Recherche de la réponse fondée sur la reconnaissance du focus de la question

Olivier Ferret, Brigitte Grau, Martine Hurault-Plantet, Gabriel Illouz, Laura Monceaux, Isabelle Robba, Anne Vilnat

Recherche de la réponse fondée sur la reconnaissance du focus de la question

Olivier Ferret, Brigitte Grau, Martine Hurault-Plantet, Gabriel Illouz, Laura Monceaux, Isabelle Robba, et Anne Vilnat

LIMSI-CNRS BP 133, 91403 Orsay
{ferret, bg, mhp, gabrieli, monceaux, robba, anne}@limsi.fr

Résumé – Abstract

Le système de question-réponse QALC utilise les documents sélectionnés par un moteur de recherche pour la question posée, les sépare en phrases afin de comparer chaque phrase avec la question, puis localise la réponse soit en détectant l'entité nommée recherchée, soit en appliquant des patrons syntaxiques d'extraction de la réponse, sortes de schémas figés de réponse pour un type donné de question. Les patrons d'extraction que nous avons définis se fondent sur la notion de focus, qui est l'élément important de la question, celui qui devra se trouver dans la phrase réponse. Dans cet article, nous décrirons comment nous déterminons le focus dans la question, puis comment nous l'utilisons dans l'appariement question-phrase et pour la localisation de la réponse dans les phrases les plus pertinentes retenues.

The QALC question answering system we are developing uses a search engine to select documents responding to the question, matches each sentence of the selected documents with the question, then extracts the answer from the more relevant sentences, either by locating the expected named entity, or by applying extraction patterns. Patterns are based on the focus, which is the main concept in the question and is expected to be present in an answer sentence. In this paper, we will describe the way we determine the focus of the question, and the way that we use it in the question-answer pairing and answer location processes.

Mots Clés – Keywords

Recherche d'information, système de question-réponse, focus, patron d'extraction.
Information retrieval, question answering system, focus, extraction pattern.

1 Introduction

Nous présentons dans cet article le système de question-réponse QALC qui participe depuis trois ans à la tâche Question-Réponse des campagnes d'évaluation de la conférence TREC (Text REtrieval Conference). Dans son principe, un système de question-réponse permet de trouver la réponse précise à une question dans un corpus de documents. La réponse doit être

concise - c'est l'information demandée qui doit être fournie et non un document dans son entier -, et le corpus doit être suffisamment vaste pour espérer y trouver une réponse. La tâche Question-Réponse de TREC consiste à trouver des réponses de taille limitée (50 caractères maximum) à un ensemble d'environ 500 questions de type factuel ou encyclopédique. Les réponses doivent être recherchées dans un corpus d'environ un million d'articles de journaux tels que le Wall Street Journal et le Los Angeles Times.

Deux étapes sont essentielles dans la recherche de la réponse : tout d'abord restreindre le champ de recherche par la sélection d'un sous-ensemble de textes pertinents par rapport à la question, puis ensuite localiser la réponse dans cet ensemble restreint. Pour sélectionner des textes pertinents, on peut utiliser un moteur de recherche classique qui sélectionnera les documents pertinents pour la question posée (Alpha et al., 2001), ou bien adapter un moteur à la recherche de passages pertinents dans les documents comme le fait le système LCC (Harabagiu et al., 2001). En dernier ressort, c'est chaque phrase du document ou du passage (ou une fenêtre plus petite, de 50 caractères par exemple) qui sera comparée à la question pour localiser la réponse. L'appariement entre la phrase du document et la question fait en général intervenir les termes simples et les multi-termes de la question ainsi que leurs variantes morphologiques et sémantiques. Mais la localisation de la réponse dans la phrase demande des informations supplémentaires. Ainsi, tous les systèmes de question-réponse déterminent si la réponse attendue correspond à une entité nommée (par exemple une Personne ou une Ville), ou à une entité numérique (montant financier, longueur, poids …). Dans ce cas en effet, la localisation de la réponse revient à la détection de la présence dans la phrase de l'entité nommée recherchée. Une autre information importante est la connaissance des dépendances syntaxiques entre les différents groupes syntaxiques de la question. La détection des mêmes dépendances dans la phrase du document aidera la localisation de la réponse (Hovy et al., 2001). Mais cette détection demande un bon analyseur syntaxique, traitant en particulier les questions. Une solution plus robuste consiste dans l'utilisation de patrons d'extraction de la réponse, sortes de schémas figés de réponse pour un type donné de question. Certains systèmes utilisent exclusivement cette solution (Soubbotin et al., 2001), d'autres ne l'utilisent que pour les types de questions où les dépendances syntaxiques sont peu utilisables. Le système LCC (Harabagiu et al., 2001) par exemple utilise des patrons d'extraction pour les questions qui sont des demandes de définition d'un terme, comme « *Qu'est-ce qu'un micron ?*», mais utilise les dépendances syntaxiques pour d'autres types de question.

