Apports de la linguistique dans les systèmes de recherche d'informations précises

Pierre Zweigenbaum, Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Sophie Rosset, Xavier Tannier, Anne Vilnat, Patrice Bellot

APPORTS DE LA LINGUISTIQUE DANS LES SYSTÈMES DE RECHERCHE D'INFORMATIONS PRÉCISES

Pierre Zweigenbaum, Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Sophie Rosset, Xavier Tannier, Anne Vilnat et Patrice Bellot

Pub. linguistiques | « Revue française de linguistique appliquée »

2008/1 Vol. XIII | pages 41 à 62

ISSN 1386-1204

Rev. franç. de linguistique appliquée, 2008, XIII-1 (41-62)

Apports de la linguistique dans les systèmes de recherche d'informations précises

Pierre Zweigenbaum, Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Sophie Rosset, Xavier Tannier, Anne Vilnat, CNRS-LIMSI
Patrice Bellot, Université d'Avignon

Résumé : *La recherche de réponses précises à des questions, aussi appelée « questions-réponses », est une évolution des systèmes de recherche d'information : peut-elle, comme ses prédécesseurs, se satisfaire de méthodes essentiellement numériques, utilisant extrêmement peu de connaissances linguistiques ? Après avoir présenté la tâche de questions-réponses et les enjeux qu'elle soulève, nous examinons jusqu'où on peut la réaliser avec très peu de connaissances linguistiques. Nous passons ensuite en revue les différents types de connaissances linguistiques que les équipes ont été amenées à mobiliser : connaissances syntaxiques et sémantiques pour l'analyse de phrases, rôle de la reconnaissance d'« entités nommées », prise en compte de la dimension textuelle des documents. Une discussion sur les contributions respectives des méthodes linguistiques et non linguistiques clôt l'article.*

Abstract: *Searching for precise answers to questions, also called "question-answering", is an evolution of information retrieval systems: can it, as its predecessors, rely mostly on numeric methods, using exceedingly little linguistic knowledge? After a presentation of the question-answering task and the issues it raises, we examine to which extent it can be performed with very little linguistic knowledge. We then review the different kinds of linguistic knowledge that researchers have been using in their systems: syntactic and semantic knowledge for sentence analysis, role of "named entity" recognition, taking into account of the textual dimension of documents. A discussion on the respective contributions of linguistic and non-linguistic methods concludes the paper.*

1. Introduction

La recherche d'information, aussi appelée recherche documentaire, est maintenant familière à tous les usagers de l'informatique, qui emploient quotidiennement des moteurs de recherche sur le web. La recherche de réponses précises à des questions est une évolution de cette fonctionnalité. Là où les systèmes de recherche d'information recherchent des documents qui contiennent les mots de la requête de l'utilisateur (p.ex. *capitale Chili*), les systèmes de recherche de réponses précises à des questions, aussi appelés systèmes de questions-réponses, s'attendent à recevoir une question (*Quelle est la capitale du Chili ?*) et doivent trouver non seulement les documents, mais surtout en extraire la réponse elle-même[1].

[1] D'où le nom de « système de recherche d'informations précises » utilisé dans Grau & Chevallet

Alors que les méthodes utilisées en recherche d'information sont essentiellement numériques, utilisant extrêmement peu de connaissances linguistiques (voir par exemple Gaussier & *al.* 2000, ou encore Moreau & *al*. 2007), qu'en est-il des systèmes de questions-réponses ?

Cet article commence par présenter plus précisément le contour de la tâche de questions-réponses et les enjeux qu'elle soulève, et pose l'architecture typique d'un système de questions-réponses (section 2). Il examine ensuite ce qu'il est possible de faire avec très peu de connaissances linguistiques, en prolongeant par exemple les méthodes de la recherche d'information (section 3). Il passe ensuite en revue les différents types de connaissances linguistiques que les travaux sur les systèmes de questions-réponses ont été amenés à mobiliser : connaissances syntaxiques et sémantiques pour l'analyse de phrases (section 4), rôle de la reconnaissance d'« entités nommées » (section 5), prise en compte de la dimension textuelle des documents (section 6). Une discussion sur les contributions respectives des méthodes linguistiques et non linguistiques clôt l'article.

2. Systèmes de questions-réponses

2.1. Tâche et enjeux

Les enjeux de la recherche de réponses précises à des questions peuvent être illustrés par la question suivante, reprise de Grau (à paraître) :

> *Quel est le nom du joueur de football qui a reçu le ballon d'or en 1994 ?*

pour laquelle une réponse peut être trouvée sur le web dans la phrase[2] :

> L'attaquant bulgare du FC Barcelone, Hristo Stoïchkov, est le lauréat du ballon d'or 1994 de l'hebdomadaire « France Football » récompensant le meilleur joueur européen de l'année sur la base des votes de journalistes représentant les 49 pays européens.

La difficulté pour trouver cette réponse est double : d'une part, il faut repérer un tel passage de texte, et d'autre part en extraire la réponse précise (« Hristo Stoïchkov »). La recherche de passages contenant la réponse se fait sur la base des mots principaux de la question : *joueur de football, recevoir, ballon d'or, 1994*. La phrase ci-dessus contient les mots recherchés sauf le verbe « recevoir ». Pour s'accommoder de cette différence, il faut connaître l'implication entre « être le lauréat de » et « recevoir », et tenir compte de cette implication pour extraire la réponse précise : il est utile de reconnaître le sujet de « être le lauréat de », et il vaut mieux vérifier que ce sujet est un « joueur de football ». Cette information peut être obtenue par l'apposition « L'attaquant bulgare du FC Barcelone, Hristo Stoïchkov », si l'on sait de plus qu'« attaquant » est une position qui peut être jouée par un « joueur de football ». On voit donc qu'interviennent des connaissances syntaxiques et sémantiques sans lesquelles il est généralement difficile de mettre en relation expression de la question et expression de la réponse. L'ensemble du présent article a pour objectif d'illustrer l'intervention de ces connaissances.

(2007) comme alternative au calque « système de questions-réponses » de l'anglais « question-answering system ». Le présent article utilise indifféremment les diverses appellations habituelles de ces systèmes.

[2] http://www.humanite.fr/popup_imprimer.html?id_article=714555. Visité le 11/3/2008.

2.2. Un domaine marqué par les campagnes d'évaluation

Après quelques travaux précurseurs (Lehnert 1978), les campagnes d'évaluation TREC QA lancées à partir de TREC8 (Voorhees & Harman 1999) ont suscité une montée en puissance des travaux sur les systèmes de questions-réponses. Ces campagnes ont cristallisé le contour de la tâche autour de questions dites « factuelles », qui portent principalement sur des noms de personnes (*Qui ?*), de lieux (*Où ?*) et sur d'autres « entités nommées » (voir la section 5). Ces campagnes visent à évaluer de façon objective les résultats des systèmes de questions-réponses qui y sont présentés par les équipes participantes. Leur principe consiste à fournir aux participants une collection de textes, sur laquelle ils peuvent préparer leurs systèmes, puis un ensemble de questions auxquelles les systèmes doivent chercher des réponses dans ces textes. Ces réponses sont jugées par des évaluateurs, et les résultats sont présentés à travers différentes mesures, dont la proportion de questions ayant obtenu une bonne réponse.

Les collections de textes de TREC sont principalement constituées d'articles ou de dépêches de presse (par exemple, la collection AQUAINT qui comprend des dépêches du *New-York Times*, des agences *Xin Hua* et *Associated Press*). Des tâches de recherche dans l'encyclopédie *Wikipedia* ont également été organisées. Les questions, dont le nombre a varié de 200 (TREC8, 1999) à 693 (TREC9, 2000), ont été soit créées de toutes pièces, soit reprises à partir de questions soumises à des moteurs de recherche (« logs » d'Excite, MSNSearch, Ask Jeeves, AOL...). Pour donner une idée de ces questions, nous en listons ci-dessous un échantillon pris au hasard parmi les 2393 questions cumulées des campagnes TREC8 à TREC12 :

What instrument is Ray Charles best known for playing?
Who wrote the book, "The Grinch Who Stole Christmas"?
Where did yoga originate?
Where is Procter & Gamble headquartered in the U.S.?
What is fibromyalgia?
What is the electrical output in Madrid, Spain?
What kind of dog was Toto in the Wizard of Oz?
What was the first spaceship on the moon?
What is the name of the heroine in "Gone with the Wind"?
What does "E Pluribus Unum" mean?
In which U.S. states have there been fatalities caused by snow avalanches?
When was the Red Cross founded?

2.3. Architecture typique d'un système de questions-réponses

La figure 1, reprise de (Grau & Chevallet 2007), décrit les étapes habituelles d'un système de questions-réponses, réparties en trois grandes phases : analyse de la question, recherche et traitement des documents, extraction de la réponse. Elle reflète une approche qui a recours à des méthodes linguistiques pour l'analyse de la question, délègue à un moteur de recherche d'information la recherche des documents susceptibles de contenir des réponses, puis effectue une analyse de cet ensemble circonscrit de documents pour en extraire des réponses précises. Nous examinerons également dans la section 3 des approches qui minimisent le recours aux connaissances linguistiques.

