Troisième Atelier Recherche d'Information SEmantique RISE, Avignon 15 mars 2011

Associé à la 8ème édition de la Conférence CORIA

ACTES DE L'ATELIER RECHERCHE D'INFORMATION SEMANTIQUE RISE 2011

Édité par
Catherine ROUSSEY LIRIS Lyon, Cemagref Clermont Ferrand (France)
Jean-Pierre CHEVALLET LIG Grenoble (France)

Troisième édition de l'atelier Recherche d'Information SEmantique

Atelier Recherche d'Information
SEmantique RISE, Avignon 15 mars 2011

Associé à la Conférence CORIA 2011

1. Introduction

Les documents produits actuellement sont essentiellement numériques. Une frénésie de numérisation est en passe de rendre accessible les ouvrages les plus anciens. La communication est également massivement numérique et voit l'émergence de nouvelles pratiques (blogs, SMS, réseaux sociaux), en plus des média textuels numériques bien implantés (email). Cette tendance s'intensifie avec la nomadisation de l'accès à l'information (téléphone portable, ultra-portables, iPad). Les objectifs à court terme sont alors une connexion ubiquitaire pour tous au réseau internet. Toutefois, même si cette masse d'informations est disponible, la difficulté majeure réside dans l'accès à de l'information ciblée, c'est à dire réellement en adéquation avec un besoin personnel et ponctuel. Cet accès se fait par filtrage, sélection, navigation ou interrogation.

Les systèmes de Recherche d'Information (RI) ont proposé une première réponse à ce problème d'accès à l'information pertinente. Les modèles développés en RI sont maintenant largement utilisés, par exemple dans les moteurs de recherche du Web. Les technologies actuelles sont basées sur des modèles statiques qui manipulent des informations de bas niveau. Par exemple, la plupart des moteurs de recherche sont basés sur le comptage des mots ou des liens sur les pages. Les dernières avancées de la recherche en RI ont concerné essentiellement l'amélioration des modèles statistiques d'appariement de documents, comme les modèles de langue statistiques. De nouvelles pistes de recherche consistent à ajouter de la sémantique pour obtenir des modèles statistiques intelligents. La sémantique permet d'améliorer la précision des résultats d'un système de RI en évitant les problèmes liés à l'ambiguïté ou au manque d'expressivité des mots simples. Même s'il ne semble pas nécessaire qu'un système de RI "comprenne" le document qu'il indexe, traiter le besoin de l'utilisateur au niveau sémantique permet plus de précision dans les réponses.

Nous pensons donc que l'avenir des systèmes de Recherche d'Information passe par la prise en compte de la sémantique du contenu des documents, permettant à un utilisateur de mieux maîtriser le flux d'information pour cibler l'information dont il a réellement besoin. Une façon d'atteindre cet objectif est de coder explicitement des

connaissances associées aux termes, par exemple dans des ontologies. Le but de cet atelier est de discuter de ce nouveau terrain de recherche: les systèmes de "concierge d'information" où le flux d'information est enrichi par une interprétation de son contenu. Nous appellerons ce nouveau paradigme: Recherche d'Information Sémantique. Cet atelier est dédié à tous les types de Recherche d'Information sans contrainte sur le mode de stockage de cette information. Par exemple la Recherche d'Information peut s'appliquer sur des documents textuels, des images, des vidéos, des flux XML etc...

2. Objectifs

Les travaux sur les ontologies ou les ressources sémantiques existent et sont actifs dans les différentes communautés informatique comme : le Web, la bio-informatique ou les systèmes d'information géographiques. Ainsi, les ressources sémantiques comme les ontologies, les bases de données lexicales, les thésaurii, se développent et sont maintenant disponibles. Cet atelier est spécialement dédié à l'usage des ressources sémantiques dans les systèmes de Recherche d'Information Multimedia et/ou Multilingue.

Des systèmes de Recherche d'Information Multilingue cherchent à retrouver des documents qui correspondent à un thème indépendamment de leur langue d'écriture. Dans le cas de documents non textuels (Multimédia), des données textuelles peuvent être extraites de leur contenu, apparaître dans le voisinage du document ou être issues d'annotations manuelles. Malheureusement, la nature peu structurée et le volume important d'information rendent difficilement accessible l'information pertinente aux utilisateurs. Pour résoudre ce problème, les travaux en Recherche d'Information (RI) se sont orientés vers les technologies issues du Web Sémantique et plus précisément sur l'usage des ressources sémantiques comme les ontologies, les thésaurii ou les bases de données lexicales.

L'atelier RISE a pour but de proposer un lieu de rencontre entre des chercheurs issus de différentes communautés comme la Recherche d'Information, le Web Sémantique, le TALN, le Multimedia, l'Ingénierie des Connaissances.

3. Thèmes

Les principaux thèmes abordés peuvent être (liste non exhaustive, d'autres thèmes pouvant être traités par les auteurs) :

- Indexation Conceptuelle et Indexation Sémantique,
- Recherche d'Information Multimedia
- Recherche d'Information Multilingue

- Extraction d'Information Multilingue et Multimedia
- Annotation Sémantique
- Web Sémantique
- Ontologies Multilingues et Multimedia,
- Alignement d'Ontologie et Correspondance pour la Recherche d'Information,
- Graphes Conceptuels, Logiques de Description, Langages de Représentation des connaissances pour la Recherche d'Information.
- Utilisation des Distances Sémantiques pour la Recherche d'Information

4. Comité de Programme

- AUSSENAC Nathalie, IRIT Toulouse (France)
- CHEVALLET Jean-Pierre, LIG, Grenoble (France)
- DAMAS Luc, LISTIC, Annecy (France)
- GRAU Brigitte, ENSIIE (France)
- METAIS Elisabeth, CNAM Paris (France)
- ROCHE Christophe, LISTIC, Annecy (France)
- ROUSSEY Catherine, LIRIS, Lyon (France)
- RUMPLER Béatrice, LIRIS, Lyon (France)
- SCHWAB Didier, LIG-GETALP, Grenoble (France)
- SERASSET Gilles, LIG, Grenoble (France)
- SIMONET Michel, TIM-C, Grenoble (France)
- ZARGAYOUNA Haïfa, LIPN, Paris (France)
- ZWEIGENBAUM Pierre, LIMSI (France)

5. Organisation

L'après midi RISE s'est déroulé en deux temps. Tout d'abord deux travaux sur l'indexation sémantique ont été présentés. Ensuite une discussion sur l'évaluation des systèmes de recherche d'information sémantique a été encadrée par Haïfa Zargayouna.

L'atelier RISE 2011 a réuni une dizaine de chercheurs francophones venant de différents laboratoires: LIG, LIRIS, LIA, LIPN. Nous remercions tous les membres du comité de programme pour la qualité de leur travail ainsi que les auteurs des articles.

Les organisateurs de l'atelier: Catherine Roussey et Jean-Pierre Chevallet.

6. Programme

Session indexation sémantique

Session évaluation des systèmes de recherche d'information sémantique

Extraction d'information conceptuelle de textes, basée sur une annotation interlingue et guidée par une ontologie

David Rouquet, Achille Falaise

LIG-GETALP
david.rouquet@imag.fr, achille.falaise@imag.fr

RÉSUMÉ. *Nous proposons dans ce papier une méthode générique (indépendante de la langue et du domaine) permettant d'extraire des informations conceptuelles à partir de textes. Une ontologie de domaine, considérée comme un paramètre du système, détermine les informations pertinentes et guide le processus d'extraction. Les textes sont lemmatisés puis annotés par des lexèmes interlingues, ce qui permet à la majeure partie du processus de rester indépendante de la langue. Un alignement automatique entre l'ontologie et le lexique interlingue permet, ensuite, l'identification des concepts présents dans le texte. Notre méthode est implémentée suivant une architecture distribuée, orientée services. Par ailleurs, dans le cadre, du projet ANR OMNIA, elle est combinée avec des analyses visuelles pour l'indexation de documents bimodaux (images et textes).*

ABSTRACT. *We propose in this paper a generic method (language and domain independent) for conceptual information extraction from texts. A domain ontology, considered as a system parameter, determines the relevant information and guide the extraction process. The texts are lemmatized, and then annotated by interlingual lexemes, which allows most of the process to remain language independent. Then, an automatic alignment between the ontology and the lexicon allows the identification of interlingual concepts in text. Our method is implemented using a distributed, service oriented architecture. In addition, as part of ANR OMNIA, it is combined with visual analysis for indexing bimodal documents (images and text).*

MOTS-CLÉS : *Annotation interlingue, multilinguisation d'ontologie, extraction de concepts*

KEYWORDS: *Multilingual annotation, ontology multilinguisation, concepts extraction*

1. Introduction

Le but du projet OMNIA (Marchesotti *et al.,* 2010) est la mise en place d'un système de recherche d'images, accompagnées de textes multilingues (légendes, commentaires, etc.), dans de grands entrepôts de données.

