ceux qui nécessitent le moins de travail préalable) pour extraire des textes les descripteurs. Les applications de cette indexation utiliseront des méthodes statistiques qui sont habituellement utilisées sur les mots.

Linguistique

Nous avons volontairement écarté les outils d'analyse grammaticale ou syntaxique et cela pour plusieurs raisons :

- Les thésaurus (dont l'indexation en mots est le premier objectif de notre système) ont un langage grammaticalement très pauvre (absence de verbes, très peu d'articles…)
- Il est impossible de définir une grammaire couvrant la totalité d'une langue [Abeillé et Blache 2000], et le langage médical a son propre vocabulaire, sa description grammaticale et syntaxique requiert un travail d'experts dont nous ne disposons pas.
- Nous disposons déjà du dictionnaire ADM qui contient suffisamment de synonymes et de flexions pour palier aux défauts de l'absence d'analyse grammaticale.
- Les thésaurus ADM et MeSH sont suffisamment précis pour représenter la plus grande partie de l'information conceptuelle des documents.

[Abeillé et Blache 2000] reconnaissent que l'on peut se passer de syntaxe dans le cas où les études "portent sur des domaines extrêmement limités et, d'autre part, utilisent des bases de connaissances très détaillées dans lesquelles les structures sémantiques associées aux objets contiennent implicitement les informations syntaxiques". Si le domaine médical n'est sûrement pas "extrêmement limité" les thésaurus médicaux contiennent effectivement suffisamment d'informations implicites pour lever beaucoup d'ambiguïtés. [1]

Architecture du système

Nous souhaitons décrire chaque terme du thésaurus par les mots de référence qu'il contient, afin de permettre la reconnaissance du terme quels que soient les flexions ou synonymes utilisés. Ensuite nous essayerons de reconnaître les termes utilisés dans une phrase à l'aide de l'indexation précédente. Des termes nous extrayons ensuite les concepts.

Premier outil : indexation en mots de référence

La première réalisation est d'exploiter le dictionnaire ADM pour résumer chaque phrase par un ensemble de mots de référence. Ainsi la phrase "Néphrite glomérulaire lupique" sera indexée par les deux mots "Lupus" et "Glomérulonéphrite" (car on a reconnu le mot composé "Néphrite glomérulaire" synonyme de "Glomérulonéphrite"). Un peu à la manière d'un lemmatiseur, à la différence qu'un lemmatiseur se limite souvent aux flexions sans gérer les dérivations ni la synonymie. Le dictionnaire ADM contenant d'emblée les flexions, dérivations, synonymes et quasi-synonymes des mots du vocabulaire, nous n'avons pas non plus eu besoin d'utiliser un lemmatiseur. Il aurait cependant été utile d'en disposer pour gérer les ajouts de mots dans le dictionnaire, mais le vocabulaire médical étant assez spécifique, le

[1] Par exemple, le mot "PORTE" est extrêmement ambigu, s'agit t'il de la porte ("d'entrée"), de la veine porte, du verbe porter… Mais ici, l'ambiguité est aussitôt levée car, dans le thésaurus, ce mot n'est employé que dans les concepts : "veine porte", "thrombose veine porte"…

travail de configuration aurait été très fastidieux. Des outils comme [Zweigenbaum et Grabar 2000] pourraient permettre d'enrichir le dictionnaire par de nouvelles formes à partir d'un thésaurus, ou d'aider à créer un nouveau dictionnaire (pour une autre langue par exemple).

Cette première indexation permet de diminuer le silence lors de la recherche d'information, ainsi, quels que soient les flexions, dérivations ou synonymes des mots utilisés en recherche, on retrouvera les mêmes informations. Notamment grâce aux mots composés et associés.

Cette fonctionnalité est ensuite exploitée pour indexer le thésaurus cible, on crée ainsi une base d'indexation des termes. Notons que le système fonctionnera différemment selon que l'on utilise l'ADM ou le MeSH comme thésaurus cible.

