Validation syntaxique de relations sémantiques pour la RI

Loïc Maisonnasse

Laboratoire CLIPS-IMAG
BP 53
38 041 Grenoble cedex 9
loic.maisonnasse@imag.fr

RÉSUMÉ. *Avec l'objectif d'améliorer la précision des systèmes de recherche d'information, c'est-à-dire les premiers résultats retrouvés par le système, des travaux se sont basés sur des indexations structurées des documents, à base d'arbres ou de graphes. La plupart de ces travaux utilisent comme index des structures uniques et certaines. Les décisions qui ont amené à la sélection de certaines informations lors de la création de la structure à partir du texte ne sont plus disponibles et ne sont pas utilisées. Ce type d'information nous parait pourtant essentiel pour obtenir des résultats précis. Nous proposons ici une méthode permettant de donner un poids d'extraction à des relations sémantiques à partir des éléments syntaxiques qui les composent dans le texte. Pour valider ce poids, nous intégrerons cette pondération dans un modèle de recherche d'information basé sur des graphes de concepts et nous évaluerons ce modèle sur la collection CLEF-Image 2005.*

ABSTRACT. *In the purpose to improve precision of information retrieval systems, that mean in improving first results find by the system, some work used structured index of document based on tree or graph. Most of this work use a certain and unique structure. Decisions that have leaded the building process are no more available. This kind of information must be used to improve precision. We propose here a method that gives an extraction weight on semantic relation by using their syntactic component. We integrate this weight in an information retrieval model based on graph of concepts. We evaluate this model on the collection CLEF-Image 2005.*

MOTS-CLÉS : *recherche d'information, analyse syntaxique, dépendance syntaxique, relation sémantique*

KEYWORDS: *information retrieval, shallow parsing, syntactic dependency, semantic relation*

1. Introduction

L'amélioration de la précision des systèmes de recherche d'information (RI) textuelle passe par l'ajout d'informations qu'elles soient endogènes (issues du document lui-même) ou exogènes (nécessitant l'utilisation de bases de connaissances). Dans de nombreux travaux, l'intégration de ces informations au sein de l'index est obtenue en utilisant des descripteurs plus complets que les simples mots clefs, tels que des concepts. Un concept étant défini comme l'abstraction d'un ensemble de termes. L'information peut aussi être intégrée dans l'index en structurant ces éléments, par exemple par l'intégration de relations typées entre les concepts. Même si de tels index peuvent améliorer les résultats, la complexité du traitement de la langue et de la modélisation des informations d'un domaine a pour conséquence que ces index structurés peuvent contenir des informations imprécises ou erronées. Nous promouvons ici l'utilisation d'un score de certitude sur les éléments de l'index qui rend compte de l'imprécision. Nous utilisons un index structuré constitué de concepts et de relations sémantiques (RS) entre concepts extraits à partir du texte. Sur cet index nous proposons d'utiliser un score de certitude sur les RS. Le processus d'extraction des RS ne fournit pas d'informations sur leur validité. Nous basant sur les résultats de travaux en question réponse (QR), nous proposons ici de fournir un score de certitude sur les RS en validant chacune par son instanciation syntaxique, c'est à dire le lien syntaxique qui relie ses concepts.

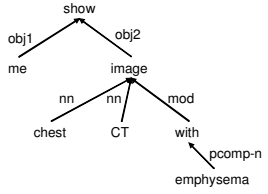
2. L'extraction de relations basées sur la syntaxe

L'extraction de RS basées sur des informations syntaxiques a été étudiée en QR pour sélectionner les passages contenant la réponse. Cette phase du QR est très importante car elle affecte les résultats du système. Elle a pour but de déterminer les passages dont le contenu sémantique est similaire à celui de la question. Retrouver les termes de la question n'est pas suffisant, la plupart des passages non pertinents trouvés par le système contenant les termes mais avec des RS différentes de celle de la question. Pour améliorer la correspondance entre les relations de la question et celle des passages, les systèmes de QR détectent les relations dans les phrases. Pour cela, des systèmes se basent sur l'extraction de relations à base de patrons sur les mots et de règles produites manuellement, c'est notamment le cas de (Jacquemin, 1997). Cependant d'autres approches ont tenté d'extraire les relations de manière automatique. Dans (Lin, 1998), l'auteur propose une méthode pour inférer automatiquement les relations utiles pour le QR à partir des relations contenues dans les textes. Pour cela l'auteur utilise l'analyseur syntaxique en dépendance Minipar. Cet analyseur produit une analyse en dépendance qui représente la phrase sous la forme d'un arbre où les nœuds sont les mots et les liens des relations syntaxiques (tel que sujet). Cet arbre est utilisé par l'auteur pour extraire des chemins syntaxiques entre des termes. Un chemin syntaxique est l'ensemble des lemmes et

des relations syntaxiques qui relie deux mots dans l'arbre de dépendance d'une phrase. L'apprentissage des relations est ensuite effectué en se basant sur l'hypothèse de distribution suivante : si deux chemins syntaxiques relient les mêmes ensembles de mots alors ces chemins sont identiques et représentent une même relation. Les auteurs proposent ensuite un apprentissage basé sur l'information mutuelle. L'algorithme est évalué en comparant les règles d'inférence produites par le système avec des paraphrases proposées par un humain, pour 15 questions de la tâche de QR de TREC-8. Le nombre de paraphrases en commun ces ensembles est faible, cependant beaucoup d'éléments dans les deux semblent corrects. Dans (Hang *et al.*, 2005), les auteurs proposent de comparer les chemins syntaxiques liant les éléments de la requête et les chemins liant ces mêmes éléments dans les phrases réponses. Les auteurs se basent sur une correspondance floue entre les chemins et testent deux apprentissages, leurs résultats sur TREC-12 montrent une forte amélioration de la sélection des phrases candidates.