À l'aide de traitements automatiques du contenu textuel et visuel, les images sont classées par rapport à une hiérarchie de concepts (ontologie) exprimée en OWL. Les utilisateurs peuvent exprimer des requêtes courtes en langue naturelle ou soumettre comme requête l'ensemble d'un texte qu'ils souhaitent illustrer. L'ensemble du projet OMNIA a été présenté en 2010 à RISE (Rouquet *et al.,* 2010). Ici, nous détaillons les composants textuels, en particulier l'annotation interlingue et l'extraction de contenu, qui étaient encore embryonnaires à l'époque.

Afin de construire les descripteurs des images ou des requêtes dans l'ontologie, nous procédons à une extraction de contenu multilingue qui ne requiert pas la traduction des textes. Il a été montré dans (Daoud, 2006) que l'annotation des textes par des lexèmes interlingues est une approche valide pour débuter ce processus : nous n'avons ainsi pas recours à une analyse syntaxique coûteuse et obtenons rapidement des données indépendantes de la langue des textes. Cela permet d'effectuer le reste du traitement grâce à des processus génériques.

Notre méthode est testée sur les 500.000 images et textes compagnons de la base Belga-News utilisée lors de la campagne de recherche d'images CLEF09.

2. Architecture générale

Dans notre scénario, nous avons deux types de données textuelles à considérer : les textes compagnons de la base (légendes d'images), et les requêtes des utilisateurs. Ces deux types de textes sont analysés de façon similaire.

Combiner dans une même ressource, à la fois une représentation linguistique multilingue (le signifiant) et une représentation conceptuelle (le signifié), est une tâche complexe (nous revenons sur ce problème en section 4.2, page 8), que nous simplifions en utilisant principalement deux ressources distinctes :

– un lexique interlingue, lié aux lemmes (ou termes) de chaque langue ;

– une ontologie de domaine.

Notre approche se veut générique et nous pouvons en principe utiliser n'importe quelle ressource lexicale et ontologique de ce type, indépendamment l'une de l'autre. Cette séparation en deux ressources, permet en outre de palier aux manque de ressources dédiées à l'extraction multilingue d'information, en réutilisant des lexiques et des ontologies développés pour d'autres problématiques. Nous relions entre eux lexique et ontologie indépendantes, par un processus automatique décrit en section 4.3.

L'architecture générale de l'extraction de contenu est décrite dans la figure 1. Ses

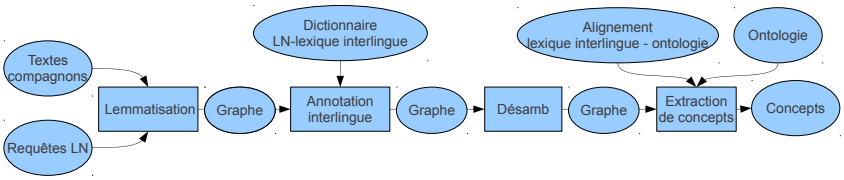


Figure 1. *Architecture générale de l'extraction de contenu multilingue dans le projet OMNIA*

principaux composants seront décrits en détail dans la suite et peuvent être résumés comme suit.

1) Les textes (compagnons et requêtes) sont lemmatisés par un logiciel dépendant de la langue. Les ambiguïtés sont préservées dans une structure de graphe ;

2) Les graphes sont enrichis avec des descripteurs interlingues. Cela ajoute de nombreuses ambiguïtés, puisque plusieurs acceptions sont possibles pour une occurrence dans le texte ;

3) Enfin, l'information conceptuelle contenue dans les textes est extraite en utilisant un alignement entre les descripteurs interlingues et les descripteurs de la hiérarchie de catégorisation des images (ontologie).

Chaque étape du traitement génère des ambiguïtés (lexicales, de segmentation, etc.). Ces ambiguïtés sont conservées dans un graphe. Nous avons choisi d'utiliser le formalisme des graphes-Q (graphes étiquetés par des arbres), manipulables grâce à des règles de réécritures, les systèmes-Q (Colmerauer, 1970). L'utilisation de ce formalisme permet de nombreux développements et expérimentations (analyses syntaxiques locales, etc.).

L'information conceptuelle extraite peut prendre différentes formes : un vecteur de couples concept-score (vecteur conceptuel), des déclarations dans l'ontologie (*A-box*), des requêtes formelles (SQL, SPARQL, etc.), etc. Dans le cadre d'OMNIA, l'information conceptuelle extraite des textes compagnons est stockée dans une base de données, alors que les requêtes des utilisateurs sont transformées en requêtes formelles (SQL).

3. Annotation interlingue

Cette section présente le processus d'annotation des textes par des lexèmes interlingues (UW) (Rouquet *et al.,* 2009a). Ce processus peut être qualifié de "lemmatisation interlingue" dont les ambiguïtés sont conservées dans une structure adéquate. Nous commençons par décrire la nature des lexèmes interlingues utilisés ainsi que que la structure de graphes de chaînes (graphes-Q) utilisée. Le processus d'annotation est détaillé ensuite.

3.1. *Ressources et structures de données*

3.1.1. *Universal Networking Language (UNL)*

Les textes sont annotés avec des *lexèmes interlingues* dits UW (*Universal Words*) constituent le vocabulaire du language UNL (*Universal Networking Language*)[1] (Boitet *et al.,* 2009). Celui-ci est un langage pivot "linguistico-sémantique" qui représente le sens d'un énoncé par une structure abstraite (un hyper-graphe) d'un énoncé anglais équivalent. Chaque UW est constitué d'un *mot vedette*, dérivé si possible de l'anglais, qui peut être un mot, des initiales, une expression ou même une phrase entière, et d'une *liste de restrictions* dont le but est de préciser sans ambiguïté le concept auquel l'UW renvoie.

Un UW est une étiquette pour un concept associé au sens d'un mot (simple ou composé) dans au moins une langue naturelle. Par exemple :

book(icl>thing) : correspond au substantif (restriction *icl>thing*) anglais *book*, sans autre précision ;

book(icl>do, agt>human, obj>thing) : correspond au verbe (restriction *icl>do*) anglais *to book*, dont l'agent est un humain (restriction *agt>human*) et l'objet une chose (restriction *obj>thing*) ;

ikebana(icl>flower_arrangement) : correspond au substantif japonais *ikebana*, dans le sens d'arragnement floral (*icl>flower_arrangement*).

Les restrictions sont composées d'une étiquette de 3 lettres suivie du signe '>' puis de la valeur. Par exemple, *icl* signifie "include" et indique une restriction de spécification.