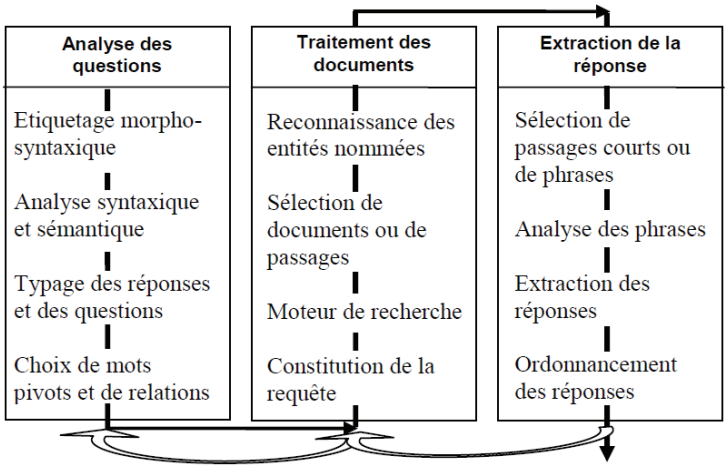

Figure 1. *Architecture classique d'un système de questions-réponses.*

L'*analyse de la question* vise à en extraire des indicateurs clés pour la recherche de réponses : type de réponse attendu, objet sur lequel porte la question, termes qui seront utilisés pour préparer une requête à un moteur de recherche d'information. Elle sera examinée dans la section 4.2. Dans la phase de *traitement des documents*, la motivation du recours à un moteur de recherche est que ce type de système est conçu pour accéder rapidement au contenu de grandes masses de textes, dont l'analyse linguistique était jusqu'à récemment considérée comme trop longue pour être envisageable[3]. Les documents renvoyés peuvent faire l'objet d'une réindexation par leurs « entités nommées » (section 5) en préalable à la phase d'*extraction de la réponse*. Celle-ci repose sur une analyse des phrases (section 4.3) qui fournit des informations plus précises sur lesquelles se fonde l'extraction des réponses proprement dite.

2.4. Connaissances linguistiques, méthodes numériques

Les systèmes de traitement automatique des langues ont besoin de connaissances sur la langue traitée pour fonctionner. Selon les tâches à réaliser et les méthodes employées, ces connaissances peuvent être plus ou moins précises, plus ou moins complètes, plus ou moins explicites. À un extrême, la modélisation linguistique, sous forme de connaissances explicites les plus précises et complètes possibles, a sous-tendu le paradigme dominant du traitement automatique des langues dans les années 1980 : lexiques et grammaires y ont tenu une part prépondérante, avec des représentations manipulant des « symboles », comme les catégories syntaxiques ou les règles grammaticales. On peut considérer que ces représentations visent à modéliser des connaissances sur le fonctionnement de la langue.

[3] Cette situation est en train de changer avec l'arrivée d'analyseurs robustes et plus rapides. Par exemple, le système Qristal (Laurent & *al*. 2006) a analysé l'ensemble de la collection EQueR, ce qui lui a permis de s'affranchir de l'utilisation d'un moteur de recherche et d'obtenir un accès plus pertinent aux phrases des documents de la collection.

À un autre extrême, les méthodes dirigées par les données se fondent sur l'observation des productions langagières pour construire des modèles numériques par apprentissage automatique. La part de connaissances linguistiques explicites y est généralement plus réduite, mais rarement totalement absente : une segmentation en mots est presque toujours effectuée[4], une lemmatisation et une catégorisation morphosyntaxique très souvent. Par ailleurs, les données qui servent à l'apprentissage sont la plupart du temps annotées (apprentissage supervisé : corpus étiquetés morphosyntaxiquement, banques d'arbres, etc.), ce qui constitue une autre façon d'introduire des connaissances linguistiques a priori.

Entre ces deux extrêmes, les approches hybrides sont nombreuses : lorsqu'une modélisation linguistique précise est disponible, elle peut fournir la base du traitement, quitte à la compléter ou à la moduler par des facteurs numériques, qui peuvent être obtenus par un apprentissage sur des données. Le rôle des connaissances linguistiques dans les systèmes de traitement automatique des langues, et en particulier dans les systèmes de recherche de réponses précises, peut ainsi prendre des formes variées. Dans cet article, les méthodes utilisant peu de connaissances linguistiques sont présentées en section 3, les autres dans les sections 4-6.

3. Recherche de réponses sans connaissances linguistiques

Cette section examine les approches qui utilisent peu ou pas de connaissances linguistiques. La section 3.1 montre ce que peuvent faire des approches issues de la recherche d'information. La section 3.2 présente une méthode qui s'appuie sur la forte redondance de la description d'informations sur le web pour simplifier l'appariement question-réponse. Enfin la section 3.3 recense les lieux où, dans un système de questions-réponses, des méthodes d'apprentissage automatique peuvent être mises en œuvre.

3.1. Approches inspirées des modèles numériques de la recherche d'informations

Comme indiqué dans l'introduction, une différence essentielle entre la recherche d'informations (RI) et la recherche de réponses précises réside dans la nature de la requête. Dans un cas, il s'agit d'une recherche de similarité (la requête est l'expression la plus probable du besoin en information de l'utilisateur, *i.e.* elle est une forme de « réponse ») ; dans l'autre cas, il s'agit d'une question qui appelle une réponse précise (la question ne contient justement pas les mots recherchés : la réponse). Malgré cette différence, les approches de la RI (Bellot & Boughanem 2008) peuvent s'avérer efficaces pour la tâche de questions-réponses (Lin 2007) car les mots employés dans les questions ont tendance à se trouver à proximité de la réponse au sein des documents. Par exemple, la réponse à la question « *combien d'habitants a le Lesotho ?* » se trouve très certainement dans un document où *Lesotho* apparaît (cela n'est pas obligatoire, on pourrait aussi trouver « *au centre de l'Afrique du Sud se trouve un pays peuplé de 2,3 millions d'habitants* »). D'autre part, la présence du mot *habitants* et de *Lesotho* dans un même document - *a fortiori* dans un même paragraphe ou une même phrase - augmente la probabilité qu'il contienne l'information recherchée.

Les approches numériques de la recherche d'informations consistent à ordonner les documents contenant des mots communs avec la requête en fonction de leurs poids et de leurs distributions dans le corpus cible. Les employer revient à considérer prioritairement

[4] Ce n'est pas toujours le cas pour les langues sans séparateurs de mots, comme le chinois ou le japonais.

les documents qui évoquent « le plus » le Lesotho et ses habitants pour la recherche approfondie de la réponse. Bien sûr, ceci ne suffit pas pour répondre à la question et d'autres traitements sont nécessaires. Dans l'exemple précédent, il s'agit notamment d'identifier que l'information recherchée est numérique et qu'un nombre est, au minimum, présent dans le document trouvé. Parmi les multiples nombres pouvant se trouver dans les documents retenus, ceux qui se trouvent le plus à proximité des mots *Lesotho* et *habitants* seront ceux qui seront considérés comme réponses candidates prioritaires. Pour ordonner les réponses candidates et les proposer à l'utilisateur, il est ainsi possible de se baser sur plusieurs critères tels que la proximité (par rapport aux mots de la question et par rapport aux autres réponses candidates et concurrentes) ou la redondance (souvent constatée sur de grands corpus, elle permet de favoriser une réponse candidate citée plusieurs fois).

Si pour la partie sélection de documents, les approches « sac de mots » paraissent suffisantes - il est encore théoriquement possible de répondre à plus de 95% des questions après filtrage des corpus à 1000 documents par question -, la question se pose plus nettement dès que l'on arrive à la sélection de phrases. Sur les corpus en français de la campagne Technolangue-EQueR (Ayache & *al*. 2005), il n'est plus possible de répondre à environ 50% des questions après avoir utilisé des méthodes d'extraction de passages de type « sac de mots » pures : 3 groupes de 3 phrases contiguës par question et extraction suivant une similarité de type *cosinus* ou selon une fonction de densité des mots de la question dans les passages (Gillard & *al*. 2006a). Des résultats similaires ont été observés à partir de différentes méthodes de segmentation non supervisées sur les corpus en anglais de TREC (Tellex & *al*. 2003). En ce qui concerne la sélection finale de la réponse, des techniques basées sur la compacité des mots de la question autour de la réponse candidate permettent de répondre correctement à près de 40% des questions factuelles EQueR sans ajout de connaissances linguistiques spécifiques ni de lexiques étendus (Gillard & *al*. 2006b).

L'étude récente sur les données des campagnes TREC 2004 et 2005 (Lin 2007) montre que le moteur de recherche documentaire Lucene donne en moyenne des résultats compétitifs par rapport aux moteurs de question-réponse dédiés (il est proche des meilleurs sur les questions de type « *other* » - destinées à recueillir des compléments d'informations et des définitions -, mais nettement moins bon sur les factuelles). La conclusion de ces analyses est que les techniques de traitement automatique des langues ont prouvé leur apport pour la tâche question-réponse factuelle mais ne sont pas encore vraiment à maturité pour les autres types de questions. Dans ces conditions, les approches traditionnelles de la RI s'en sortent donc plutôt bien (Kelly & Lin 2007).

