Identification, interprétation et représentation de relations sémantiques entre concepts.

Florence Le Priol

LaLIC (Langage Logique Informatique Cognition)
Université Paris-Sorbonne - 96 bd Raspail - 75006 Paris
Florence.Le-Priol@paris4.sorbonne.fr
http://www.lalic.paris4.sorbonne.fr/~lepriol

Résumé – Abstract

SEEK-JAVA est un système permettant l'identification, l'interprétation et la représentation de connaissances à partir de textes. Il attribue une étiquette aux relations et identifie automatiquement les concepts arguments des relations. Les résultats, capitalisés dans une base de données, sont proposés, par le biais d'une interface, soit sous forme de graphes soit sous forme de tables. Ce système, intégré dans la plate-forme FilText, s'appuie sur la méthode d'exploration contextuelle.

SEEK-JAVA is a identification, interpretation and representation system of knowledge extracted from texts. It gives label to the relation between concepts and automatically identify arguments of the relations. The results, capitalized in a data base, are proposed by an interface, either in the form of graphs or in the form of tables. This system, integrated in the FilText platform, is based on contextual exploration method.

Mots-clés

Identification, interprétation, représentation des connaissances, exploration contextuelle, graphes, base de données

1 Introduction

Que ce soient les entreprises, les organisations, les personnes, chacun aujourd'hui engrange des masses phénoménales d'informations de toute sorte. Un des principaux problèmes, tant dans le milieu industriel que dans celui de la recherche est de transformer ces collections de données en connaissances nouvelles, intelligibles, utiles et intéressantes dans l'environnement où l'on se trouve. SEEK-JAVA, en identifiant et en interprétant les relations entre concepts peut répondre à ce défi dans de nombreux domaines. La représentation des résultats (triplets concept-relation-concept) sous forme de graphes améliore l'accès aux informations en en

donnant une représentation visuelle facilement lisible. La base de données où sont capitalisées les relations entre concepts qui accompagne les graphes vient compléter la représentation des connaissances du texte en permettant diverses opérations pour les structurer.

SEEK-JAVA se situe à la frontière de plusieurs domaines : informatique, extraction d'information, acquisition de connaissances, terminologie, linguistique. A partir de conceptualisations sémantiques élaborées dans le cadre d'un modèle général de traitement des langues en rapport avec la cognition, le modèle de la Grammaire Applicative et Cognitive (GAC), un premier système automatique a été réalisé en 1993 par C. Jouis (Jouis, 1993) : SEEK. La présente étude a repris certaines des analyses déjà présentées dans SEEK mais notre effort a porté d'une part, sur une réanalyse et une réorganisation des connaissances linguistiques et d'autre part, sur une nouvelle modélisation de façon à proposer directement une représentation des connaissances sous forme de graphes et sous-graphes et d'intégrer ce modèle dans la plate-forme d'Exploration Contextuelle FilText (Minel & al. 2001).

2 Identification et interprétation des relations entre concepts

L'identification et l'interprétation des relations sémantiques entre concepts s'appuient sur la méthode d'exploration contextuelle (Desclés & al., 1991, Jouis, 1993). Cette méthode se base sur l'hypothèse suivante : les textes contiennent des unités linguistiques spécifiques qui sont des indicateurs pertinents pour résoudre une tâche précise. Cependant, l'identification de ces indicateurs n'est pas suffisante. L'analyse d'une unité linguistique identifiée dans un contexte fait nécessairement appel à d'autres indices linguistiques complémentaires qui doivent être co-présent dans le contexte, ces indices participent directement à la résolution de la tâche. Elle aboutit à une mise en œuvre informatique : les systèmes d'exploration contextuelle constitués d'une base de connaissances linguistiques comprenant les indicateurs déclencheurs et les indices contextuels complémentaires et des règles d'exploration contextuelle.

Prenons comme exemple l'énoncé (i) *les armes et les outils sont constitués de poignards et lames, hallebardes, faucilles, haches, armes diverses, claviformes*.