Indexation de documents en concepts

Un document médical est découpé en phrases, chaque phrase est ensuite analysée par le processus précédent pour en extraire les mots de référence. On effectue une recherche des termes du thésaurus qui sont inclus dans la phrase (qui s'apparente à une recherche sur un modèle booléen). C'est-à-dire que l'on extrait les termes dont tous les mots sont présents dans la phrase. On retiendra ensuite les concepts représentés par ces termes. Et cela constituera l'indexation de la phrase. Par exemple, dans la phrase "Néphrite glomérulaire lupique", on reconnaîtra les concepts "Glomérulonéphrite lupique" et " Glomérulonéphrite" dans le MeSH. Dans l'ADM, on reconnaîtra les concepts "Néphrites glomérulaires" et "Lupus".

Une option du processus d'indexation permet de générer automatiquement les concepts pères des concepts extraits (en utilisant l'information du treillis de concepts). Ce qui permettra, à partir du concept "Glomérulonéphrite lupique" de générer les concepts "Glomérulonéphrite", "Néphrite", "Rein, maladies", "Appareil urinaire, maladies" et "Maladies urologiques et appareil génital male" (selon la relation "est_un" - "is_a" - de l'UMLS).

Nous avons fait le choix d'attribuer un poids à chaque phrase, afin de privilégier les concepts apparaissant dans le titre du document (ou un entête de paragraphe, ou une phrase en gras…). *Le résultat de cette seconde indexation sera stocké dans une base de données.*

