

# Extraction et structuration des relations multi-types à partir de texte

LE Thi Hoang Diem, Jean-Pierre CHEVALLET

{Thi-Hoang-Diem.Le, Jean-Pierre.CHEVALLET}@imag.fr

**Abstract**— Les relations entre termes jouent toujours un rôle très important dans la représentation de contenu du texte. Ces relations qu’elles soient de nature statistique, syntaxique ou sémantique, leur information abordée reflète le contexte au niveau local ou global des termes et contribuent donc à la compréhension du texte. Dans le cadre de ce travail, nous proposons une méthode hybride d’extraction, de structuration et de filtration des relations multi types dans le texte. La base de connaissances extraite pourra éventuellement être appliquée à l’expansion de requête en vue d’améliorer la performance de recherche d’information. Nous avons également évalué lors d’une expérimentation les différentes relations extraites.

**Mot clés**— Relation, extraction d’information, base de connaissances, expansion de requête

## I. INTRODUCTION

Nous partons du contexte général des Systèmes de Recherche d’Information (SRI): problème de recherche avec termes - mot clés. Les limitations des mots clés sont claires ; ils ne sont pas assez précis au niveau de description du contenu du texte ; Les documents pertinents ne partagent pas toujours les mêmes mots clés avec ceux de la requête. De ce fait, les connaissances d’un plus haut niveau de compréhension du texte comme connaissances syntaxiques et sémantiques sont exploitées et intégrées dans le SRI, notamment par l’utilisation à l’indexation de syntagmes ou de terme structuré [1], [2] et pour expansion de requête [3].

Un système d’extraction d’information ou de connaissances est un système qui produit une représentation de l’information textuelle pertinente dans un domaine particulier et pour une application particulière. Les connaissances extraites doivent représenter le contenu des textes sous une forme compréhensive par l’ordinateur et riche en information. Cette base de connaissances extraite peut être utilisée dans les deux phrases principales d’un SRI. Premièrement, pour effectuer une indexation à vocabulaire contrôlé, les termes d’indexations sont alors plus complets et plus précis, ils permettent d’atteindre une meilleure performance. Deuxièmement, en consultant ce base de connaissances, les termes similaires avec ceux de la requête sont retrouvés et utilisés dans reformulation de la requête.

Notre méthode d’extraction de connaissances s’appuie concrètement sur l’extraction des termes significatifs et représentatifs du contenu informationnel du corpus et les relations entre les couples de termes.

## II. TRAVAUX SIMILAIRES

Les travaux dans la littérature du domaine d’extraction de connaissances textuelles foisonnent d’idées et de techniques. On peut les grouper en trois approches principales :

### A. Approche statistique

Cette approche observe la régularité des termes dans un contexte déterminé d’une grande masse de collection de texte. Elle est basée sur hypothèse dans [4] : « *L’emploi de deux termes en cooccurrence est l’expression d’une relation sémantique entre eux* ». Afin d’exploiter la relation entre les termes, de nombreuses mesures de similarité ont été utilisées dans la littérature statistique. L’avantage de cette approche est qu’elle est facile à mettre en œuvre et indépendante du corpus. Parmi des travaux de cette approche, [5] contribue au domaine de l’extraction de connaissances par une technologie d’extraction des contextes des mots à partir des corpus textuels pour produire la liste des mots reliés à n’importe quel mot apparu dans le corpus. Ces mots reliés utilisés dans l’expansion de requêtes donnent de meilleurs résultats que ceux obtenus par un seul caractère de cooccurrence. Cette technologie utilise une analyse syntaxique de surface sans aucune autre connaissance linguistique.