Nous utilisons les 207 000 UW construites à partir des *synsets* de WordNet dans le cadre du consortium U++[2]. Wordnet est une base lexicale sémantique qui relie des sens lexicaux à des « sacs de lemmes », quasi-synonymes. Chaque ensemble sens lexical - liste de lemmes est appelé un synset, et les synsets sont reliés entre eux par des relations sémantiques, notamment d'hyperonymie, et se définit par rapport à ces relations. Par contre, dans UNL, un UW se veut plus fin que les quasi-synonymes des synsets de Wordnet ; par conséquent, en général, chaque quasi-synonyme correspond à une entrée du dictionnaire d'UW. Les entrées d'un dictionnaire d'UW sont donc des lexèmes, et non des concepts. De plus, bien qu'il soit possible de retracer les relations entre UW en passant par Wordnet, ces relations ne sont pas présentes formellement dans le dictionnaire, qui n'offre pas les capacités de raisonnement d'une ontologie, c'est pourquoi on considère les UW comme des lexèmes interlingues et non comme des concepts.

1. http ://www.unld.org
2. http ://www.unl.fi.upm.es/consorcio/index.php

Les UW sont reliés aux langues naturelles par des dictionnaires bilingues. Ceux-ci sont créés et maintenus par les membres du projet UNL. Notre équipe est par exemple en charge du dictionnaire français-UW. Il contient à ce jour environ 47 000 lemmes français reliés à des UW. Les autres langues dont les dictionnaires sont les plus développés sont l'espagnol, le japonais, le russe et l'hindi.

Tous ces dictionnaires sont stockés au sein d'une *base de données lexicales multilingues* sur la plate-forme Jibiki-PIVAX (Nguyen *et al.,* 2007). Les articles de chaque langue (y compris les UW) constituent des volumes monolingues. Ensuite, les sens de mots sont reliés entre les différentes langues par des acceptions interlingues (*axie*).

3.1.2. *Les systèmes-Q*

Lors du traitement des textes à des fins de recherche (ou d'extraction) d'information, la conservation d'une ambiguïté est toujours préférable à sa mauvaise résolution. Pour représenter les ambiguïtés présentes dans les textes (lexicales, segmentation, etc.), nous utilisons le formalisme du *langage-Q* (Colmerauer, 1970). Ce formalisme représente les énoncés dans une structure de graphe (un *graphe-Q*) dont les arcs sont décorés par des expressions parenthésées (des arbres). Des opérations sont possibles sur ces structures grâce à un système de réécriture de graphes (les *règles-Q*). Dans notre traitement, une règle-Q représente la traduction d'un UW dans une langue. Un ensemble (non ordonné) de règles-Q est appelé *traitement-Q*. Un *système-Q* est une séquence de traitements-Q. Un exemple de graphe-Q et de règle-Q est donné dans la figure 3 du paragraphe 3.2.3.

De notre point de vue, l'utilisation du langage-Q a trois avantages principaux :

– il fournit une structure de représentation formelle pour les textes qui facilite le portage linguistique (Hajlaoui *et al.,* 2007) ;

– les traitements sur les textes sont unifiés grâce à un puissant système de règles de réécritures ;

– les textes représentés sont facilement interprétables et manipulables par des non-informaticiens (linguistes, etc.)

Nous utilisons une version du langage-Q réimplémentée en 2007 par Hong-Thai Nguyen (Nguyen, 2009).

3.2. *Les étapes du processus d'annotation*

3.2.1. *Vue d'ensemble*

Le processus d'annotation comporte les étapes suivantes :

1) lemmatisation avec un logiciel adapté et dépendant de la langue ;

2) transcription des textes lemmatisés en graphes-Q ;

3) création de dictionnaires bilingues locaux sous forme de systèmes-Q (langue source - UW) pour chaque texte (ou fragment) ;

4) exécution de ces dictionnaires sur les graphes-Q.

3.2.2. *Lemmatisation*

Pour l'annotation interlingue (décrite plus loin), nous utilisons des dictionnaires dont les entrées sont des lemmes. La première étape du traitement est donc la lemmatisation des textes (i.e. l'annotation de chaque occurrence avec les lemmes possibles). Il est important de noter que le lemmatiseur doit conserver toutes les ambiguïtés dans un réseau de confusion (un simple "tagger" (baliseur) ne convient pas). Plusieurs logiciels peuvent être utilisés pour couvrir les langues souhaitées ; leurs sorties devront être transformées en graphes-Q.

Pour le français et l'anglais, nous avons développé des lemmatiseurs, dont les sorties sont des graphes-Q, basés sur les dictionnaires morphologiques DELA[3] . Ces dictionnaires sont disponibles sous licence LGPL.

L'algorithme de lemmatisation peut être résumé comme suit.

1) Le texte est d'abord segmenté au maximum. Par exemple, pour l'anglais et le français, les segments sont séparés par des espaces et des signes de ponctuation ; mais pour une langue sans espace comme le chinois, on peut considérer chaque caractère comme un segment.

2) On initialise le graphe en créant un nœud par séparateur ainsi trouvé.

3) Ensuite, les arcs sont construits. Tous les regroupements de segments contigus possibles, dans les limites d'une fenêtre de taille paramétrable (3 segments par défaut), sont testés. S'ils sont présents dans le dictionnaire, l'arc correspondant est ajouté au graphe.

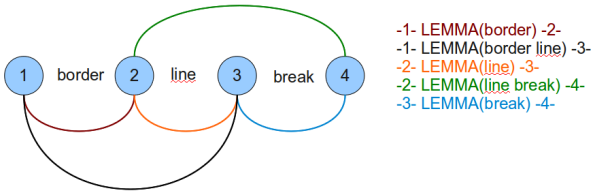


Figure 2. *Résultat simplifié de la lemmatisation de la séquence "border line break"*

Cela permet de gérer les ambiguïtés morphologiques (un segment peut être identifié comme plusieurs lemmes), de segmentation (plusieurs segments peuvent correspondre individuellement à des lemmes, mais aussi former un lemme à part entière lorsqu'ils sont combinés, par exemple "line break"), et de recouvrement de lemmes multisegment (par exemple, dans "border line break", les lemmes "border line" et "line break" sont possibles). Un exemple simplifié de sortie est donné dans la figure 2.

3. http ://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html

3.2.3. *Export de dictionnaires locaux sous forme de systèmes-Q*

Une fois le texte lemmatisé sous forme de graphe-Q, nous utilisons les possibilités de réécriture des systèmes-Q pour enrichir ce graphe-Q avec des UW, comme illustré par la figure 3. Le texte en entrée est dans un premier temps lemmatisé et converti en graphe-Q, pour représenter les ambiguïtés de lemmatisation (segmentation et identification du lemme). Le dictionnaire langue naturelle - UW a préalablement été compilé sous forme de règles-Q, qui sont appliquées au graphe-Q du texte lemmatisé. Chaque règle-Q correspond à une entrée de dictionnaire bilingue et transforme un lemme d'une langue en un UW (plusieurs UW en cas d'ambiguïté).

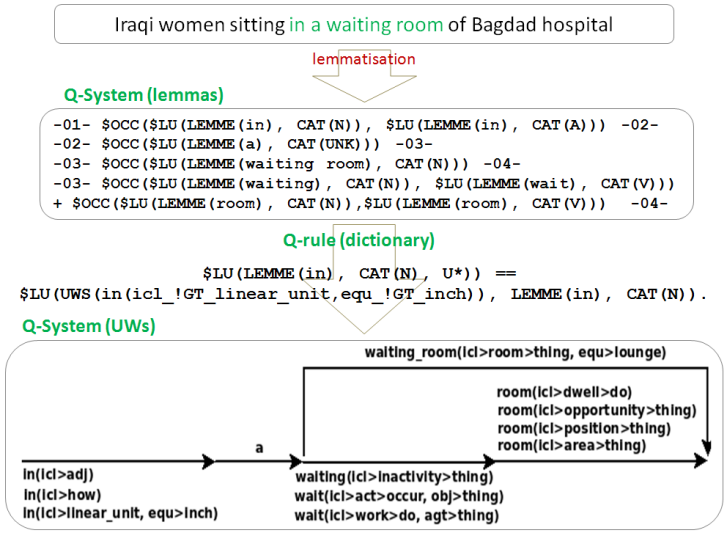


Figure 3. *Création et exécution d'un Système-Q*

Le nombre d'ambiguïtés exhibées dans les textes annotés est conséquent : jusqu'à douze UW pour une occurrence, auxquelles s'ajoutent les ambiguïtés de segmentation. Des procédés de désambiguïsation automatique sont utilisés pour assigner des scores aux interprétations possibles d'un mot suivant leur vraisemblance dans le contexte, mais sans sélectionner une interprétation en particulier. Ces processus de désambiguïsation ne sont pas détaillés ici.