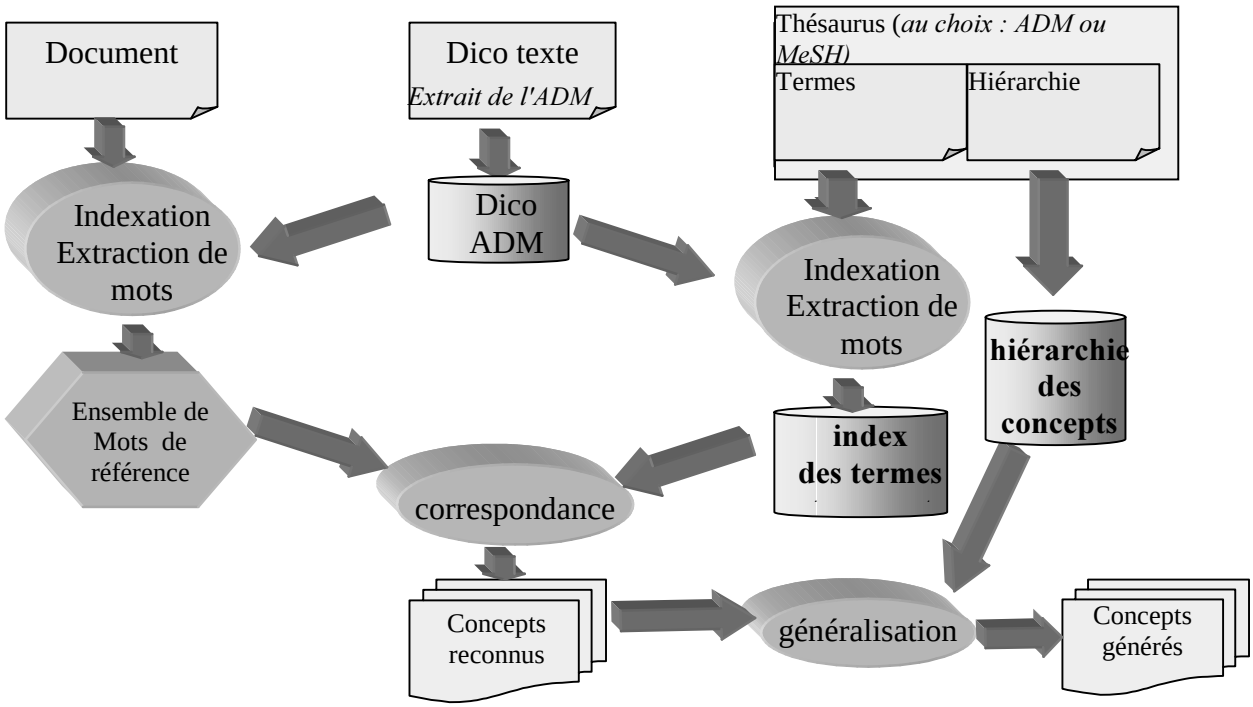

Figure 2 : Architecture du système

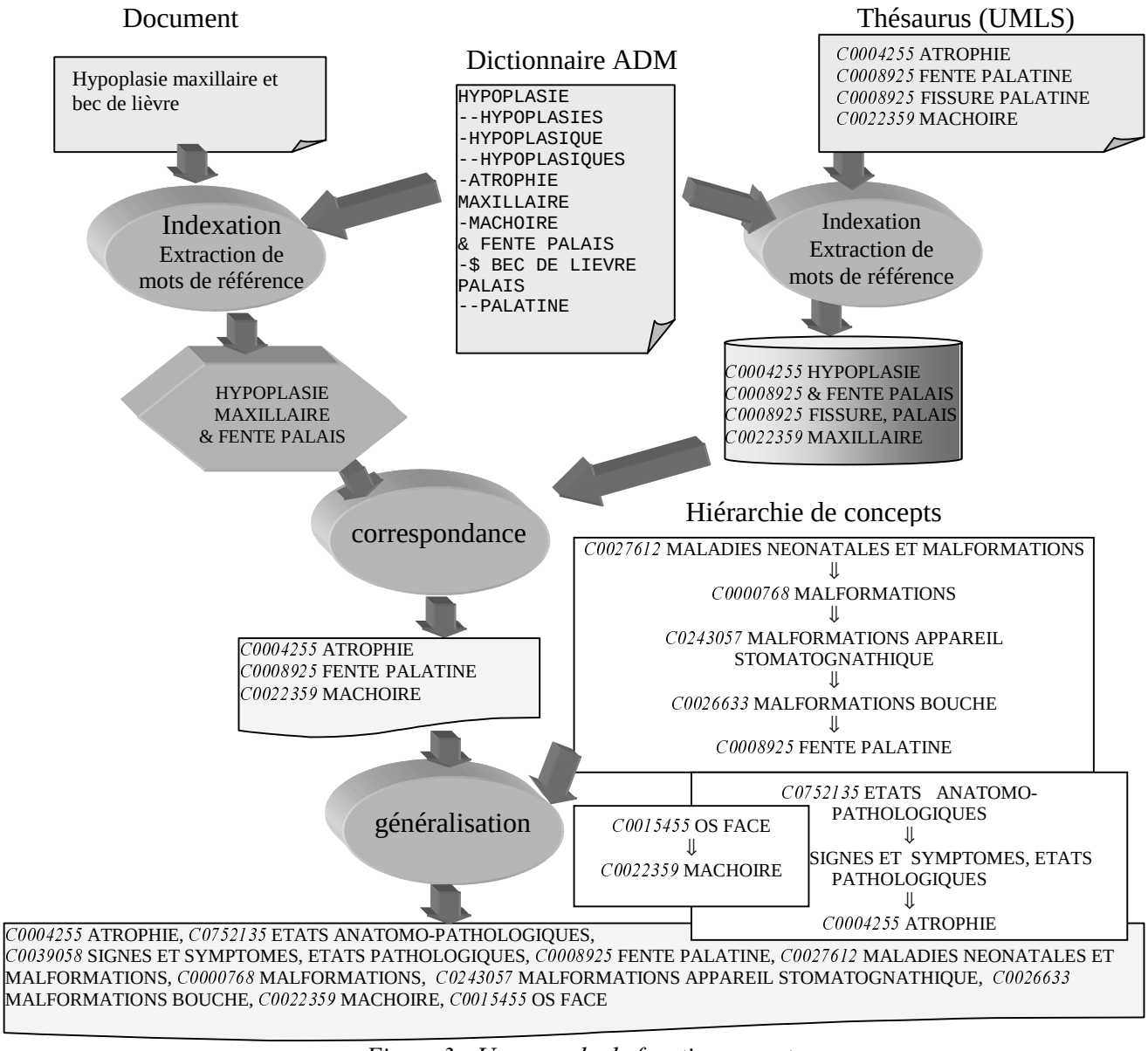

Figure 3 : Un exemple de fonctionnement

4. Utilisations de l'indexation

Modèle choisit

À ce niveau se posait le problème du choix du modèle de recherche d'information, nous avons choisi le modèle vectoriel qui nous a semblé plus adapté que le modèle booléen. La raison principale est qu'il paraît simpliste d'appliquer une logique binaire à une recherche d'information (un document correspond ou ne correspond pas). Le modèle booléen a l'inconvénient de privilégier les longs documents (un document de 100 pages contient beaucoup de concepts différents et risque donc de correspondre très souvent aux requêtes) contrairement au modèle vectoriel qui pondère le résultat par les autres concepts du document. De plus le modèle vectoriel permet de calculer des scores de similarité entre documents.