Le système QALC utilise les documents sélectionnés par un moteur de recherche, les sépare en phrases afin de comparer chaque phrase avec la question, puis localise la réponse soit en détectant l'entité nommée recherchée, soit en appliquant des patrons syntaxiques d'extraction de la réponse. Les patrons d'extraction que nous avons définis se fondent sur la notion de focus de la question. Cette notion a été introduite par Wendy Lehnert (1979) puis largement reprise et redéfinie par la suite. Pour Lehnert, le focus de la question est le concept dans la question qui englobe les attentes d'information exprimées par la question. Pour nous, c'est le mot (ou le groupe nominal) de la question représentant l'entité sur laquelle on désire une information et qui généralement se trouve dans les phrases contenant la réponse.

Dans cet article, après une brève présentation du système QALC, nous détaillerons l'analyse de la question et les informations qu'elle fournit, puis nous décrirons les processus d'appariement question-phrase et de localisation de la réponse qui s'appuient sur le focus déterminé.

2 Architecture du système QALC

Le système QALC est composé de trois modules principaux : le premier est dédié au traitement des questions, le deuxième sélectionne et traite les documents du corpus et le dernier module est chargé de produire la réponse. Chacun de ces modules est composé de plusieurs processus (cf. Figure 1).

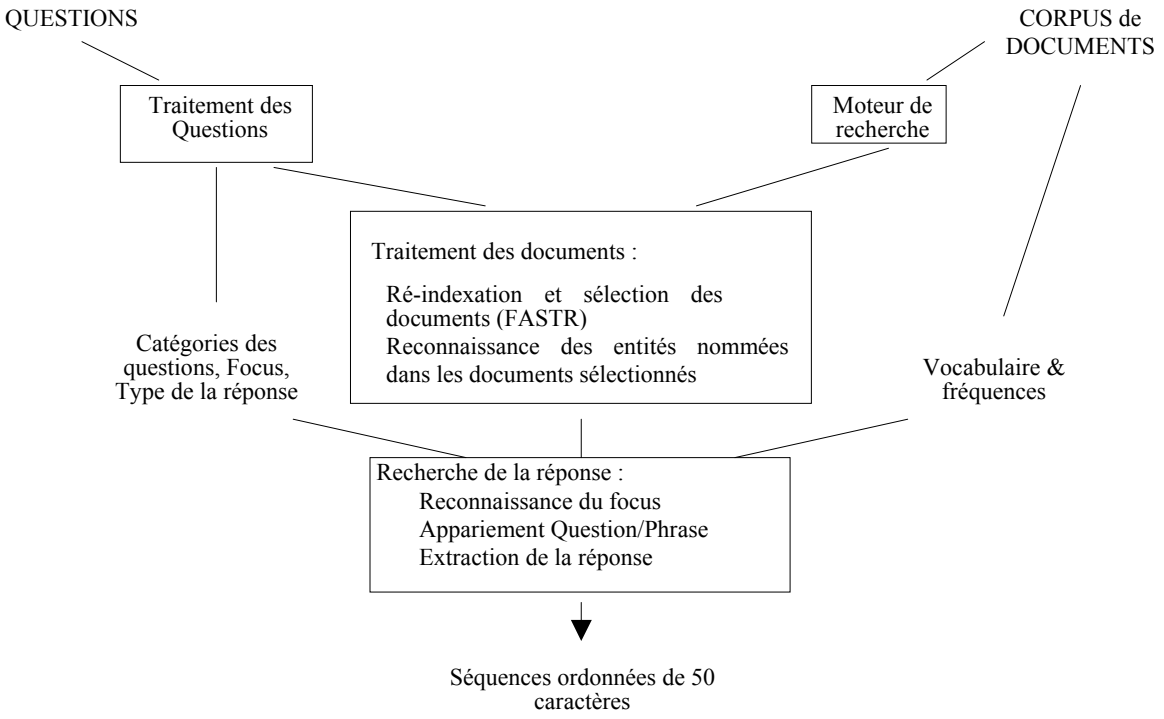

Figure 1 : Le système QALC

Le module de traitement des questions regroupe un module d'analyse de la question et un extracteur de termes. L'extracteur de termes est basé sur des patrons syntaxiques qui décrivent les groupes nominaux. Le groupe de longueur maximale est produit ainsi que tous les groupes plausibles de moindre longueur qui le forment.

L'analyse de la question se fonde sur l'utilisation d'un analyseur robuste dans le but d'extraire plusieurs informations de la question, informations utiles pour la sélection de phrases ou pour l'extraction de la réponse. Cette partie sera développée dans la section suivante.

Le module de traitement des documents utilise les sorties fournies par le NIST[1], résultat de l'application d'un moteur de recherche de type vectoriel sur le corpus de documents pour l'ensemble des questions de l'évaluation TREC. Les 200 meilleurs documents sont ré-indexés par le système FASTR (Jacquemin, 1999), analyseur transformationnel de surface qui reconnaît les occurrences et les variantes des termes produits par le module d'extraction de

[1] Le NIST est l'organisateur des conférences TREC.

