Analyse sémantique dans un système de question - réponse.

Laura Monceaux

LIMSI-CNRS, BP 133, Bat 508, 91403 Orsay Cedex
Laura.Monceaux@limsi.fr

Résumé

Dans cet article, nous présentons le système QALC (Question Answering Langage Cognition) qui a participé à la tâche Question Réponse de la conférence d'évaluation TREC. Ce système a pour but d'extraire la réponse à une question d'une grande masse de documents. Afin d'améliorer les résultats de notre système, nous avons réfléchi à la nécessité de développer, dans le module d'analyse, le typage des questions mais aussi d'introduire des connaissances syntaxico-sémantiques pour une meilleure recherche de la réponse.

Abstract

In this paper, we present the QALC system that participated to the Question Answering track of the TREC evaluation conference. This system extracts the answer to the question from a large amount of documents. In order to improve the result of our system, it is necessary to develop, in the question analysis module, the search of question type and to introduce a semantic knowledge.

Mots Clés : Analyse syntaxico-sémantique, système question – réponse

Key Words : Semantic analysis, question – answering system

1 Introduction

Dans le cadre de la conférence d'évaluation TREC (Text REtrieval Conference), le groupe LIR (Langues, Information, Représentations) a présenté, depuis deux années, un système pour la tâche Question Answering, QALC (Ferret et al, 2000). Ce système a pour but d'extraire la réponse à une question d'une grande masse de documents. Là où un système de recherche d'informations renvoie une liste de documents pouvant contenir la réponse, un système de question – réponse renverra une liste courte de réponses d'une ou deux phrases. A la conférence TREC 2000, le système QALC a réalisé les scores suivants :

- TREC 2000 – réponses en 250 caractères : 55 % des questions résolues – 6ème système sur 25
- TREC 2000 – réponses en 50 caractères : 26,7 % des questions résolues – 19ème système sur 24

L'un des buts que nous proposons pour améliorer le système QALC pour des réponses en 50 caractères est d'introduire une analyse des questions plus poussée en utilisant des connaissances sémantiques. L'analyse d'une question dans un système, tel que QALC, doit fournir classiquement deux types d'information :

- le type attendu de la réponse (un nom de personne, un nombre, par exemple)
- les termes importants de la question à retrouver dans la réponse.

En effet, comme on peut le voir dans l'architecture du système QALC (figure 1), les différents termes extraits de l'analyse sont utiles pour la ré-indexation et la sélection des documents réalisées par FASTR (Jacquemin, 1999) et le type attendu de la réponse ainsi que les termes permettent de tester l'appariement entre la question et la réponse afin de sélectionner la meilleure réponse.