4. Ontologie comme paramètre de l'extraction de concepts

L'extraction de contenu doit être guidée par une "base de connaissances" qui contient les types d'information que l'on recherche. Cette extraction de contenu est le processus qui fait le lien entre les descripteurs de la hiérarchie de classes du projet OMNIA (ontologie) et les descripteurs interlingues qui enrichissent les textes (UW).

4.1. *Contexte : travaux antérieurs de l'équipe en extraction de contenu*

Notre approche vient de projets de traduction automatique comme C-STAR II (1993-1999) (Blanchon *et al.,* 2000) ou Nespole ! (2000-2002) (Metze *et al.,* 2002), pour la traduction à la volée de dialogues dans le domaine du tourisme. Dans ces projets, le transfert sémantique passait par un IF (Interchange Format), c'est-à-dire un pivot sémantique dédié au domaine. Ce format d'échange permet non seulement de stocker l'information extraite des textes, mais aussi de guider l'extraction de contenu en fournissant une représentation formelle des informations pertinentes à extraire. L'IF est un pivot non pas linguistique ou linguistico-sémantique, mais sémantico-pragmatique (il contient en effet des informations sur l'acte de parole et éventuellement la force illocutoire).

L'IF de Nespole ! contenait 123 concepts du domaine touristique, associés à différents arguments et constructions linguistiques (patterns).

4.2. *Intégration d'ontologies existantes comme paramètres du domaine*

Dans le projet OMNIA, la base de connaissances prend la forme d'une ontologie peu contrainte pour la classification d'images (les instances de l'ontologie sont les images à classer). L'usage d'une ontologie est conforme avec les spécificités du format d'échange précédent, et présente les avantages suivants :

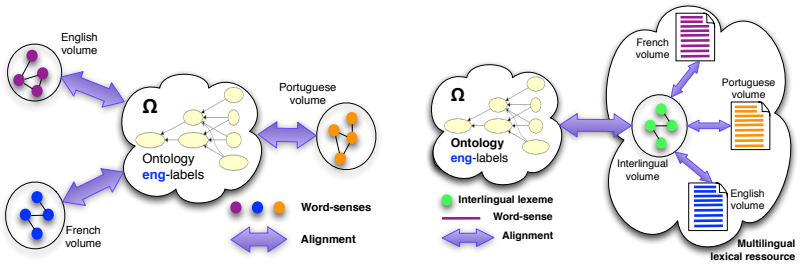
– Les ontologies donnent une description axiomatique du domaine, basée sur des logiques (en général des logiques de description (Baader *et al.,* 2003)) avec une sémantique explicite et formelle. Ainsi, les informations qu'elles contiennent peuvent être utilisées par des logiciels.

– Les structures ontologiques sont proches de l'organisation des idées dans l'esprit humain sous forme de réseaux sémantiques (Aitchenson, 2003) et sont étiquetées avec des items dérivés d'une langue naturelle. Ainsi, des humains peuvent les utiliser (navigation, contribution, etc.) de façon plutôt naturelle.

– Enfin, avec les récentes avancées du Web Sémantique et des initiatives de standardisation comme le W3C[4], les ontologies sont équipées de nombreux outils partagés pour l'édition, les requêtes, l'alignement, etc.

Dans le cadre d'OMNIA, on pourrait envisager d'utiliser l'ontologie comme pivot linguistique, à la manière de l'IF dans Nespole !. L'ontologie devrait alors être directement reliée aux langues naturelles visées, comme illustré dans la figure 4(a). Cette idée peut sembler naturelle, mais elle mène à de nombreux problèmes, bien connus en lexicographie multilingue (Mangeot *et al.,* 2003). Il a ainsi été avancé, dans (Tze, 2009), que les ontologies ne sont pas des structures adaptées au rôle de pivot linguistique. De plus, une situation idéale comme celle de la figure 4(a) ne peut être atteinte que si l'on dispose au préalable de ressources multilingues suffisantes. Séparer l'ontologie

4. http ://www.w3.org/



(a) Comme pivot sémantique (b) Liée à un pivot linguistique

Figure 4. *Places possibles pour une ontologie dans une architecture multilingue*

et le pivot linguistique nous permet donc à la fois de nous affranchir des problèmes d'intégration de ces deux représentations formelles, et d'utiliser des ontologies et des pivots linguistiques déjà existants.

Dans notre approche, illustrée par la figure 4(b), nous avons choisi d'utiliser les UW comme lexique pivot et l'ontologie comme un paramètre du domaine, qui peut être changé pour améliorer l'extraction de contenu sur des données spécifiques. Nous avons donc à faire à deux types de symboles :

– d'une part, les étiquettes de l'ontologie qui représentent des concepts ou des relations entre ces concepts ;

– d'autre part, les UW qui représentent des acceptions (sens de mots) dans plusieurs langues.

Il est donc nécessaire de relier ces deux ensembles de symboles en tenant compte des contraintes suivantes :

– **La création manuelle d'une telle correspondance étant coûteuse** à cause de la taille des ressources, des procédés automatiques sont nécessaires.

– **Les ontologies et les lexiques évoluent** : un alignement doit donc pouvoir s'adapter à des évolutions incrémentales des ressources.

– **La correspondance doit être manipulée aisément par les utilisateurs** humains ou logiciels.

4.3. *Spécification et calcul d'un alignement entre ontologie et lexique*

La construction et le maintien d'un alignement entre une ontologie et un lexique est une tâche délicate (Rouquet *et al.,* 2009b, Prévot *et al.,* 2005). Afin de bénéficier au mieux des technologies du Web Sémantique, nous utilisons des données formatées suivant les recommandations du W3C. Ainsi, nous considérons des ontologie expri-

mées en OWL[5] et notre lexique d'UW est présenté dans le format SKOS[6] dérivé de OWL. Les ressources utilisées et produites (ontologies, dictionnaires et alignements) sont disponibles sur le site Web Kaiko [7].

Nous utilisons les définitions suivantes, adaptées de celles trouvées dans (Euzenat *et al.,* 2007) pour les alignements entre ontologies.

Définition 1 (Correspondance) *Étant donnés une ontologie O et un lexique L, une correspondance est un quadruplet : $\langle e, e', r, n\rangle$ où $e \in O$ est une entité (e.g., formules, termes, classes, individus) de l'ontologie et $e' \in L$ est une entrée du lexique ; r est la relation entre e et e', parmi l'ensemble des relations d'alignement (e.g., $\equiv$, $\sqsubseteq$, ou $\sqsupseteq$) ; et $n \in [0\ 1]$ est le degré de confiance associé à la relation.*

Définition 2 (Alignement) *Un alignement A est un ensemble de correspondances.*

Dans les expériences préliminaires, nous utilisons deux méthodes d'alignement automatique développées à l'aide de l'API d'alignement décrite dans (Euzenat, 2004). Elles sont basées sur des méthodes simples de comparaison de chaîne et seront améliorées de deux façons :

1) en utilisant des synonymes (par exemple les synsets de WordNet) pour trouver plus de correspondances ;

2) en adaptant des méthodes classiques de TALN à la désambiguïsation des alignements (pour le calcul des scores de confiance).

5. Processus d'extraction générique

Dans OMNIA, les résultats de l'extraction de contenu textuelle doivent pouvoir être interprétés dans une perspective multimodale, conjointement avec les résultats d'analyse visuelle réalisés par les partenaires, qui ne reposent pas sur une ontologie. Par conséquent le résultat final de l'extraction de contenu consiste en une liste autonome (non liée à l'ontologie) des concepts identifiés dans le texte, associées à un score de vraisemblance. Les concepts extraits peuvent au besoin être présentés dans différents formalismes : CSV, axiomes de l'ontologie (*A-box*), requêtes SQL ou SPARQL etc.

