Évaluation des systèmes de question réponse

Aspects méthodologiques

Karine Lavenus*,— Guy Lapalme*****

**MoDyCo (Modèles, Dynamiques, Corpus)*
UMR 2329 CNRS / Paris X
200 avenue de la République
92 000 Nanterre

***LexiQuest*
261 rue de Paris
93 556 Montreuil cedex
karine.lavenus@lexiquest.fr

****RALI (Recherche Appliquée en Linguistique et Informatique)*
DIRO, Université de Montréal
CP 6128, Succ. Centre ville
Montréal (Québec), Canada, H3J 3J7
lapalme@iro.umontreal.ca

RÉSUMÉ. *Dans cet article, nous traitons des problèmes épistémologiques rencontrés lors de l'évaluation des systèmes de question réponse. A partir d'une étude s'inscrivant dans la cadre de la « Text Retrieval Conference » (TREC), nous nous attardons plus particulièrement sur le choix des questions, l'évaluation des réponses et sur les critères de pertinence dans l'appariement du couple question réponse.*

ABSTRACT. *This paper is about epistemological problems encountered while evaluating question answering systems. In the context of TREC, we focus on the choice of the questions, on the answers evaluation and on the relevance criteria in the question answer pairing.*

MOTS-CLÉS : *systèmes de question réponse, évaluation, épistémologie, traitement automatique de la langue.*

KEYWORDS: *question answering systems, evaluation, scaling, epistemology, natural language processing.*

1. Introduction

Bien qu'ils soient apparus au début des années 70, les systèmes de question réponse (QR) n'ont provoqué un réel engouement qu'à partir du début des années 90. Cet engouement est directement lié à la vulgarisation d'internet, et s'explique aussi par le fait que les systèmes de QR viennent combler certaines lacunes propres à la recherche d'information (RI). L'utilisateur passe en effet trop de temps à rechercher précisément l'information qu'il convoite dans la masse de documents proposés par un moteur de recherche. Plutôt que de parcourir un, voire plusieurs documents, l'utilisateur préfèrerait obtenir une réponse précise à la question qu'il se pose. Cependant, si les systèmes de QR répondent mieux à la demande et aux besoins des utilisateurs, ils sont plus difficiles à évaluer que les moteurs de recherche. On s'accorde en effet sur les mesures d'évaluation des systèmes de recherche d'information, principalement la précision et le rappel[1]. Évaluer les systèmes de QR est problématique car on ne possède pas de référentiel permettant de calculer le rappel, et il est difficile de définir ce qu'est une bonne réponse. Les mesures dont nous disposons ne sont pas assez fines et évaluent plutôt une quantité, alors qu'une approche qualitative serait nécessaire. A travers cet article, nous tenterons de comprendre plus précisément pourquoi l'évaluation d'un système de QR est difficile alors que de manière intuitive on pourrait penser que c'est très simple, voire binaire lorsqu'on a des questions à la « Trivial Pursuit » : il suffit de compter le nombre de bonnes réponses…

Dans un premier temps, nous verrons quels sont les éléments à prendre en compte lors de l'évaluation de systèmes de question réponse, ainsi que leur interdépendance. Prenant pour prétexte une tâche destinée au suivi de la tâche QR de TREC et à partir du petit corpus fourni dans ce cadre, nous avons décidé de nous focaliser sur trois de ces éléments. Nous nous sommes penchés tout d'abord sur les questions TREC, que nous avons tenté de caractériser notamment en les comparant à de vraies questions d'utilisateurs. Nous avons dans le même temps tâché de les analyser et de les classifier afin de proposer *aussi* une échelle d'évaluation en entrée des systèmes de QR, puisque l'évaluation des systèmes à TREC ne se fait qu'en tenant compte de la réponse. Ensuite, en menant une étude comparative entre l'évaluation des réponses proposée par TREC et la nôtre, nous avons soulevé différents problèmes relatifs à l'appréciation de la validité d'une réponse. Enfin, à la jonction de ces deux pôles, question et réponse, il nous a paru intéressant d'aborder les mécanismes d'appariement. Nous nous sommes attardés plus particulièrement sur la recherche de la réponse basée sur la correspondance avec les mots-clefs contenus dans la question, technique issue de la recherche d'information.

1. La précision (pourcentage de documents pertinents parmi les documents rapportés) mesure le bruit et le rappel (pourcentage de documents retrouvés parmi tous les documents pertinents) mesure le silence.

Dans la mesure où de nombreuses réponses valides contiennent des mots-clefs, nous nous sommes demandés, si, afin d'évaluer les systèmes de QR (et en particulier leur mécanisme d'appariement) de manière plus drastique, il ne valait pas mieux creuser la distance entre la formulation de la question et celle de la réponse.

2. Contexte

2.1. Protocole d'évaluation de TREC

La tâche QR (« *QA track* ») est apparue en 1999 dans le cadre de la huitième conférence d'évaluation en recherche d'information, la TREC-8. Cette conférence a lieu tous les ans depuis 1992 et est organisée par une instance du gouvernement américain, le *National Institute of Standards and Technology* (NIST). Les candidats (entreprises ou laboratoires de recherche pour la plupart) doivent proposer un système de QR capable de répondre à des questions de culture générale, triviales en apparence, du type : How far is it from Denver to Aspen? What county is Modesto, California in? Who was Galileo? What is an atom? When did Hawaii become a state? How tall is the Sears Building?[2] Pour ce faire, ils ont accès lors de la compétition à une base composée d'une collection de documents issus de journaux de langue anglaise. Lors des campagnes d'évaluation de 1999 et 2000, il était demandé aux candidats de fournir une réponse de 250 ou 50 caractères [VOO 01]. Par exemple, pour la question What are the animals that don't have backbones called? on obtient en guise de réponses des bribes de 50 caractères (ou moins) du type : *backbones -- collectively called invertebrates -- ; Invertebrates* ; *invertebrates-seem to have the kind of immune sys.*