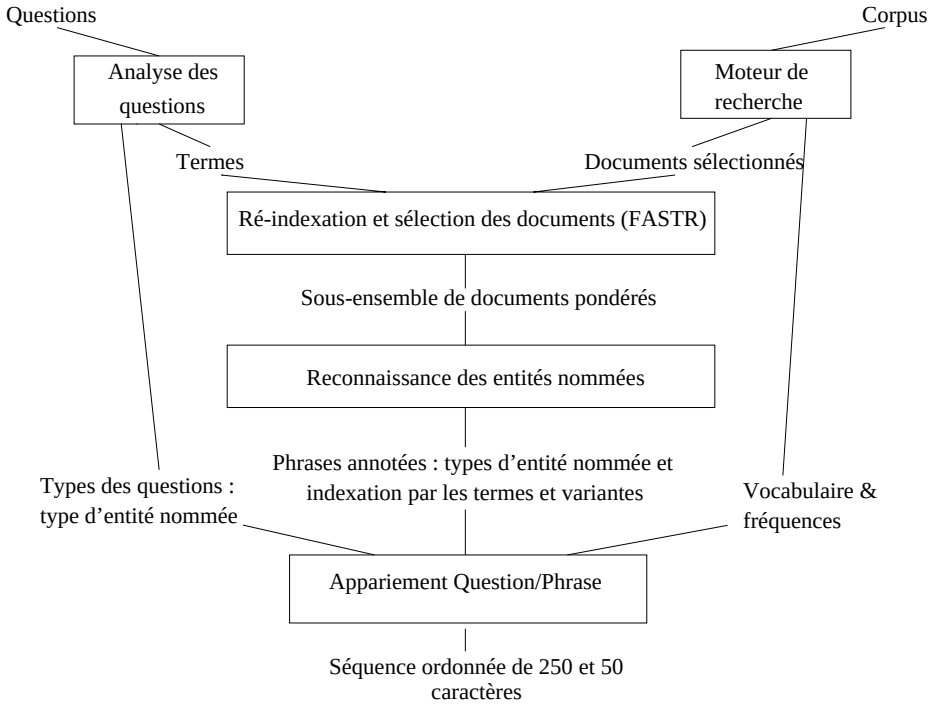

Figure 1 : Architecture du système QALC

Ainsi actuellement, pour la question « What was the name of the first Russian astronaut to do a spacewalk ? », la recherche de la réponse est faite par appariement entre les termes de la question (« first », « Russian », « astronaut », « spacewalk ») et le type attendu de la réponse (une Personne) d'une part, et les phrases des documents d'autre part. Mais les liens syntaxiques et sémantiques qui existent entre les différents termes ne sont pas du tout pris en compte.

Notre but est donc d'étendre l'analyse des questions d'une part en améliorant le module de typage quand le type de réponse attendu ne correspond pas à une entité nommée décelable et d'autre part d'utiliser les liens entre les différents termes obtenus. Ainsi dans notre exemple, savoir que l'on recherche le nom d'une personne de nationalité russe, qui a pour métier d'être astronaute, et qui a été le premier à faire une sortie dans l'espace, ne peut que nous aider à sélectionner plus facilement la meilleure réponse dans les documents sélectionnés. On a pu remarquer d'ailleurs que les meilleurs systèmes, dès la première évaluation de TREC, comportaient un module d'analyse des questions et que l'année précédente, ils ont introduit des connaissances sémantiques à celui ci.

2 Recherche de la réponse

2.1 Typage des questions

Le premier objectif de notre analyse des questions est de définir le type attendu de la réponse. Selon la question, le type recherché sera plus ou moins facile à détecter.

2.1.1 Réponses du type « entité nommée »

On essaye, dans un premier temps, d'attribuer à la question une ou plusieurs étiquettes (catégories sémantiques) correspondant au type attendu de la réponse grâce à la reconnaissance d'un certain nombre d'indices présents dans la phrase. Une hiérarchie des catégories sémantiques des types de réponses possibles a été construite manuellement (Figure 2).

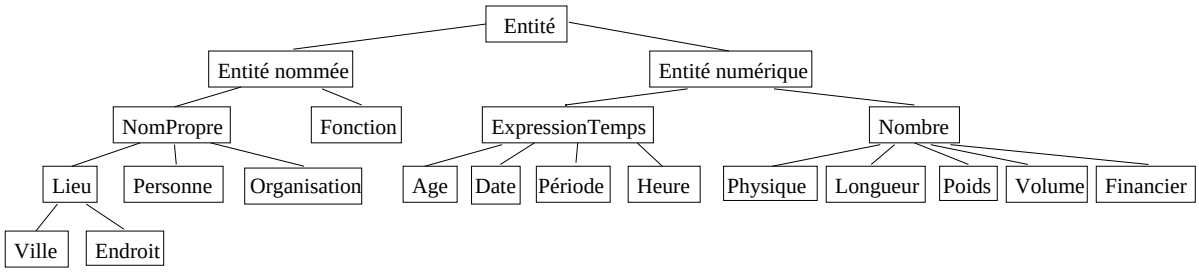

Figure 2 : Hiérarchie des types de réponses et catégories sémantiques

On essaye donc d'associer à la question une ou plusieurs étiquettes sémantiques :

- PERSONNE : Who is the richest person in the World ?
- VILLE, ENDROIT : Where is Taj Mahal ?

Pour déterminer ces étiquettes, plusieurs règles décrivant les différentes formes que peuvent prendre les questions ont été écrites, en voici quelques exemples :

- le type de la réponse ne dépend que du pronom interrogatif : where (VILLE, ENDROIT), who / whom / whose (PERSONNE), when (DATE).

  Exemple : Who is Martin Luther King ? (PERSONNE)

- le type de la réponse est déterminé grâce à la catégorie sémantique du nom de tête du groupe nominal pour les questions de la forme :

  - what / which … be GNSem

  - what / which GNSem

- what / which is the name of GNSem

- how much / many GNSem

où GNSEM est de la forme : (Det) (Adj | N | Vbg | Vbc)* Nsem.

Exemple : What city in Florida is Sea World in ? (VILLE)

- le choix de l'étiquette est déterminé en fonction de la catégorie auquel appartient l'adjectif : how AdjSem …

  Exemple : How old are you ? (DATE)

On essaye donc dans un premier d'appliquer l'une de ces règles à partir du début de la question. Ces différentes règles ont été implémentées dans QALC et permettent d'étiqueter environ 60 % des 700 questions de TREC9. Cependant certains GNSem ne peuvent être associés à une catégorie sémantique ; ainsi pour certaines questions, l'association d'une étiquette (entité nommée) est difficile.