Tête : Rinc08
Tâche : Relation_Statique
Classes d'indicateurs déclencheurs : &Ietre={être, est, sont, était…}, &Iverbeinc1={constituer, former…}
Espaces de recherche
E1 := espace de recherche situé entre le début de la phrase et l'indicateur déclencheur
E2 := espace de recherche situé entre l'indicateur déclencheur et la fin de la phrase
Classes d'indices contextuels
L1 = &ppeinc = {constitué, divisé, …}
L2 = &LIngPrep2 = {avec, de, en, …}
Conditions
Il existe un indice ind1 appartenant à E2 tel que ind1 appartient à L1
Il existe un indice ind2 appartenant à E2 tel que ind2 appartient à L2
ind1 et ind2 se suivent
Action
1. attribuer l'étiquette "est_inclu" à la phrase
2. rechercher arg1 dans E1 et de arg2 dans E2, l'orientation de la relation est "arg1←arg2"

Figure 1 : la règle d'exploration contextuelle Rinc08

Une occurrence de la classe d'indicateurs déclencheurs *&Ietre* (*sont*) est repérée dans le texte et permet de déclencher un certain nombre de règles. Les indices contextuels complémentaires *constitués* et *de* permettent la validation de la règle d'inclusion Rinc08

(Figure 1) : attribution de l'étiquette sémantique d'inclusion *est_inclus* et identification des arguments *les armes et les outils, poignards, lames, hallebardes, faucilles, haches, armes diverses* et *claviformes*. L'identification des arguments est basée sur une liste des termes du corpus constituée soit à l'aide d'un thesaurus existant, par exemple sur le web, soit constituée à la main et, sur des règles élaborées à partir de l'étude de corpus et en s'appuyant sur l'hypothèse que la position des arguments dans la phrase et l'orientation de la relation sont liés à la position de l'indicateur et à la nature des classes d'indices contextuels permettant d'identifier la relation.

Dans les quatre exemples ci-dessous (ii, iii, iv, v), où les indicateurs déclencheurs sont soulignés en trait plein, les indices contextuels en trait pointillé et où les arguments sont en italique, une relation d'inclusion est identifiée entre « chats » et « mammifères ». L'orientation de chaque relation est toujours la même (chats est inclus dans mammifères) mais la position des arguments dans la phrase est différente.

(ii) Il y a différentes sortes de *mammifères* : les *chats*, les chiens	mammifères (arg1) ← chats (arg2)
(iii) On classe les *chats* dans les *mammifères*	chats (arg1) → mammifères (arg2)
(iv) Les *mammifères* forment une classe plus générale que celle des *chats*	mammifères (arg1) ← chats (arg2)
(v) Tous les *chats* sont des *mammifères*	chats (arg1) → mammifères (arg2)

Dans les exemples (ii) et (iii), les deux arguments sont du même coté de l'indicateur. Le premier argument est le premier dans la phrase en allant dans le sens de la lecture. La relation est orienté de l'argument2 vers l'argument1 dans l'exemple (ii), les indices que l'on doit trouver dans la phrase pour l'identification de la relation ne permettent pas que la relation soit orienté dans l'autre sens. On ne peut pas avoir « il y a différentes sortes de chat : (deux-points) les mammifères ». Dans l'exemple (iii), la relation est orienté de l'argument1 vers l'argument2. L'élément le plus spécifique est donné avant l'élément le plus générale. L'énoncé « on classe les mammifères dans les chats » s'il est syntaxiquement acceptable ne l'est pas sémantiquement. Dans les exemples (iv) et (v), l'argument1 est à gauche de l'indicateur, l'argument2 est à droite de l'indicateur. La classe la plus générale, dans un exemple du type de l'exemple (iv) sera toujours le premier argument. La relation est donc orienté de l'argument2 vers l'argument1. Dans les exemples du genre de l'exemple (v), c'est la classe la plus spécifique qui sera toujours le premier argument. La relation est donc orienté de l'argument1 vers l'argument2.