Les systèmes de QR ne sont en fait qu'une étape qui doit conduire vers des travaux ultérieurs, utilisant notamment la technique du résumé, afin de répondre à des questions d'experts. TREC augmente donc graduellement ses exigences. En 2000, il a été décidé de suivre les cinq années suivantes les orientations consignées dans un cahier des charges [BUR 00]. Pour la compétition TREC-10 (2001), la présence d'une réponse - limitée à 50 caractères - dans la collection de documents n'étant plus garantie, il fallait être à même de constater ce fait. Les participants devaient chercher à regrouper des bribes éparses de réponses réparties au sein de deux ou plusieurs documents afin de générer une réponse complète. Cette directive devait faire avancer la recherche en motivant les participants à extraire des bribes de réponses de plusieurs documents et à les fusionner en vue de générer une réponse unique ; à repérer l'information redondante et contradictoire ; à prendre en compte des données temporelles ; à reconnaître les stades

[2]. Nous utilisons une fonte particulière pour distinguer les questions et les réponses du reste du texte.

du déroulement d'une histoire ou d'un événement. On peut d'ores et déjà affirmer que cette directive n'a pas abouti de la manière souhaitée. Elle a eu pour effet de bord que certains candidats alignent des mots extraits ça et là pour augmenter statistiquement leur chance de réussite. De ce fait, lors de la conférence TREC-10 il a été décidé que la prochaine évaluation serait basée sur les réponses *stricto sensu*, quelque soit leur taille. Il semble donc que se fasse progressivement un glissement de la forme vers le contenu, de la quantité vers la qualité.

Pour chaque question, chaque participant a droit à cinq propositions de réponses. On assigne un score qui est l'inverse du rang du premier fragment qui contient une réponse correcte. Si aucun fragment ne contient de réponse, la question reçoit un score de zéro. On calcule ensuite la moyenne des scores pour obtenir la moyenne du rang inverse (*Mean Reciprocal Rank*). Ce calcul signifie qu'on considère un résultat donné par rapport à cinq résultats fournis. L'évaluation est faite manuellement par des « experts »[3] qui connaissent les réponses et vérifient s'ils les retrouvent (sans ambiguïté possible), dans les bribes retournées. De plus, les réponses doivent être supportées par le document dont elles sont extraites. Malgré tout, certaines réponses incomplètes ou ambiguës sont validées. Par exemple, à la question What New York City structure is known as the Twin Towers? la réponse suivante, incomplète, a été validée : *as a result of a bomb that rocked the World Trad*. Dans l'exemple suivant Who invented the slinky? des bribes issues de différents documents ont été concaténées afin de multiplier les chances de succès. La réponse doit contenir les noms *Richard James* ou *James* seul. De nombreuses chaînes contiennent *Betty James* qui n'est pas l'inventeur du Slinky mais sa femme. La réponse suivante contient les deux prénoms (*Betty* a été ajouté à l'extrait initial : *husband Richard* Betty *James mechanical engineer*), elle est donc ambiguë et n'aurait pas dû être validée.

2.2. Résultats

Lors de la *QA Track* de TREC-8, pour la compétition concernant l'extraction d'une réponse de 50 caractères, le meilleur système était celui de Cymfony [VOO 00]. La compétition concernant l'extraction d'une réponse de 250 caractères avait été remportée par une équipe de la SMU (Southern Methodist University). C'est cette même équipe (moyennement quelques changements) qui est ensuite arrivée en tête pour les deux types de compétition de la TREC-9 [VOO 01], puis en 2ème position à la TREC-10, avec 51% de réponses correctes[4], cette fois-ci pour le compte de la compagnie Langage

[3]. Il s'agit de personnes âgées qui veulent bien consacrer de leur temps et se prêter au jeu de l'évaluation.
[4]. Parmi les cinq réponses proposées à chaque fois pour chaque question.

Computer Corporation (LCC), [HAR 01]. La première place a été remportée par l'équipe russe d'InsightSoft [SOU 01], avec 77% de réponses correctes. Viennent ensuite dans le classement Oracle et l'Université de Pennsylvanie, avec 40% de réponses correctes. L'Information Science Institute (ISI), dépendant de l'University of Southern California, a fourni 38% de réponses correctes [HOV 01]. La majorité des participants en tête des classements sont américains. Pour TREC-8, il s'agissait surtout de sociétés. L'intérêt pour la tâche QR, le nombre de participants et les retombées économiques croissent de plus en plus. Il est donc important d'examiner les différents aspects de l'évaluation des systèmes de QR.

Les résultats de TREC-10, avec InsightSoft en tête [SOU 01], tendent à prouver que les concurrents et les juges devront faire preuve d'imagination, les uns pour augmenter leurs performances, les autres pour suivre l'état de l'art. En proposant des patrons afin d'extraire des réponses potentielles d'après le type de question, le système d'InsightSoft comble assez bien le fossé entre question et réponse. Il est plus précis que celui des autres concurrents mais les patrons d'extraction s'appuient toujours sur la recherche de mots-clefs qui doivent se trouver à proximité. Ce système fonctionne bien dans la mesure où, évidemment, la réponse colle au patron prédéfini. Oracle [ALP 01], en utilisant la méthode « traditionnelle» de recherche par mots-clefs suivie d'un zoom sur la réponse parvient à un résultat tout à fait honorable avec 40% de bonnes réponses. Mais qu'en sera-t-il lors de la TREC-11, lorsqu'il faudra fournir une seule réponse exacte? Notons que sur les 36 participants, la plupart obtiennent des résultats s'inscrivant dans une fourchette allant de 10 à 30% de réponses correctes. Le cahier des charges [BUR 00] concernant les futures TREC est particulièrement ambitieux, et les résultats actuels nous font douter de l'optimisme qui s'en dégage. Il se peut qu'au vu des résultats de la TREC-11, plus difficile parce que plus restrictive, les objectifs soient revus à la baisse.