Le modèle vectoriel [Salton 1971] propose de représenter un document sur les dimensions représentées par les mots. Nous l'avons adapté pour représenter un document par un vecteur de concepts. Et, plutôt que de le représenter en fonction de la fréquence du concept dans le document, nous utilisons le score TFIDF [Salton et Buckley 1988]. Ce score permet de

donner une importance au concept en fonction de sa fréquence dans le document (TF = Term Frequency) pondérée par la fréquence d'apparition du concept dans tout le corpus (IDF = Inverse Document Frequency). Ainsi un concept très spécifique au document (n'apparaissant que dans ce document) aura un score correspondant à sa fréquence d'apparition, par contre, un concept apparaissant dans tous les documents du corpus aura une pondération maximale.

Après la phase d'indexation du corpus de textes, nous calculons donc, pour chaque concept dans un document, son score TFIDF. Nous verrons que, dans toutes les applications de l'indexation, nous utiliserons ce score TFIDF comme métrique de l'importance du concept dans le document.

$$TFIDF_{c,d} = TF_{c,d}.((\log_2 \frac{N}{DF_c}) + 1)$$

avec : c : un concept, d : le document, $TF_{c,d}$: la fréquence d'apparition du concept dans le document et DF_c : le nombre de documents du corpus contenant le concept

Par contre, l'ajout d'un nouveau document dans le système nécessite de recalculer tous les scores TFIDF. Il s'avère néanmoins, que, lorsque le nombre de documents est élevé, l'ajout d'un nouveau document ne modifie pas beaucoup les autres scores TFIDF. Le recalcul complet peut donc être différé.

Recherche d'information

L'application la plus intéressante de ce modèle vectoriel est de pouvoir calculer le score de similarité entre un ensemble de concepts (extraits d'une phrase ou d'un document) et les autres documents. La première application a donc été de créer un moteur de recherche. Ce moteur de recherche permet à l'utilisateur d'entrer une phrase en langage naturel, il indexe cette phrase en mots de référence, recherche les concepts du thésaurus correspondants et, pour chaque document du corpus, calcule le score de similarité avec cet ensemble de concepts. Le système présentera les documents ayant le score le plus élevé comme résultat de la recherche.

La mesure de similarité utilisée est la formule *Cosine* qui calcule le cosinus de l'angle entre le vecteur représentant la requête de l'utilisateur et chaque document du corpus [Salton 1983].

$$COSINE(d,r) = \frac{\sum_{c \in d \cap r} TFIDF_{c,d}.TFIDF_{c,r}}{\sqrt{(\sum_{c \in d} TFIDF_{c,d}{}^2).(\sum_{c \in r} TFIDF_{c,r}{}^2)}}$$

avec : c : un concept, d : le document, r : la requête de l'utilisateur, $TFIDF_{c,d}$: le score TFIDF du concept c dans le document d

D'autres mesures de similarité existent : nombre de concepts communs entre la requête et le document, somme des produits TFIDF ou encore la mesure *Okapi*. Parmi ces mesures *Cosine* est souvent celle qui donne les meilleurs résultats [Bellot 2000]. En effet, elle pondère la somme des produits TFIDF par la taille de la requête et, surtout, par la taille des documents. Elle sera donc élevée si le document contient principalement les concepts de la requête.

Similarité de documents

L'autre utilisation de l'indexation consiste à calculer, pour chaque document, son score de similarité par rapport à tous les autres, et ainsi de créer un réseau de proximité de documents. L'interface consistera à afficher les documents ayant le score le plus élevé.

L'analyse factorielle des correspondances [Benzécri et al. 1973], permet également de représenter synthétiquement les documents sur un graphique à deux dimensions, et donc d'évaluer les distances, selon certains axes qu'il faut interpréter. Cette méthode, utilisée le plus souvent sur une matrice documents-mots a été appliquée sur une matrice documents-concepts. L'avantage de cette méthode est de mettre en évidence très rapidement les proximités sémantiques des documents. L'analyse factorielle des correspondances est une méthode statistique qui commence à donner de bons résultats quand le corpus de textes est suffisamment étendu pour que les co-occurrences de mots soient suffisamment significatives, cela fonctionne d'autant mieux que les unités textuelles sont courtes, comme en bibliographie (cf. [Kerbaol et Bansard 1999]). Par contre, lorsque les textes sont de longueur importante, les co-occurrences de mots perdent de leur signification (ce qui impose de segmenter les textes). Par contre les co-occurrences de concepts sont moins sensibles à la taille des textes.