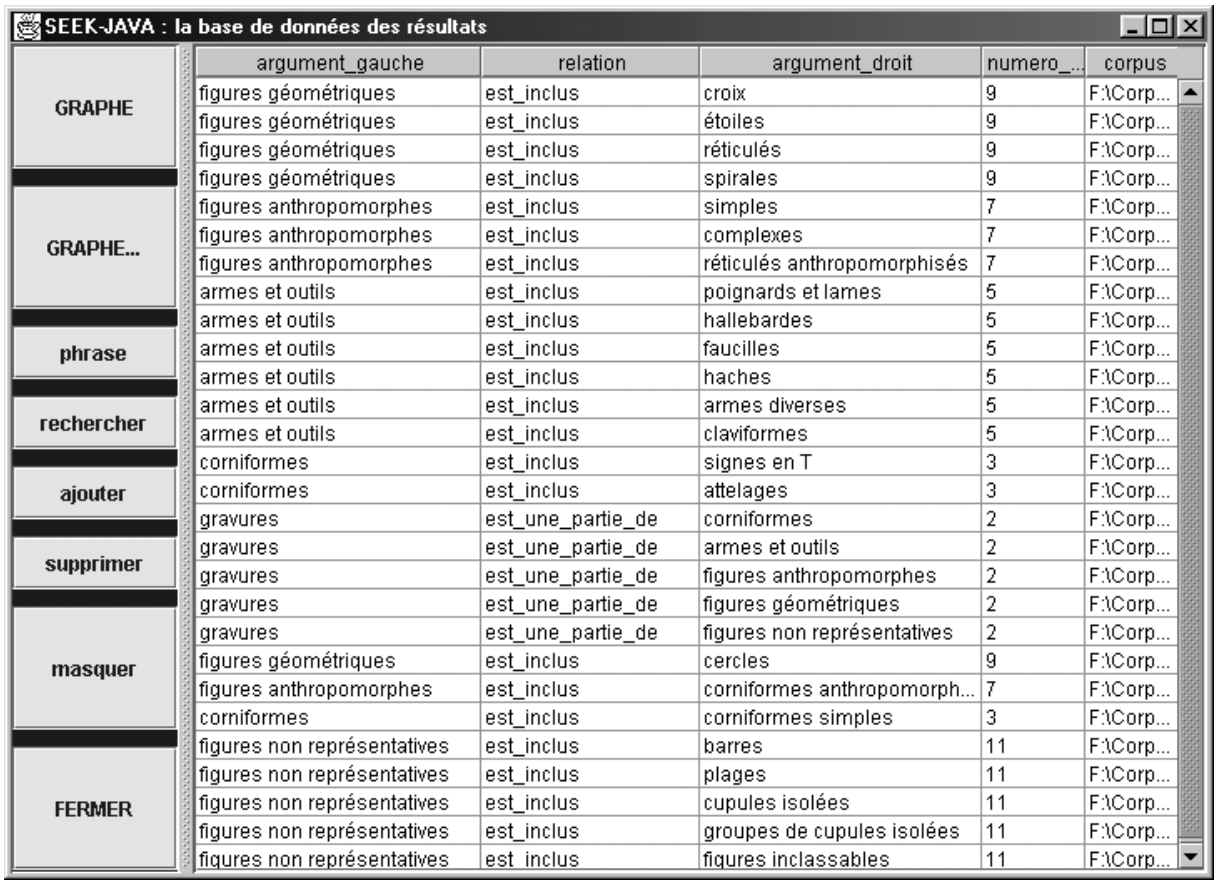
3 Représentation des connaissances

Les relations sémantiques entre concepts identifiées dans le texte sont stockées dans une base de données sous la forme d'un triplet argument1-relation-argument2. Ces connaissances peuvent être représentées soit sous forme de tables soit sous forme de graphes ou de sous-graphes. Prenons l'exemple du texte (Figure 2) dont est issu l'énoncé (i).