3.2. Réécriture de la question et redondance du web

Un exemple typique de tâche illustrant la mise en œuvre de stratégies simples et donnant de bons résultats est la recherche de réponses sur le web, stratégies qui ne peuvent s'appliquer en l'état sur des corpus de taille plus limitée.

L'un des premiers systèmes de QR conçus pour interroger le web en utilisant un moteur existant est MULDER (Kwok & *al*. 2001). Avec la création de la piste QA à TREC, beaucoup de systèmes ont utilisé le web comme source de réponse dès les premières évaluations. Ces systèmes effectuent une réécriture spécifique des questions afin de ramener les documents pertinents dans les premiers, ou afin de n'extraire que les extraits donnés par le moteur (les *snippets*). Tous exploitent l'idée que la grande redondance des informations présentes sur le web permet de trouver des documents pertinents même avec une requête très spécifique et qu'une telle requête permet d'obtenir les documents susceptibles de

contenir la réponse dans les premières positions. Alors que MULDER se fonde sur une réécriture syntaxique des questions pour formuler des requêtes, Brill & *al*. (2001) exploitent la taille du web dans le système AskMSR et émettent l'hypothèse qu'*il doit exister un document contenant la réponse donnée sous une forme identique à la question* : mêmes termes et même syntaxe, hors la forme interrogative, et qu'il suffit d'une réécriture basique des questions pour trouver des réponses. Nous allons développer ce travail afin de montrer les performances d'une telle approche, et ses limites.

AskMSR garde les mots de la question dans leur ordre original et déplace le verbe dans toutes les positions possibles, en supposant que l'une des formes sera correcte, et que celles qui sont incorrectes ne ramèneront pas de documents, puisque le moteur effectue une comparaison entre chaînes de caractères lors de la recherche. Avec les réécritures, le système précise la place potentielle de la réponse, soit à gauche de la forme, à sa droite ou indifférente.

Par exemple, à la question "*Who created the character of Scrooge?*" correspondent les réécritures :

"created +the character +of Scrooge" LEFT, 5
"+the character +of Scrooge +was created +by", RIGHT, 5
"created" AND "+the character" AND "+of Scrooge" NULL, 2
"created" AND "character" AND "Scrooge", NULL, 1

Les auteurs exploitent aussi la redondance du web afin d'extraire les réponses sans avoir besoin d'utiliser des connaissances sur la langue, qu'elles soient syntaxiques ou sémantiques.

La technique d'extraction consiste à collecter les unigrammes, bigrammes et trigrammes[5] situés à droite ou à gauche de la requête, selon la réécriture opérée. Le système retient ensuite la meilleure chaîne de caractères en tenant compte de sa fréquence dans les extraits (*snippets*) renvoyés par le moteur de recherche et de son poids selon les patrons de réécriture, après application de quelques filtres, largement fondés sur des critères de surface, associés aux types de questions. Ces filtres permettent de typer les chaînes de caractères, et de ne sélectionner que celles qui peuvent être une réponse à la question : existence de majuscules pour les noms de personne, de nombres pour les entités numériques par exemple.

Les réponses ainsi collectées, pondérées et ramenées à la taille la plus plausible pour la requête concernant le créateur de Scrooge sont les suivantes, avec leur nombre d'occurrences :

Charles Dickens 117	A Christmas Carol 78	Walt Disney's uncle 72
Carl Banks 54	uncle 31	

Afin de quantifier cette stratégie, Dumais & *al.* (2002) ont appliqué l'approche mise au point pour le web sur la collection de référence de TREC, ou du moins se sont ramenés à une stratégie analogue. Le système, appliqué au web avec les questions de TREC9, trouve 61 % des réponses dans les 5 premiers rangs. Appliqué à la collection AQUAINT (voir la section 2.2), le système, après quelques adaptations dans l'extraction de courts passages, trouve 24 % de réponses correctes, contre 56 % pour le système précédent appliqué au web dans les mêmes conditions. On voit bien ici les limites d'une telle stratégie : en l'absence d'une abondante formulation de réponses, il devient indispensable de mettre en œuvre des processus de résolution modélisant plus finement la langue.

[5] Un *n*-gramme de mots est une séquence constituée de *n* mots contigus.

3.3. Approches par apprentissage automatique

De nombreux composants des systèmes de question-réponse peuvent faire appel à des méthodes d'apprentissage automatique afin de :

- diminuer le temps nécessaire à l'adaptation des systèmes à différentes langues ou domaines de spécialité (avantage à pondérer toutefois par le besoin de collecter des exemples en nombre suffisant) ;
- permettre la prise en compte simultanée d'un grand nombre de critères (morphologiques, syntaxiques, lexicaux, sémantiques…, mais aussi positionnels ou contextuels) ;
- associer à chaque prise de décision des fonctions de score utiles pour les traitements effectués par les autres composants du système ou pour indiquer un taux de confiance à l'utilisateur.

De manière spécifique à la tâche question-réponse, ce sont les modules d'analyse de la question et de sélection finale de la réponse qui utilisent le plus souvent des approches d'apprentissage automatique supervisé ou non.

L'analyse de la question comprend la capacité à déterminer la forme de la réponse : oui/non pour les questions booléennes, une liste pour les questions à réponses multiples (*Quels sont les trois pays les plus peuplés d'Europe ?*), une définition (*Que signifie… ?*), etc. Lorsque la réponse est factuelle, il s'agit de déterminer le type de l'entité nommée attendue. La phase d'apprentissage correspond à l'analyse de couples (question annotée / type de réponse) et à la détermination de marqueurs morphologiques, syntaxiques ou lexicaux (unigrammes, bigrammes ou trigrammes, entités nommées ou syntagmes) qui conduisent au typage correct. À cet effet, les systèmes emploient des approches de classification automatique bien connues dont les arbres de décision ou les machines à vecteurs supports (Hacioglu & Ward 2003). Sur les données des campagnes d'évaluation TREC ou CLEF, les taux de succès sont très élevés (de l'ordre de 90 %) et permettent de typer finement les questions en plusieurs dizaines de classes hiérarchisées (Sekine 2004).

Une autre manière d'utiliser l'apprentissage automatique consiste à apprendre des fonctions d'ordonnancement. Usunier (2006) propose et expérimente ainsi une méthode d'apprentissage par *boosting* pour ordonner les passages des textes contenant les réponses candidates qui est plus souple que les méthodes « sac de mots » classiques (Tellex *& al*. 2003) tout en obtenant des résultats qualitativement comparables.

À l'autre bout de la chaîne de traitement, l'apprentissage automatique peut être utilisé afin de sélectionner une ou plusieurs réponses parmi les candidates. Dans ce cas ce sont des couples (question, réponse(s)) qui servent de base d'apprentissage (Miliaraki & Androutsopoulos 2004) et, comme pour l'analyse de la question, des critères de différentes natures sont pris en compte : caractéristiques syntaxiques (Ittycheriah *& al*. 2001), contexte d'apparition de la réponse candidate (Ravichandran *& al*. 2003), bases de connaissance externes, etc. À l'inverse, un modèle probabiliste basé sur des modèles de langage a été proposé par Whittaker *& al*. (2006) où, à partir des seuls documents trouvés sur le web dans lesquels apparaissent les mots des questions, les auteurs ont constitué un système question-réponse multilingue (espagnol, français, japonais, anglais, russe, chinois et suédois) sans aucun traitement linguistique : ni analyse morphosyntaxique, ni même reconnaissance d'entités nommées. Les auteurs obtiennent des résultats plus qu'honorables sur les questions factuelles des campagnes TREC ou CLEF : 25,1 % de bonnes réponses, significativement au-dessus des performances du système médian mais en deçà du meilleur qui obtient un

score de 57,8 %. Ce système constitue une bonne indication des performances qu'il est possible d'obtenir actuellement sans aucun traitement linguistique.

Pour plus d'informations concernant l'apprentissage automatique dans les systèmes de question-réponse, le lecteur pourra se référer à Usunier & *al*. (2008). Au-delà des résultats des systèmes basés sur des approches numériques, nous devons nous interroger sur leur potentiel à intégrer de la connaissance linguistique afin d'exploiter au mieux, par exemple, la résolution des références (anaphores notamment, voir la section 6.1) ou l'identification de la portée des marqueurs de négation. Il s'agit enfin de déterminer quels sont les phénomènes linguistiques que chacun des modèles numériques est capable de capter le plus efficacement.

4. L'analyse de phrases dans la recherche de réponses précises

Dans cette section, nous commençons par passer en revue les connaissances syntaxiques utilisées par les systèmes de questions-réponses (section 4.1). Nous examinons ensuite plus particulièrement leur rôle dans l'analyse des questions (section 4.2) et dans l'extraction des réponses (section 4.3). Nous abordons ensuite rapidement le type d'analyse sémantique effectué par les systèmes de questions-réponses (section 4.4).

