Les analyseurs syntaxiques : atouts pour une analyse des questions dans un système de question-réponse ?

Laura Monceaux, Isabelle Robba

LIMSI / CNRS – Université Paris XI
Bat 508, 91403 Orsay Cedex
{Laura.Monceaux, Isabelle.Robba}@limsi.fr

Résumé – Abstract

Cet article montre que pour une application telle qu'un système de question – réponse, une analyse par mots clés de la question est insuffisante et qu'une analyse plus détaillée passant par une analyse syntaxique permet de fournir des caractéristiques permettant une meilleure recherche de la réponse.

In this paper, we show that in a question-answer application, a light parsing using only keywords is not sufficient. A syntactico-semantic parsing of the question must be used and must be completed with some others characteristics, which will a best search for the good answer.

Mots Clés – Keywords

Analyse syntaxique de questions, connaissances sémantiques, détection du focus
Syntactico-semantic parsing of questions, semantic knowledge, focus detection

1 Introduction

Pour le système de question-réponse QALC (Ferret et al., 2001), que nous mettons au point au LIMSI dans le groupe LIR, le module d'analyse des questions est un module essentiel, car il a pour rôle de déterminer avec précision les caractéristiques qui permettent de guider finement la recherche de la bonne réponse. Dans cet article, nous exposons tout d'abord un exemple où nous soulignons toute l'importance d'une analyse syntaxique et sémantique de la question (§ 2). Puis nous présentons les différentes catégories d'analyseurs robustes qui existent aujourd'hui (§ 3), et nous montrons à travers quelques exemples les problèmes que peut poser leur utilisation (§ 4). Finalement (§ 5), nous décrivons en détails le fonctionnement de notre module d'analyse de la question en définissant les caractéristiques qui sont calculées par ce module (focus, type attendu de la réponse), et en indiquant en quoi elles peuvent contribuer à une meilleure recherche de la réponse.

2 D'une analyse par mots clés à une analyse syntaxico-sémantique des questions

Dans un système de question-réponse, le module d'analyse des questions a pour rôle de fournir aux autres modules des informations afin de leur faciliter la sélection des phrases candidates et d'en extraire la réponse. Généralement, ce module d'analyse de la question retourne deux types d'information :

- Le type attendu de la réponse sous forme d'entité nommée : une Personne, un Nombre, un Lieu …
- Les termes les plus importants de la question : ceux qui devraient idéalement se trouver dans la réponse.

Cependant, après l'étude des résultats obtenus par notre système QALC à TREC-9, il s'est avéré que la détection seule des termes importants de la question ne suffisait pas toujours pour sélectionner les phrases contenant la bonne réponse. Et que pour sélectionner celles-ci, les relations entre les mots, syntaxiques voire sémantiques, jouaient un rôle primordial. Lorsque le domaine d'application, dans lequel on travaille, est ouvert, une analyse par mots clés est généralement insuffisante. Prenons l'exemple issu de TREC-9 (figure 1), si une analyse par mots clés est utilisée, le système sélectionnera indifféremment les deux phrases candidates pour extraire la réponse. En revanche si le système utilise une analyse plus détaillée, c'est-à-dire tenant compte des relations entre les mots, seule la première phrase sera sélectionnée. En effet, le module d'analyse détermine, dans ce cas, que l'on cherche une personne et que celle-ci doit avoir les caractéristiques suivantes : être astronaute, être de nationalité russe et être la première à avoir fait un voyage dans l'espace. Ainsi, si ces éléments sont pris en compte, la deuxième phrase candidate est rejetée par le système à cause de l'adjectif *US* du groupe nominal *the first US woman*, en effet des connaissances sémantiques et pragmatiques peuvent être utilisées pour détecter que les deux adjectifs *US* et *Russian* sont contradictoires. La prise en compte des liens entre les mots permet donc de rejeter des phrases candidates : celles contenant des liens contradictoires par rapport aux liens exprimés dans la question.