Le processus est décomposé en trois phases, comme présenté dans la figure 5.

5. http ://www.w3.org/2004/OWL/
6. http ://www.w3.org/TR/skos-reference/
7. http ://kaiko.getalp.org

5.1. *Annotation conceptuelle*

À cette étape, nous disposons d'une part d'un graphe-Q annoté par des UW, et d'autre part d'un alignement entre le lexique des UW et les concepts de l'ontologie. Dans un premier temps, suivant cet alignement, on marque, pour chaque UW du graphe, l'éventuel concept correspondant.

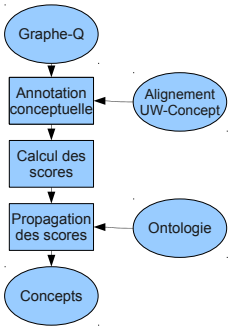


Figure 5. *Étapes de l'extraction de contenu*

Un certain nombre de concepts sont ainsi identifiés, mais pas tous ; l'alignement se fait sur les *headwords* des UW, mais les restrictions de ces UW ne sont pas utilisées. Or, les restrictions de type *icl*, *iof* et *equ*, qui indiquent respectivement une relation de spécification, instanciation, équivalence, et peuvent être utilisées à défaut d'un alignement explicite. Par conséquent, dans un second temps, pour les UW auxquelles on n'a pas pu assigner de concept précédemment, on identifie un concept si l'un des headwords auquel il est associé est présent dans l'une des restrictions de type *icl*, *iof* et *equ*.

Cela est particulièrement utile pour les entités nommées. Par exemple, si l'on a identifié l'UW *Grenoble(iof>city>thing)* dans un texte, mais que l'ontologie n'a pas de concept pour *Grenoble*, la recherche dans l'alignement UW-Concept ne donnera rien. Mais si l'ontologie contient un concept pour *city*, il pourra être identifié grâce à la restriction *iof>city>thing*.

5.2. *Calcul des scores et propagation*

Lors des étapes d'annotation et d'extraction, nous calculons des scores de confiance sur la qualité des données produites. En particulier, le score de confiance calculé pour un concept extrait d'une légende (score de l'UW $\times$ score de la correspondance UW-concept) quantifie notre certitude quand à la présence du concept dans l'image associée. Dans l'ontologie de classification des images du projet OMNIA, on dira que l'image est une instance du concept. Afin d'exploiter ces scores de façon cohérente en lien avec des ontologies, nous avons choisi d'utiliser la théorie des ensembles flous (Zadeh, 1965) comme modèle de nos ontologies. Ainsi, un score est

interprété comme un degré d'appartenance flou (d'une image) à un concept de l'ontologie.

Pour l'indexation d'images dans le projet OMNIA, nous souhaitons obtenir des résultats autonomes (interprétables indépendamment de l'ontologie utilisée en paramètre). Les degrés d'appartenance sont donc propagés dans la hiérarchie de l'ontologie en utilisant des opérateurs ensemblistes flous.

5.3. *Le cas des requêtes*

Dans le cas de l'analyse des textes, nous utilisons toutes les étapes décrites précédemment. Mais dans le cas de l'analyse des requêtes en langage naturel de l'utilisateur, le traitement fait l'économie de la propagation des concepts, afin de limiter la généralisation. Par exemple, si dans un texte on extrait le concept HOSPITAL, et que l'ontologie le fait dépendre du concept BUILDING, il serait hasardeux de faire porter la recherche sur ce concept BUILDING. Par contre, la propagation de concepts effectuée au préalable sur les textes est intéressante, car elle permet par exemple, pour une requête mentionnant explicitement le concept plus général de BUILDING, de retrouver des textes où le concept de BUILDING n'est pas explicitement mentionné, mais seulement ses dérivés comme HOSPITAL, HOUSE, etc.

6. Premières expérimentations

Nous avons développé deux environnements d'expérimentation. Le premier propose une interface informatique REST pour l'analyse de corpus entiers, permettant ensuite de récupérer les résultats intermédiaires et finaux pour chaque texte ; il permet d'évaluer la qualité de l'analyse des textes. Le second présente une interface graphique en ligne, permettant à un utilisateur d'effectuer des requêtes, et d'afficher les images correspondantes ; il permet une évaluation des requêtes en contexte.

6.1. *Implémentation*

La chaîne de traitements textuels est implémentée suivant une architecture orientée services (SOA) dans laquelle chaque processus correspond à un service Web. Les données passent d'un service à l'autre et les résultats intermédiaires peuvent être consultés au besoin.

Nous pouvons ainsi utiliser des ressources existantes, appelées par des interfaces REST (Fielding, 2000) ou de simples formulaires HTML et les changer au besoin, de façon modulaire. Un "superviseur" a été développé pour gérer ces interfaces Web hétérogènes et les problèmes de normalisation des données (encodages, *cookies*, etc.).

De plus, cette architecture est capable de gérer plusieurs tâches en parallèle, ce qui est intéressant pour le traitement des requêtes des utilisateurs en temps réel.

6.2. *Premier environnement : analyse et indexation de textes*

À l'aide de cet environnement, nous avons effectué une première expérimentation, portant sur 4.046 textes du corpus Belga-News, choisis par le coordinateur du projet. L'ontologie utilisée[8] comporte 732 classes dans les domaines suivants : animaux, politique, religion, armée, sports, monuments, transports, jeux, divertissements, affects, etc. Les classes de l'ontologie sont liées à environ 2.000 UW[9].

Sur un processeur Athlon II 240 (2x2,8GHz), le temps de calcul complet (annotation interlingue et conceptuelle) pour un texte d'une cinquantaine de mots est de 4832 ms, avec une fenêtre de 5 segments pour la lemmatisation et l'utilisation d'un cache pour les accès à Jibiki. Ce temps prend en compte la lemmatisation (259 ms), l'annotation interlingue (174 ms), la désambiguïsation (4 292 ms) et l'annotation conceptuelle (116 ms), ce qui, pour prendre l'exemple du projet OMNIA, ouvre notamment la voie au traitement de flux textuels (dans cet exemple, légendes d'images) ainsi qu'au traitement de requêtes utilisateur en temps réel.

Lemmatisation	259 ms
Annotation interlingue	174 ms
Désambiguïsation	4292 ms
Annotation conceptuelle	36 ms
Calcul et propagation des scores	80 ms
Total	4832 ms

Tableau 1. *Temps de calcul des différents services, par texte de 48,5 mots en moyenne.*

On trouve en moyenne 6 concepts par texte, et 23% des textes sont "muets" (aucun concept trouvé), mais cela concerne surtout des textes très courts, voire comportant juste le nom d'un lieu ou d'une personne, ou une date. En complément, nous avons défini un protocole d'évaluation subjective selon deux critères :

1) **L'adéquation visuelle** considère qu'un concept trouvé est correct si il est porté par au moins un élément de l'image. Par exemple le concept SPORT sera considéré comme correct pour une image montrant un ministre des sports, même si l'image ne le montre pas en train de pratiquer un sport.

2) **L'adéquation textuelle** considère qu'un concept trouvé est correct si il est effectivement porté par le texte, indépendamment de sa présence dans l'image ; ce peut être par exemple un élément de contexte.

Une première étape de validation de cette approche a porté sur un échantillon 30 textes. 124 concepts ont été trouvés dans 23 textes (pour 7 textes, aucun concept n'a été trouvé). 99 concepts (80%) étaient visuellement adéquats, 110 (89%) l'étaient textuellement.

8. http ://kaiko.getalp.org/kaiko/ontology/OMNIA/100606_OMNIA_v6.owl
9. http ://kaiko.getalp.org/kaiko/link/Kaiko_align_UWpp-OMNIAv6_StringEq.rdf

À titre d'exemple, nous avons extrait les concepts suivants de la légende d'image de la figure 6.