4.1. Connaissances syntaxiques pour les systèmes de questions-réponses

L'utilisation de la syntaxe commence avec le recours à un étiqueteur morphosyntaxique afin de déterminer la catégorie morphosyntaxique des mots de la question, pour choisir les mots « importants » et procéder à l'interrogation *via* un moteur de recherche. La reconnaissance des lemmes est aussi une étape déterminante pour le français (le *stemming*, c'est-à-dire l'élision des dernières lettres d'un mot, étant souvent considéré comme satisfaisant pour l'anglais). Le fait de pouvoir reconnaître des unités plus importantes telles que les syntagmes, et connaître leurs liens de dépendance, est l'étape suivante, à la fois pour affiner l'analyse de la question et pour améliorer la recherche de la réponse dans les documents. C'est l'application plus ou moins importante de ces différentes composantes que nous allons détailler dans cette section. Pour plus de détails, le lecteur pourra consulter Poibeau & Vilnat (2007).

4.1.1. Étiquetage morphosyntaxique pour les systèmes de questions-réponses

Un étiqueteur morphosyntaxique vise à attribuer à chaque occurrence de mot dans un énoncé la catégorie morphosyntaxique qu'il possède dans le contexte de cet énoncé. De nombreux étiqueteurs morphosyntaxiques existent et fournissent des résultats assez satisfaisants pour l'étiquetage de gros corpus de textes, que ce soit l'étiqueteur de Brill (1995) ou encore le *TreeTagger* (Schmid 2004).

Cependant, les règles appliquées par ces étiqueteurs, qu'elles soient apprises par entraînement sur un corpus ou fournies manuellement, ont été mises au point sur des phrases majoritairement affirmatives. De ce fait les résultats obtenus sur des questions sont nettement moins bons. Bien évidemment les mots connus et non ambigus posent peu de problèmes, mais les règles appliquées, que ce soit pour lever les ambiguïtés ou pour fournir des éléments partiels en cas de mots n'appartenant pas au lexique, sont souvent mises en échec dans ce cadre.

Concernant les documents dans lesquels les réponses sont recherchées, l'étiquetage pose moins de problèmes, surtout quand ces documents sont des textes « propres » comme les

collections d'articles de journaux ou les dépêches (campagnes TREC ou CLEF). Élargir ces collections à des ressources telles que l'encyclopédie *Wikipedia* ne modifie pas les résultats obtenus, du moins au stade de l'étiquetage morphosyntaxique. En revanche quand la recherche concerne l'ensemble des documents qui peuvent être retrouvés sur la toile, la quantité de scories rencontrées dans ces documents peut fortement perturber les résultats des étiqueteurs. Les caractères indus tels que ceux qui proviennent de mauvaises conversions d'encodage provoquent de nombreuses erreurs : l'étiqueteur ne reconnaît plus les limites des mots ou des phrases. Le grand nombre de fautes de frappe ou d'erreurs orthographiques rendent également la tâche d'étiquetage difficile sur ces textes.

4.1.2. Analyse syntaxique

Le recours à une analyse syntaxique (constituants, dépendance, relations grammaticales) est de plus en plus fréquent pour trouver des informations précises. En effet, l'élément précis attendu en réponse rend nécessaire le recours à une analyse syntaxique aussi complète que possible pour l'isoler dans un passage, et la recherche d'une réponse valide nécessite également d'obtenir des informations précises sur la façon dont la réponse est exprimée. On peut considérer que l'on recherche un passage qui constitue une paraphrase de la question, sous forme déclarative et avec l'élément réponse en plus. La section 3.1 a présenté l'emploi de méthodes numériques pour approximer cette identification. A l'opposé, de nombreux travaux s'appuient sur la comparaison de structures syntaxiques pour ce faire. Soit cette comparaison porte sur la structure complète des phrases par appariement d'arbres syntaxiques (Tanev & *al.* 2005) ou par la recherche du taux de recouvrement entre les relations de dépendance de la question et du passage réponse (Bouma & *al.* 2005). Soit les systèmes vérifient des appariements de sous-phrases : appariement de syntagmes nominaux par repérage des termes complexes de la question sous forme de variantes, vérification de la présence de certaines relations, entre la réponse par exemple et les éléments correspondants de la question. Ainsi, lors de l'analyse de la question : « *Qui a tué John Kennedy ?* », on peut savoir que l'on cherche quelqu'un qui est sujet de *tuer* quand *Kennedy* en est l'objet. On peut ainsi reconnaître que dans la phrase « *Jack Ruby a tué l'assassin de Kennedy* », *Jack Ruby* n'est pas la réponse recherchée : si *Jack Ruby* est bien sujet de *tuer*, l'objet est *l'assassin de Kennedy* et non pas *Kennedy* lui-même. En revanche, en utilisant des connaissances sur des paraphrases, on pourra retenir *Oswald, l'assassin de Kennedy* comme une bonne réponse, à condition de savoir passer de *X tuer Y* à *X, assassin de Y*. La plupart des analyseurs syntaxiques sont en mesure de reconnaître des relations telles que la relation *sujet-de* avec un grand taux de succès dans des textes de type journalistique. C'est ce que montrent les évaluations de la campagne EASy (Vilnat & *al*. 2004). Toutefois il reste encore de nombreux cas où soit la relation cherchée est plus complexe à trouver, soit l'analyse de la phrase pose des problèmes et ne permet pas à un analyseur d'en construire une représentation complète. Mais on peut souvent s'appuyer sur une analyse partielle pour reconnaître des patrons de réponses, comme dans l'exemple précédent où il est inutile d'analyser toute la phrase pour trouver juste une forme telle que *X, assassin de Y*. L'analyse syntaxique pour l'appariement entre question et réponse est développée en section 4.3.

Nombre de systèmes de questions-réponses visent ainsi une analyse syntaxique la plus complète possible à la fois des questions et des documents pour garantir une réponse adéquate. Certains sont même en mesure de résoudre des anaphores au sein des documents-réponses (voir la section 6.1) afin de valider comme réponse correcte une forme telle que *Oswald l'assassina en 1964*, lorsque l'on sait que ce document relate la vie de Kennedy, et donc que le pronom *l'* y fait référence.

4.2. Analyse syntaxique de la question

Dans le cadre d'un système de question-réponse, la demande d'accès à l'information se traduit par une question comme « *Quelle est la capitale du Chili ?* » ou une requête en langage naturel, par exemple « *Citez les religions les plus pratiquées en Hongrie* », qui peut être assimilée à une question. La nature de l'information recherchée est explicitée, de même que ses relations avec les termes de la question, ce qui n'est pas le cas dans le cadre d'une requête au sens de la recherche d'information : dans un moteur de recherche classique, la requête est constituée de mots-clés, comme « *capitale Chili* », l'information recherchée n'étant pas précisée (cherche-t-on à connaître le nom de cette ville ? sa population ? sa situation géographique ?). Formuler la demande sous forme de question fournit des indices sur l'information attendue, et il est primordial de pouvoir analyser la question non plus en termes de mots-clés uniquement, mais comme une phrase complète.

L'analyse des questions dans un système de question-réponse a pour objectif d'identifier ces caractéristiques de la question, afin de les utiliser ensuite dans le processus de recherche de réponse, en permettant de sélectionner des documents ou des passages de documents pertinents, et d'extraire des réponses possibles. Cette étape devant guider l'ensemble du processus, différents types d'information sont retenus. Tout d'abord, des mots-clés de la question sont sélectionnés, afin de construire une requête (contrairement à ce qui se passe dans les moteurs de recherche classiques, où l'utilisateur doit lui-même effectuer ce travail de sélection). Puis, pour pouvoir extraire la réponse des documents, des informations sur la réponse précise recherchée sont nécessaires.

Burger & *al*. (2001) ont identifié la classification et la compréhension des questions comme les deux premières problématiques de recherche en question-réponse. Il est possible de dégager des tendances générales pour l'analyse des questions. Ainsi, la sélection de mots-clés de la question s'appuie principalement sur un niveau d'étiquetage morphosyntaxique des mots de la question : certaines catégories comme les noms propres seront privilégiées dans la recherche des documents. Les mots de la question jouant des rôles sémantiques particuliers peuvent également être privilégiés, comme l'objet sur lequel porte la question (*focus*). Pour l'extraction de réponses à partir des documents, deux informations principales sont utilisées : le type de réponse attendu, et les relations attendues sur la réponse. Le type de réponse peut être rattaché à une hiérarchie d'entités nommées : par exemple, pour la question « *Dans quelle région se situe Angers ?* », la réponse doit être un lieu, ou une région, selon la granularité de la hiérarchie d'entités nommées utilisée. Ce type de catégorisation est commun à la plupart des systèmes de question-réponse, et utilise généralement des listes d'hyponymes des catégories d'entités nommées. Les relations attendues sur la réponse peuvent être représentées soit sous forme de catégories de questions, comme la typologie de questions définie par Hovy & *al.* (2002) fondée sur des classes d'équivalence de types de réponses, soit sous forme plus formelle, par exemple sous la forme logique adoptée par Pasca (2003). Cette modélisation est déduite des relations syntaxiques et d'autres informations comme l'étiquetage en entités nommées.