2.1.2 Autres types de réponses

Pour la phrase « What flower did Vincent Van Gogh paint ? », aucune catégorie sémantique ne peut être associée au GNSem « flower » : ce n'est ni une personne, ni une date, ni un nombre, ni un lieu… Dans ce cas, le type de réponse attendu n'est plus une entité nommée mais plutôt un nom commun voire une phrase. Ainsi le typage de cette réponse est beaucoup plus complexe. Pour palier ce problème, on envisage d'introduire la notion de focus qui a été proposée par Wendy Lenhert (1979) : elle définit le focus de la question comme un concept qui exprime les attentes exprimées dans la question. Pour notre exemple, notre focus correspondra au concept « flower ». Dans notre système, on considère que le focus peut être soit un nom soit un groupe de mots voire éventuellement un ensemble de mots que l'on devrait retrouver dans la réponse. Plusieurs règles pour déterminer ce focus seront écrites comme pour les questions attendant une entité nommée :

- pour les phrases : What GNSem …, le focus sera le concept tête du GNSem voir le GNSem entier.

  Exemple : What video format was an alternative to VHS ? (VIDEO FORMAT)

- pour les questions de la forme : Who Aux GNSem ?, le focus correspondra au concept associé au GNSem

  Exemple : Who is the prime minister of France ? (PRIME MINISTER OF FRANCE)

- etc.

On pourra par la suite pour l'appariement du type attendu et des réponses candidates utiliser les connaissances de WordNet notamment la hiérarchie des mots (hyponymes) mais aussi les synonymes afin de voir si les réponses candidates sont compatibles avec le focus et le type attendu de la réponse.

2.2 Appariement question / réponse

Ce module permet de choisir parmi une liste de réponses retenues par FASTR la meilleure réponse à notre question. Actuellement, dans chacun des documents sélectionnés, les entités nommées ainsi que les occurrences des termes de la question et leurs variations ont été repérées et marquées. La comparaison entre la question et les réponses candidates est réalisée grâce à un vecteur possédant 3 types d'éléments : les mots pleins, les termes de la question et les entités nommées, chaque élément étant pondérés en fonction de son importance par rapport aux autres (Ferret et al., 2000).

> *Exemple :*
>
> *Question : How many people watch network television ?(NUMBER)*
>
> *Réponse : <b_numex_type = NUMBER> 2100 <e_numex> people watch network television.*

La recherche de la meilleure réponse est donc faite par appariement entre, d'une part les mots des phrases candidates et d'autre part les termes de la question et le type attendu de la réponse. Nous envisageons, pour le système de cette année, de comparer notre question et nos phrases candidates grâce à leur représentation sémantique, c'est à dire prendre en compte les liens sémantiques entre les différents termes. Ainsi pour la phrase « What was the name of the first Russian astronaut to do a spacewalk ? », nous déterminerons les termes importants ("first", "Russian", "astronaut", "spacewalk") mais nous aurons aussi des liens entre ces termes pour éliminer les mauvaises réponses : on devra trouver le nom d'un astronaute qui a pour caractéristique d'être russe et d'être le premier à avoir fait une sortie dans l'espace. Par exemple, grâce aux liens sémantiques, la réponse « l'astronaute américain X a participé à la première sortie russe spatiale de cette fusée » sera éliminée, car on ne parle pas d'une première sortie russe mais d'un premier astronaute russe. En effet nous avons constaté que pour certaines questions l'appariement question – réponse sur les liens auraient permis de répondre correctement à la question et donc d'éliminer les phrases candidates qui ne contiennent pas la réponse attendue. Pour prendre en compte les liens entre les termes des questions, nous avons commencé à développer un analyseur syntaxico-sémantique.

3 Notre analyseur syntaxico-sémantique

3.1 Quel analyseur syntaxique ?

Avant de travailler sur la représentation sémantique d'une phrase, nous avons choisi d'en étudier la syntaxe. Grâce au recueil des questions des années précédentes, nous avons commencé à étudier les sorties obtenues par différents analyseurs syntaxiques comme l'analyseur IFSP de Xerox (Aït-Mokhtar S and Chanod J, 1997), l'analyseur Link Grammar, l'analyseur Sylex, l'analyseur IPS de Genève (Wehrli, 1992), etc … Il s'agit ici pour nous de regarder si la segmentation des phrases et les dépendances extraites peuvent nous aider à construire une représentation sémantique correcte. Cette évaluation est en cours, mais pour des raisons pratiques, nous avons décidé d'utiliser l'analyseur de Xerox. Cependant, nous envisageons de réaliser un analyseur sémantique utilisable avec n'importe quel analyseur syntaxique. Voici un exemple de sorties que nous aurons à utiliser, il s'agit ici des sorties obtenues par l'analyseur IFSP (Aït-Mokhtar and Chanod, 1997) de Xerox :