termes. Chaque occurrence ou variante constitue un index qui est ensuite utilisé dans le processus de classement des documents. En effet, ces index permettent à QALC de réordonner les documents et de sélectionner les plus pertinents (Ferret et al. 2001). Le module de reconnaissance des entités nommées est ensuite appliqué sur les documents sélectionnés.

Le module de recherche de la réponse réalise deux opérations : la sélection des phrases candidates et l'extraction de la réponse, en utilisant les informations extraites de la question ainsi que les documents étiquetés retenus.

3 L'analyse de la question

L'analyse de la question est réalisée afin d'extraire des questions certaines caractéristiques qui pourront être utilisées dans le module de recherche de la réponse. Les caractéristiques de la question que nous détectons sont le type attendu de la réponse, le focus, et la catégorie de la question. Pour les trouver, nous utilisons des connaissances syntaxiques et sémantiques. Les informations syntaxiques sont fournies par un analyseur syntaxique robuste (Aït-Mokhtar et al., 1997) appliqué à toutes les questions. Cet analyseur renvoie une segmentation en groupes pour chaque question ainsi qu'un ensemble de relations syntaxiques entre ces groupes. Les règles pour trouver le focus, le type de la question et le type attendu de la réponse ont été écrites à partir de ces représentations syntaxiques. Les connaissances sémantiques, provenant de WordNet (Fellbaum, 1998), permettent de compléter les lexiques correspondant à chaque type d'entité nommée et d'améliorer la détermination du type attendu de la réponse.

- Le type attendu de la réponse

Dans un premier temps, le module d'analyse essaie de trouver si le type attendu de la réponse correspond à une ou plusieurs entités nommées classées par ordre d'importance. Les entités nommées détectées font partie d'une hiérarchie organisée en 22 classes sémantiques (Organisation, Ville, Poids...). Les règles de reconnaissance utilisent essentiellement le type de l'interrogatif et l'appartenance du nom sur lequel il porte à l'une des listes constituées pour chaque catégorie. Pour la conférence d'évaluation TREC10, notre module de question trouve environ 90,5 % de types d'entités nommées attendus corrects.

Exemple : Question : Who developed the Macintosh Computer ?
(Qui a conçu l'ordinateur Macintosh ?)
Type attendu EN de la réponse = { PERSONNE , ORGANISATION }

Ensuite, si le type de la réponse n'est pas une entité nommée, QALC essaiera de déterminer un type sémantique plus général correspondant au nom sur lequel porte l'interrogatif si ce nom est une entrée de WordNet. Pour la conférence d'évaluation TREC10, notre module de question a trouvé 87 % de types attendus généraux corrects.