Pour donner un exemple concret d'analyse de la question et des traitements mis en jeu, considérons la question suivante : « *Dans quel pays ont eu lieu les Jeux Olympiques en 1992 ?* », et son traitement par l'analyse de la question du système FRASQUES, décrite dans Ligozat & *al.* (2006).

Tout d'abord, la question est annotée par un étiqueteur morphosyntaxique, puis la segmentation en constituants syntaxiques est calculée par un analyseur syntaxique. Les caractéristiques suivantes sont ensuite déduites de l'analyse syntaxique :

- le type de réponse attendu est « *pays* » (mot qui suit l'interrogatif), qui correspond à une catégorie d'entité nommée « *pays* », sous-catégorie de « *lieu* » ;
- l'objet de la question est « *Jeux Olympiques* » (groupe nominal sujet de la question). Ce terme sera donc privilégié au cours du processus de recherche et d'extraction de la réponse ;
- le contexte temporel est « *1992* » (selon l'étiquetage en entités nommées de la question) ;
- les mots-clés de la question sont « *pays* », « *Jeux Olympiques* » (terme composé reconnu par des patrons morphosyntaxiques), et « *1992* ».

Toutes ces informations seront utilisées dans la suite du processus de recherche de la réponse dans la collection de documents.

4.3. Analyse syntaxique pour l'extraction de réponses

En dehors de l'analyse de la question, l'analyse syntaxique de phrase est aujourd'hui utilisée par les systèmes de questions-réponses au moment de l'extraction de la réponse précise. Ce module d'extraction prend en entrée un ensemble de phrases candidates et produit en sortie pour chacune de ces phrases la réponse la plus précise possible. Ces réponses précises sont ensuite ordonnées selon leur fiabilité et celle d'entre elles qui obtient le meilleur score est finalement retournée.

Pour réaliser cette extraction, cette étape s'appuie sur les caractéristiques obtenues lors de l'analyse de la question (voir ci-dessus, section 4.2). Cependant, sans analyse syntaxique, il lui est difficile d'extraire de façon certaine la bonne réponse. En particulier, les relations de dépendance entre les syntagmes sont précieuses car elles permettent au module d'extraction de chercher à apparier les relations observées dans la question avec les relations qu'il extrait d'une phrase candidate. Suivant la fiabilité des relations obtenues le mécanisme peut être modulé et ne tenter d'apparier que les relations dont il est sûr (les relations sujet, objet ou modifieur de nom par exemple). Cette approche est proposée par Ligozat dans sa thèse (2006).

Prenons par exemple la question « *Qui est le maire de Bastia ?* » pour laquelle une des phrases candidates obtenue par le module d'interrogation pourrait être « *Il est le père d'Emile Zuccarelli, ex-ministre des postes et des télécommunications dans le gouvernement Pierre Bérégovoy (1992-1993), aujourd'hui député de Haute-Corse, maire de Bastia et président délégué du Parti radical-socialiste* ». L'analyse de la question nous indique que l'entité cherchée est de type personne et que cette entité doit avoir pour caractéristique « *maire de Bastia* ». La réponse précise à extraire pourra être apposée à cette caractéristique (comme dans la phrase proposée) ou encore être liée à elle par une relation plus explicite de type « *X est le maire de Bastia* », ou « *X est élu maire de Bastia* »… Sans une analyse fine des relations de la phrase, le système de questions-réponses a tout autant de chances de répondre « *Pierre Bérégovoy* » que la bonne réponse « *Emile Zuccarelli* ». Seule une analyse détectant les appositions et les rattachant correctement nous indique que l'apposition « *maire de Bastia* » se rapporte bien à « *Emile Zuccarelli* ». Sans cette information le système ne peut faire qu'un choix plus approximatif et donc risqué.

Les systèmes de questions-réponses qui obtiennent les meilleurs résultats lors des campagnes d'évaluation utilisent des analyseurs syntaxiques et aussi parfois d'autres modules de TAL. Ainsi, pour leur système QRISTAL, Laurent & *al*. (2006) ont recours non seulement à l'analyse syntaxique, mais ils mettent également en œuvre différents

traitements tels que la désambiguïsation sémantique, la recherche des référents des anaphores (voir la section 6.1), ou encore une analyse conceptuelle (voir la section 4.4).

Lorsqu'ils ne disposent pas de tels analyseurs, ou lorsque ceux dont ils disposent ne sont pas assez fiables, les systèmes de questions-réponses ont recours à des connaissances syntaxiques plus limitées que l'on peut qualifier de locales. Ces connaissances sont exprimées sous la forme de patrons d'extraction appliqués à la phrase candidate ; ils expriment la proximité possible de la réponse avec des syntagmes contenant les éléments importants de la question tels que le focus ou le type général. Cette méthode est moins fiable qu'une analyse syntaxique complète mais présente néanmoins des avantages : elle est rapide à développer et à maintenir, elle peut être adaptée sans difficulté à d'autres langues que le français, dans le cadre de systèmes multilingues par exemple.

4.4. Analyse sémantique pour les systèmes de questions-réponses

Au-delà de l'analyse syntaxique, les systèmes les plus avancés effectuent une analyse sémantique pour construire une représentation des questions et des textes. Cette représentation est souvent de nature logique, comme dans Ahn & *al.* (2004), Durme & *al.* (2003), ou fondée sur des schémas prédicatifs (Narayanan & Harabagiu 2004).

Le système QED de Ahn & *al.* (2004) analyse la question et les phrases des textes à l'aide d'une grammaire catégorielle combinatoire puis produit une représentation « DRS » fondée sur la DRT (Discourse Representation Theory (Kamp & Reyle 1993)). Cette représentation contient une description des événements mentionnés, sous forme de prédicats, arguments et relations, et traite entre autres les alternances actif-passif. L'extraction de réponse apparie la question et les phrases réponses sur leur représentation DRS.

De même, Narayanan & Harabagiu (2004) produisent des structures prédicat-argument et des cadres sémantiques en utilisant les ressources de PropBank (Kingsbury & *al.* 2002) et FrameNet (Baker & *al*. 1998). Ces structures sont construites aussi bien pour la question que pour les phrases des documents. La recherche de réponses se fait par inférence probabiliste sur ces représentations.

Moldovan & *al.* (2003) observent que les résultats des systèmes de questions-réponses stagnent et ne pourront progresser qu'avec l'aide de mécanismes de compréhension de la langue plus poussés. Dans leur approche, les sorties de l'analyse syntaxique des questions et des phrases candidates sont transformées en représentations logiques, et les connaissances sur le monde représentées dans les gloses de *WordNet* sont également représentées sous forme logique. Un démonstrateur logique, COGEX, est alors utilisé pour extraire la réponse puis vérifier à l'aide d'inférences que celle-ci est correcte.

Des traitements sémantiques ont également lieu à des paliers plus larges que la phrase ; ils sont décrits dans la section 6.

5. Analyse sous-phrastique : le cas des « entités nommées »

La reconnaissance des « entités nommées » est la tâche par laquelle des entités spécifiques sont détectées. Ces entités recouvrent des expressions désignant par exemple des personnes, des lieux ou des organisations. Dans le cadre des systèmes de questions-réponses, les entités nommées sont utilisées pour sélectionner les documents et, à l'intérieur de ceux-ci, les passages contenant potentiellement une réponse et pour sélectionner les réponses possibles, quand le type de celle-ci a été identifié par l'analyse de la question (section 4.2). Toutes les entités correspondant au type de réponse attendu sont considérées comme des réponses

potentielles. Nous commençons par présenter plus précisément ce que recouvre la notion d'entités nommées (section 5.1) et les différentes méthodes utilisées pour les détecter (section 5.2) et nous finissons avec la présentation de leur utilisation dans les systèmes de questions-réponses (section 5.3).

5.1. Les entités nommées

Les conférences MUC (DARPA 1998) sur l'extraction d'information ont mis en avant des tâches génériques dont la reconnaissance d'« entités nommées ». La définition la plus utilisée est héritée de celles de MUC (Grishman 1995) qui comprennent trois catégories : les expressions de noms propres (personne, lieu, organisation), les expressions temporelles (date, heure) et les expressions numériques (valeurs monétaires et pourcentages).

Ces définitions sont parfois étendues à l'intérieur d'une catégorie, comme les organisations, afin d'en affiner le contour référentiel, par exemple les définitions adoptées dans le cadre de la campagne d'évaluation ACE (NIST 2000a). Afin de couvrir de nouveaux besoins ou de nouvelles tâches, ces définitions et hiérarchies ont été étendues. Sekine (2004) propose ainsi jusqu'à 200 éléments, incluant d'autres catégories comme les prix. En effet, pour répondre à une question du type « *Quel prix Nobel a gagné Albert Fert cette année ?* », il faut savoir reconnaître un prix Nobel et plus encore une catégorie de prix Nobel. La phrase suivante comporte des exemples d'entités nommées traditionnelles et étendues, marquées par des balises XML indiquant leur catégorie :

<pers>Albert Fert</pers> a obtenu le <prix>prix Nobel</prix> de <type_prix>physique</type_prix> en <date>2007</date>.