Une autre utilisation de l'Analyse Factorielle des Correspondances sur les concepts détectés serait de classifier automatiquement les textes, mais, comme nous l'avons vu en introduction, c'est un vaste sujet que nous n'avons pas abordé.

Synthèse automatique

Notre processus d'indexation utilise la phrase comme unité d'indexation, nous connaissons donc l'ensemble des concepts d'une phrase d'un document. Notre outil de synthèse automatique calcule un score de similarité entre chaque phrase et le document dans son entier. L'interface consiste simplement à afficher les phrases ayant le score le plus élevé. Le nombre de phrases étant paramétrable. Le résultat fera apparaître les phrases les plus importantes dans le même ordre que dans le texte original.

Extraction de mots-clé

Les scores TFIDF, qui représentent l'importance d'un concept dans un document pondérée par la présence de ce concept dans les autres documents, sont une bonne métrique pour déterminer les mots-clé d'un document. D'autant plus que nous avons augmenté le poids de chaque concept apparaissant dans une phrase importante (titre, entête de paragraphe…). Les concepts dont le score TFIDF est le plus fort sont souvent les plus représentatifs d'un document. Il faut cependant utiliser cet outil comme une aide à l'indexation (beaucoup de concepts mots-clé proposés ne sont pas adaptés, par exemple : le concept "rat" proposé sur un document parlant des cancers, parce que les expérimentations portaient sur des rats).

5. Résultats

Les différents outils développés ont été évalués sur plus de 100 documents médicaux avec le MeSH comme thésaurus cible. Le résultat de l'indexation est visible à l'adresse : http://www.med.univ-rennes1.fr/nomindex/

Le système d'indexation, d'un point de vue quantitatif, répond aux besoins exprimés, car on peut indexer 100 documents en moins d'une demi-heure, quant aux diverses utilisations, elles sont toutes assez rapides pour être faites "en ligne" (par exemple, le temps de réponse d'une recherche sur tout le corpus est inférieur à trois secondes).

L'évaluation de la qualité de l'indexation elle-même est une tâche énorme, pour ne pas dire impossible. Nous avons choisi de comparer le résultat d'une méthode purement statistique (l'Analyse Factorielle des Correspondances) tantôt sur l'indexation en texte-intégral, tantôt sur l'indexation en concepts. Par contre les utilisations de cette indexation sont plus proches du raisonnement humain, et donc, plus faciles à évaluer. Les critères utilisés sont le silence et le bruit (critères à minimiser) ou leurs compléments : le rappel et la précision (critères à maximiser). Nous utilisons les mêmes critères pour évaluer l'attribution automatique de mots-clé. Par contre, la similarité de documents, ou la synthèse automatique sont plus difficiles à quantifier, l'évaluation sera donc très subjective.

Indexation

Afin d'évaluer la pertinence de l'information sémantique contenue dans l'indexation, nous avons utilisé la méthode statistique d'analyse des correspondances [Benzécri et al. 1973], se basant sur les co-occurrences de mots dans les textes. Nous avons essayé cette méthode avec les mots contenus dans 100 documents. Il s'est avéré que le nombre de documents était insuffisant pour extraire une information sémantique intéressante avec si peu de co-occurences. Nous l'avons expérimentée avec cette fois-ci des concepts extraits de notre indexation, et l'information sémantique apparaît clairement. Cela tend à démontrer que notre indexation est sémantiquement pertinente.

Si les méthodes statistiques, ignorant les informations linguistiques, ne deviennent performantes que lorsque le corpus est suffisamment important, leurs performances peuvent être très sensiblement améliorées par une extraction préalable de concepts importants.