Exemple : Question : What metal has the highest melting point ?
(Quel métal a le plus haut point de fusion ?)
Type attendu général de la réponse = métal

- Le focus de la question

Le focus de la question correspond à un nom de la question qui idéalement devrait être présent dans la phrase réponse. Pour chaque question, nous déterminons donc le focus, ainsi que ses modifieurs (adjectif, complément du nom...) qui interviendront également dans la recherche de la réponse. Les règles déterminant le focus dépendent essentiellement de la forme de la question, et celui-ci correspond souvent au sujet. Pour le corpus de la conférence d'évaluation TREC10, notre module de question trouve 85 % de focus corrects.

Exemple : Question: Who was the first governor of Alaska?
(Qui a été le premier gouverneur de l'Alaska ?)

Groupe nominal du focus = the first governor of Alaska
Focus = governor
Modifieurs du focus = ADJ first, COMP Alaska

- La catégorie de la question

La détection de la catégorie de la question nous permet de différencier les patrons syntaxiques d'extraction à appliquer aux phrases candidates à la réponse. La catégorie de la question correspond à la forme syntaxique de la question, plus ou moins détaillée. Par exemple :

Question: What do bats eat? (Que mangent les chauves-souris ?)
Catégorie de la question = What-Do-GN-VB[2]

Question: When was Rosa Park born? (Quand Rosa Park est-elle née ?)
Catégorie de la question = When-Be-NP-born

Après avoir étudié les questions des conférences TREC8 et TREC9 accompagnées des différentes phrases contenant les réponses, nous avons trouvé 82 formes syntaxiques de questions.

4 La recherche de la réponse

4.1 La reconnaissance du focus dans les documents

Pour chaque question, le module d'analyse nous a fourni le nom représentant le focus et ses modifieurs. Le système QALC cherche à les localiser dans les phrases des documents sélectionnés, en détectant d'abord le focus, puis en identifiant le groupe nominal qui le contient. Nous avons défini une grammaire locale pour détecter les frontières du groupe nominal. Cette grammaire se fonde sur l'étiquetage produit par le Tree-Tagger (Smidt, 1999).

Exemple : Question: Who is the creator of the Muppets?
(Quel est le créateur des Muppets ?)

Focus = creator
Modifieur du focus = COMP Muppets

---

[2] GN signifie groupe nominal, VB signifie verbe, et NP, nom propre.

Pour cette question, nous avons trouvé dans l'un des documents sélectionnés le groupe nominal *late Muppets creator Jim Henson*, qui correspond à l'expression :

Adjectif + Nom Pluriel + Nom + Nom Propre + Nom Propre

QALC recherche aussi les groupes nominaux contenant des synonymes du focus, synonymes déterminés par FASTR. Lorsqu'on ne trouve le focus dans aucune des phrases des documents sélectionnés pour la question, QALC recherche dans les phrases des groupes nominaux contenant les noms propres de la question, s'il y en a. Les noms propres font partie des éléments qu'une phrase réponse a de grande chance de mentionner.

QALC associe un poids à chaque groupe nominal trouvé. Ce poids prend en compte l'origine du groupe nominal (focus, synonyme du focus, ou nom propre), ainsi que les modifieurs trouvés dans la question. Lorsque le groupe nominal contient un ou plusieurs modifieurs, le poids est augmenté en proportion des modifieurs trouvés. Le poids le plus fort est obtenu lorsque tous les modifieurs sont présents. Dans l'exemple décrit plus haut, le poids est maximal : le groupe nominal a été obtenu à partir du focus *creator*, et le modifieur *Muppets* est présent. Lorsque le groupe nominal est obtenu à partir d'un synonyme du focus, le poids est légèrement baissé, il l'est encore plus lorsque le groupe nominal est obtenu à partir d'un nom propre. Par exemple, le poids attribué au groupe nominal *their copy of the 13th century Magna Carta* trouvé pour la question *Which king signed the Magna Carta ?*, n'aura pas le poids maximal car il n'a pas été obtenu à partir du focus *king*, mais à partir du nom propre *Magna Carta*. D'une manière générale, l'algorithme de pondération tient toujours compte du rapport entre le nombre de mots significatifs de la question, et le nombre de ces mots retrouvés dans le groupe nominal sélectionné dans le document.

Pour chaque phrase du document sélectionné, le système QALC étiquette tous les groupes nominaux pertinents suivant les critères décrits, avec leurs poids. Seuls les groupes nominaux obtenant les plus forts poids sont gardés, fournissant un critère d'évaluation de la phrase.