### B. Approche linguistique

A l’inverse de l’approche statistique, l’approche linguistique n’observe pas la régularité des termes dans le corpus. Ce qui est important est les informations linguistiques exploitées à partir du texte. Autrement dit, cette approche vise à extraire les dépendances ou les relations entre les termes grâce aux phénomènes langagiers. Des travaux récents utilisent ces technologies:

- Exploitation des marqueurs linguistiques dont [6] est un travail profond en compréhension du langage naturelle. Elle exploite les marqueurs de causalité (par exemple : « à cause de », « *parce que* » ...) afin d’extraire la relation causale pour l’intégrer dans le système de question-réponse KALIPSO. [7] est un travail similaire sur ce type de relation en s’appuyant en plus sur le rôle que joue cette relation dans les SRI.
- Exploitation des patrons syntaxiques: travaux de [8] et [9] exploitent des patrons syntaxiques dans l’extraction de relation d’hyponymie. Exemple des patrons d’hyponymie: «SN tel que LIST », « SN comme LIST »,...

Même si cette approche atteint un niveau plus élevé de compréhension du texte, et même si l'on obtient donc une meilleure capacité de représentation de contenu du corpus, elle n'est pas facile à mettre en œuvre car il faut beaucoup de connaissances pour résoudre les ambiguïtés de la langue naturelle.

### C. Approche hybride

Les travaux récents se centrent sur la combinaison de l'approche statistique et linguistique, afin de profiter de l'avantage des deux approches. Cette approche hybride utilise souvent des données statistiques pour filtrer les données linguistiques. SEXTANT [10] et IOTA [11] sont les systèmes qui effectuent cette approche et montrent des résultats très encourageants. Plus récent, [12] vise à construire une base de connaissances grâce à une technologie de fouille de donnée, appelée règles d'association et grâce à une extraction des syntagmes nominaux.

## III. PROPOSITION

L'approche hybride d'extraction de connaissances montre son efficacité dans l'augmentation de la performance de RI. D'ailleurs, les travaux en profondeur d'extraction d'un seul type de relation sémantique sont pertinents mais ces approches sont coûteuses et adaptées seulement à un certain type de requêtes. L'approche statistique n'est plus utilisée toute seule car elle est trop limitée [13]. En effet, les SN sont de bonnes représentations du texte. Les éventuelles relations entre SN extraites sont intéressantes à exploiter. Inspirés des constatations sur ces travaux récents, nous sommes intéressés aux apports bilatéraux entre les informations sémantiques et statistiques :

- Apports d'informations sémantiques aux informations statistiques : information linguistique représentée sous forme de relations sémantiques précise le sens d'association entre deux termes qui co-occurrent. Ces relations sémantiques sont plus compréhensibles et représentatives du contenu textuel que les statistiques.
- Apports d'informations statistiques aux informations sémantiques : Quant à elles, les informations statistiques ont un caractère global et indépendant du corpus de textes. Aussi, elles sont faciles à mettre en œuvre, elles sont donc souvent utilisées pour filtrer les relations sémantiques et pour donner un poids sur l'importance d'une relation.

A partir des apports conjoints entre informations sémantiques et statistiques, nous proposons une idée sur la coopération de ces deux types d'information dans l'extraction de relation de type syntaxique, statistique et sémantique, ainsi que dans la représentation du contenu du texte avec les termes et leurs relations extraites. Les relations statistiques sont filtrées par la combinaison de la statistique ou de la sémantique. L'idée est basée sur les deux hypothèses suivantes:

*H1: Si deux termes co-occurrent souvent, et s'ils sont reliés*

*au moins une fois par un type de relation sémantique donné, alors toutes ses relations de cooccurrences sont de ce type.*

*H2: La relation sémantique (synonymie, hyponymie et causale) entre deux termes ne co-occurrent pas très souvent est aussi importante et doit être extraire. L'importance des termes dépend non seulement de leurs fréquences, de leurs quantités d'information mais aussi de leurs liens sémantiques avec les autres.*

Nous proposons donc une approche hybride en fusionnant des associations entre les couples de termes extraits de manière statistique avec les relations sémantiques extraites par approche linguistique.