3. Proposition

Nous utilisons un graphe sémantique pour indexer des documents. Le contenu d'un document est alors représenté par un graphe unique composé de concepts et de RS. Nous considérerons un concept comme l'abstraction de plusieurs termes de la langue naturelle. Ces concepts sont définis par une base de connaissance et sur un domaine précis. Les RS sont les relations binaires typées (ex : location of (chest, emphysema)), elles sont définies comme reliant des classes de concepts. Une classe de concepts est un ensemble de concepts possédant des caractéristiques communes. Dans la majorité des bases de connaissances, le lien entre les concepts et les termes qu'ils abstraient est bien défini. L'indexation des concepts d'un document est possible à l'aide de méthodes simples telles que de la détection de termes (Radhouani *et al*., 2006). Pour les RS, les bases de connaissances ne définissent que les concepts reliés par la relation. Elles ne permettent pas de savoir si la relation apparait ou non dans un document. Ici nous proposons de construire un graphe sémantique qui contient toutes les RS définies entre deux concepts et nous ajoutons un poids de certitude qui représente la validité d'une relation par rapport au contenu du document, en nous inspirant des méthodes utilisées en QA..



relation	Chemin syntaxique
T135 (C0817096, C0034067) location of(chest, emphysema)	(nn)image(mod)With (pcomp-n)
T135 (C0202823, C0817096) location of (chest, chest CT)	(ident)

Figure 1. *Extraction des chemins syntaxique*

3.1. *Pondération des relations sémantiques*

Nous construisons un modèle de relations pour chaque RS définie dans la base de connaissances. Nous émettons l'hypothèse que la validité d'une RS peut être évaluée en connaissant son instance syntaxique. Cette instance syntaxique étant le chemin syntaxique qui relie les deux concepts de la relation (voir Figure 1). Nous considérons pour une relation *Rs* la probabilité que le chemin syntaxique *ch* qui lui correspond soit généré par le modèle générique de cette relation M_{Rs}. Nous créons deux modèles différents :

Modèle 1 : dans le premier, un chemin syntaxique est considéré comme un ensemble d'éléments indépendants (modèle unigramme) $ch = \{elem\}$, où les *elem* sont indifféremment des lemmes ou des relations. La probabilité d'un chemin *ch* pour une relation *tr* est alors :

$$P(ch|M_{tr}) = \prod_{elem \in ch} P(elem|M_{tr}) = \prod_{elem \in ch} \frac{N(elem, tr)}{N(*, tr)} \quad (1)$$

où *N(elem,tr)* est le nombre de fois que *elem* apparaît dans la relation *tr* et *N(*,tr)* est le nombre d'éléments de la relation *tr*.

Modèle 2 : dans le deuxième, nous construisons un modèle bigramme où le chemin est considéré comme un ensemble de couples $ch' = \{(lemme, Rsyn)\}$ constitué de *lemmes* et de relations syntaxiques *Rsyn*, la probabilité d'un chemin *ch'* pour une relation *tr* est alors :

$$P(ch'|M_{tr}) = \prod_{(lemme,Rsyn) \in ch'} (\lambda) P((lemme, Rsyn)|M_{tr}) + (1-\lambda) P(lemme|M_{tr}) P(Rsyn|M_{tr})$$
$$P((lemme, Rsyn)|M_{tr}) = \frac{N((lemme, Rsyn), tr)}{N((*,*), tr)} \quad (2)$$

où *N((lemme,Rsyn),tr)* est le nombre de fois où le couple *(lemme,Rsyn)* apparaît dans la relation *tr* et *N((*,*),tr)* est le nombre total de couples de la relation *tr*.

En regard de ces modèles, nous considérerons qu'un chemin syntaxique est valide pour une RS si la probabilité de générer ce chemin à l'aide du modèle de cette relation est supérieure à la probabilité de générer ce chemin à l'aide d'un modèle de relations génériques. Nous calculons donc le score de vraisemblance suivant pour établir la validité d'une relation :

$$cert(ch, tr) = \frac{P(ch, M_{Rs})}{P(ch, M_{Rg})}$$

où $P(ch, M_{Rg})$ est la probabilité du chemin syntaxique *ch* pour le modèle de relations génériques. Ce modèle est obtenu sur l'ensemble des RS du corpus.