Extraction de mots-clé

Notre outil a été utilisé pour comparer l'indexation automatique avec l'indexation manuelle de documents médicaux sur le portail de site médicaux francophones du CISMEF [Darmoni et al. 2000]. Les résultats sont très encourageants. Il a été comparé avec deux autres outils commerciaux, qui utilisent des méthodes statistiques de co-occurrences de mots, l'analyse est toujours en cours mais les premiers résultats montrent la supériorité de notre outil : beaucoup de bruit, mais beaucoup de concepts pertinents sont proposés (parmi ceux-ci beaucoup de concepts "oubliés" dans l'indexation manuelle, en moyenne notre système propose un nouveau concept pertinent dans un document sur deux). Une évaluation sur 6000 documents (extraits du CISMEF), sera bientôt disponible (sur http://www.med.univ-rennes1.fr/nomindex/).

Recherche d'information

Nous avons extrait des traces de notre serveur web toutes les interrogations (qui étaient faites sur un moteur de recherche en texte intégral). Et nous les avons systématiquement reformulées sur le nouveau moteur. Les résultats sont incontestablement meilleurs (moins de bruit; moins de silence), sur 27 interrogations fréquemment posées, 485 documents sont trouvés par notre système contre 146 par une recherche en texte intégral.

Synthèse automatique et similarité de documents

Aucune évaluation précise n'a pu être faite sur ces deux applications. L'évaluation d'une synthèse automatique de document n'est pas aisée, quels sont les critères, quel est le but à atteindre ? [Desclès et Minel 2000]. Nous pouvons juste affirmer que le résultat de la synthèse automatique paraît très souvent décousu (notre système propose des phrases hors de leur contexte) mais quasiment toujours pertinent (l'information sémantique essentielle apparaît

bien dans les phrases résultat). De même, nous n'avons pas évalué de manière objective les résultats de la similarité de documents. Mais les résultats sont jugés pertinents.

6. Conclusion

Les résultats sont jugés très satisfaisant en comparaison avec l'existant (système de recherche en texte intégral), la plupart des défauts constatés peuvent être corrigés en intervenant sur le dictionnaire, ou, parfois, en intervenant sur certains termes du thésaurus. Les défauts inhérents à la construction des thésaurus sont une limite du système, mais gageons que le meta-thésaurus UMLS corrigera peu à peu ces défauts…

D'autres défauts ne pourront être corrigés qu'avec le recours à la linguistique, notamment le recours à un étiquetage lexical semble nécessaire, mais cela nécessite un travail considérable sur le dictionnaire ADM (afin d'ajouter l'information grammaticale), le recours à des outils statistiques ou à d'autres dictionnaires existants permettrait de faciliter la tâche.

Le dictionnaire ADM, sur lequel repose notre système d'indexation, est uniquement en français, ce qui exclut toute utilisation sur des thésaurus anglophones non traduits. Cependant, afin de pouvoir néanmoins indexer des textes multilingues, nous avons expérimenté une traduction préalable des termes en français, traduction faite à partir des traductions de termes de l'UMLS, l'outil se révèle pratique, mais engendre beaucoup de bruits. Il existe une autre possibilité pour l'indexation multilingue : fournir en entrée du système un dictionnaire différent par langue, le dictionnaire ne comporte aucune information grammaticale, il serait possible de générer un dictionnaire médical anglais depuis l'UMLS.

Ce système serait applicable à d'autres domaines que la médecine, à la condition de construire un dictionnaire approprié. Ce qui n'est pas tâche facile, mais des outils existent pour constituer un dictionnaire de termes à partir, soit du thésaurus (comme [Zweigenbaum et Grabar, 2000]), soit du corpus de textes (comme [Jacquemin et Tzoukermann, 1999]). La simplicité de la construction du dictionnaire peut alors devenir un atout, car il ne sera pas nécessaire de définir les informations syntaxiques ou sémantiques de chaque mot du dictionnaire.

Les limites et défauts du système sont, dans le cadre de l'utilisation qui en a été faite, compensés par ses performances quantitatives et qualitatives.

Références

Abeillé A. et Blache P. (2000), Analyse syntaxique, *Ingénierie des langues*, Hermes.

Baeza-Yates A., Ribeiro-Neto B. (1999), *Modern information retrieval*, ACM Press books, Addison-Wesley,

Bellot P. (2000) Méthodes de classification et de segmentation locales non supervisées pour la recherche documentaire, *Thèse de doctorat*, Université d'Avignon,

Benzécri J.-P. & al., (1973), *La taxinomie*, Vol. (1); *L'analyse des correspondances*, Vol. (2), Dunod, Paris.