AWA05 - 20020924 - BAGHDAD, IRAQ : Iraqi women sit under a portrait of Iraqi President Saddam Hussein in a waiting room in Baghdad's al-Mansur hospital 24 September 2002. Saddam Hussein is doggedly pursuing the development of weapons of mass destruction and will do his best to hide them from UN inspectors, the British government claimed in a 55-page dossier made public just hours before a special House of Commons debate on Iraq. Iraqi Culture Minister Hamad Yussef Hammadi called the British allegations "baseless." EPA PHOTO AFPI AWAD AWAD

Figure 6. *Image et légende extraits de la base Belga-News.*

CONCEPT	SCORE
BUILDING	0.098
HOSPITAL	0.005
HOUSE	0.043
MINISTER	0.016
OTHER_BUILDING	0.005
PEOPLE	0.142
PERSON	0.038
POLITICS	0.032
PRESIDENT	0.016
RESIDENTIAL_BUILDING	0.043
WOMAN	0.005

6.3. *Second environnement : analyse de requêtes*

Le second environnement permet d'étudier les requêtes, et il est prévu de l'utiliser pour une évaluation adaptée à la tâche, selon deux scénarios : recherche d'image «classique» (par mots clefs ou courte requête en langage naturel), et prépresse (recherche d'images pouvant illustrer un texte donné).

Il permet d'effectuer des recherches selon trois modes, soit en utilisant uniquement les concepts, pour une meilleure précision, soit en utilisant uniquement les UW, pour une meilleure couverture, soit en combinant les deux, suivant un facteur de pondération paramétrable. Les requêtes sont en langage naturel, en anglais ou en français,

et permettent de rechercher des images, indépendamment de la langue de la légende (anglais ou français). L'algorithme de recherche en lui-même est très simple, dans la mesure où il sort du cadre du projet ; il consiste simplement à cumuler les scores des éléments trouvés (concepts ou UW, selon le mode), et à classer les images de la base en fonction de ce score. Un module permet en outre d'explorer possibilités de désambiguïsation interactive offertes par notre approche.

7. Conclusion

Nous avons présenté dans cet article une approche originale pour la multilinguisation de la RI basée sur des ontologies, disinguant formellement l'information linguistique (le signifiant), de l'information ontologique (le signifié). Cette approche permet la réutilisation de ressources linguistiques et ontologiques existantes. On conserve ainsi les possibilités d'inférence de l'ontologie, et l'expressivité de la ressource linguistique. Notre système est générique à deux niveaux :

– il est indépendant de la langue, dans la mesure où il repose sur une représentation interlingue,

– le contenu à extraire peut être spécifié, en passant une ontologie de domaine en paramètre du système.

Nous analysons entièrement une légende d'image de 50 mots en moins de 5 secondes, à l'aide d'une machine à base d'Athlon II, un processeur de bureau d'entrée de gamme, soit plus de 20 000 légendes par jour. Ce temps de calcul est suffisant pour traiter à la volée de grands flux d'images légendées, comme celles de la base *Wikimedia Commons*, qui croissait d'environ 6 000 images par jour en 2010[10].

8. Bibliographie

Aitchenson J., *Words in the Mind. An Introduction to the Mental Lexicon*, Blackwell Publishers, 2003.

Baader D. F., Calvanese D., McGuinness D., Patel-Schneider P., Nardi D., *The Description Logic Handbook*, Cambridge University Press, 2003.

Blanchon H., Boitet C., « Speech Translation for French within the C-STAR II Consortium and Future Perspectives », *Proc. ICSLP 2000*, Beijing, China, p. 412-417, 2000.

Boitet C., Boguslavskij I., Cardeñosa J., « An Evaluation of UNL Usability for High Quality Multilingualization and Projections for a Future UNL++ Language », *Computational Linguistics and Intelligent Text Processing*, p. 361-373, 2009.

Colmerauer A., « Les systèmes-q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur », *département d'informatique de l'Université de Montréal, publication interne*, September, 1970.

10. http ://fr.wikipedia.org/wiki/Wikimedia_Commons

Daoud D., Il faut et on peut construire des systèmes de commerce électronique à interface en langue naturelle restreints (et multilingues) en utilisant des méthodes orientées vers les sous-langages et le contenu, PhD thesis, UJF, September, 2006.

Euzenat J., « An API for Ontology Alignment », *Proceedings of the 3rd International Semantic Web Conference*, Hiroshima, Japan, p. 698-7112, 2004.

Euzenat J., Shvaiko P., *Ontology matching*, Springer, Heidelberg (DE), 2007.

Fielding R. T., Architectural styles and the design of network-based software architectures, PhD thesis, University of California, 2000.

Hajlaoui N., Boitet C., « Portage linguistique d'applications de gestion de contenu », *TOTh07*, Annecy, 2007.

Mangeot M., Lafourcade M., « Collaborative building of a multilingual lexical database : Papillon project », , vol. Electronic dictionaries : for humans, machines or both ?, n° 44 :2/2003, p. 151-176, 2003.

Marchesotti L., et al., « The Omnia Project (accessed on may 2010) », http ://www.omnia-project.org, May, 2010.

Metze F., McDonough J., Soltau H., Waibel A., Lavie A., Burger S., Langley C., Levin L., Schultz T., Pianesi F., Cattoni R., Lazzari G., Mana N., Pianta E., « The Nespole ! Speech-to-Speech Translation System », *Proceedings of HLT-2002 Human Language Technology Conference*, San Diego, USA, march, 2002.

Nguyen H., Boitet C., Sérasset G., « PIVAX, an online contributive lexical data base for heterogeneous MT systems using a lexical pivot », *SNLP*, Bangkok, Thailand, 2007.

Nguyen H.-T., « EMEU_w, a simple interface to test the Q-Systems (accessed on september 2009) », http ://sway.imag.fr/unldeco/SystemsQ.po ?localhost= /home/nguyenht/SYS-Q/MONITEUR/, September, 2009.

Prévot L., Borgo S., Oltramari A., « Interfacing ontologies and lexical resources », *Workshop on Ontologies and Lexical resources (OntoLex2005)*, 2005.

Rouquet D., Falaise A., Schwab D., Blanchon H., Bellynck V., Boitet C., Dellandréa E., Liu N., Saidi A., Skaff S., Marchesotti L., Csurka G., « Classifiation multilingue et multimédia pour la recherche d'images dans le projet OMNIA », *atelier Recherche d'Information SÉmantique (RISE)*, Marseille, France, 2010.

Rouquet D., Nguyen H., « Interlingual annotation of texts in the OMNIA project », *4th Language and Technology Conference (LTC09)*, Poznan, Poland, 2009a.

Rouquet D., Nguyen H., « Multilinguïsation d'une ontologie par des corespondances avec un lexique pivot », *TOTh09*, Annecy, France, May, 2009b.

Tze L. L., « Multilingual Lexicons for Machine Translation », *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*, Kuala Lumpur, Malaysia, p. 732-736, 2009.

Zadeh L., « Fuzzy sets », *Information and Control*, vol. 8, n° 3, p. 338-353, 1965.