4.2 La sélection des phrases

La méthode que nous avons adoptée dans cette version de QALC pour sélectionner les phrases susceptibles de contenir une réponse à la question posée reprend globalement celle définie pour ses versions précédentes : grâce à une fonction de comparaison entre phrases déterminant leur proximité relative par rapport à une question, QALC parcourt phrase par phrase les documents sélectionnés pour une question et retient les *N* phrases les plus proches de celle-ci[3]. Cette fonction de comparaison repose sur les caractéristiques de la question considérée qui sont identifiées dans les phrases des documents par les modules linguistiques exposés précédemment. Ces caractéristiques sont 1) les termes de la question, 2) le focus et ses modifieurs, 3) le type de la réponse attendue, que l'on retrouve au niveau des phrases des documents sous la forme d'entités nommées.

[3] N est au moins égal à 5. Les phrases sélectionnées sont classées par ordre de similarité décroissante avec la question.

Un score spécifique est associé à chacun de ces traits et le score de la phrase correspond à une combinaison linéaire de ces poids. Le principal critère de sélection repose sur la présence d'un maximum de mots de la question dans la phrase, sachant que l'on privilégie les phrases contenant le nom focus. Le critère portant sur la présence des entités nommées est utilisé pour renforcer l'importance d'une phrase, mais non comme critère premier de sélection. Un dernier critère, utilisé pour départager les phrases *ex æquo*, favorise les phrases dont le taux de dispersion des termes de la question est le plus faible. Ce taux est donné par la taille de la plus petite partie de la phrase contenant tous les termes reconnus de la question. Le détail de la méthode de sélection peut être trouvée dans (Ferret et al., 2001).

4.3 L'extraction de la réponse

Le processus d'extraction est différent selon que le type attendu de la réponse est une entité nommée ou non. En effet, lorsque la réponse est une entité nommée, l'extraction consiste à localiser l'entité nommée dans la phrase candidate à la réponse. Elle repose donc essentiellement sur les résultats du module de reconnaissance des entités nommées dans les documents. En revanche, lorsque la réponse n'est pas une entité nommée, le processus d'extraction repose alors sur la reconnaissance du focus de la question et consiste à reconnaître, dans la phrase, les patrons d'extraction de la réponse.

4.3.1 Les réponses caractérisées par une entité nommée

Lorsque le système QALC peut prédire le type de la réponse en terme d'entité nommée, le processus d'extraction recherche dans la phrase candidate toutes les expressions étiquetées par le type attendu. Lorsque plusieurs expressions sont trouvées, QALC sélectionne la plus proche, en terme de distance, du focus, si celui-ci a été reconnu dans la phrase, ou la première rencontrée sinon. Si aucune expression du type attendu n'a été trouvée, QALC généralise le type recherché en utilisant la hiérarchie des entités nommées que nous avons définie. Par exemple, le type *PERSONNE* est généralisé par le type *NOM PROPRE*, et le type *LONGUEUR* par le type *NOMBRE*.

4.3.2 Les réponses de type « noms communs » et « groupe verbal »

Lorsque le type attendu de la réponse n'est pas une entité nommée, le système QALC détermine la position de la réponse dans la phrase à l'aide de patrons syntaxiques d'extraction de la réponse. Un patron syntaxique d'extraction est une structure qui comporte le groupe nominal du focus dans la phrase candidate et le groupe nominal de la réponse, éventuellement séparés par d'autres éléments tels que virgules, guillemets, verbe ou préposition. Le patron d'extraction comporte toujours le focus de la question. Pour pouvoir reconnaître dans la phrase candidate un patron d'extraction, il faut donc que le focus de la question ait été préalablement déterminé par le module d'analyse des questions. Prenons par exemple la question suivante :

Question : What do Knight Ridder publish? (Que publie Knight Ridder ?)

Cette question a une structure syntaxique du type : What-do-GN-VB, *Knight Ridder* étant le groupe nominal (GN) et *publish* le verbe (VB). Le focus, déterminé par les règles du module d'analyse de la question, est *Knight Ridder*.