### A. Chaîne de traitements d'extraction d'informations:

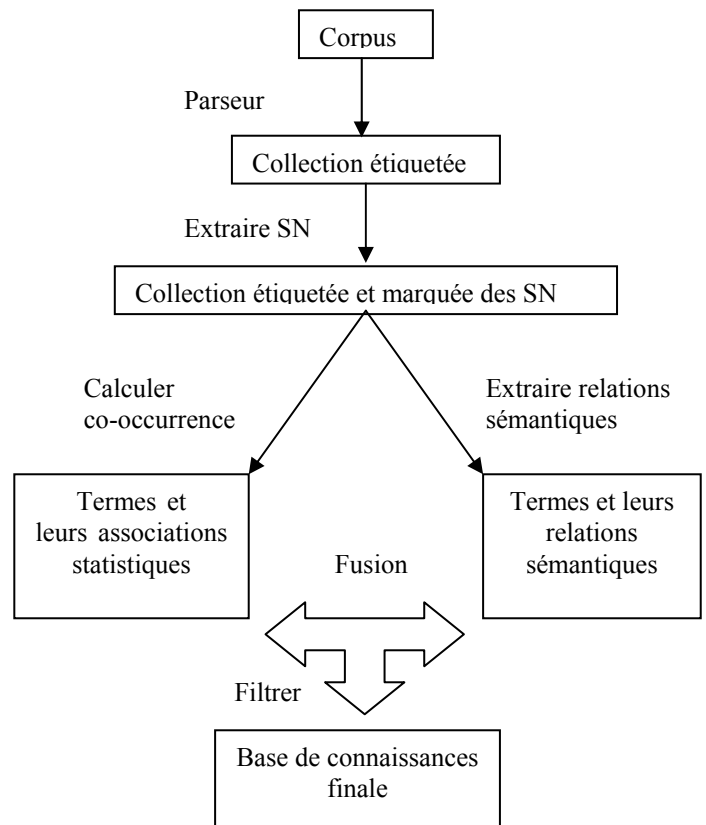


Fig. 1. Chaîne de traitements d'extraction de connaissances

#### 1) Étiquetage

A cette étape, une technique de traitement automatique de langage naturel est appliquée pour donner une étiquette de catégorie grammaticale pour chaque unité linguistique dans le texte. Nous utilisons un parseur pour cette tâche.

Ex : « SystèmeSUBC dePREP informationSUBC estVBCJ unARTD nouveauADJQ domaineSUBC .PNTF »

#### 2) Extraction de SN à l'aide de patrons syntaxiques

Un patron syntaxique est une règle sur l'ordre d'enchaînement des catégories grammaticales formées un SN.

Ex : ADJQ SUBC

« Premier ministre, grande école », etc.

SUBC PREP SUBC

« job d'été », « guerre en Irak », etc.

On utilise un module d'extraction de SN de IOTA pour réaliser cette tâche.

### 3) Calcul de fréquence et de Cooccurrence entre termes :

A cette étape, nous proposons de calcul la cooccurrence (terme, terme) par une variation de la formule Cosinus [13] d'origine dans [14].

4) Extraction de relations sémantiques: Relation synonyme, hyponyme et causale sont extraites grâce aux patrons lexico\_syntaxique :

Ex : Patron lexico\_syntaxique : SN1 comme LISTE (SN2, SN3,...)

Notre objectif n'est pas d'extraire des relations avec tous les patrons lexico-syntaxique possibles, mais avec seulement quelques patrons principaux pour ensuite, fusionner avec les données statistiques et enfin filtrer, évaluer la base de connaissances finales.

### 5) Fusion des deux graphes des termes

Nous proposons d'ajouter des nouveaux termes et relations, ou renommer les relations au nom des relations sémantiques a priori si elles coïncident.