4. Expérimentations

Nous présentons ici les résultats que nous avons obtenus sur la collection Image-CLEFmed 2005 (Clough *et al.* 2005). Le corpus Image-CLEFmed 2005 est composé de rapports médicaux en différentes langues associés à des images médicales. Les 25 requêtes de la base ImageCLEFmed 2005 ont été formulées avec des images-exemples et de courtes descriptions textuelles. Pour chaque requête, l'ensemble des images pertinentes est fourni. Nous procédons à nos expérimentations seulement sur la partie textuelle et anglaise du corpus et nous évaluons directement la pertinence des rapports en considérant un rapport pertinent si une de ces images est pertinente.

Nous avons utilisé le méta thésaurus UMLS comme ressource externe pour l'indexation conceptuelle et pour la détection des relations sémantiques. La méthode de détection des concepts est similaire à celle de (Radhouani *et al.,* 2006), mais utilise l'analyseur Minipar. Une fois les concepts détectés, nous détectons les relations entre ces concepts. UMLS établit au sein d'un réseau sémantique des relations sémantiques entre des types de concepts. Deux concepts entretiennent une relation sémantique *Rs* s'ils sont dans la même phrase et si leurs types entretiennent la relation *tr* dans le réseau sémantique. Pour chaque relation sémantique détectée, le chemin syntaxique est extrait en se basant sur l'analyse syntaxique fournie par Minipar. Sachant que deux concepts d'une relation sémantique peuvent avoir la même tête, nous ajoutons la relation syntaxique *ident* pour de tels cas.

A l'aide du système expérimental XIOTA (Chevallet, 2004), nous indexons les graphes et nous calculons une correspondance entre les concepts et les relations de la requête et celles des documents. Pour chaque relation sémantique le score de validité est calculé. Au sein de l'index le poids des relations et des concepts est calculé par un *tf*. Pour les relations, si la validité est prise en compte, ce score est multiplié avec la somme des scores de validité de chaque relation du document. Les résultats obtenus par nos scores de certitude sont présentés dans les tableaux 2 et 3 et peuvent être comparés aux résultats sans certitude présentés dans le tableau 1.

	Précision moyenne	Précision à 5 documents
Concepts	0.2363	0.4
Relations	0.2056	0.33
Graphe	0.2535	0.384

Tableau 1. *Résultat sans certitude*

	Précision moyenne	Précision à 5 documents
Relations	0.2185	0.4300
Graphe	0.2635	0.4480

Tableau 2. *Résultat avec certitude selon le modèle 1*

	Relations		Graphe	
λ	Précision moyenne	Précision à 5 documents	Précision moyenne	Précision à 5 documents
0	0.2635	0.4480	0.2185	0.4300
0.2	0.2612	0.4320	0.2107	0.4000
0.5	0.2590	0.4320	0.2139	0.4100
0.7	0.2567	0.4400	0.2133	0.4000
1	0.2515	0.4240	0.2083	0.3900

Tableau 3. *Résultat avec certitude selon le modèle 2*

Les résultats montrent que l'utilisation de certitude améliore la précision moyenne de l'indexation à base de graphes d'environs 3,9 %. Cette amélioration semble provenir de l'amélioration de la précision à 5 documents sur les relations qui passent de 0,33 à 0,43. Par comparaison aux concepts, seuls les résultats obtenus sur les graphes avec nos mesures de certitude montrent une amélioration de la précision moyenne mais aussi de la précision à 5 documents, ce qui n'était pas le cas auparavant. La comparaison des deux modèles montre que le modèle prenant en compte les bigrammes est moins efficace que le modèle unigramme. Notre corpus n'est pas assez grand pour que les bigrammes soit discriminant.

5. Conclusion

Ce travail montre l'intérêt d'intégrer un score de certitude sur les relations sémantiques en RI. Nous avons proposé une méthode pour obtenir ce score qui se base sur l'instance syntaxique de la relation sémantique à pondérer. Cette méthode, basée sur des modèles de chemins syntaxiques, donne des résultats encourageants. Afin d'obtenir des résultats plus probants, l'utilisation de données d'apprentissage semble nécessaire, notamment un corpus contenant des relations annotées. Les modèles de chemins syntaxiques pourraient eux aussi être améliorés en prenant compte les étiquettes syntaxiques des lemmes du chemin. En effet un verbe sera plus discriminatif d'une relation qu'une préposition.

6. Bibliographie

Chevallet J.P., « X-iota: An open xml framework for IR experimentation. » *Computer Science (LNCS)*, AIRS'04 Beijing, p. 263-280, 2004

Hang C., Renxu S., Keya L., Min-Yen, K., Tat-Seng, C., « Question answering passage retrieval using dependency relations », *ACM SIGIR*, Salvador, Brazil, p 400-407, 2005

Clough P., Müller H., Deselaers T., Grubinger M., Lehmann T.M., Jensen J.R., Hersh W.R., « The CLEF 2005 Cross-Language Image Retrieval Track. » *CLEF 2005*, p 535-557, 2005

Jacquemin C., *Variation terminologique : reconnaissance et acquisition automatique des termes et de leurs variantes en corpus*, Habilitation à diriger des thèses, Université de Nantes, Nantes, 1997.

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

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