> *Exemple :*
> *INVSUBJ (be, valley), NNPREP (valley, of, king) …*
> *[SC Where :v is SC] [NP the valley NP] [PP of the king PP] ?*

Notre but ne sera pas forcément de réaliser une analyse syntaxico-sémantique complète de nos phrases mais éventuellement de proposer des morceaux de représentation qui nous aideront tout de même à rechercher la meilleure réponse. Ainsi nous avons décidé d'utiliser un analyseur syntaxique partiel qui a l'avantage, même si la phrase est mal formulée, d'extraire des informations syntaxiques, plutôt qu'un analyseur complet qui, si la phrase est mal formulée ou trop complexe, arrêtera son analyse. Dans le cadre de notre participation à la tâche question - réponse de la conférence TREC, nous allons travailler sur des questions – réponses en anglais. Le travail que j'ai effectué précédemment sur l'élaboration d'un analyseur syntaxico-sémantique pour le français, pour des phrases issues de dialogue intégrant du langage et des gestes (Monceaux, 2000), pourra être réutilisé pour atteindre notre objectif. En effet certaines règles ont été écrites à partir des connaissances syntaxiques obtenues : par exemple, dans la plupart des cas, le sujet de la phrase correspond à l'agent du verbe. Nous avions démontré alors l'avantage d'utiliser un analyseur partiel pour des phrases issues de dialogue, or l'une des évolutions envisagée pour la conférence TREC est d'introduire la notion de dialogue.

3.2 Quel formalisme de représentation sémantique ?

Lors de notre précédente étude, nous avions décidé de représenter nos phrases, que ce soit des questions, des réponses ou tout autre phrase, au moyen du formalisme des graphes conceptuels (Sowa, 1983). Les graphes conceptuels sont des graphes bipartites, connexes, finis possédant deux sortes de nœuds : des nœuds concept et des nœuds relation. Un nœud concept se représente sous la forme suivante [type : référent] où le type est la classe caractérisant les individus et où le référent spécifie les individus du monde de référence désigné par le concept. Les nœuds relations permettent quant à eux de savoir comment les nœuds concepts peuvent être inter-connectés. Ainsi pour une phrase comme « What was the name of the first Russian astronaut to do a spacewalk ? », on doit obtenir la représentation suivante :

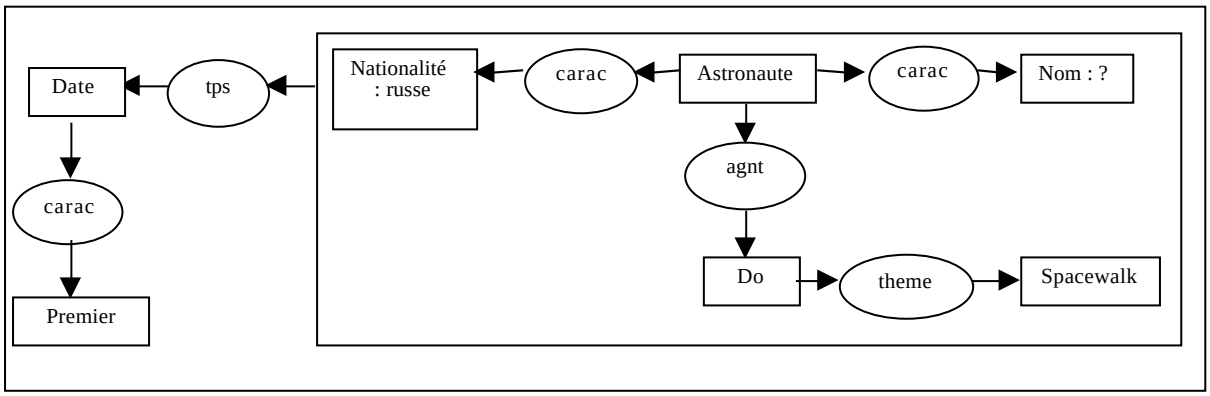

Figure 3 : Représentation sémantique idéale de la phrase :
« What was the name of the first Russian astronaut to do a spacewalk ? »