Le corpus iconographique.
Cinq grandes catégories de gravures rupestres peuvent être distinguées : les corniformes, les armes et les outils, les figures anthropomorphes, les figures géométriques et les figures non représentatives. Les corniformes sont composés des corniformes simples, signes en T, attelages. Les corniformes représentent 46 % des gravures. Les armes et les outils sont constitués de poignards et lames, hallebardes, faucilles, haches, armes diverses, claviformes. Les armes et les outils représentent 4 % des gravures. Les figures anthropomorphes : simples, complexes, corniformes anthropomorphisés, réticulés anthropomorphisés. Les figures anthropomorphes représentent 0.2 % des gravures. Les figures géométriques regroupent par exemple,les cercles, croix, étoiles, réticulés, spirales. Les figures géométriques représentent 7 % des gravures. Les figures non représentatives : cupules isolées, groupes de cupules isolées, plages, barres, figures inclassables. Les figures non représentatives représentent 42.8 % des gravures.

Figure 2 : extrait d'un texte du web sur les gravures rupestres

Une interface Homme-Machine permet de représenter ces connaissances soit dans une table (Figure 3), soit sous forme d'un graphe ou de sous-graphes (Figure 4).



argument_gauche	relation	argument_droit	numero_...	corpus
figures géométriques	est_inclus	croix	9	F:\Corp...
figures géométriques	est_inclus	étoiles	9	F:\Corp...
figures géométriques	est_inclus	réticulés	9	F:\Corp...
figures géométriques	est_inclus	spirales	9	F:\Corp...
figures anthropomorphes	est_inclus	simples	7	F:\Corp...
figures anthropomorphes	est_inclus	complexes	7	F:\Corp...
figures anthropomorphes	est_inclus	réticulés anthropomorphisés	7	F:\Corp...
armes et outils	est_inclus	poignards et lames	5	F:\Corp...
armes et outils	est_inclus	hallebardes	5	F:\Corp...
armes et outils	est_inclus	faucilles	5	F:\Corp...
armes et outils	est_inclus	haches	5	F:\Corp...
armes et outils	est_inclus	armes diverses	5	F:\Corp...
armes et outils	est_inclus	claviformes	5	F:\Corp...
corniformes	est_inclus	signes en T	3	F:\Corp...
corniformes	est_inclus	attelages	3	F:\Corp...
gravures	est_une_partie_de	corniformes	2	F:\Corp...
gravures	est_une_partie_de	armes et outils	2	F:\Corp...
gravures	est_une_partie_de	figures anthropomorphes	2	F:\Corp...
gravures	est_une_partie_de	figures géométriques	2	F:\Corp...
gravures	est_une_partie_de	figures non représentatives	2	F:\Corp...
figures géométriques	est_inclus	cercles	9	F:\Corp...
figures anthropomorphes	est_inclus	corniformes anthropomorph...	7	F:\Corp...
corniformes	est_inclus	corniformes simples	3	F:\Corp...
figures non représentatives	est_inclus	barres	11	F:\Corp...
figures non représentatives	est_inclus	plages	11	F:\Corp...
figures non représentatives	est_inclus	cupules isolées	11	F:\Corp...
figures non représentatives	est_inclus	groupes de cupules isolées	11	F:\Corp...
figures non représentatives	est_inclus	figures inclassables	11	F:\Corp...