Exploiting and Extending a Semantic Resource for Conceptual Indexing

Karam ABDULAHHAD* — Jean-Pierre CHEVALLET — Catherine BERRUT***

** UJF-Grenoble 1, ** UPMF-Grenoble 2, LIG laboratory, MRIM group*

{karam.abdulahhad,jean-pierre.chevallet,catherine.berrut}@imag.fr

ABSTRACT. *Information Retrieval Systems that compute a matching between a document and a query based on terms intersection, cannot reach relevant documents that do not share any terms with the query. The objective of this study is to propose a solution to this problem in the context of conceptual indexing. We study an ontology-based matching that exploits links between concepts. We propose a model that exploits the weighted links of an ontology. We also propose to extend the links of the ontology to reflect the structural ambiguity of some concepts. A validation of our proposal is made on the test collection ImagCLEFMed 2005 and the external resource UMLS 2005.*

RÉSUMÉ. *Les Systèmes de Recherche d'Information qui calculent la correspondance entre un document et une requête à base d'intersection de termes, ne peuvent pas atteindre les documents pertinents qui ne partagent aucun termes avec la requête. L'objectif de ce travail de master est alors de proposer une solution à ce problème dans le cadre d'une indexation par concepts. Nous étudions une correspondance basée sur une ontologie qui exploite les liens entre les concepts. Nous proposons un modèle de correspondance qui exploite la pondération des liens de l'ontologie. Nous proposons également d'étendre les liens de l'ontologie pour tenir compte de l'ambigüité de structure de certains concepts. Une validation de notre proposition est effectuée sur la collection de test ImagCLEFMed 2005 et la ressource externe UMLS 2005.*

KEYWORDS: *term mismatch, concept mismatch, Bayesian matching, conceptual indexing*

MOTS-CLÉS : *variation terminologique, correspondance Bayésien, indexation conceptuelle*

1. Term and concept mismatch

Information Retrieval Systems IRSs based on term [1] intersection to compute a matching between a document and a query, suffer from *term mismatch* problem. This problem appears when users write a query using terms different from terms in relevant document. For example, the following two terms '*Skin Cancer*' and '*melanoma*' have a close meaning in a medical context. In less than 20% of cases, two people use the same term to describe the same meaning (Crestani, 2000). So without an external resource that links these two terms, we cannot retrieve a document containing '*Skin Cancer*' as a response to a query containing '*melanoma*'.

To solve the term mismatch problem the first step is: *using concepts* [2] *instead of terms* (Chevallet *et al.,* 2007). Using concepts solves a part of the problem when different terms correspond to the same concept, e.g. the two terms "*Atrial Fibrillation*" and "*Auricular Fibrillation*" correspond to the same concept "*C0004238*" in UMLS [3]. However, when two terms corresponds to two concepts, and these two concepts have a relation, this relation must be used for matching. For example, the two terms "*B-Cell*" and "*Lymphocyte*" correspond to the two concepts "*C0004561*" and "*C0024264*" respectively, and there is a relation of type "*isa*" between these two concepts. Here, and without exploiting the relations between concepts, we get the same problem but at the conceptual level "*concept mismatch*".

The previous problem can be solved by using conceptual relations during the matching process (Le, 2009).

Using concepts and conceptual relations supposes the existence of external resources that encompass them. However, external resources are *incomplete*. We found out that many potential relations based on the syntax of terms are missing. For example, in UMLS there are five concepts containing the word "*spirochaete*", so we have twenty pairs of concepts that potentially have a linguistic relation, but we did not find any relation (see Table 1).

In this work, we propose to enrich the external resource by adding more relations between concepts, and in this way we hope to enhance system's recall. We must, however, be careful in building and using relations, because building too many relations may decrease precision.

---

1. A term is a noun phrase that has a unique meaning in a specific domain (e.g. medical domain) and that belongs to a terminology (Baziz, 2005) (Chevallet, 2009).

2. "Concepts" can be defined as "Human understandable unique abstract notions independent from any direct material support, independent from any language or information representation, and used to organize perception and knowledge" (Chevallet *et al.,* 2007). In IR domain, to achive the conceptual indexing, each concept is associated to a set of terms that describe it (Baziz, 2005) (Chevallet, 2009).

3. Unified Medical Language System. It is a meta-thesaurus in medical domain. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls

Table 1. *Statistics from UMLS*

word	# concepts	# all concepts pairs	# pairs with relation
device	86,985	7,566,303,240	161,660
activity	22,395	501,513,630	380,052
sedum	98	9,506	122
spirocheate	5	20	0

2. Proposed model

Our model bases on using concepts as indexing elements, enriching external resource by new relations, and then using these relations at matching time. The model consists of three main components:

1) External Resource: The external resource is used in conceptual indexing to map text to concepts. It contains terms $T = \{t_1, t_2, \ldots\}$, concepts $C = \{c_1, c_2, \ldots\}$ linked by relations $R = \{r | r \subseteq C \times C\}$. Each concept corresponds to several terms, then a function can be defined: $\zeta : C \rightarrow T^*$ where T^* is the set of all subsets of T. We enrich the external resource by:
a) Adding relations between concepts: e.g. "*shared-words*": this relation means that there are words in common between two concepts. Formally, there is a shared-words relation between two concepts $C_1 = \{t_1^{c1}, t_2^{c1}, \ldots, t_k^{c1}\}$ and $C_2 = \{t_1^{c2}, t_2^{c2}, \ldots, t_l^{c2}\}$ iff $NSW > 0$, where: $NSW = |C_1 \cap C_2|$ the number of shared words between the two concepts. NSW could be calculated because each concept corresponds to terms and each term corresponds to a sequence of words.
b) Defining a Certainty property to distinguish relations already defined in the external resource R_C from relations added by us $R_{\neg C}$. The Certainty represents how much we are sure that there is a semantic relation between two concepts. The hypothesis here is: if there is a document d contains a concept c_d, a query q contains a concept c_q, and if there is a relation of type R_C (e.g. *isa*) between c_d and c_q. Then it is more probable that d is relevant document for q than if the relation between c_d and c_q is of type $R_{\neg C}$ (e.g. *shared-words*). $R_C \cap R_{\neg C} = \emptyset$.
c) Defining the notion of '*Strength of relation*' which represents the ability of a relation to retrieve relevant documents of a query. In other words, the strength assigned to a relation between two concepts C_1 and C_2 measures the extent to which if a document talks about C_1, it also talks about C_2 (Nie, 1992). We calculate the strength of a relation by using the following formula $\forall r \in R, \forall (c_i, c_j) \in r$:

$$Strength_r(c_i, c_j) = sim_r(c_i, c_j) \times certainty(r) \quad [1]$$

Where:

$$\forall r \in R, \quad certainty(r) = \begin{cases} 1 & r \in R_C \\ x \in]0,1[& r \in R_{\neg C} \end{cases} \quad [2]$$

$sim_r(c_i, c_j)$ represents the semantic similarity between two concepts.

Finally we define the conceptual indexing function $Index$: suppose there are a query q and a collection of documents D then:

$$Index : D \cup \{q\} \rightarrow C^* \quad [3]$$

where, C^* is the set of all subsets of C

2) Bayesian Network: To compute the matching between a document and a query, we use a Bayesian network (Murphy, 1998)(Le, 2009). The network in our model contains three types of nodes: documents D, concepts C, and query q. Nodes are connected by using three types of weighted links:
(1) $L_{DC} = \{(d, c) \,|d \in D, c \in Index\,(d)\}$: links from documents to their concepts, weighted by the importance of concept in its document.
(2) $L_{CQ} = \{(c, q) \,|c \in Index\,(q)\}$: links from concepts to their query, weighted by the importance of concept in the query.
(3) $L_{CC} = \{(c_i, c_j) \,|\exists d \in D, c_i \in Index\,(d)\,, c_j \in Index\,(q)\,, \exists r \in R, (c_i, c_j) \in r\}$: links from documents' concepts to query's concepts, represent relations between concepts, weighted by the *strength* of the relation.