Pour obtenir une telle représentation, j'ai écrit un algorithme qui utilise notamment les différentes opérations qui peuvent être réalisées sur les graphes conceptuels : copie, restriction, simplification et jointure. Mais cet algorithme requiert un certain nombre de connaissances comme des graphes canoniques (qui donne des contraintes sur les liens sémantiques entre deux concepts), le treillis des concepts (qui propose une hiérarchie des concepts) ... Pour l'instant, nous ne possédons pas de base complète de ces connaissances ni en français ni en anglais : cette étude des connaissances est longue et fastidieuse à réaliser. D'ailleurs, de nombreuses études permettant l'apprentissage automatique de ces connaissances sont en cours. Ainsi, pour une application concernant un domaine limité, l'élaboration des connaissances nécessaires est possible mais pour une application telle que la tâche question – réponse proposée à la conférence TREC, cela s'avère difficile car les questions proposées couvrent de nombreux domaines. Nous

essayerons donc grâce à des connaissances limitées mais disponibles d'élaborer une représentation sémantique sous forme de graphes conceptuels et donc de modifier l'algorithme écrit pour les phrases en français issues de dialogue.

3.3 Quelles connaissances à notre disposition ?

Pour intégrer un analyseur syntaxico-sémantique dans le système proposé cette année à la conférence TREC, nous avons donc décidé d'utiliser des connaissances moins complètes mais accessibles. Pour l'anglais, une base de données lexicales a été élaborée par un ensemble de psycholinguistes et d'informaticiens à l'Université de Princeton : WordNet (Fellbaum, 1998) (resp. EuroWordNEt pour le français). Les relations d'hyponymie et de synonymie développées dans cette base permettront notamment de nous aider dans l'élaboration de notre analyseur pour par exemple étendre notre question grâce aux synonymes des différents termes de la question, de regarder si un terme appartient à une des catégories sémantiques de notre hiérarchie, etc.

Exemple : What actress starred in « The Lion in Winter » ?
Type attendu : PERSONNE car la relation d'hyperonymie de WordNet renvoie : Actress

- ➔ *Actor, player, role player*
- ➔ *Performer, Performing artist*
- ➔ *Entertainer*
- ➔ *Person, individual, someone …*

En utilisant les connaissances de WordNet et des connaissances que nous ajouterons, comme par exemple les différentes règles de typage des questions, mais aussi en utilisant les sorties de l'analyseur syntaxique partiel choisi, nous élaborerons une représentation adéquate de nos phrases.

4 Conclusion

Nous proposons donc dans cet article d'améliorer le système QALC proposé l'année dernière à la conférence d'évaluation TREC en améliorant le module d'analyse par l'ajout de connaissances sémantiques. Dès maintenant, se pose la question suivante : jusqu'où doit on aller ? Quelles sont les connaissances suffisantes et nécessaires pour pouvoir répondre à notre question ? Ainsi un travail important devra être réalisé sur les informations que l'on doit fournir grâce à notre analyseur pour une meilleure recherche de la réponse. Dans cet objectif, on envisage de créer un analyseur sémantique en plusieurs phases : la première proposant une analyse sémantique minimale et une deuxième phase déclenchée selon les besoins spécifiques de l'application. Pour la tâche question - réponse, des connaissances spécifiques, comme la forme de la question, pourront nous permettre d'orienter cette analyse dans la bonne direction. Par la suite, on envisage d'étendre cette analyse non plus seulement à un système de question - réponse mais également à un système de dialogue. Cela consistera en une autre phase où les connaissances pragmatiques, c'est à dire les connaissances sur le contexte dans lequel a lieu le dialogue, pourront être utiles pour la compréhension de la phrase.

Références

Aït-Mokhtar S. and Chanod J. (1997), Incremental Finite-State Parsing, In Proceedings of *ANLP-97*, Washington.

Fellbaum C. (1998) *WordNet: An Electronic Lexical Database,* Cambridge, MA, Mit Press

Ferret O., Grau B., Hurault-Plantet M., Illouz G., Jacquemin C., Lecuyer P., Masson N. (nov 2000), QALC- the Question-Answering program of the Language and Congnition group at LIMSI-CNRS, Actes de *la conférence TREC-8*, Gaithersburg MN, pp.465-474.

Jacquemin C. (1999) Syntagmatic and paradigmatic representations of term variation, Actes d'*ACL'99*, 341-348.

Monceaux L. (2000) Analyse syntaxico-sémantique de phrases en vue de dialogues intégrant geste et langage, Actes de *la conférence ERGO-IHM*, Biarritz.

Lehnert W. (1979) *The Process of Question Answering* Lawrence Erlbaum Associates, 1979

Sowa J. F. (1983) *Conceptual Structures: Information processing in mind and machine*, Edition Addison- Wesley Publishing Company, Inc.

Wehrli E (1992) *The IPS system*. Actes du quinzième colloque international en linguistique informatique, COLING-92