Exemple de fusion entre le graphe des termes et leurs associations statistiques/syntaxique avec celui des termes et leurs relations sémantiques :

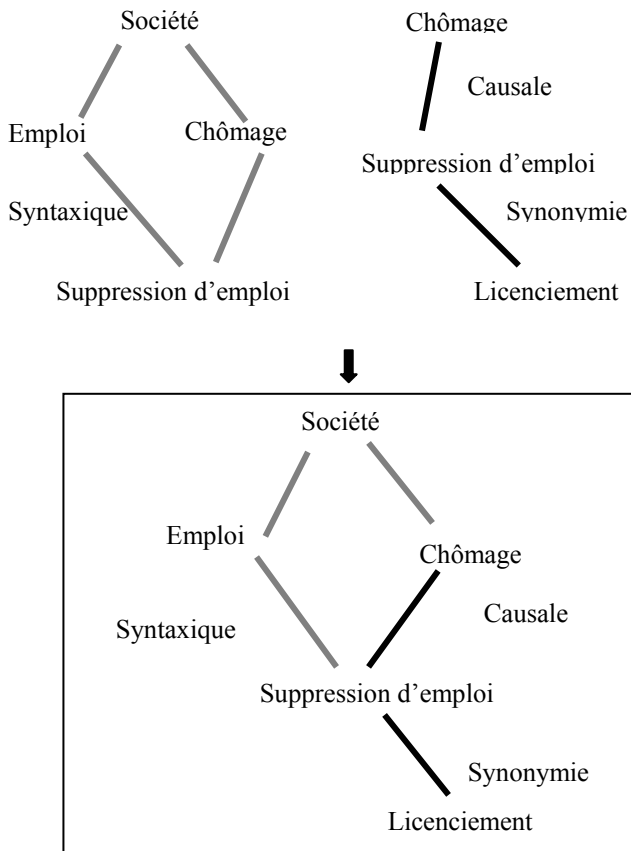


Fig. 2. Exemple de fusion des graphes de termes

### 6) Le filtrage des termes

Pour éliminer les termes qui ne sont pas importants, on utilise un seuil de filtrage. Les termes d'une valeur de filtrage inférieure au seuil prédéfini sont éliminés. Le seuil est défini

expérimentalement afin d'obtenir le meilleur taux de précision/rappel.

## B. Structuration de connaissances

Le résultat du travail d'extraction de connaissances est représenté sous la forme d'une liste de termes avec leurs relations, qui doivent être présentées aux utilisateurs tel qu'ils puissent les consulter selon différents critères et fournir les suggestions lors de la phrase d'interrogation pour faciliter ou augmenter la qualité de la recherche. C'est la raison pour laquelle nous avons besoin de la structuration des connaissances extraites. Avec une approche fusion de l'information statistique, syntaxique et sémantique, la structuration doit refléter ces trois informations pour pouvoir les consulter de façon flexible. Nous adoptons la structuration en réseau de dépendance syntaxique (tête et expansion) en ajoutant les associations statistiques et sémantiques. Notre hypothèse de structuration est :

*Les SN qui partagent la même tête, co-occurrent souvent ou sont reliés ensemble par une des relations sémantiques (Synonyme, hyperonyme, causale) représentent le même thème.*

Cette structuration permet à la fois de situer sur les relations à l'intérieur d'un SN (relations entre expansions), ainsi que relations à l'extérieur des SN. L'objectif est de récupérer en même temps dépendance syntaxique, statistique et sémantique pour fournir une vue cohérence sur le contenu de corpus.

La structuration des SN en tête et expansion permet de mettre en évidence non seulement la dépendance syntaxique entre SN mais aussi la relation hyponymie entre un SN et ses variations (Ex : système, système d'information, système d'information multimédia) et co-hyponymie entre SN qui partage la même tête (Ex: « Système automatique », « système informatique ») Cette structuration en tête et expansion se base sur des patrons syntaxiques. Par exemple:

Patrons syntaxiques	Tête	Expansion
SUBC1 SUBC2	SUBC1	SUBC2
SUBC ADJQ	SUBC	ADJQ
ADJQ SUBC	SUBC	ADJQ
SUBC1 PREP SUBC2	SUBC1	SUBC2