Pour ce type de question, un des patrons que nous avons retenu correspond à la séquence grammaticale suivante :

GNfocus VB GNréponse

Le GNfocus est le groupe nominal du focus dans la phrase-réponse. Il est suivi d'un séparateur qui est ici le verbe *publish*, puis d'un groupe nominal qui est supposé contenir la réponse à la question, appelé GNréponse. La réponse suivante, résultant de l'application de ce patron, a été trouvée dans le corpus :

Knight Ridder publishes 30 daily newspapers ... , extraite de la phrase :

Knight Ridder publishes 30 daily newspapers, including the Miami Herald and the Philadelphia Inquirer, owns and operates eight television stations and is a join venture partner in cable television and newsprint manufacturing operations.

(*Knight Ridder publie 30 quotidiens, dont le Miami Herald et le Philadelphia Inquirer,il possède et gère 8 chaînes de télévision et est partie prenante dans des opérations liées à la télévision cablée et à la fabrication de papier journal.*)

Dans le paragraphe 3.3, nous avons vu qu'environ 80 catégories de questions ont été répertoriées. Parmi celles-ci, 45 n'attendent pas en réponse une entité nommée. Pour chacune de ces 45 catégories de question, nous avons construit des patrons syntaxiques d'extraction. Les différents patrons, ainsi que les différentes catégories de question, ont été déterminés par l'étude des corpus de questions et de réponses fournis par les campagnes d'évaluation des conférences TREC8 et TREC9. En tout, nous avons retenu 24 patrons d'extraction. Le nombre de patrons par catégorie de question va de 2 à 20, avec une moyenne de 10 par catégorie de question. Les différentes catégories de questions ont donc plusieurs patrons en commun.

Les patrons d'extraction sont plus difficiles à établir pour certaines catégories de question que pour d'autres. Cela peut être en raison de leur rareté dans le corpus ou en raison de la variété grammaticale des réponses. Les questions de type *Why* (*Pourquoi*), comme *Why can't ostrishes fly* ? (*Pourquoi les autruches ne peuvent-elles voler ?*) sont rares dans le corpus (seulement 4) ainsi que les questions en *How VB* (*Comment Verbe*) (4 également), comme *How did Socrates die ?* (*Comment Socrate est-il mort ?*). De plus, les réponses à ces questions sont difficiles à modéliser sous la forme d'un patron syntaxique. Il est difficile également de trouver des régularités grammaticales dans les réponses aux questions du type *What-GN-be-GN* , comme *What format was VHS's main competition* ? (*Quel était le principal concurrent du format VHS ?*) ou encore *What nationality was Jackson Pollock ?* (*Quelle était la nationalité de Jackson Pollock ?*). Suivant les cas, c'est le premier GN (*format*) ou le deuxième (*Jackson Pollock*) qui joue un rôle prépondérant dans le patron d'extraction.

5 Analyse des résultats

Nous avons réalisé une évaluation des réponses que nous trouvons sur le corpus de la conférence TREC10, suivant que la question appartient à une catégorie définie par une entité nommée ou non. Cela nous permet d'évaluer indépendamment l'un de l'autre le processus de localisation de l'entité nommée attendue en réponse et le processus qui utilise les patrons d'extraction. L'évaluation a d'abord été faite sur les phrases candidates à la réponse puis sur les réponses à 50 caractères, afin de mesurer la performance des deux modules constituant la recherche de la réponse dans les documents sélectionnés. Le tableau de la Figure 2 montre que QALC a de meilleures performances sur les catégories définies par une entité nommée. En effet, ce type de question comporte en général plusieurs termes, ce qui facilite la sélection des phrases candidates par rapport aux questions du type *demande de définition* qui ne comportent généralement qu'un seul terme. Par ailleurs, l'extraction de la réponse est facilitée par l'étiquetage des entités nommées dans les documents.

Questions	Nombre de questions	Nombre de réponses correctes (phrases)	Nombre de réponses correctes (50 car.)	Extraction correcte (phrase -> 50 car.)
Catégorie définie par une entité nommée	229	96 (42 %)	82 (36 %)	85 %
Autre catégorie	263	103 (39 %)	63 (24 %)	61 %

Figure 2 : Les réponses suivant la catégorie de question

QALC a obtenu 0.18 à l'évaluation TREC10. La version TREC10 du système QALC a obtenu de meilleurs résultats que la version précédente pour les catégories de question qui ne sont pas définies par une entité nommée. En effet, on obtient une réponse correcte pour 24% de ces questions, contre seulement 10% dans la version précédente.