Berrut C. (1998), Une étude d'indexation fondée sur l'analyse sémantique de documents spécialisés. *Thèse de doctorat en informatique*. Université Joseph Fourier. Grenoble.

Bouaud J., Bachimond B., Charlet J., Zweigenbaum P., Boisvieux J.-F., (1994) Structuration and acquisition of an ontology within conceptual graphs, In *: ICCS 94 Workshop of Knowledge Acquisition using Conceptual Graph Theory*,

Bourigault D. et Jacquemin C. (2000), Construction de ressources terminologiques, *Ingénierie des langues*, Hermes.

[CIM 1977] *Classification Internationale des Maladies - 9ème Révision* (1977) , OMS, Genève.

Darmoni S., Leroy J.-P., Thirion B., Baudic F., Douyere M., Piot J. (2000) CISMeF : a structured Health resource guide, *Information in Medicine*, Janvier 2000; 39(1) 30-35

Desclès J.-P. et Minel J.-L. (2000) Résumé automatique et filtrage sémantique de textes, *Ingénierie des langues*, Hermes.

Fluhr C. (2000), Indexation et recherche d'information textuelle, *Ingénierie des langues*, Hermes, 2000

Fresnel A. et al. (1997), A first evaluation of a pedagogical network for Medical Students at the university hospital of Rennes, *MEDNET 97 The world Congress on the internet in Medicine* Brighton, Abstract Book Ed Arvanitis TN, Watson D page 60

Gross M. (1986), Lexicon-grammar, The representation of compound words, *proc of COLING-86*, Bonn, pages 1-6

Jacquemin C. et Tzoukermann E. (1999) NLP for Term Variant Extraction : Synergy between Morphology Lexicon and Syntax, *Natural Language Information retrieval*, T. Strzalkowski (ed.), Kluwer,

Kerbaol M. et Bansard J.-Y., (1999), Pratique de l'analyse des données textuelles en bibliographie; Ecole MODULAD SFdS, INRIA, *Bases de données et statistiques*, Editeur sous presse.

Le Beux P., Duff F., Fresnel A., Berland Y., Beuscart R., Burgun A., Brunetaud JM, Chatellier G., Darmoni S., Duvauferrier R., Fieschi M., Gillois P., Guille F., Kohler F., Pagonis D., Pouliquen B., Soula G., Weber J. (2000) The French Virtual Medical University, *Stud Health Technol Inform.* 2000;77:554-62

Lenoir P. , Michel J.-R., Frangeul C. et Chales G. (1981): Réalisation, développement et maintenance de la base de données A.D.M., *Médecine informatique*, vol. (6), N° 1, pages 51-56

Lindberg DAB, Humphreys BL., Mc Cray AT., et al. (1993) The Unified Medical Language System *Meth Inform Med*; 4 (32) : pages 281-91

[MeSH 1986] NATIONAL LIBRARY OF MEDICINE (1986) *Medical Subject Headings* Bethesda, Maryland

Pouliquen B., Riou C., Denier P., Fresnel A., Delamarre D., Le Beux P. (1995) Using World Wide Web Multimedia in Medicine, *Proc of MEDINFO'95*, IMIA, Eds Greenes, Peterson, Protti, pages 1519-1523

Salton G. (1971) *The SMART retrieval system. Experiment in automatic document processing*. Prentice Hall. Englewood Cliffs. New Jersey.

Salton G. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill.

Salton G. et Buckley C. (1988), Term weighting approaches in automatic text retrieval, *Information Processing and Management*, (l) 24, n° 5, 1988, pages 513 à 523.

Sowa J.-F. (1984) *Conceptual structures. Information processing in mind and machine*. Readings, Massachusetts: Addison-Wesley.

Steinberger Ralf, Pouliquen Bruno, Hagman Johan (2002). Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus Eurovoc. *Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'2002).* Mexico-City, 17-23 Février 2002 (En cours de publication)

Zweigenbaum P. et Grabar N. (2000), Expériences d'acquisition automatique de connaissances morphologiques par amorçage à partir d'un thésaurus, *Actes du 12 congrès Reconnaissance des Formes et Intelligence Artificielle*, Paris, pages II-101-II-110

Zweigenbaum P., Bachimont B., Bouaud J., Charlet J. and Boisvieux J.-F. (1995) Issues in the structuring and acquisition of an ontology for medical language understanding, *Methods of Information in Medicine*, (34): pages 15-24