Tab.1. Exemple des patrons syntaxiques de SN

Nous adoptons la structuration des SN dans [12] et en ajoutons une mesure de qualité sémantique. Un SN est donc défini comme :

$$SN : [T, Freq, Qinf, Q_{sem}]$$

**T** : la tête de SN

Ex : *Système d'information* => Tête : *Système*

**Freq** : La fréquence dans le corpus entier.

**Qinf** : La quantité d'information

Cette mesure a pour but de résoudre le problème de basse fréquence des termes complexes. Même si les termes complexes sont de bonnes représentations du contenu textuel,

ils n'apparaissent pas aussi souvent dans corpus que les termes simples. Pour équilibrer l'importance des SN par rapport aux termes simples, la quantité d'information a pour but d'évaluer l'importance de l'information qu'apporte un SN :

$$Q_{\text{inf}}(\text{SN}) = \sum_{a \in \text{SN}} Q_{\text{inf}}(\text{Cat}(a))$$

Où **Cat(a)** : Catégorie grammaticale de chaque mot qui forme le SN (Substantif, adjectif,...)

**Qinf(Cat(a))** : Valeur prédéfinie pour chaque catégorie grammaticale. Plus la catégorie est porteuse d'information plus cette valeur est grande. Dans ce cas, les substantifs ont la valeur plus élevée tandis que les prépositions, conjonctions et articles ont une quantité d'information nulle.

On prédéfinit heuristiquement la valeur  $\alpha$  pour exprimer la quantité d'information de la catégorie subjonctifs et  $\beta$  pour les catégories adjectif, verbe à l'infinitif, participe passé et adverbe avec  $\alpha \gg \beta$ .

Ex : SN = nouveau système d'information

$$\begin{aligned} Q_{\text{inf}}(\text{SN}) &= \text{inf}(\text{Cat}(\text{nouveau})) + Q_{\text{inf}}(\text{Cat}(\text{système})) \\ &\quad + Q_{\text{inf}}(\text{Cat}(\text{information})) \\ &= Q_{\text{inf}}(\text{Cat}(\text{Adjectif})) + Q_{\text{inf}}(\text{Cat}(\text{Substantif})) \\ &\quad + Q_{\text{inf}}(\text{Cat}(\text{Substantif})) \\ &= \beta + 2\alpha \end{aligned}$$

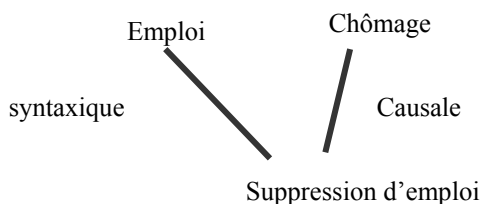
$Q_{\text{sem}}$  est la *qualité sémantique* que le SN apporte. Nous proposons cette mesure en supposant que les relations sémantiques sont les connaissances les plus fines et les plus fortes au sens sémantique. Dans la phase d'extraction du contenu textuel, on exploite au fur à mesure toutes les informations possibles afin de trouver les relations. Lorsque l'on a trouvé une relation, nous pensons qu'il faut la prendre en compte dans l'évaluation de l'importance des termes qui sont reliés par cette relation. Nous basons sur l'hypothèse que : la qualité sémantique des relations sémantiques est beaucoup plus forte que celle de relation syntaxique et statistique. Cela reflète l'importance des termes qui sont reliés par une relation sémantique. Plus un terme possède des relations sémantiques, plus il obtient une forte qualité sémantique. De ce fait, nous définissons  $Q_{\text{sem}}$  d'un terme SN par rapport aux relations auxquelles il est relié.

$$Q_{\text{sem}}(\text{SN}) = \sum_{i=1}^n Q_{\text{sem}}(\text{Rel}_i)$$

$Rel_i$  est le type de relation qui relie terme T avec un autre terme.

$Q_{\text{sem}}(\text{Rel}_i)$  est une valeur prédéfinie pour évaluer la force sémantique de la relation  $Rel_i$ . On définit une valeur  $\gamma > 1$  pour chaque relation d'un terme avec une autre par relation synonyme, hyperonyme ou causale, et  $\gamma > \delta > 1$  pour exprimer la qualité sémantique des termes reliés avec les autres par dépendance syntaxique ou statistique.