Question : " What was the name of the first Russian astronaut to do a spacewalk?"
Phrase candidate n°1 : The broad - shouldered but paunchy Leonov, who in 1965 became **the first man** to walk in space, signed autographs. He drew a dove beside his name for Amanda Clark, 8, of Altadena and pinned on his lapel a **Russian**-language button from a well-wisher reading « Let us …
Phrase candidate n°2 : Sullivan was **the first US woman** to perform a **spacewalk** as she and fellow crewman David Leetsma demonstrated satellite refueling …

Figure 1 : Phrases candidates pour une question donnée

Pour obtenir des connaissances sur les liens entre les mots, la première étape consiste à utiliser un analyseur syntaxique. En effet celui-ci permet d'obtenir, dans un premier temps, un découpage en groupes syntaxiques de la phrase : détection des groupes nominaux (NP), des groupes prépositionnels (PP) … mais il peut également parfois identifier des relations syntaxiques de dépendance entre ces différents groupes de mots ou directement entre les mots de la phrase comme les relations Sujet, Objet, Rattachement prépositionnel (voir la figure 2).

Les résultats de l'analyse syntaxique vont nous permettre de compléter les informations obtenues par le module d'analyse de la question pour une meilleure sélection de la réponse.

Question: " What was the name of the first Russian astronaut to do a spacewalk?"
Segmentation: *[SC [NP* What *NP] : v* was *SC] [NP* the name *NP] [PP* of the first Russian astronaut *PP] [IV* to do *IV] [NP* a spacewalk *NP]* ?
Relations de dépendance: SUBJECT (what, was), ADJECTIVE (first, astronaut), ADJECTIVE (Russian, astronaut), etc.

Figure 2 : Résultats de l'analyse syntaxique d'une question

Les résultats obtenus par l'analyseur syntaxique peuvent également être utilisés pour faciliter l'extraction de la réponse dans les phrases candidates. En effet, l'analyse de la question nous donne parfois des indices sur la fonction syntaxique de la réponse attendue. Prenons l'exemple de la figure 3 (tiré de Hovy 2001), les informations obtenues par l'analyse syntaxique de la question *Who killed Lee Harvey Oswald*? permettent non seulement de déterminer le type attendu de la réponse : une Personne, mais également de savoir que la personne cherchée est l'agent (qui est souvent le sujet) du verbe *kill* (ou un synonyme de celui-ci) et que l'objet de ce verbe doit être *Oswald.* Cela permet donc de déterminer qu'ici la réponse est *Jack Ruby* et non pas *J. F. Kennedy.*

Question : " Who killed Lee Harvey Oswald ? "
Analyse syntaxique de la question : Subject (who, kill) Object (kill, Oswald) → Answer's Function = Subject

Réponse : "*Jack Ruby*, who killed *J.F Kennedy* assassin Lee Harvey Oswald"
Analyse syntaxique de la réponse : Subject (Ruby, kill), Object (kill Oswald), Noun-Modifier (assassin, Oswald), Noun-Modifier (Kennedy, assassin), etc.

Figure 3 : Utilisation des relations de dépendances pour l'extraction de la bonne réponse

L'analyse de la question nécessite donc des connaissances syntaxiques et sémantiques et passe par une première étape qui est l'analyse syntaxique de la question. Ainsi, notre premier objectif est de choisir quel type d'analyseur est le mieux adapté pour une telle application et surtout dans un premier temps quel type d'analyseur syntaxique.

3 Les analyseurs syntaxiques robustes

Depuis quelques années, les analyseurs syntaxiques ont connu un important développement. Alors que les premiers analyseurs avaient été développés dans le but de reconnaître des phénomènes linguistiques dans les différentes phrases étudiées, aujourd'hui les analyseurs syntaxiques sont dits plus « réalistes », dans le sens où ils s'attaquent à l'analyse de corpus réels, c'est-à-dire qui peuvent contenir des phrases agrammaticales mais aussi des phénomènes spécifiques comme des marques d'hésitations (issues de dialogues), des liens HTML, des tags XML… La priorité de ces nouveaux analyseurs est donc la robustesse, d'où leur appellation d'analyseurs robustes. La robustesse est la capacité de retourner une analyse syntaxique (même minimale) pour tout type d'entrées. Ces analyseurs robustes sont déterministes, incrémentaux et se différencient par les résultats qu'ils retournent (segmentation minimale voire plus complexe et parfois détection de relations syntaxiques) et par les heuristiques qu'ils utilisent pour extraire l'analyse la plus probable. Cependant, certains analyseurs peuvent produire plusieurs solutions, c'est par exemple le cas lors d'ambiguïté dans les rattachements prépositionnels.