Certains systèmes distinguent en outre les prénoms des noms pour les réunir ensuite dans une entité de niveau supérieur (*pers* par exemple). On peut aussi trouver *Nobel* représenté comme un spécifieur de prix. Des expressions comme les noms de couleur, de forme ou encore de langue peuvent aussi être considérées comme des entités nommées (Turmo & *al.* 2007).

5.2. Détection des entités nommées

Les méthodes de détection des entités nommées peuvent être classées en deux grandes familles : les approches statistiques et les approches symboliques.

L'analyse linguistique, *stricto sensu*, est nécessaire pour chacune de ces familles d'approches mais à des moments différents. Les approches statistiques, comme celles décrites dans Bikel & *al.* (1997) à base de Modèles de Markov Cachés, ou dans Isozaki & Kazawa (2002) qui utilisent un classifieur à vecteurs support, nécessitent de grands corpus préalablement annotés en entités nommées. Annoter de grands corpus est une opération qui nécessite la rédaction d'un guide d'annotation précis (par exemple, le guide d'annotation de ACE (NIST, 2000b)) indiquant aux annotateurs non seulement les types d'entités, mais également les règles permettant de désambiguïser les entités. Le plus souvent, ces guides donnent de nombreux exemples plutôt que des règles trop précises (Sekine 2004).

Les systèmes de détection d'entités nommées qui reposent sur des approches symboliques utilisent, entres autres, des listes (listes de prénoms, de noms, d'organisations, de villes, etc.) et des grammaires. Ces grammaires partent soit du texte brut, soit du résultat d'analyses préalables, comme l'étiquetage morphosyntaxique ou l'analyse en dépendance, et prennent des formes différentes. Elles sont généralement constituées par des automates ou des expressions régulières. Les analyses préalablement faites, lorsqu'elles existent,

consistent le plus souvent en une segmentation en mots, une segmentation du texte en phrases, puis une analyse en parties du discours. Le système ANNIE (Cunningham & *al.* 2007) utilise un ensemble de règles sous forme de transducteurs à états finis pour les étapes de segmentation. L'annotation en parties du discours est faite en utilisant un étiqueteur proche de celui proposé par Brill (1995), qui se fonde sur des listes de mots et un ensemble de règles apprises sur des corpus manuellement annotés. L'analyseur utilisé par Rosset & *al.* (2006) est constitué d'expressions régulières basées sur des mots. Il utilise également des listes et des classes de mots (*mot commençant par une majuscule*, *mot ne contenant que des majuscules*, *chiffre*). Il est utilisé après une étape d'accès au lexique, et dispose ainsi de toutes les catégories morphosyntaxiques non désambiguïsées pour chacun des mots qui peuvent servir d'indices pour l'analyse. L'exemple suivant illustre le texte tel que ce système le reçoit :

Albert|NP Fert|NP a|V|Ax|N reçu|V|K|N|A le|PRO|DET prix|N Nobel|NP...

où NP = nom propre, V = verbe, Ax = auxilliaire, N = substantif, K = participe passé, A = adjectif, PRO = pronom, DET = déterminant.

À partir des phrases annotées (mais non analysées) de cette façon, le système applique des règles qui s'appuient sur des listes et des définitions de contextes d'application. Un exemple de règle d'analyse est :

loc : ((<= &prep_loc) %caps | %pays | %ville) ;

où *loc* est le nom de la règle (le type d'entité nommée), *&prep_loc* fait référence à une définition de contexte qui décrit des prépositions ou groupes prépositionnels précédant un lieu, *%caps* la classe des mots commençant par une majuscule et *%pays* et *%ville* les classes des mots appartenant à un dictionnaire de pays et de villes (des ressources propres). Cette règle dit que *est un lieu, tout mot en majuscule précédé de prépositions locatives, ou tout mot appartenant à la liste des pays et des villes connues*.

Dans ce système, le linguiste/développeur peut définir ses propres listes, par exemple celles des verbes de mouvement, des verbes de perception, ou encore des prépositions spatiales.

5.3. Utilisation dans des systèmes de questions-réponses

Les entités nommées peuvent être utilisées à deux moments dans un système de questions-réponses : (1) lors de la sélection des documents et passages pertinents, et (2) lors de l'extraction des candidats réponses.

Certains systèmes (par exemple Laurent & *al.* 2006) analysent la collection de documents et procèdent ensuite à l'indexation des documents. Cette indexation peut alors se faire notamment sur les entités nommées. Ainsi, lorsqu'il faut rechercher les documents et passages pouvant contenir la réponse, ceux-ci sont sélectionnés si et seulement si ils contiennent une entité du type recherché. Par exemple, à la question « *Quand Thomas Mann a-t-il reçu le prix Nobel ?* », le système de Rosset & *al.* (2006) sélectionne les documents et passages contenant les entités nommées *pers(Thomas Mann) prix (prix Nobel)* et *date()*, avec la valeur de la date non précisée.

Tous les éléments du passage observé correspondant au type attendu sont perçus comme des réponses possibles. Si l'ontologie des entités nommées est fine, il est possible d'attribuer aux candidats des poids (degrés de confiance) variables en fonction de cette typologie. Par exemple, le système décrit dans Rosset & *al*. (2006) peut attribuer des poids différents selon le niveau du type d'entité. Ainsi pour la question « *Quelle église a ordonné*

des femmes prêtres en 1994 ? », le type de la réponse possible pourra être une *organisation religieuse* ou, mais avec un degré de confiance moindre, une *organisation* non spécifiée.

6. Prise en compte de la dimension textuelle des documents

Les traitements décrits jusqu'à présent se situent au niveau local, avec pour limites maximales les frontières de la phrase, car il s'agit du domaine le mieux maîtrisé. Néanmoins la tâche de questions-réponses gagne à étendre son analyse pour explorer le niveau du discours.

Cela est bien entendu vrai en ce qui concerne la collection de documents. Une sélection efficace du passage candidat et de la réponse peut difficilement faire l'économie, sinon d'une « compréhension » plus globale du texte, du moins de la mise en évidence d'un certain nombre de liens entre ces passages et le reste du texte. Ces liens entre texte et cotexte doivent également être considérés dès lors que l'on souhaite mettre en place une interaction (à défaut d'un véritable dialogue) avec l'utilisateur du système. Il faut alors analyser une question nouvelle en fonction de celles qui ont été posées auparavant. Récemment, les campagnes d'évaluation des systèmes de questions-réponses ont proposé des questions « groupées », c'est-à-dire un ensemble de questions sur le même sujet qui simulaient une interaction avec l'utilisateur et illustraient ainsi certains des phénomènes décrits ci-dessus (Giampiccolo & *al.* 2007 ; Dang & *al*. 2006).

Nous examinons dans cette section deux des aspects primordiaux à prendre en compte dès lors que l'on souhaite dépasser l'étape « mot-clé » du processus de recherche. L'anaphore (section 6.1) est un phénomène trop courant pour être négligé, et les contraintes temporelles (section 6.2) sont extrêmement présentes dans les questions posées aux systèmes de question-réponses.

6.1. Anaphore

Une étude réalisée par Vicedo et Ferrández (2000) montre l'importance de l'anaphore en questions-réponses. Se concentrant sur l'anaphore pronominale exclusivement, les auteurs sélectionnent 93 requêtes et, pour chacune de ces requêtes, un passage contenant la bonne réponse justifiée par le texte. Sur ces 93 passages, seuls 37 contiennent la réponse et sa justification sans aucune référence pronominale. Dans 25 documents, la réponse elle-même est représentée par un pronom (comme dans l'exemple 1 ci-dessous). Enfin, 31 documents contiennent une référence pronominale pour au moins un élément de la requête (exemple 2).

Question : « *Qui a atteint l'Everest pour la première fois ?* »

1. « *Sir Edmund Percival Hillary* est un alpiniste néo-zélandais né le 20 juillet 1919 à Tuakau. *Il* est le premier homme à avoir gravi l'Everest ».
2. « Les sept premières tentatives d'escalade de *l'Everest* furent des échecs. Edmond Hillary *l'*atteignit finalement en 1953. »

De nombreuses équipes tentent donc d'incorporer à leur système de questions-réponses un module d'analyse des anaphores. Les algorithmes mis en place s'appliquent généralement en trois phases :

- Pour une expression anaphorique, recherche de l'ensemble des antécédents possibles dans les *n* phrases précédentes.
- Sur chacun de ces candidats, application de contraintes de filtrage (Lappin & Leass 1994) :
  - accords en genre, en nombre, en personne ;

  - non-appartenance au même groupe nominal complexe (dans la phrase « *Edmond Hillary gravit l'Everest avec l'aide de son sherpa* », « *son* » ne peut référer à « *aide* ») ;
  - etc.
- Si plus d'un antécédent reste possible après le filtrage, sélection d'un seul d'entre eux par des contraintes de préférences, comme par exemple la correspondance de structures ou de fonctions syntaxiques : un pronom non réflexif en position sujet aura de préférence un antécédent en position sujet. Pour une anaphore nominale, le sujet référera plutôt à un antécédent objet, etc .
- En dernier critère, toutes choses égales par ailleurs, on préférera le candidat le plus proche.