Ex :



$$\begin{aligned} \text{On a : } Q_{\text{sem}}(\text{suppression d'emploi}) &= \\ &Q_{\text{sem}}(\text{Dépendancesyntaxique}) + Q_{\text{sem}}(\text{Causale}) \\ &= \delta + \gamma \end{aligned}$$

$$\begin{aligned} Q_{\text{sem}}(\text{emploi}) &= Q_{\text{sem}}(\text{Dépendance syntaxique}) \\ &= \delta \end{aligned}$$

$$\begin{aligned} Q_{\text{sem}}(\text{emploi}) &= Q_{\text{sem}}(\text{Causale}) \\ &= \gamma \end{aligned}$$

### C. Filtrage des SN

Le nombre de termes extraits à partir d'un corpus volumineux est très important. De ce fait, il ne faut pas tous les conserver. On garde seulement ceux qui sont importants pour leur capacité à représenter le contenu du corpus. Cette capacité dépend non seulement de la fréquence et de la quantité d'information mais aussi de la qualité sémantique ( $Q_{\text{sem}}$ ) via leurs associations avec les autres.  $Q_{\text{sem}}$  reflète la force sémantique d'un terme par rapport à sa corrélation sémantique avec les autres termes. Le filtrage est donc défini par une fonction qui doit refléter cette dépendance multiple de la fréquence, la quantité d'information et de la qualité sémantique. Nous proposons la fonction de filtrage des SN :

$$F(\text{SN}) = \text{Freq}(\text{SN}) \times Q_{\text{inf}}(\text{SN}) \times Q_{\text{sem}}(\text{SN})$$

Via un seul prédéfini expérimentalement, terme SN n'est gardé si  $F(\text{SN})$  supérieur au seuil.

## IV. EXPERIMENTATION

Dans un premier temps, nous avons effectué une expérimentation d'extraction des relations sémantiques et de filtrage du résultat obtenu. En utilisant simplement des patrons lexico-syntaxique prédéfinis, nous pouvons extraire du corpus des relations de synonymie, hyperonymie et causalité. Le corpus utilisé est le corpus « le Monde » de taille 150Mo de la campagne d'évaluation CLEF.

Relation	Patrons
Synonymie	SN1, appelé aussi SN2 SN1, ou SN2
hyperonymie	SN1 comme LISTE (SN2, SN3,...) SN (LISTE (SN2, SN3,...))
Causalité	SN1 à cause de SN2

Tab.2. Relations et leurs patrons

Exemple des informations intéressantes extraites :

« Pays d'Europe centrale comme Pologne, Irlande, Hongrie, Slovaquie, République Chèque »

« Organisation de défense de droit de l'homme comme Human Right Watch, Human Right In China,... »

« Comédien comme Jack Nicholson »

« Statut universitaire (professeur, maître de conférence) »

Exemples de bruits de relation hyperonyme avec patron «SN COMME LISTE» :

- A cause de la structure complexe de la phrase :

«Son arrière-grand-père a officié dans l'armée française comme son grand-père, son père et ses trois beaux-frères »

On extrait des relations bizarres:

armée\_français -COMME- grand-père  
 armée\_français -COMME- père  
 armée\_français -COMME- beau-frère

- A cause d'ambiguïté du sens :

«Rapide, silencieux, stable, la navette se présente *en fait* comme un train tiré de sorte de caisson container dans lequel peut prendre place environ 90 voitures. »

On extrait : fait -COMME- train

Il faut d'ailleurs reconnaître l'expression avec « en fait » pour éliminer ce bruit.