3. Facettes d'un système de QR

3.1. Approche théorique et analyse systématique

Un système de QR prend en entrée la question en langage naturel posée par un utilisateur. Au cœur du système, des mécanismes d'appariement permettent de trouver des documents contenant des éléments pertinents et de proposer en sortie une réponse. Celle-ci aide, le cas échéant, l'utilisateur à effectuer une tâche.

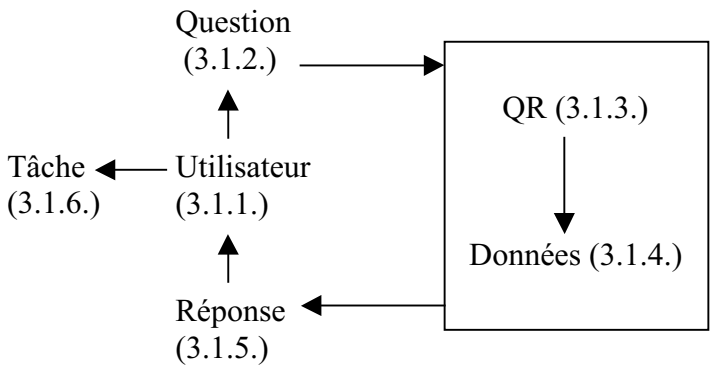

Figure 1. *Éléments à prendre en compte lors de l'évaluation des systèmes de question réponse ; chaque élément est développé dans la sous-section mentionnée entre parenthèses*

Chacun des différents éléments représentés ci-dessus doit être pris en considération lors de l'évaluation. En effet, dans la mesure où ils sont tous plus ou moins interdépendants, on ne peut se focaliser sur le système lui-même en faisant abstraction du reste.

3.1.1. L'utilisateur

Différents facteurs influent sur le fond et la forme des questions que l'utilisateur est susceptible de poser. Citons-en quelques-uns : son âge, son niveau de langue, son milieu socio-culturel, sa culture générale, son expérience, son degré d'expertise dans le domaine visé par la question, etc. Gardons cependant à l'esprit que l'objet d'une évaluation n'est pas d'évaluer les utilisateurs mais les systèmes.

3.1.2. La question

Il faut distinguer entre la question posée par un utilisateur et la question élaborée en vue de tests. Dans le premier cas on a affaire à une « vraie » question, c'est-à-dire à une question émanant d'un utilisateur qui a réellement besoin d'une information, éventuellement en vue d'accomplir une tâche. Dans le second cas, la question peut être fabriquée de toutes pièces, en vue de tester une compréhension, en jouant sur les finesses de la langue. Comme le maître vis-à-vis de l'élève, l'utilisateur connaît déjà la réponse, il n'a pas besoin de l'information fournie en tant que telle, il en a besoin uniquement pour la vérifier et évaluer le système. Dans les deux cas, il existe différents degrés de

complexité. Et, en ce qui concerne la réponse, les attentes varient d'un utilisateur à l'autre. Il y a donc une première interdépendance entre la nature de la question posée ; le profil de l'utilisateur ; le type de réponse attendu (degré de granularité par exemple). L'entrée qui va nous permettre d'évaluer les systèmes de QR, c'est la question. On peut se demander, de manière triviale en apparence : qu'est-ce qu'une question simple, qu'est-ce qu'une question compliquée? (*cf*. section 4) Dans certains cas, le système trouve sans problème la réponse à une question qui nous semble complexe. Il suffit par exemple que la question ait été prévue dans une *base de FAQs*[5], ou que la réponse soit formulée de manière très proche de la question dans la collection de documents. Ici encore, on remarque qu'il y a interdépendance puisque l'on ne peut parler de la complexité de la question sans parler de la nature du système et des données.

3.1.3. Le système de question réponse

Les systèmes varient en fonction de plusieurs paramètres : 1) le domaine d'application : n'importe quel sujet peut-il être abordé? Ou y-a-t-il un ou plusieurs domaines circonscrits? 2) le type d'application : commerce, support, accueil, etc. 3) les bases dans lesquelles ils vont puiser les réponses ; 4) leur fonctionnement (analyse de la question ; recherche de la réponse : mécanismes d'appariement (*cf*. section 6)).

3.1.4. Les données

Taille de la collection et hétérogénéité des données sont des facteurs qui peuvent avoir un impact sur les résultats. En effet, la nature des données et des réponses sont intimement liées. Si la réponse à la question posée n'est pas contenue dans les documents, on ne pourra pas y répondre. Dans ce cas, on ne peut incriminer le système. De plus, les bribes trouvées dans les documents pouvant constituer des réponses sont plus ou moins éparses, complètes et leur formulation est plus ou moins proche de celle de la question.

3.1.5. La réponse

La réponse dépend de la question mais aussi de la base de données. Même si la question est simple *a priori*, s'il n'y a pas de réponse dans la base de données, le système ne fournira aucune réponse ou pire, une réponse erronée. On constate que la validité de la réponse et la formulation de la question sont aussi intrinsèquement liées ; dans l'exemple suivant How many pounds in a ton ?, les réponses comportant uniquement un chiffre sont acceptées puisque l'unité de mesure est indiquée dans la question. En revanche, pour la question What is the average body temperature?, des réponses comportant uniquement un chiffre ont été validées : *is 98.6 degrees* ; *normal body temperature is*

[5]. Une base de FAQs (*Frequently Asked Questions*, ou questions fréquemment posées) est composée d'ensembles de variantes d'une même question rattachées à une réponse standard.

98.6 degrees . Nous ne sommes pas tout à fait d'accord avec ce jugement étant donné que l'unité de mesure n'est pas précisée dans la question (*cf.* section 5) ; de ce fait, si l'utilisateur n'a aucune notion de ce que représentent les unités de mesure Celcius et Fahrenheit, il peut hésiter (voire se tromper) entre elles. Cet exemple soulève aussi le problème suivant : quelles sont les limites de l'appel aux connaissances du monde? Est-il évident pour tous que la température moyenne du corps humain est de 98,6 degrés, *en Fahrenheit*? Pour poser une telle question, peut-être pas…

3.1.6. L'adéquation à la tâche

Les besoins de l'utilisateur peuvent être les suivants : 1) combler un manque de connaissances par simple curiosité ou intérêt pour un sujet donné ; vérifier des connaissances ; 2) combler un manque d'information pour mener à bien une action. Pour chacun de ces besoins, on peut encore distinguer des degrés d'exigence différents en fonction du type d'utilisateurs. Il y a donc à nouveau interdépendance entre la réponse attendue et l'utilisation qui en sera faite, qui dépend à la fois de l'utilisateur et de la tâche qu'il doit accomplir.