3) Matching function: To calculate RSV (Relevance Status Value), we use the calculation rules of the conditional probability in Bayesian network, according to the following steps:
a) choosing a document $d_{selected}$ from document collection D, then:

$$\forall d \in D, \quad P\,(d) = \begin{cases} 1 & d = d_{selected} \\ 0 & \text{else} \end{cases} \quad [4]$$

b) for concepts that belong to the selected document $\{c_i|\,(d_{selected}, c_i) \in L_{DC}\}$:

$$P\,(c_i|L_{DC}) = \frac{weight_{DC}\,(d_{selected}, c_i)}{\sum_{(d_j, c_i) \in L_{DC}} weight_{DC}\,(d_j, c_i)} \quad [5]$$

c) for concepts that belong to the query and don't belong to the selected document and that are linked to a concept of the selected document $\{c_i|c_i \in Index\,(q)\,, c_i \notin Index\,(d_{selected})\,, \exists c_j \in Index\,(d_{selected})\,, (c_j, c_i) \in L_{CC}\}$:

$$P\,(c_i|L_{CC}) = \frac{\sum_{(c_j, c_i) \in L_{CC}} weight_{CC}\,(c_j, c_i) \times P\,(c_j|L_{DC})}{\sum_{(c_j, c_i) \in L_{CC}} weight_{CC}\,(c_j, c_i)} \quad [6]$$

d) now for the query node $RSV\,(d_{selected}, q) = P\,(q|L_{CQ})$:

$$P\,(q|L_{CQ}) = \frac{\sum_{(c_i, q) \in L_{CQ}} weight_{CQ}\,(c_i, q) \times P\,(c_i|L_{CC})}{\sum_{(c_i, q) \in L_{CQ}} weight_{CQ}\,(c_i, q)} \quad [7]$$

3. Model validation context

We validated the proposed model by applying it to the test collection: ImageCLEFMed2005, and by using the UMLS 2005 as an external resource. We used

MetaMap (Aronson, 2006) tool to identify concepts from raw text, we program a tool to build Bayesian network and calculate correspondence value, and we use the $tf.idf$ measure to calculate the importance of a concept in its document.

The goal of these experiments is showing that by enriching the external resource, more relevant documents could be retrieved. We have tested three variants of the model:
(1) Basic: there is no relations between concepts, i.e. it depends on the shared concepts between a document and a query to find matching.
(2) ISA: the *isa* relation (*isa*: this relation is predefined in UMLS) is used to link documents' concepts and query's concepts. Here,
$isa \in R_C$ then we have $certainty\,(isa) = 1$
$\forall\,(c_i, c_j) \in isa, \quad sim_{isa}\,(c_i, c_j) = \frac{1}{minLen(c_i, c_j)}$ where $minLen\,(c_i, c_j)$ is the path of minimum length between c_i and c_j according to isa.
b) ISA_SW: another relation (*shared-words*: this relation is added by us to UMLS) is added to ISA. Here,
$shared - words \in R_{\neg C}$ then we have $certainty\,(shared - words) = 0.1$ (10% is the value that gives the best result in our experiments)
$sim_{shared-words}\,(c_i, c_j) = mutual\_information\,(c_i, c_j) = \frac{NSW_{ij}}{NW_i \times NW_j}$
Where:
NSW_{ij}: number of shared words between c_i and c_j
NW_i: number of words in c_i
NW_j: number of words in c_j

We got the following results (see Tables 2, 3).

Table 2. *MAP of Basic, ISA, ISA_SW*

	MAP
Basic	0.1240
ISA	0.1395
ISA_SW	**0.1408**

Table 3. *Number of relevant, retrieved, and retrieved-relevant documents of Basic, ISA, ISA_SW*

	# Relevant documents	# Retrieved documents	# Retrieved-Relevent
Basic	2217	58037	1234
ISA	2217	101182	1464
ISA_SW	2217	128342	**1698**

From the previous results we can notice that, by exploiting relations between concepts (ISA), we could retrieve more relevant documents for a query (see Table 3) and

at the same time we gain a small enhancement in the precision of the system (see Table 2).

Also we can notice that, the enrichment of the external resource by adding a very simple relation (ISA_SW), allows us to retrieve more relevant documents (see Table 3) and also gain more enhancement in the precision (see Table 2).

4. Conclusion

We have presented in this paper our model to solve term and concept mismatch problems. In this model, documents and queries are represented by concepts, and we have also modeled the different relations between concepts.

This model also depends on the techniques of Bayesian Network to compute the matching value between a document and a query.

We show in this work that conceptual indexing is insufficient to solve the term mismatch problem. The use of relations from the conceptual resource increase the MAP, but we think that in UMLS, too many potential relation between concepts are missing. When we add these relations, we show an interesting increase in the MAP. In conclusion, these research tend to show that existing resources even very larges ones like UMLS, are not totally adapted to IR because of lack of relations between concepts. This lack can be partly compensated by analysis of terms associated to concepts. Finally there are many points in this work, that need more study, like studying the influence of adding other relations to the model, using properties other than Certainty to describe relations, and validation the model by using another test collections and another external resources.

A detailed version of this paper with different and more comprehensive experiments could be found in (Abdulahhad *et al.,* 2011).

5. References

Abdulahhad K., Chevallet J.-P., Berrut C., « Solving Concept mismatch through Bayesian Framework by Extending UMLS Meta-Thesaurus », *la huitième édition de la COnférence en Recherche d'Information et Applications (CORIA 2011), Avignon, France*, March 16–18, 2011.

Aronson A. R., « Metamap: Mapping text to the umls metathesaurus », 2006.

Baziz M., Indexation conceptuelle guidée par ontologie pour la recherche d'information, Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre, 2005.

Chevallet J.-P., « endogènes et exogènes pour une indexation conceptuelle intermédia », Mémoire d'Habilitation a Diriger des Recherches, 2009.

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.

Crestani F., « Exploiting the similarity of non-matching terms at retrieval time », *Journal of Information Retrieval*, vol. 2, p. 25-45, 2000.

Le T. H. D., Utilisation de ressource externes dans un modèle Bayesien de Recherche d'Information: Application a la recherche d'information médicale multilingue avec UMLS, PhD thesis, Université Joseph Fourier, Ecole Doctorale MSTII, 2009.

Murphy K., « A Brief Introduction to Graphical Models and Bayesian Networks », 1998.

Nie J.-Y., « Towards a probabilistic modal logic for semantic-based information retrieval », *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, ACM, New York, NY, USA, p. 140-151, 1992.

Quel cadre d'évaluation pour la RI sémantique ?

Haïfa Zargayouna

Laboratoire d'Informatique de l'université Paris-Nord (LIPN) - UMR 7030
Université Paris 13 - CNRS
haifa.zargayouna@lipn.univ-paris13.fr

RÉSUMÉ. *La recherche d'information sémantique (RIS) a pour but de dépasser les limites d'une recherche classique par mots-clés. De plus en plus de travaux s'intéressent à l'exploitation de ressources sémantiques (ontologies, terminologies, thésaurii, etc.) pour améliorer l'accès à l'information. Ces travaux sont à l'intersection de deux communautés: recherche d'information et Web sémantique. Ils s'intéressent entre autres à des questions d'annotation et indexation sémantique, désambiguïsation, extraction d'information (termes, entités nommées), expansion de requêtes, etc.*

Malgré les expériences accumulées, il est encore difficile de dresser un bilan. Ceci est principalement dû à l'absence d'une base de test commune qui permettrait de comparer entre eux les différents algorithmes et méthodes.

Cette présentation fait un état des lieux des différentes expériences d'évaluation qui peuvent être réutilisées et montre les spécificités de l'évaluation de l'apport d'une « sémantique » à la recherche d'information.

Une première expérience d'évaluation sur un corpus de recettes de cuisine est présentée. Elle permet de dresser un premier bilan.

Un effort commun est nécessaire pour pouvoir établir un cadre ouvert et partagé pour l'évaluation et qui aiderait à mesurer les avancées mais aussi à mettre en commun les algorithmes et les ressources. La prise en compte d'une sémantique est très liée à la qualité des ressources, au domaine, à la langue, etc. L'évaluation de l'ensemble reste une tâche difficile et il est nécessaire d'identifier les briques de base d'un moteur de recherche sémantique pour évaluer leurs apports respectifs.

A l'instar d'autres bancs de test d'évaluation, cette évaluation devrait suivre un cycle de vie itératif (spécifications, mise en place et recalibrage) qui permettrait d'enrichir incrémentalement la base de tests .

MOTS-CLÉS : *évaluation, recherche d'information sémantique, ressources, mesures, campagnes*