- A cause d'étiquetage incorrect

«Un autre roman d'amour nous est rapporté dans les trente derniers pages de livre et *produitSUBC* comme un *coup\_de\_théâtre* »

On extrait : produitSUBC -COMME- coupSUBC\_dePREP\_théâtreSUBC

La méthode d'extraction des relations sémantiques donne des résultats intéressants mais aussi des limites comme nous venons de montrer. Pour évaluer nos résultats d'extraction de relations, à cause de la voluminosité du résultat et la nature sémantique des relations, dans ce premiers temps, nous avons mesuré le taux de précision en comptant manuellement le nombre de relation correcte parmi une centaine des relations extraites sélectionnées au hasard. Même si le taux de précision du résultat n'est pas encore élevé (moyenne de 60%), des informations intéressantes ont été trouvées par une analyse simple du texte.

Afin d'améliorer la précision, nous avons filtré les termes par rapport à leur fréquence et leur quantité d'information. Le résultat obtenu après dans le tableau montre une petite augmentation de précision. Pourtant, sur un petit corpus Le Monde (150M), le filtrage des SN basé sur leurs fréquences n'est pas efficace car presque tous les termes ont la fréquence égale 1.

Filtrage 1	Filtrage 2
Freq >= 2 1 < Qinf < 7	1 < Qinf < 7

Tab.3. Les filtres des termes

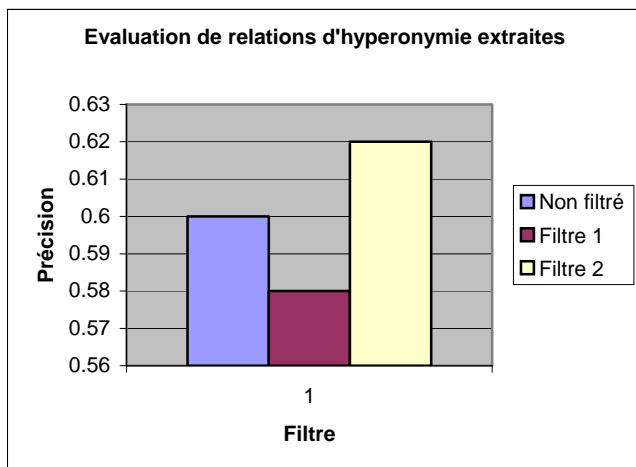


Fig.4. Statistique du taux de la précision avant et après le filtrage

	Nombre des relations extraites	Précision
Non filter	2331	0.6
Filtre 1	67	0.58
Filtre 2	1890	0.62

Tab.4. Statistique du taux de précision des relations extraites avant et après les filtrages

Grâce au filtre 2, le taux de précision des relations extraites a été augmenté de +2%. Pourtant, cette évaluation de précision n'est correcte que pour les premières centaines des relations. Un jugement sur le total des relations extraites est donc prévu. Les éventuelles améliorations raisonnables du parseur de langue française sur un plus large corpus et de la fonction de filtrage pourraient donner un meilleur résultat.

## V. CONCLUSION

Dans le cadre de ce travail, nous avons proposé une méthode d'extraction de connaissances et un filtrage qui fusionne à la fois l'approche linguistique et statistique. Pour l'approche statistique, un calcul de la fréquence de termes et leurs occurrences est effectué. L'approche linguistique est effectuée à partir d'une analyse de surface du texte afin d'obtenir des syntagmes nominaux et d'autre part grâce à l'exploitation de patrons lexico-syntaxique afin de repérer les relations sémantiques entre les termes (mot simple ou syntagmes nominaux). Ces deux types d'informations sur les termes sont fusionnés et filtrés à l'aide d'informations de nature statistique, syntaxique ainsi sémantique.

Nous avons proposé une méthode d'extraction de relations en fusionnant les relations statistiques, syntaxiques et sémantiques. Une structuration des connaissances extraites avec les trois sources d'informations différentes donne une vue générale d'un terme et de son contexte. L'aspect statistique avec la fréquence, l'aspect syntaxique avec la structure tête - expansion, la mesure de la quantité d'information et l'aspect sémantique avec les relations sémantiques. La mesure de la qualité sémantique permet d'évaluer l'importance d'un terme et sa contribution à la représentation du contenu du corpus. Un filtrage des termes basé sur ces trois types d'informations a été proposé pour l'adapter à ce contexte.