3.2. Les données de l'évaluation

Afin de mener à bien notre évaluation – que l'on pourrait qualifier de méta-évaluation, puisque c'est en partie l'évaluation de TREC que nous allons évaluer - et pour comparer ce qui est comparable, nous avons choisi de travailler sur des données relativement homogènes. Le cadre fourni par la tâche QR proposée par TREC nous a paru adéquat. Même si les questions ne sont pas de « vraies » questions telles que nous définirons ce concept en 4.1., les organisateurs de TREC qui les conçoivent s'inspirent largement de questions effectivement posées sur internet par de « vrais » utilisateurs. Ces questions n'appartiennent pas à un domaine en particulier, ce qui répond à l'objectif premier des systèmes de QR (pallier les lacunes des moteurs de recherche) : ils vont puiser leurs informations dans une collection hétérogène de documents et sont censés répondre à des questions ouvertes (*i.e.* qui ne sont pas limitées à un domaine). Quel que soit leur fonctionnement et leur mécanisme d'appariement, les systèmes concurrents sont soumis aux mêmes règles de participation, ce qui homogénéise les données : ils doivent répondre aux mêmes questions ; ils doivent fournir une réponse de 50 caractères au plus issue de la même collection de documents montée par les organisateurs de TREC. Cette restriction empêche certains systèmes de concourir, tels les systèmes basés sur des FAQs, qui de toute façon ne fonctionnent qu'au sein de domaines bien circonscrits et donc restreints. Notons cependant que cette homogénéisation globale des conditions de participation est entachée du fait que les concurrents peuvent choisir un moteur de recherche autre que celui proposé par TREC.

Pour cet article, nous avons donc utilisé un petit corpus issu des données de TREC. Voici le contexte dans lequel s'est inscrite l'étude que nous avons menée. Quelques temps avant la dixième conférence TREC organisée par NIST en novembre 2001, dans le cadre de la tâche QR, Hellen Voorhees, l'organisatrice, donna 20 questions - ainsi que les réponses proposées par les différents participants - à des bénévoles, afin que ceux-ci fournissent des patrons permettant de retrouver les réponses valides. Notre contribution nous a donné l'occasion de nous pencher d'un point de vue épistémologique sur les questions (section 4), puis sur les réponses (section 5) et enfin sur l'appariement (section 6).

4. En entrée des systèmes : la question

4.1. Questions TREC versus « vraies » questions

La comparaison des questions TREC aux « vraies » questions d'utilisateurs nous permettra dans un premier temps de les critiquer et de mieux en cerner l'essence. Ensuite, deux propositions de classification des questions nous donneront l'occasion de mieux les analyser.

Essayons tout d'abord de définir les questions TREC non pas par ce qu'elles sont mais par ce qu'elles ne sont pas. Contrairement aux « vraies » questions d'utilisateurs que nous avons puisées dans différents corpus[6], les questions TREC sont ciblées : il n'y a pas de questions floues (qui équivalent à des demandes d'informations générales pouvant être satisfaites par la RI), du type : je voudrais des informations ou des photos de BMX et de skateboard. On attend effectivement une réponse concise, qui tient dans 50 caractères ou moins. De plus, les questions TREC sont des interrogatives à la syntaxe canonique et convenue (de type : pronom interrogatif – V - Cplt), alors que la plupart du temps les internautes expriment des demandes, pas nécessairement à la forme interrogative, souvent sous forme de constat ce qui rend la question implicite. Elles ne contiennent qu'une interrogation à la fois alors que les questions d'utilisateurs sont souvent multiples et, de ce fait, longues et anaphoriques : Je cherche à savoir s'il existe encore des championnats de courses de motos sur glace, en France ? Pouvez-vous m'aiguiller sur un site Web et me donner quelques indications sur les dates des prochaines courses ? Souvent, demandes d'informations générales et demandes précises

[6]. Notamment le corpus fournit par *Voilà* en 2000. *Voilà* est un moteur de recherche français, qui offre aux internautes la possibilité de poser des questions par courriel à des opérateurs, domaine par domaine. Le corpus de questions dont nous disposons concerne le sport, les activités et nouvelles sportives.

sont d'ailleurs combinées : Je vous serais reconnaissant de m'envoyer une documentation complète sur votre organisation des courses et me faire connaître les conditions de participation, les éléments requis pour joindre votre organisation des courses et les services que vous proposez ? Beaucoup de détails anecdotiques apparaissent lorsque la question n'est pas délimitée par une fenêtre très étroite. Ces détails servent généralement à introduire et justifier la demande, à l'inscrire dans un contexte. Cependant ils sont généralement superflus et généreraient du bruit au sein du traitement automatique. Il y a aussi des demandes qui ne sont pas très claires, dont le focus est difficile à déterminer. D'autre part, très souvent les demandes posées par les utilisateurs le sont en vue d'effectuer une action que celle-ci soit déclarée explicitement ou non : Je recherche des places pour le tournoi des six nations. Savez-vous où je peux en trouver sur le net ? Ce sont les demandes précises avec action à la clef (le fait qu'elles soient précises s'expliqueraient par le but, l'objectif) qui ressemblent le plus aux questions de TREC à la forme canonique : Comment envoyer un mail aux joueurs de tennis français ? ; Comment réserver des places pour voir un grand prix de F1 ? ; Quelle est l'adresse du site pour réserver des places pour le match France - Angleterre ? Notons ici que les questions de TREC sont exemptes de fautes d'orthographe, de syntaxe ou de frappe, ce qui n'est pas le cas des vraies questions d'utilisateurs. Enfin, les questions de TREC ne sont pas des questions polémiques, qui entraînent un jugement subjectif, une prise de position, et qui nécessitent en guise de réponse une argumentation.