Nous pouvons distinguer deux types d'analyseurs robustes : les analyseurs linguistiques, fondés sur des formalismes grammaticaux, et les analyseurs probabilistes, fondés sur l'apprentissage à partir de corpus. Dans un premier temps, nous avons choisi d'étudier les analyseurs linguistiques par rapport aux analyseurs probabilistes, car si ces derniers ont l'avantage d'être entraînés sur des corpus, il n'en reste pas moins qu'il faut construire ces corpus de référence, et que des questions se posent alors : comment annoter ces corpus, de quelle taille doivent–ils être ?

Au début de leur développement, les analyseurs linguistiques se répartissaient en trois catégories en fonction des résultats qu'ils produisaient :

1. Les analyseurs fondés sur les constituants (SCOL d'Abney, 96 – IPS de Wehrli, 92…) qui retournent essentiellement une segmentation en groupes. Dans un premier temps, la phrase est segmentée en unités lexicales, puis de façon incrémentale en constituants.

2. Les analyseurs fondés sur les dépendances (Link Grammar de Sleator et Temperley, 91…) qui retournent les dépendances entre les mots. Le Link Grammar Parser (LGP par la suite) utilise un dictionnaire spécifique dans lequel chaque mot correspond à une formule logique qui spécifie les liens que ce mot peut avoir avec les autres mots. Le but de cet analyseur est de construire un « chemin de liens » entre les mots de la phrase en respectant certaines propriétés : les liens entre les mots ne doivent pas se croiser et l'ensemble de ces liens doit être connexe.

3. Les analyseurs fondés sur les constituants et les dépendances (IFSP d'Aït-Mokthar et Chanod 97 développé chez Xerox – Vergne, 97…) qui retournent une segmentation en groupes et des relations de dépendances entre ces groupes. Ces analyseurs ne sont pas monotones : certaines décisions peuvent être affinées ou remises en cause. Différentes stratégies sont utilisées, mais généralement ces analyseurs cherchent à obtenir, dans un premier temps, une segmentation minimale de la phrase et des relations syntaxiques simples (comme le sujet) puis des informations plus complexes sur la segmentation et sur des relations syntaxiques plus délicates à trouver.

Aujourd'hui la plupart des analyseurs robustes peuvent fournir la segmentation de la phrase et certaines relations de dépendances. Mais ce en quoi ils diffèrent est le format de leurs résultats et surtout les processus mis en œuvre pour obtenir ces résultats. Afin de choisir quel type d'analyseur est le plus adéquat pour notre application, nous avons étudié les analyseurs cités ci-dessus sur un corpus de questions.

4 Les analyseurs robustes sont-ils adaptés au traitement des questions ?

Pour la conférence d'évaluation TREC-10, nous avons choisi d'utiliser l'analyseur IFSP pour notre module d'analyse de la question. Comme la plupart des analyseurs robustes, IFSP a été développé essentiellement pour des applications manipulant de grandes bases de données textuelles comme la recherche d'informations ou l'extraction terminologique. Et de ce fait, comme les autres analyseurs robustes, il n'est pas particulièrement adapté pour analyser des formes interrogatives.

4.1 Exemples de problèmes rencontrés lors de l'analyse des questions

Dans les exemples suivants, nous présentons certaines formes syntaxiques de questions pour lesquelles l'analyse réalisée par IFSP est erronée et nous comparons cette analyse avec les résultats obtenus par d'autres analyseurs. Pour une question de temps, mais aussi de complexité, nous n'avons pas élaboré une évaluation approfondie sur tout le corpus[1] de questions de TREC. Cette étude d'exemples a été faite à partir d'erreurs d'analyse d'IFSP, il n'est donc pas surprenant que les résultats, comparés à ceux des autres analyseurs, n'apparaissent pas comme étant les meilleurs. Voici quelques exemples d'erreurs :

- Le verbe reconnu comme un nom :

Dans l'exemple *What year did the Titanic sink?*, les quatre analyseurs testés (IFSP, IPS, LGP et SCOL) commettent la même erreur : ils reconnaissent *sink* comme un nom alors qu'il s'agit ici d'un verbe. Mais dans l'exemple *Why does the moon turn orange?,* seul LGP ne se trompe pas et reconnaît bien *turn* comme un verbe.

- Le superlatif reconnu comme un nom :

Dans l'exemple *What metal has the highest melting point?,* seul IFSP produit une analyse erronée : il segmente la phrase en 3 groupes de mots (un pour *highest*, un autre pour *melting* et un dernier pour *point*).