Du côté expérimental, nous avons effectué un travail d'extraction de relations sémantiques à l'aide de patrons lexico-syntaxique. Parmi les relations extraites, nous avons trouvé des informations intéressantes mais aussi des bruits dus à : l'ambiguïté du langage naturel, la structure complexe des phrases et les étiquettes incorrectes. Afin d'éliminer ces bruits, nous avons filtré les termes par leur fréquence et leur quantité d'information.

Le travail présenté peut être complété et étendu pour différentes applications. La plus importante est l'utilisation de cette base de connaissances dans l'expansion automatique de requête :

- Elargir le champ de la recherche : En ajoutant les termes qui relient sémantiquement ceux de la requête, en particulier les termes synonymes. Dans le cas d'une requête donnée par utilisateur si les termes ajoutés ont un faible pouvoir de discrimination, les termes ajoutés plus discriminants permettront une meilleure recherche.
- Focaliser le besoin d'information de l'utilisateur: La relation hyperonyme est très efficace dans la focalisation de l'information demandée. Dans ce cas si l'utilisateur ne précise pas assez son besoin d'information. Cela entraîne des bruits dans la recherche. Les termes proposés à l'utilisateur à l'aide de cette base de connaissances focaliseront donc la recherche.

D'ailleurs, la base de connaissance peut également servir à d'autres tâches comme: présenter le contenu du corpus à l'utilisateur pour lui permettre de mieux percevoir le contenu effectif de l'ensemble des documents, de mieux exprimer son besoin d'information dans construction de la requête ; filtrer et classer les documents du Web, etc.

#### REFERENCES

- [1] A.T. Arampatzis, Th.P. van der Weide, P. van Bommel et C.H.A. Koster, *Linguistically Motivated Information Retrieval*, Encyclopedia of Library and Information Science, Marcel Dekker Inc, New York, Basel, 2000
- [2] Jean-Pierre Chevallet, Hatem Haddad, *Proposition d'un modèle relationnel d'indexation syntagmatique:mise en oeuvre dans le système IOTA*, INFORSID 2001, Genève-Martigny, pp. 465-483, 2001
- [3] E.M.Voorhees. *Query expansion using lexical-semantic relations*. In SIGIR 94, page 61-69.ACM,1994.
- [4] C.J. Van. Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.
- [5] Gregory Grefenstette. *Use of syntactic context to produce terms association list for text retrieval*. In Conference in Recherche and Developement in Information Retrieval (SIGIR'92), Copenhagen, Danmarke, pages 89-97, juin 1992.
- [6] Adeline Nazarenko. *Compréhension du langage naturel: le problème de la causalité*. Thèse de doctorat, 1994 [Mor99] Emmanuel Morin, extraction de liens sémantiques entre termes à partir de corpus de textes techniques, Thèse de doctorat, 1999.
- [7] Christopher Soo-Guan Khoo. *Automatic identification of causale relations in text and their use for improving precision in information retrieval*. P.D thesis 1995.
- [8] M.A.Hearst. *Automatic acquisition of hyponyms from large text corpora*. In Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, pages 539-545, juillet 1992.
- [9] Emmanuel Morin. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse de doctorat, 1999.
- [10] Gregory Grefenstette. *Automatic thesaurus generation from raw text using knowledge-poor technique*. In Rapport de recherche Xerox MLTT - 001, Sept 1993.
- [11] M.F.Bruandet. *Construction automatique d'une base de connaissances du domaine dans un système de recherche d'information*. Habilitation à diriger des recherches, université Joseph Fourier, 1990
- [12] Mohamed Hatem HADDAD. *Extraction et l'impact des connaissances sur les performances des systèmes de recherche d'information*. Thèse de Doctorat, Université Joseph Fourier 2002.
- [13] Helen J. Peat and Peter Willett. *The Limitation of terme Co-occurrence data for query expansion in document retrieval systems*. In Journal of the American society for information science,42(5):378-383,1991.
- [14] Gerard Salton, Michael J. MacGill. *Introduction to Information Retrieval*.