4.2. Classification de Lehnert

Avant de procéder à notre propre classification, observons d'abord celle de Wendy Lehnert [LEH 78], qui regroupe les questions sous treize catégories conceptuelles, déterminées par le focus de la question. Cette classification - une des seules dont nous disposons - a inspiré les organisateurs de TREC pour créer leur ensemble de questions. Cependant, il nous a semblé que cette classification ne reflétait que partiellement le type de questions posées dans le cadre des systèmes de question réponse. Tout d'abord il semble que la différence entre les « *yes/no questions* » et les disjonctives ne mérite pas d'être soulignée : les deux types sont semblables au fond, sauf que la deuxième proposition propre aux disjonctives n'est pas explicite dans les « *yes/no questions* ». La catégorie « *verification* », tout comme la catégorie « *completion* » ne nous paraissent pas être des catégories en soi, parce qu'elle ne sont pas assez discriminantes : presque toutes les questions peuvent être considérées comme des vérifications, ou comme un besoin de complément d'information, qui indique justement le focus de la question. Ensuite les questions appelant une dénomination, une définition ou une explication sont absentes de

cette classification, ainsi que les « *why famous person* »[7] et les questions portant sur la fonctionnalité d'un objet. Enfin on observe que la plupart des catégories proposées par Lehnert s'inscrivent dans le cadre du dialogue ; certaines catégories de questions n'ont d'ailleurs pas été retenues pour la compétition TREC car elles impliquent une réponse subjective (*judgmental*) ou une action (*request*). Cette classification est en fait liée à l'application pour laquelle elle a été initialement créée : Lehnert a proposé en 1978 un système de QR (QALM) permettant à l'utilisateur de tester la compréhension d'histoires, ce qui expliquerait l'emphase sur la cause au sein de la catégorisation. Les exemples de questions portant sur la cause ou le but (« causal antecedent », « goal orientation », « causal consequent », « expectational ») ne semblent d'ailleurs pas toujours pertinents, la distinction entre la cause et le but, la cause et la manière, la cause et la conséquence n'étant pas toujours flagrante. Dans TREC, les questions portant sur la cause relèvent du domaine scientifique, afin que la réponse soit factuelle, objective et irréfutable.

4.3. Premier essai de classification « in abstracto »

Les questions de TREC semblent de prime abord factuelles et triviales. Nous verrons section 5 que cette première impression doit être nuancée, car dans de nombreux cas les réponses proposées diffèrent et nous donnent à penser que les questions sont pas aussi simples qu'elles en ont l'air. Nous avons pris la liste des 20 questions que les organisateurs de TREC nous avaient attribué au hasard et avons essayé de les classer en fonction d'un degré de difficulté supposé, au vu de chaque question (des plus simples *a priori* aux plus complexes). Cette classification devrait nous permettre de proposer une grille d'évaluation graduelle des systèmes en tenant compte des types de questions et des difficultés linguistiques de compréhension.

[7]. Questions du type « Who is Margaret Atwood ? » visant à connaître la raison pour laquelle telle personnalité est célèbre.

rg	%RV	No	Questions	F	H	tal	def	den	EN	log
1	37%	1110	What is sodium chloride?	X			X			
2	26%	1112	How many pounds in a ton?	X					X	
3	23%	1113	What is influenza?		X		X			
4	19%	1100	Who was the 23rd president of the United States?	X					X	
5	17%	1101	What is the average body temperature?	X					X	
6	16%	1095	What city's newspaper is called "The Enquirer"?	X					X	
7	15%	1094	What is the name of the satellite that the Soviet Union sent into space in 1957?	X				X		
8	14%	1102	What does a defibrillator do?		X					X
9	14%	1106	What province is Montreal in?	X					X	
10	12%	1105	How fast is the speed of light?	X					X	
11	7%	1107	What New York City structure is also known as the Twin Towers?			X				
12	5%	1097	What are the animals that don't have backbones called?			X		X		
13	4%	1109	What is the most frequently spoken language in the Netherlands?			X		X		
14	3%	1096	Who invented the slinky?	X					X	
15	2%	1104	What year did the United States abolish the draft?	X					X	
16	2%	1103	What is the effect of acid rain?		X					X
17	1%	1099	Where is the volcano Olympus Mons located?			X			X	
18	1%	1108	What is fungus?	X			X			
19	0%	1098	What is the melting point of copper?	X					X	
20	0%	1111	What are the spots on dominoes called?	X				X		

Figure 2. *Classification des questions de l'échantillon TREC : les questions numérotées (colonnes 3 et 4) apparaissent par ordre décroissant de succès (colonne 1), en fonction du nombre de réponses valides (RV) trouvées (colonne 2) ;* **F** *désigne les questions factuelles,* **H** *les questions plus hétérogènes,* **tal** *les questions qui posent des problèmes d'analyse,* **def** *les questions portant sur une définition,* **den** *sur une dénomination,* **EN** *visant des entités nommées,* **log** *des liens pseudo-logiques*

Dans un premier temps, en fonction de la nature de la réponse, nous distinguerons deux types de questions : d'une part les questions factuelles à réponses unique dans le fond (mais pas nécessairement dans la forme, notées **F** dans le tableau et analysées en **4.3.1.1**.) ; d'autre part les questions factuelles à réponses de degrés de granularité divers (notées **H** dans le tableau et analysées en **4.3.1.2**.).

Nous nous pencherons ensuite sur les problèmes qu'un système peut rencontrer lors de l'étape d'analyse et de compréhension des questions (notées **tal** dans le tableau et analysées en **4.3.2.**).

Puis nous présentons une autre classification des questions, en fonction de critères sémantiques cette fois-ci : questions qui visent à obtenir une définition (**def, 4.4.1.**), une dénomination (**den, 4.4.2.**), une entité nommée (**EN, 4.4.3.**) ou d'autres éléments pouvant être représentés sous forme logique (**log , 4.4.4.**).