Figure 3 : les relations identifiées dans le texte présentées dans une table

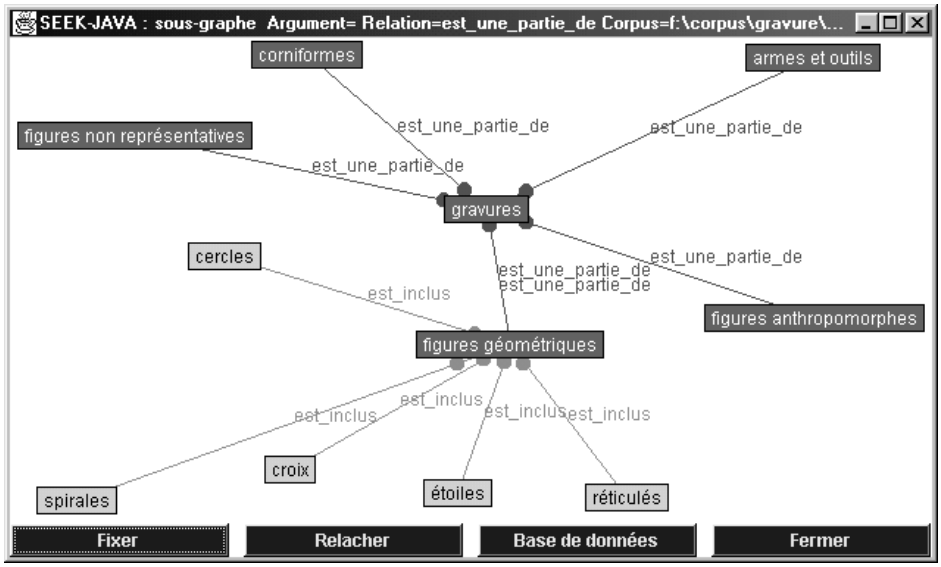


Figure 4 : sous-graphe des relations partie-tout (*est_une_partie_de*) augmenté des relations centrées autour de *figures géométriques*

La table présente toutes les relations en indiquant le numéro de la phrase dont elles sont extraites. L'utilisateur peut, en fonction de ses besoins, afficher le contexte de la relation, rechercher, ajouter ou supprimer une relation et ses arguments. Le graphe ou les sous-graphes sont affichés dans une fenêtre active : les arguments des relations se placent automatiquement les uns par rapport aux autres de manière à ne pas se superposer et à rester à l'intérieur de la fenêtre. Néanmoins, l'utilisateur peut déplacer les éléments à l'aide de la souris de manière à opérer les regroupements qu'il souhaite. L'affichage de sous-graphes donne la possibilité de visualiser un type de relation, par exemple toutes les relations partie-tout (*est_une_partie_de*), ou de n'afficher que les relations autour d'un concept donné. L'utilisateur pourra alors développer le graphe en fonction de ses besoins en cliquant sur le concept qu'il souhaite spécifier, par exemple *figures géométriques* (Figure 4).

4 Applications

Le thésaurus d'un domaine est un outil en construction permanente, car il reflète un état de structuration liée aux connaissances d'un domaine or, nos connaissances évoluent, le domaine s'élargit, se transforme, s'interconnecte avec d'autres domaines. Les relations établies ne sont pas figées une fois pour toutes, les spécialités évoluent rapidement et leur textualisation expriment souvent cette évolution. Les lacunes des thésaurus actuels sont importantes : le coût de l'élaboration et de gestion (dû, entre autres, au besoin de disposer d'une équipe de spécialistes de plusieurs disciplines, ce qui entraîne un certain ralentissement dans leur production) ; l'absence d'étiquettes sémantiques attachées aux relations entre concepts ; la spécialisation étroite des thésaurus entraînent un manque de thésaurus multidisciplinaires. Ainsi, il est déconseillé d'envisager de se servir d'un thésaurus existant pour un domaine précis si on traite d'un autre domaine (même s'il est similaire) ou si on appartient à d'autre univers culturel (types de métiers par exemple). SEEK-JAVA offre la possibilité, en donnant une étiquette sémantique aux relations entre concepts, de complexifier un thésaurus ou même une terminologie, habituellement constituée de relations non étiquetées entre concepts. Par exemple, le couplage de IOTA (Chiaramella & al. 1986) et SEEK-JAVA ou de LEXTER (Bourigault 1994) et SEEK-JAVA permet la création, de manière automatique, de thésaurus dont les relations sont étiquetées permettant ainsi de palier les lacunes qui leur sont encore attribuées. Le terminologue pourra donc utiliser SEEK-JAVA d'une part pour enrichir une terminologie, d'autre part pour visualiser, grâce à son interface, les liens entre concepts sous forme de graphes ou sous forme de table.