- L'adjectif reconnu comme un nom :

Dans l'exemple *Who is the Prime Minister of Canada?,* l'utilisation d'une majuscule pour l'adjectif *Prime* entraîne une erreur pour IFSP et SCOL : l'adjectif n'est pas reconnu et le groupe nominal *Prime Minister* est segmenté en deux groupes nominaux distincts (un pour *Prime,* un autre pour *Minister*).

Pour cette raison, notre module d'analyse de la question doit rattraper les erreurs provenant de l'analyse syntaxique, ce qui a été fait dans une certaine mesure par l'écriture de règles spécifiques. Par exemple, concernant la première erreur – le verbe reconnu comme un nom – nous avons écrit la règle suivante :

> Si la forme de la phrase est *What/Why/When … auxiliary NP?*,
> et si aucun verbe autre que l'auxiliaire n'a été reconnu
> alors le verbe est probablement le dernier élément du NP

Cette règle a pu être appliquée à un certain nombre de formes telles que : *What year did [NP the Titanic sink NP] ?*, *When did [NP the Hindenberg crash NP] ?*, mais elle s'est aussi parfois révélée inefficace comme dans le cas de *Why does [NP the moon turn orange NP]?* pour lequel une étude plus approfondie permettra d'écrire de nouvelles règles.

Malgré ces exemples, IFSP présente un atout important : il retourne toute l'information nécessaire sans que l'on ait besoin de modifier ou d'ajouter de nouvelles connaissances ou de nouveaux processus. En effet, IFSP retourne d'une part la segmentation en constituants, et d'autre part des relations syntaxiques entre les termes, alors que IPS, SCOL ou LGP retournent une structure unique, dans laquelle les relations de dépendances ne sont pas

[1] Au LIMSI, dans le groupe LIR, un projet sur l'évaluation des analyseurs syntaxiques (Gendner et al, 02) pour le français est actuellement en cours de développement et une partie du corpus retenu est constitué de 500 formes interrogatives.

étiquetées. Certes, il est possible d'obtenir de SCOL ou de LGP l'étiquetage de ces relations, mais cela demande un travail de modification, c'est pourquoi dans une première approche, IFSP convenait mieux.

4.2 Vers une meilleure analyse syntaxique

Dans la mesure où les résultats obtenus par notre module d'analyse des questions sont satisfaisants (les résultats sont présentés § 5.4), nous pourrions choisir de continuer dans la voie choisie jusqu'à présent et donc de compléter les règles qui permettent le rattrapage des erreurs provenant de l'analyse. D'autant plus, que chez Xerox, XIP, un nouvel analyseur est aujourd'hui disponible pour le français et est en cours de développement pour l'anglais. XIP devrait être plus performant que IFSP, notamment pour traiter les formes interrogatives. Mais d'autres solutions peuvent aussi être étudiées :

- Modification des analyseurs syntaxiques

Une possibilité est de choisir un analyseur pour lequel il est facile et rapide d'écrire les règles syntaxiques nécessaires à l'analyse de formes interrogatives. SCOL, par exemple, a l'avantage d'avoir une grammaire facilement modifiable. De plus, les erreurs d'analyse peuvent parfois être dues à un mauvais étiquetage morpho-syntaxique. C'est en effet le cas du Tree Tagger qui ne reconnaît pas par exemple *turn* ou *die* comme des verbes dans les phrases suivantes : *Why does the moon turn orange? How did Janice Joplin die?*. Dans l'analyseur SCOL, le marquage morpho-syntaxique n'est pas intégré dans le processus d'analyse, aussi il est possible de séparer les tests et de localiser les modules qui posent problème.

- Utilisation d'un analyseur par apprentissage

Contrairement à notre premier choix, nous pouvons adopter la solution d'un analyseur probabiliste. Il faudra alors l'entraîner sur un corpus annoté de formes interrogatives, celles de TREC (1300 tournures sont disponibles), mais également d'autres que l'on extraira d'Internet – qui a pour intérêt de proposer un échantillon très large de questions non artificielles. Pour leur système Webclopedia, Hovy et al. (2001) ont entraîné leur analyseur, CONTEXT, sur environ 1150 questions obtenant ainsi une précision et un rappel de 89 %.