4.3.1. Questions classées en fonction du degré de granularité des réponses possibles

4.3.1.1. Série de questions factuelles

Pour cette série de questions, on attend en guise de réponse une ou deux chaînes de caractères exprimant un fait de nature objective et indiscutable. Même si la formulation de la réponse diffère, le fait dont on parle est un et univoque. Par exemple, pour la question Who was the 23rd president of the United States?, la réponse est *Benjamin Harrison* (*Harrison* seul peut aussi être éventuellement accepté). Comme c'est le cas pour la question What city's newspaper is called The Enquirer ? le fait que la question soit factuelle n'empêche pas cependant qu'il puisse y avoir plusieurs réponses : *Battle Creek ; Yorkville ; Cincinnati*.

4.3.1.2. Questions qui impliquent des réponses plus hétérogènes

Les questions qui suivent sont factuelles, cependant les réponses varient dans la forme et surtout dans le fond : 1) What is influenza? : il y a plusieurs degrés de granularité possibles au niveau de la réponse : maladie (virale, respiratoire), virus ; 2) What does a defibrillator do? : cette question est ambiguë malgré une apparente simplicité : cherche-t-on à connaître la fonction d'un défibrillateur, ses effets, les buts de son utilisation? 3) What is the effect of acid rain? : plusieurs réponses sont possibles, il n'y a pas qu'un effet. Pour ce type de question, on risque d'ailleurs d'avoir des réponses qui ne correspondent pas à la forme canonique d'une réponse, puisque la réponse ne tient pas nécessairement dans un ou deux mots (contrairement aux entités nommées), et peut se présenter soit sous forme de verbe, soit sous forme de substantif, etc. On obtient ainsi pour cette question les quatre réponses suivantes : *rain threatens lakes and streams and their aquatic ; destroys aquatic life in lakes and streams ; modernization. 1. Destruction of Forests. Before ; damage to trees - it can be hard to distinguish t.*

4.3.2. Questions qui soulèvent des problèmes dans le cadre du traitement automatique de la langue (TAL)

Nous avons soulevé les points qui pourraient poser problème au niveau de l'analyse et la compréhension de certaines questions lors du traitement automatique. Le rappel des résultats trouvés (figure 2, p.12) permet d'établir – ou non – un lien entre les difficultés annoncées et les résultats. Il nous semble que le focus de la question est l'élément le plus important mais aussi le plus difficile à déterminer pour orienter la recherche afin de trouver la réponse. Très souvent, dans ses questions l'utilisateur fait d'une part état de ses connaissances sur un sujet donné, et d'autre part état de ses lacunes [KEY 96]. Le focus correspond à ce qu'il cherche à savoir. Les informations mentionnées dans la question sont des présupposés qui aiguillent la recherche de la réponse. Cependant, une fois le focus repéré, il faut voir de quelle manière la réponse peut être exprimée : l'information associée à la réponse n'est pas toujours la même que celle donnée dans la question. La plupart du temps, la réponse se présente sous forme de bribes ou de données implicites, à partir desquelles l'*humain* peut reconstituer ou déduire une réponse compacte.

Dans l'exemple suivant What are the animals that don't have backbones called?, le focus n'est pas facile à circonscrire. De manière générale, il est difficile de trouver automatiquement le focus de questions commençant par What et de plus celle-ci possède une syntaxe particulière, enchâssée en quelque sorte. La comparaison entre les deux interrogatives : What are the animals that don't have backbones called? / What are the animals that don't have backbones doing? met en exergue cet enchâssement. Dans cette question, on cherche le terme adéquat correspondant à une définition, une description, le nom qui vaut pour cette représentation en particulier, le couple signifiant/signifié ramassé en une seule entité pour désigner un concept déployé ici. La description donnée sert en quelque sorte d'indice pour une devinette, comme quand on cherche un mot qu'on a « sur le bout de la langue ». S'il n'y a pas de définition utilisant le terme « backbones », même si le terme apparaît dans la collection de documents, il va être difficile de retrouver le mot recherché. La forme négative ajoute un degré de complexité supplémentaire à l'analyse de la question : l'indice qui doit permettre de retrouver la réponse est à la forme négative : les animaux ne sont pas décrits par ce qu'ils sont mais par ce qu'ils ne sont pas… Or, dans la plupart des cas, la négation n'est pas – ou est mal - traitée en TAL! On risque donc de construire une requête représentant la question à partir des termes suivants : animal(s) + backbone(s) + call(ed), ce qui aboutirait à une représentation inverse de celle souhaitée. Le même type de problème se pose lors du traitement de la voie passive : celle-ci n'est pas nécessairement prise en compte lors de l'analyse de la question, s'il n'y a pas, par exemple, d'analyse syntaxique fine. On peut donc sélectionner en guise de réponse des phrases du type *The animals call (to) their young* si dans le voisinage se trouve « backbone », ou si on utilise la « dégradation de

clauses[8] ». Se pose enfin le problème de la désambiguïsation syntaxique et sémantique : le terme « call » est ambigü. Il peut être nom ou verbe (transitif ou intransitif), et dans tous les cas possède plusieurs sens.

L'exemple suivant What New York City structure is also known as the Twin Towers? soulève d'autres types de problèmes. Le focus n'est pas indiqué directement, explicitement : tel bâtiment est connu *sous le nom de* « tours jumelles ». De plus, en tant que mot vide, le terme "as" a de fortes chances de faire partie d'une liste de mots interdits et donc de ne pas être pris en compte, alors qu'il joue un rôle sémantique important. Contrairement à l'exemple précédent, il ne s'agit pas ici de trouver le terme correspondant à une définition mais l'équivalent, le synonyme. *A priori*, il y a peu de chances pour que le terme recherché et son synonyme soient cooccurrents : pour désigner le concept, on va choisir l'un ou l'autre terme. Mais l'utilisation de synonymes pour éviter la répétition peut favoriser une certaine proximité. Enfin, il faut être en mesure de reconnaître *New York City structure* et *Twin Towers* comme des composés, le second étant figé, afin de ne pas rechercher chaque terme séparément, ce qui engendrerait énormément de bruit.