6.2. Aspects temporels

Une question temporelle n'est pas seulement une question pour laquelle la réponse est une expression temporelle (exemple 1 ci-dessous). Elle peut également contenir une date visant à restreindre les possibilités de réponse (exemples 2 et 3). Mais les contraintes temporelles ne sont pas nécessairement des dates, comme le montrent les exemples 4 et 5.

1. *Quand* la déclaration universelle des droits de l'Homme a-t-elle été adoptée ?
2. Quel journal a été fondé à Kiev *en 1994* ?
3. Qui était le président des États-Unis *entre 1976 et 1980* ?
4. Qui jouait le rôle de Superman *avant d'être paralysé* ?
5. De quelle maladie ont souffert certains soldats américains *après la guerre du Golfe* ?

Harabagiu & Bejan (2005) dressent la classification des questions temporelles. À ces questions, la plupart des systèmes réagissent par un traitement des entités nommées et des mots-clés uniquement : pour la première question, ils rechercheront un élément de type « *date* » (voir la section 5). Pour la seconde, le nombre « *1994* » sera cherché comme tout autre mot-clé dans le texte. Les deux derniers exemples sont rarement traités du point de vue du temps, car cela nécessite, comme le font Ahn & *al*. (2006) par exemple, de rechercher les dates des événements suggérés dans des bases de connaissances. Saquete & *al*. (2004) essaient pour leur part de décomposer les questions complexes en plusieurs questions ; pour le dernier exemple, cela donnerait les questions « *Quand a eu lieu la guerre du Golfe ?* » puis « *De quelle maladie ont souffert certains soldats américains ?* », avec pour contrainte temporelle le résultat de la première question.

D'un certain point de vue, on peut considérer la composante temporelle du texte et des questions comme un problème de coréférence comparable à celui de la résolution d'anaphore. Il s'agit en général de déterminer si le contexte temporel des événements ciblés par les passages retrouvés est compatible avec celui demandé (explicitement ou pas) par la question. Ce travail se fait par la détection de dates absolues (ex : « *le 21 juin 2007* »), qui sont assez rares, ou par l'interprétation d'expressions relatives (« *mardi dernier* », « *le 21 juin* », « *plus tard* »...) permettant le positionnement des événements par rapport à la date de création du document ou par rapport à un autre événement.

Pustejovsky & *al*. (2002) montrent l'intérêt du traitement du temps en questions-réponses, et distinguent quatre pistes de recherche à explorer pour progresser dans le domaine :

- l'attribution d'une date aux événements lorsque c'est possible ;
- l'affectation de relations temporelles entre les événements ;
- le raisonnement sur les changements apportés par un événement ;

- le raisonnement sur la durée d'un événement.

Les deux premiers aspects étaient notamment au centre de la campagne d'évaluation TempEval en 2007 (Verhagen & *al*. 2007). Les deux autres semblent dans l'immédiat hors de portée des systèmes généralistes.

7. Synthèse

Cet article a distingué d'un côté des approches évitant (quasiment) le recours aux connaissances linguistiques (méthodes numériques ou fondées sur l'apprentissage automatique, section 3), et d'un autre côté des approches reposant sur des connaissances linguistiques les plus complètes possibles (sections 4-6). Les premières sont plus facilement généralisables (à différents domaines de spécialité, à d'autres langues) si les données sur lesquelles elles s'appuient sont riches ; les secondes sont plus précises si les connaissances qu'elles exploitent sont assez complètes.

Les premières, inspirées de la recherche d'information, supposent qu'une similarité suffisante pourra être trouvée entre la formulation de la question et la formulation de la réponse dans les textes, quitte à effectuer des transformations minimes sur la question pour la rapprocher de la formulation des réponses. Ces approches fonctionnent d'autant mieux que les données sont redondantes, au point de contenir suffisamment de reformulations de la même information pour que l'une d'elles ait une forme comparable à la question. Elles s'appliquent donc de façon privilégiée au web et aux très grandes collections, mais sont inopérantes sur des corpus de taille plus modérée.

Les secondes s'appuient sur des connaissances syntaxiques, sémantiques ou sur le traitement du discours pour construire des représentations (typage de la question et de la réponse attendue, contraintes sur la réponse) et leur appliquer une normalisation (morphologique, syntaxique, sémantique) qui rendra la question et les réponses possibles formellement plus proches, éventuellement même déductibles l'une de l'autre par inférence logique. La qualité des résultats obtenus par ces méthodes à base de connaissances dépend bien sûr de la complétude des ressources linguistiques employées. Les méthodes d'apprentissage automatique peuvent justement aider à compléter ces connaissances.

Il ressort ainsi que, du point de vue de l'ingénierie des systèmes de questions-réponses, là où des ressources linguistiques sont disponibles en taille et en précision suffisantes, les méthodes de type linguistique sont indiquées ; et qu'à l'inverse, si des données vastes et redondantes permettent de trouver facilement la réponse, ou si elles permettent d'entraîner facilement un système à base d'apprentissage, on aurait tort de se passer de ces données et de ces méthodes. Les méthodes par apprentissage automatique fonctionnent cependant presque toutes après un premier niveau d'analyse linguistique symbolique, que ce soit pour segmenter le texte en mots, les lemmatiser ou même les étiqueter morphosyntaxiquement.

La conclusion à laquelle on arrive est que les seules méthodes numériques ou par apprentissage restent limitées en dehors de situations où l'on a pléthore de données, et qu'il est donc très utile de préparer des ressources linguistiques (lexique, grammaire, cadres de sous-catégorisation, etc.). Mais on doit reconnaître en même temps que les approches linguistiques sont plus longues à mettre en place et peuvent être rejointes (voire dépassées) dans certaines situations, comme indiqué en fin de section 3.1 à propos des questions non factuelles. Une voie différente peut alors consister à tourner les efforts linguistiques vers la caractérisation et l'annotation de données qui serviront à entraîner un système fonctionnant par apprentissage : un corpus annoté selon le niveau que l'on veut obtenir pour l'analyseur (catégories morphosyntaxiques, dépendances, etc.). À la limite, si la collection cible est très

grande et redondante, il n'est plus nécessaire de faire autant d'efforts pour obtenir une analyse précise. Mais, comme souvent, l'issue peut être à la combinaison des méthodes, les approches linguistiques formant des parties sûres du traitement (voir par exemple Habert & Zweigenbaum 2002), les méthodes numériques intervenant pour prendre en charge des parties dont la modélisation linguistique est absente ou incomplète ou pour aider à acquérir les connaissances linguistiques nécessaires, ou encore pour aider à la prise de décision lorsqu'elle n'est pas fondée sur des conditions catégoriques auxquelles la modélisation par règles nous a habitués.

Pierre Zweigenbaum, Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba,
Sophie Rosset, Xavier Tannier, Anne Vilnat
CNRS-LIMSI, BP 133, F-91403 Orsay
<{prenom.nom}@limsi.fr>

Patrice Bellot, Université d'Avignon, LIA, F-84911 Avignon
<patrice.bellot@univ-avignon.fr>

Références

Ahn, K., Bos, J., Clark, S., Curran, J.R., Dalmas, T., Leidner, J.L., Smillie, M.B., Webber, B. (2004). Question-Answering with QED and Wee at TREC-2004. In E.M. Voorhees & L.P. Buckland (dir.), *13th Text REtrieval Conference (TREC 2004)*, Gaithersburg.

Ahn, D., Schockaert, S., De Cock, M., Kerre, E. (2006). Supporting temporal question answering: strategies for online data collection. In *Proceedings of Inference in Computational Semantics (IcoS-5)*, Buxton.

Ayache, C., Grau, B., Vilnat, A. (2005). Campagne d'évaluation EQueR-EVALDA : Évaluation en question-réponse.In *TALN 2005*, Dourdan, 6-10.

Azzam, S., Humphreys, K., Gaizauskas, R. (1998). Evaluating a focus-based approach to anaphora resolution. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, 74-78.

Baker, C.F., Fillmore, C.J., Lowe, J.B. (1998). The Berkeley FrameNet Project. In C. Boitet & P. Whitelock (dir.), *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montréal, 86-90.

Bellot, P. & Boughanem, M. (2007). Recherche d'informations et question-réponse. In B. Grau & J.P. Chevallet (eds), 31-68.

Bikel, D., Miller, S., Schwartz, R. & Weischedel, R. (1997). Nymble: a high-performance learning name-finder. In *Proceedings of ANLP'97*.

Borthwick, A. (1999). *A Maximum Entropy Approach to Named Entity Recognition*. New York University, PhD dissertation.