6 Conclusion

Le système QALC, que nous venons de décrire, sélectionne d'abord les 10 premières phrases qui répondent à la question posée, pour en extraire ensuite une réponse précise. Ce principe est efficace lorsque les phrases trouvées ont des poids très différents les uns des autres. Dans ce cas, en effet, la réponse a de fortes chances de se trouver dans les 10 premières phrases sélectionnées. En revanche, lorsqu'un grand nombre de phrases ont le même poids, la réponse peut se trouver aussi bien dans l'une des dernières phrases que dans la première. C'est plus particulièrement le cas des questions de type définition comme *Qu'est-ce qu'un micron ?* , qui n'ont qu'un seul terme caractéristique et n'attendent pas d'entité nommée en réponse. Dans ce cas, il sera probablement plus efficace d'appliquer les patrons d'extraction à un nombre important de phrases, afin de les discriminer par ces patrons et non par les termes.

Une autre difficulté que nous avons rencontrée dans l'extraction de la réponse est la constitution même des patrons syntaxiques d'extraction. Ces patrons émergent des corpus de questions/réponses, et nécessitent donc d'abord la constitution de grands corpus. Par ailleurs, certaines catégories de questions se prêtent moins bien que d'autres à la définition de patrons. Pour ces catégories, il faudra trouver d'autres solutions. L'une d'entre elles consiste à utiliser la hiérarchie des catégories sémantiques de Wordnet pour les questions dont le type général

de la réponse attendue peut être déterminé. En effet, la connaissance du type attendu de la réponse (lorsqu'il existe) facilite la recherche de la réponse dans la phrase. Actuellement, seul le type correspondant à une entité nommée est utilisé, et non le type sémantique général. Par exemple, la question *What language is mostly spoken in Brazil* ? (*Quelle langue est la plus parlée au Brésil ?*) attend en réponse le nom d'une langue. Ce type n'étant pas étiqueté dans le corpus, QALC ne peut pas le localiser dans les phrases candidates. En revanche, nous avons testé l'utilisation de WordNet pour vérifier si un mot de la phrase candidate est du type *langue*. En effet, la réponse dans notre exemple est *le portugais* , qui fait partie des hyponymes de *langue* dans WordNet.

Nous serons donc de plus en plus amenés à adopter des stratégies différentes suivant la catégorie de la question. Nous devrons alors spécifier plus précisément les catégories de questions, ainsi que les patrons d'extraction de la réponse qui leur sont associés afin d'améliorer l'application de ces patrons.

Références

Alpha S., Dixon P., Liao C., Yang C. (2001), Oracle at TREC10, Actes de *TREC10*, 419-428.

Aït-Mokhtar S. and Chanod J. (1997), Incremental Finite-State Parsing, Actes de *ANLP-97*.

Ferret O., Grau B., Hurault-Plantet M., Illouz G., Jacquemin C. (2001), Document selection refinement based on linguistic features for QALC, a question answering system, Actes de *RANLP 2001*.

Ferret O., Grau B., Hurault-Plantet M., Monceaux, L., Robba, I., Vilnat, A., (2001), Finding an answer based on the question focus, Text retrieval conference, TREC 10.

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

Harabagiu S., Moldovan D., Pasca M., Surdeanu M., Mihalcea R., Girju R., Rus V., Lacatusu F., Morarescu P., Bunescu R. (2001), Answering complex, list and context questions with LCC's Question-Answering Server, Actes de *TREC10*.

Hovy E., Hermjakob U., Lin C. (2001), Actes de *TREC10*, 166-174.

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

Lehnert W. (1979),*The process of question answering*, Lawrence Erlbaum Associates, 1979.

Schmid H. (1999), *Improvements in Part-of-Speech Tagging with an Application To German*, Dordrecht, Editeurs Armstrong S., Church K.W., Isabelle P., Manzi S., Tzoukermann E., and Yarowsky D., *Natural Language Processing Using Very Large Corpora*, Kluwer Academic Publishers.

Soubbotin M.M., Soubbotin S.M. (2001), Patterns of Potential Answer expressions as Clues to the Right Answers, Actes de *TREC10*, 175-182.