- Développement d'un analyseur intégré

Nous pouvons aussi envisager une toute nouvelle approche consistant en l'élaboration d'un processus complet, intégrant à la fois l'analyse de la question et l'extraction des informations qui sont spécifiques à la tâche de question-réponse (type de la question, focus de la question… voir § 5). Il faut aussi noter que la résolution de la référence pourrait être réalisée dans ce même module. En effet dans la dernière session de TREC, une nouvelle tâche a été proposée : la tâche Contexte. Les systèmes qui participaient à cette tâche devaient identifier les objets du discours à travers une série de questions comme : *Which museum in Florence was damaged by a major bomb explosion in 1993? On what day did this happen? Which galleries were involved ?*. Pour mener à bien cette tâche, les systèmes ont à résoudre de nouveaux problèmes : les expressions référentielles doivent être résolues – *this* réfère à *explosion* – et les relations entre les objets du discours doivent être établies – comme la relation partie-de qui existe entre *galleries* et *museum*.

Dans le but de participer à la tâche Contexte, la dernière solution, celle qui envisage le développement d'un module d'analyse intégré, nous paraît être celle à adopter même si elle est probablement plus complexe à mettre en œuvre.

5 Analyse d'une question en langage naturel

Pour la conférence d'évaluation TREC-10, l'analyse de la question est réalisée, comme nous l'avons vu, dans le but d'extraire les caractéristiques de la question qui sont utilisées dans le module de recherche de la réponse. Grâce aux résultats obtenus par l'analyseur syntaxique, le module d'analyse de la question retourne plusieurs informations :

1. Le type de la question, auquel on peut associer une liste de patterns qui permettent une meilleure extraction de la réponse.
2. Le type attendu de la réponse qui peut être une entité nommée ou un type sémantique plus général qui permet de mieux localiser la réponse dans les phrases candidates.
3. Le focus de la question qui permet d'établir des critères pour la sélection des phrases candidates et qui aide à l'extraction de la réponse.

5.1 Type de la question

La détection du type de la question – qui correspond à sa forme syntaxique – donne un indice pour localiser la réponse dans les phrases candidates, en effet à chaque type de question correspond un ensemble de patterns de réponse (Ferret et al., 2001). Cette détection permet aussi l'extraction des autres caractéristiques de la question. Après avoir étudié les questions de TREC-8 et TREC-9 accompagnées de leurs réponses, nous avons déterminé 82 formes syntaxiques de question plus ou moins détaillées comme WhatbeGNofGN, HowADJ, WhatGNbeGN… qui nous permettent de classifier toutes les questions étudiées pour une meilleure sélection des autres caractéristiques de la question.

Dans les autres systèmes de question-réponse, la détermination du type de la question est plus ou moins complexe : par exemple dans le système d'Oracle (Alpha et al., 2001), le type correspond seulement à un pronom interrogatif (who, where, …), alors que Soubbotin et al. (2001) utilisent un ensemble d'environ 35 types. La détermination plus ou moins complexe du type de la question dépend essentiellement de l'utilisation qui en est faite dans les autres modules.

5.2 Type attendu de la réponse

Notre module d'analyse des questions assigne à chaque question un type attendu de la réponse, qui peut être une entité nommée ou un type sémantique plus général.

5.2.1 Type Entité nommée

Dans un premier temps, le module cherche si le type attendu de la réponse correspond à une ou plusieurs entités nommées ; s'il y en a plusieurs, celles-ci sont classées par ordre d'importance. Les entités nommées sont organisées en une hiérarchie qui contient 22 classes sémantiques (Ferret et al, 2001) telles que Personne, Organisation, Ville, Date, Poids, Période, Durée…, nombre qui nous parait suffisant pour une bonne extraction de la réponse sous forme d'entité nommée. Dans ce but, des lexiques[2] ont été construits à l'aide de la base

[2] Des lexiques de locutions sont également necessaires. Par exemple les locutions *life span* et *life expectancy* doivent être insérées dans un lexique associé à l'entitée nommée Durée, pour permettre de traiter des questions telles que *What is the average life span for a chicken?* et *What is the life expectancy of a dollar bill?*

lexicale WordNet (Fellbaum, 1998) : chacun de ces lexiques correspond à une liste de mots qui caractérise une entité nommée. Par exemple, le lexique LOCATION-CITY contient les mots suivants : capital, town, city, municipality…

La détection des entités nommées a été réalisée grâce à l'écriture de différentes règles. Les conditions de ces règles portent sur la structure syntaxique de la question et/ou l'appartenance d'un terme défini à un des lexiques d'entités nommées.