Finalement, hasard ou pas, sur les 4 questions jugées ***a priori*** les plus difficiles (1097, 1099, 1107, 1109), 3 se suivent dans le classement des résultats (voir figure 2, p.12). On remarque qu'elles présentent toutes moins de 10% de réponses valides et apparaissent juste en dessous du 10ème rang qui était encore dans la moyenne. Elles n'apparaissent cependant pas en dernier lieu. Quant aux deux questions auxquelles on n'a pas su répondre (1098 ; 1111), elles n'avaient peut-être pas de réponses dans les documents (ce qui est quasiment impossible à vérifier manuellement vu la taille de la collection de documents fournie par TREC[9], ou en s'appuyant sur les techniques de la RI, inadéquates ici).

4.4. Deuxième essai de classification, en fonction du type d'informations recherchées

On tâchera ici de prendre en compte la sémantique de la question. On relève tout d'abord des questions qui visent à connaître la définition d'un terme, ou à connaître la dénomination précise d'un concept exprimé sous forme de paraphrase. Suit une série de questions dont la réponse peut être trouvée à l'aide d'entités nommées (noms de

---

[8]. On lance une première requête contenant tous les mots-clefs. Si les résultats sont nuls ou insuffisants, on relance plusieurs requêtes en éliminant à chaque fois un mot-clef (en observant des règles syntaxiques afin de rendre compte des relations hiérarchiques et sémantiques de dépendance entre termes).

[9]. La collection comprend 979 000 articles, ce qui fait 3 gigabytes de texte.

personnes, lieux, chiffres) ou grâce à des modèles supportés par des liens logiques (effets, buts, conséquences).

4.4.1. Définitions

Les questions 1108, 1110, 1113 sont toutes construites sur le même modèle : What is + SN ? Elles orientent la recherche vers une définition. On n'attend pas nécessairement d'une définition une explication agrémentée de moult détails. Un simple mot peut suffire, qu'il s'agisse d'un terme en relation d'hyperonymie ou de synonymie :

- relation d'hyperonymie : dans le cas de la question What is influenza?, le générique d'influenza peut être donné en guise de réponse : influenza ***est une sorte de*** *virus* ou *desease*. On peut s'appuyer sur un réseau sémantique tel que Wordnet pour répondre à ce type de question [HOV 01] ;

- relation de synonymie : dans le cas de la question What is sodium chloride? on cherche la définition d'un terme savant qui désigne en fait quelque chose de très courant et de très banal : le sel. La correspondance des mots-clefs fonctionne bien car il y a de nombreuses mises en apposition dans les documents, visant à rappeler grâce à un simple mot la signification du composé savant « sodium chloride » :

is part of sodium chloride, ordinary table ***salt****,*
percent. Table ***salt****, or sodium chloride, is one*
of the sodium chloride _ ***salt*** *_ in the curing*
g sodium chloride (table ***salt****), aluminum oxide and*

Figure 3. *Réponses valides contenant des mises en apposition*

4.4.2. Dénominations

L'exemple What New York City structure is also known as the Twin Towers? nous montre que la frontière entre les cas de questions/définitions - qui vont trouver leur réponse dans la relation de synonymie - et les cas de questions/dénomination est très mince. Il existe aussi une sorte de relation de complémentarité entre définitions et dénominations. A l'inverse des questions de type définition, on peut avoir la définition et demander le terme qui y correspond : What are the spots on dominoes called? On aurait pu connaître le terme mais pas sa définition, et poser la question complémentaire : What are dots?

4.4.3. Entités nommées

L'utilisateur possède déjà une information sur un sujet donné, et voudrait un complément d'information. Par exemple, à un fait, un événement, une situation est

associé un nom de personne. Pour la question 1100, on sait qu'il y a eu un certain nombre de présidents aux Etats-Unis et on a besoin de savoir qui était précisément le 23ème. Ce genre de questions part de connaissances et représentations du monde : s'il est précisé « *of the United States* », c'est que le locuteur sait que cette information est nécessaire pour obtenir une réponse ciblée. On suppose aussi qu'il y a eu au moins 23 présidents. Noms de personnes (1096, 1100), lieux (1095, 1099, 1106) et données chiffrées (1098, 1101, 1104, 1105, 1112) peuvent être repérés relativement facilement et extraits des documents lorsque ceux-ci bénéficient d'un étiquetage des entités nommées.

4.4.4. Autres

Un autre type de réponses traitant des effets, buts, conséquences (1102, 1103) ne relève pas des entités nommées mais peut être trouvé en s'appuyant sur la reconnaissance de patrons relativement simples et courts qui utilisent des liens sémantiques logiques tels que : *cause [V] conséquence*, où [V] représente l'axe syntagmatique occupé ici par un verbe tiré d'un paradigme du type « provoquer, causer ».

En plus de la vingtaine de questions attribuées par TREC, nous avons analysé une centaine de questions (questions 984-1113), ce qui nous a permis de mettre à jour un nombre restreint de patrons regroupés par catégories, qui représentent bien l'ensemble. Comme de nombreuses questions sont construites sur le même modèle, on peut se demander dans quelle mesure les tests ne sont pas biaisés : il aura suffi aux participants de TREC-9 de s'inspirer des questions de TREC-8 pour concevoir des modèles [HER 01], augmentant ainsi leur chance de succès (qui devient de ce fait somme toute relatif). Lors de la compétition de la TREC-10, sur les 500 questions posées, 232, soit quasiment la moitié, relevaient du modèle des questions définitions What is + SN? Celles-ci, *in abstracto*, paraissent assez simples à traiter, surtout lorsque les cas de synonymie et d'hyperonymie (tout comme les cas de questions dénominations d'ailleurs) peuvent être facilement résolus par l'appel à un réseau sémantique tel que WordNet [HAR 01]. Les questions auxquelles on peut répondre en extrayant des entités nommées sont aussi relativement simples à traiter. L'étiquetage permettant leur repérage et leur extraction s'appuie en effet sur des patrons intégrant si besoin est un élément issu d'une liste (tels que prénoms ou titres permettant de repérer les noms de personnes). Les autres questions font appel à des représentations logiques du type cause/conséquence qui peuvent aussi être modélisées par des patrons. Finalement, les questions les plus difficiles à traiter sont celles qui posent des problèmes d'analyse et de compréhension, tels que ceux évoqués en 4.3.2.