Une des difficultés de la traduction de texte, pour avoir une traduction de qualité, est de trouver un traducteur qui soit également un expert du domaine du texte à traduire. En effet, seul un expert du domaine peut traduire correctement les termes techniques spécifiques. Il peut donc être difficile de trouver une même personne ayant la double compétence surtout si la langue destination est une langue rare. Utiliser un système comme SEEK-JAVA, dans le processus de traduction, permet d'extraire du texte en langue source les principaux concepts et les relations qui les lient. Ce sont ces concepts et relations qui seront traduits par un traducteur. Les concepts principaux du texte étant extrait automatiquement, le traducteur n'a pas besoin d'être un expert du domaine. Ainsi traduit, il restera à un expert du domaine de la langue de destination à reconstituer le texte à l'aide des concepts et des relations extraits.

5 Conclusion

D'autres perspectives s'ouvrent pour améliorer ce travail. Un des point faible de SEEK-JAVA est qu'il n'extrait pour l'instant que des relations sémantiques statiques. Les relations évolutives (cinématiques et dynamiques) ne sont pas traitées. L'ajout d'un module d'extraction des connaissances cinématique et dynamique offre la possibilité d'un travail intéressant tant d'un point de vue linguistique que d'un point de vue informatique. Une jonction avec les travaux sur la causalité et l'action (Garcia 98, Jackiewicz 98) est une autre piste intéressante. La forme du corpus oppose également une limite au système. Les textes traités par SEEK-JAVA doivent être rédigés. Ils sont présentés dans un format brut où la mise en page, les images et les indications typographiques ne sont pas prises en considération. Avec le développement de la norme XML, nous pouvons envisager d'intégrer, dans les règles d'exploration contextuelle, des indications sur la mise en page et ainsi ne pas nous priver d'indications importantes données par l'auteur du texte.

Références

Bourigault D. (1994), *LEXTER, un Logiciel d'Extraction de TERminologie. Application à l'extraction des connaissances à partir d'un texte,* Thèse de Doctorat, EHESS, Paris

Chiaramella Y., Defude B., Bruandet M-F., Kerkouba D. (1986), IOTA : a full text information retrieval system, *Proc. Of Sigir Conference on Research and development in Information Retrieval*, Pisa, Italy, pp 207-213

Condamines A., Rebeyrolle J. (1997), Utilisation d'outils dans la constitution de Bases de Connaissances Terminologiques : expérimentation, limites, définition d'une méthodologie, dans les *Actes des Journées Scientifiques et Techniques du réseau FRANCIL* (*JST'97)*, Avignon, pp 529-535

Desclés J-P., Jouis C., Oh H-G., Maire-Reppert D. (1991), Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte, *Knowledge modeling and expertise transfer (ed. D.Herin-Aime, R. Dieng, J-P. Regourd, J-P.Angoujard)*, Amsterdam, pp 371-400

Jouis C. (1993), *Contributions à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. Réalisation d'un prototype : le système SEEK*, Paris, Thèse de doctorat

Le Priol F. (2000), *Extraction et capitalisation automatiques de connaissances à partir de documants textuels. SEEK-JAVA : identification et interprétation de relations entre concepts*, Paris, Thèse de doctorat, Université Paris-Sorbonne

Minel J-L., Descles J-P., Cartier E., Crispino G., Ben Hazez S., Jackiewicz A. (2001), Résumé automatique par filtrage sémantique d'informations dans des textes, *TSI (Technique et Science Informatique)*, n° 3