Bouma, G., Fahmi, I., Mur, J., van Noord, G., van der Plas, L., Tiedermann, J. (2005). Linguistic knowledge and question answering. *Traitement automatique des langues*, 46-3, 15-39.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics* 21(4), 543-565.

Brill, E., Lin, J., Banko, M., Dumais, S., Ng, A. (2001). Data-Intensive Question Answering. In *Proceedings of TREC10*, Gaithersburg, MD.

Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S. & *al*. (2001). Issues, tasks and program structures to roadmap research in question & answering (Q & A). Rapport technique, NIST.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Ursu, C. & *al*. (2007). ANNIE: a Nearly-New Information Extraction System. In Developing Language Processing Components with GATE Version 4 (a User Guide). University of Sheffield.

Dang, H., Lin, J., Kelly, D. (2006). Overview of the TREC 2006 question answering track. *TREC 2006*, NIST Special Publication, Gaithersburg (USA).

DARPA (1998). *Proceedings of the seventh message understanding conference (MUC-7)*. Morgan Kaufmann. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.

Dumais, S., Banko, M., Brill, E., Lin, J., Ng, A. (2002). Web Question Answering : Is More Always Better ? In *Proceedings of SIGIR'02*.

Durme, B.V., Huang, Y., Kupść, A., Nyberg, E. (2003). Towards light semantic processing for Question-Answering. In *HLT-NAACL 2003 Workshop on Text Meaning*, Edmonton, 54-61.

Ferrández, A., Palomar, M., Moreno, L. (1998). Anaphor resolution in unrestricted texts with partial parsing. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, Montreal, 385-391.

Gaussier, E., Grefenstette, G., Hull, D., Roux, C. (2000). Recherche d'informations en francais et traitement automatique des langues. *Traitement Automatique des Langues*, 41-2.

Giampiccolo, D., Forner, P., Peñas, A., Ayache, C., Cristea, D. & *al*. (2007). Overview of the CLEF 2007 Multilingual Question Answering Track.

Gillard, L., Bellot, P. El-Bèze, M. (2006a). Influence de mesures de densité pour la recherche de passages et l'extraction de réponses dans un système de questions-réponses. In *3e Conférence en Recherche d'Informations et Applications (CORIA)*, Lyon, 193-204.

Gillard, L., Sitbon, L., Blaudez, E., Bellot, P., El-Bèze, M. (2006b). The LIA at QA@CLEF-2006. In *Cross Language Evaluation Forum (CLEF 2006)*, Alicante.

Grau, B. (à paraître). Question-réponse et le Web. In M. Boughanem (ed.), *Recherche d'information et le Web*. Paris, Hermes.

Grau, B. & Chevallet, J.P. (eds) (2007). *La recherche d'informations précises : apprentissage, traitement automatique de la langue et connaissances pour les systèmes de question-réponse*. Paris, Hermès.

Grishman, R. (1995). *MUC-6*. http://cs.nyu.edu/cs/faculty/grishman/muc6.html.

Habert, B. & Zweigenbaum, P. (2002). Régler les règles. *Traitement automatique des langues*, 43(3), 83-105.

Hacioglu, K. & Ward, W. (2003). Question classification with support vector machines and error correcting codes. In *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HTL-NAACL)*, vol. 2, 28-30.

Harabagiu, S. & Bejan, C.A. (2005). Question Answering Based on Temporal Inference. In *Proceedings of the Workshop on Inference for Textual Question Answering*, Pittsburg.

Isozaki, H. & Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. In *Proceedings of COLING*.

Ittycheriah, A., Franz, M., Roukos, S. (2001). IBM's Statistical Question Answering System-TREC 10. In *TREC-2001 Conference*, NIST, Gaithersburg, 258-264.

Kamp, H. & Reyle, U. (1993). *From Discourse to Logic. An Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and DRT*. Dordrecht, Kluwer.

Kelly, D. & Lin, J. (2007). Overview of the TREC 2006 ciQA task. *ACM SIGIR Forum*, 41(1), 107-116.

Kingsbury, P., Palmer, M. (2002). From Treebank to PropBank. In *3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, 1989-1993.

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer speech & language* (Print), 6(3), 225-242.

Kwok, C.C.T., Etzioni, O., Weld, D.S. (2001). Scaling Question Answering to the Web. In *Actes de WWW10*, 150-161.

Lappin, S. & Leass, H.J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4).

Laurent, D., Nègre, S., Séguéla, P. (2006). QRISTAL, le QR à l'épreuve du public. *Traitement*

Automatique des Langues, 46(3), 41–70.

Ligozat, A.L. (2006). *Exploitation et fusion de connaissances locales pour la recherche d'informations précises*. Thèse de doctorat, Université Paris Sud.

Ligozat, A.L., Grau, B., Robba, I., Vilnat, A. (2006). L'extraction des réponses dans un système de question-réponse. In *Actes de la 13e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*, Louvain.

Lin, J. (2007). Is question answering better than information retrieval? A task-based evaluation framework for question series. In *Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2007)*, 212–219.

Miliaraki, S. & Androutsopoulos, I. (2004). Learning to identify single-snippet answers to definition questions. In *Proceedings of COLING*, 1360-1366.

Moldovan, D., Clark, C., Harabagiu, S., Maiorano, S. (2003). COGEX : a logic prover for Question Answering. In *Actes de HLT-NAACL*, Edmonton.

Moreau, F., Claveau, V., Sébillot, P. (2007). Intégrer plus de connaissances linguistiques en recherche d'information peut-il augmenter les performances des systèmes ? In *Actes de la 4e Conférence en recherche d'informations et applications, (CORIA'07)*, Saint-Étienne.

Narayanan, S. & Harabagiu, S. (2004). Question-Answering based on Semantic Structures. *20th International Conference on Computational Linguistics (COLING 2004)*, Genève, 22-29.

NIST (2000a). Entity detection and tracking, phase 1, ACE pilot study. Task definition. http://www.nist.gov/speech/tests/ace/phase1/doc/summary-v01.htm.

NIST (2000b). ACE documentation. http://www.nist.gov/speech/tests/ace/phase2/doc/index.htm.

Pasca, M. (2003). *Open-Domain Question Answering from Large Text Collections*. CSLI.

Poibeau, T. & Vilnat, A. (2007). Traitement Automatique des Langues et Question-Réponse. In B. Grau & J.P. Chevallet (eds).

Pustejovsky, J., Wiebe, J., Maybury, M. (2002). Multi-perspective and temporal question answering. In *Third International Conference on Language Resources and Evaluation (LREC,) Workshop on Question Answering: Strategy and Resources*, Canary Islands.

Ravichandran, D., Hovy, E., Och, F.J. (2003). Statistical QA-Classifier versus Re-ranker : What's the difference? In *ACL Workshop on Multilingual Summarization and Question Answering - Machine Learning and Beyond*.

Rosset, S., Galibert, O., Illouz, G., Max, A. (2006). Interaction et recherche d'informations : le projet ritel. *Traitement Automatique des Langues*, 46(3).

Schmid, H. (2004). Probabilistic part of speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester.

Sekine, S. (2004). Definition, dictionaries and tagger of extended named entity hierarchy. In *LREC'04*, Lisbon.

Tanev, H., Kouylekov, M., Magnini, B., Negri, M., Simov, K. (2005). Exploiting Linguistic Indices and Syntactic Structures for Multilingual Question Answering: ITC-IRST at CLEF 2005, *Working notes.*

Tellex, S., Katz, B., Lin, J., Fernandes, A., Marton, G. (2003). Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, Toronto, ACM Press, 41-47.

Turmo, J., Comas, P., Ayache, C., Mostefa, D., Rosset, S., Lamel, L. (2007). Overview of the QAST 2007. In *Working Notes for the CLEF Workshop*, Budapest.

Usunier, N. (2006). *Apprentissage de fonctions d'ordonnancement : une étude théorique de la réduction à la classification et deux applications à la recherche d'information*. Thèse de doctorat, Université Pierre et Marie Curie.

Usunier, N., Amini, M., Gallinari, P. (2007). Apprentissage et systèmes de question-réponse. In B. Grau & J.P. Chevallet (eds), 69-104.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, J. (2007). TempEval Temporal Relation Identification. In *Proceedings of SemEval workshop at ACL 2007*, Prague.

Vicedo, J.L. & Ferrández, A. (2000). Importance of pronominal anaphora resolution in question answering systems. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics,* Hong-Kong, 555-562.

Vilnat, A., Monceaux, L., Paroubek, P., Robba, I., Gendner, V. & *al*.. (2004). Annoter en constituants pour évaluer des analyseurs syntaxiques. In *Actes de TALN* 2004, Fès.

Voorhees, E.M. & Harman D.K. (eds.) (1999). *The Eighth Text REtrieval Conference* (TREC-8). NIST Special Publication, 500-246.

Whittaker, E.W.D., Novak, J.R., Chatain, P., Dixon, P.R., Heie, M.H., Furui, S. (2006). CLEF2006 Question Answering Experiments at Tokyo Institute of Technology. In *Proceedings of the CLEF2006 Workshop*.