REGLE 1:
Si *Forme syntaxique de la question* = WhatBeGN1ofGN2?
et *la tête du GN1 appartient à un des lexiques de EN*
Alors *Entité Nommée* = EN qui correspond au lexique EN trouvé

Par exemple, la question *What is the capital of Bahamas?* a pour forme syntaxique WhatBeGN1ofGN2 ? (condition de la règle 1), son groupe nominal GN1 est *the capital* et son groupe nominal GN2 est *Bahamas*. La tête de GN1 *capital* appartient au lexique LOCATION-CITY, donc le type attendu de la réponse à cette question est LOCATION-CITY. Quand aucune entité nommée ne peut être trouvée à partir de la question, le module d'analyse de la question tente de déduire un type sémantique plus général.

5.2.2 Type sémantique plus général

Le type général que nous définissons correspond à un type sémantique appartenant à la base lexicale WordNet. Quand un type général est défini par notre système, la réponse doit être un nom ou un groupe nominal, hyponyme de ce type. Pour obtenir ce type plus général, nous avons écrit des règles qui utilisent la forme syntaxique de la question, par exemple :

REGLE 2:
Si *Forme syntaxique de la question = WhatGN1haveGN2?*
Alors *Type de la réponse = Tete du GN1*

Pour l'exemple suivant *What metal has the highest melting point?,* la règle 2 peut être appliquée car la forme syntaxique de cette question correspond à celle de la condition de cette règle ; et l'on obtient *metal* pour type attendu de la réponse. Ces différentes règles ont été écrites, après avoir étudié les questions de TREC-8 et TREC-9 ainsi que leurs réponses. Cette étude nous a permis de détecter des tournures spécifiques de l'anglais telle que *What city's newspaper is called the Enquirer?* où la présence du possessif implique l'écriture de règles supplémentaires.

5.3 Focus de la question

Le focus est un groupe nominal de la question. Celui-ci permet une meilleure sélection des phrases candidates en attribuant un poids plus fort aux phrases qui le contiennent. De plus, il permet l'écriture de patterns de réponses qui déterminent la position de la réponse par rapport à lui. Nous déterminons donc pour chaque question : le focus, la tête du focus (le nom principal) et les modificateurs de cette tête (adjectif, complément…). Ces différentes caractéristiques sont obtenues à partir d'un ensemble de règles ordonnées qui utilisent des connaissances syntaxiques et sémantiques décrites dans notre module, un lexique des mots abstraits par exemple.

Pour la forme syntaxique WhatbeGN1prepGN2? :
REGLE 3: **Si** TETE-GN1 appartient au lexique Abstrait
Alors FOCUS = GN2 et TETE-FOCUS = TETE-GN2

REGLE 4: **Si** TETE-GN1 appartient aux lexiques Personne ou Organisation
Alors FOCUS = GN1ofGN2 et TETE-FOCUS = TETE-GN1
REGLE 5: **Si** TETE-GN1 appartient à un autre lexique d'entité nommée
Alors FOCUS = GN2 et TETE-FOCUS = TETE-GN2

Si la question est *What is the abbreviation for Texas?*, la condition de la règle 3 est respectée, le focus de la question est *Texas*. En revanche pour la question *What is the diameter of a golf ball?,* la condition de la règle 3 n'est pas respectée : *diameter* n'est pas un mot abstrait, cependant il appartient au lexique LENGTH donc la règle 5 peut s'appliquer et le focus obtenu est : *a golf ball* et la tête de ce focus : *ball*.
Pour trouver si les conditions des règles sont respectées, le système utilise les résultats de l'analyseur IFSP avec nos modifications mais aussi des connaissances sémantiques grâce à la base lexicale WordNet. Notre système n'est pas le seul à utiliser la notion de focus correspondant à un ou plusieurs termes de la question : Soubbotin (2001), Alpha (2001) ou Hovy (2001) l'utilisent aussi. Mais l'originalité de notre approche est dans l'utilisation des relations syntaxiques qui existent entre la tête du focus et les autres termes de la question.

5.4 Résultats

Pour les questions de la conférence d'évaluation TREC-10, notre module d'analyse a obtenu :

- Pour le type attendu de la réponse : 90,5% d'entités nommées correctes et 87 % de types plus généraux eux aussi corrects
- Pour l'identification du focus : 85 % de focus corrects, 89,6 % de tête de focus corrects

Les erreurs peuvent être dues à des lexiques incomplets, à l'analyse syntaxique ou encore à des règles incomplètes ou manquantes.

Il est alors intéressant d'étudier les résultats en fonction du type de la question. En effet selon le type de la question, la reconnaissance du focus et du type attendu de la réponse sont plus ou moins difficiles à effectuer. On pourra prendre connaissance de résultats plus complets dans Hurault-Plantet et Monceaux (2002). De plus, cette étude du type de la question permet de détecter des règles incomplètes ou erronées. Par exemple, pour les questions commençant par *How many*, les règles permettant la détection de l'entité nommée ne sont pas satisfaisantes, car les entités nommées comme Volume, ou Poids ne sont jamais détectées, l'entité nommée Nombre est toujours retournée à leur place ; comme pour la question *How many liters in a gallon?*

6 Conclusion

Suite à l'étude des questions de TREC-8 et TREC-9 accompagnées de leurs réponses, nous avons déterminé qu'une analyse par mots clés était insuffisante pour sélectionner la bonne réponse. L'utilisation de connaissances syntaxiques et sémantiques est alors apparue primordiale pour d'une part une meilleure sélection des phrases candidates et d'autre part une meilleure extraction de la réponse. Dans un premier temps, nous avons donc étudié les analyseurs syntaxiques robustes : même si ces analyseurs ne sont pas forcément adaptés au traitement des formes interrogatives, ils apportent des informations essentielles (parfois modifiables quand elles sont erronées) pour notre module d'analyse. L'utilisation d'un

analyseur syntaxique nous a permis, pour TREC-10, de déterminer avec précision des caractéristiques telles que le focus et le type attendu de la réponse afin de guider plus finement la recherche de cette réponse.

Dans les années à venir, nous envisageons d'améliorer notre module d'analyse par l'extension de la requête grâce aux autres relations sémantiques non encore utilisées de WordNet : la synonymie et la méronymie. Nous travaillons actuellement sur les connaissances sémantiques voire pragmatiques qui sont nécessaires en fonction du type de la question. En effet nous pensons que certains types de question nécessitent une analyse plus approfondie : l'évaluation des résultats par type de question nous permettra cette étude. Par exemple, si la question est *What is the population of New York?;* il est pertinent de savoir que la réponse est un nombre et que celui-ci est de l'ordre du million plutôt que de la dizaine.

Références

Abney S. (1996), Partial Parsing via Finite-State Cascades , *Journal of Natural Language Engineering*, Vol. 2, n° 4, pp. 337-344.

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

Alpha S., Dixon P., Liao C., Yang C. (2001), Oracle at TREC-10, Actes de *la conférence d'évaluation TREC-10*, Gaithersburg.

Fellbaum Ch. (ed). (1998) *WordNet : An Electronical Database.* Cambridge : MIT Press.

Ferret O., Grau B., Hurault-Plantet M., Illouz G., Monceaux L., Robba I., Vilnat A. (2001), Finding an answer based on the recognition of the question focus, Actes de *la conférence TREC-10*, Gaithersburg MN.

Gendner V., Illouz G., Jardino M., Monceaux L., Paroubek P., Robba I., Vilnat A. (2002), A Protocol for Evaluating Analyzers of Syntax, *LREC2002*, Las Palmas, Spain.

Hovy E., Hermjakob U., Lin C-Y. (2001), The Use of External Knowledge in Factoid QA, Actes de *la conférence d'évaluation TREC-10*, Gaithersburg.

Hurault-Plantet M., Monceaux L. (2002), Cooperation between black box and glass box approaches for the evaluation of a question answering system, *LREC2002*, Las Palmas, Spain.

Sleator D.D., Temperley D. (1991), *Parsing English with a Link Grammar*, Rapport technique CMU-CS-91-196, Carnegie Mellon University, School of Computer Science.

Soubbotin M. M., Soubbotin S. M. (2001), Patterns of Potential Answer Expressions as Clues to the Right Answers, Actes de *la conférence d'évaluation TREC-10*, Gaithersburg.

Vergne J. (1997) Syntactic analysis of unrestricted french. In proceedings of the International *Conference on Recent Advances in NLP, RANLP-97*, Tzigov Chark, Bulgaria.

Wehrli E. (1992), The IPS system, Actes de *Coling-92*, Nantes, C. Boitet (éd), 870-874.