A partir de ces questions de type Quizz, les participants risquent donc de limiter leurs efforts en construisant leur système en fonction de questions « à la TREC » alors que le but de la compétition est de stimuler la recherche et de pousser à l'amélioration des

systèmes. En même temps, ces questions factuelles ne constituent qu'une étape dans le cahier des charges prévu par TREC : le degré de difficulté va aller en augmentant [BUR 00]. D'un point de vue épistémologique, nous avons constaté que les questions proposées par TREC sont relativement éloignées des « vraies » questions d'utilisateurs[10] (*cf*. 4.1), même si elles s'en inspirent. Cependant, elles ne sont pas aussi triviales qu'elles en ont l'air. Elles sont difficiles à classer, et peuvent être classées en fonction de différents critères, ce qui pose un problème au cas où nous aurions besoin d'une grille d'évaluation prenant en entrée des questions de plus en plus complexes. Elles sont aussi parfois ambiguës, comme nous allons le voir plus précisément en nous penchant sur les réponses.

5. En sortie des systèmes : la réponse

5.1. Définition

En théorie, une réponse satisfaisante doit fournir à l'utilisateur l'information dont il a besoin. En pratique, il est difficile de savoir quel degré de granularité et d'expertise est attendu par l'utilisateur. Les systèmes de QR, de toute façon, ne font pas la différence entre les utilisateurs, et proposent les réponses qu'ils trouvent, en fonction du contenu de la collection de documents. Jusqu'à la TREC-10, la réponse pouvait être présentée au sein d'une bribe de 50 caractères. Comme les questions sont factuelles, on s'attend à ce que la réponse soit courte, contenue dans quelques chaînes de caractères, voire dans un mot ou deux (nom de personne, lieu, date, etc.). Cependant, lors de la TREC-11, les candidats devront fournir la réponse exacte, débarrassée de ses scories. Le défi est donc plus grand, car les bonnes réponses trouvées par hasard seront de plus en plus rares, et les techniques d'ajustement du zoom sur la réponse devront être particulièrement fines. Les résultats risquent d'en pâtir. D'un autre côté, la tâche des juges sera encore plus délicate car jugeront-ils comme valide une réponse bonne en soi mais entourée d'éléments accessoires? Cela signifie-t-il aussi que les réponses bonnes en soi mais non validées à cause du contexte (notées 2), rejetées lors des précédentes TREC, pourraient être acceptées? Les divergences entre juges pourraient être nombreuses. Cette modification de la présentation de la réponse pourrait avoir pour conséquence de limiter le bruit (les ambiguïtés, par exemple lorsque dans une même chaîne on a deux réponses possibles), ou bien au contraire de l'augmenter si l'ajustement du zoom est mauvais. Mais elle ne résout pas le problème de granularité de la réponse attendu par l'utilisateur. Même si une réponse plus explicite, plus détaillée peut être fournie en dépassant la limite

[10]. Le focus de « vraies » questions n'est pas facile à déterminer, parce qu'il s'agit de questions - ou plutôt de demandes - bruyantes car implicites, multiples, à la syntaxe parfois fantaisiste.

des 50 caractères, on ne sait pas si la réponse sera acceptable du point de vue de l'utilisateur vu qu'on ne connaît pas son niveau ni la tâche qu'il doit éventuellement accomplir grâce à l'information contenue dans la réponse.

5.2. Méthodologie

Afin de vérifier la validité des réponses proposées, nous ne nous sommes pas fiés aveuglément à l'évaluation des juges de TREC : nous avons le cas échéant utilisé un dictionnaire ou un moteur de recherche. D'autre part, nous avons parcouru l'ensemble des réponses : nous n'avons pas seulement vérifié que les réponses jugées valides (notées 1) l'étaient effectivement, nous avons aussi cherché à savoir si parmi les réponses jugées non valides en soi ou à cause du contexte (respectivement notées -1, 2), il n'y avait pas d'erreur d'appréciation. Cette expérience nous a amenés à réfléchir sur le concept de validité des réponses, tout en remettant en cause certaines évaluations proposées.

5.3. Divergences au niveau de l'évaluation

En se penchant sur les bribes de réponses de 50 caractères validées ou non par les juges de TREC, on arrive à la conclusion qu'il est possible de proposer d'autres types de jugements :

- **mauvaises réponses validées** : nous considérons que le contenu de ces réponses n'est pas valide, pas acceptable même si ces réponses ont été validées par les juges de TREC. Par exemple, pour la question What is the name of the satellite that the Soviet Union sent into space in 1957 ?, la réponse suivante n'aurait pas dû être validée (la bonne réponse est *Sputnik I*) *: Soviet Union launched Sputnik II , the second* .

- **réponses inexactes** : les bribes de réponses validées ne sont pas fausses, mais elles ne nous semblent pas assez précises. L'appréciation dépend en fait du degré attendu de granularité de la réponse, et donc du type d'utilisateur qui se sert du système. Par exemple, on constate un manque de précision concernant des données chiffrées. La question suivante How fast is the speed of light ?, factuelle, implique une réponse unique dans le fond. Les réponses validées à plusieurs reprises *186.000 miles per second ; 300.000 km per second* ne correspondent pas aux réponses *exactes* qui sont : *186.451 miles per second ; 299.792 km per second.*

- **réponses incomplètes** : la chaîne de caractères constituant la bribe de réponse est coupée au début ou à la fin de ce qu'on pourrait considérer comme une réponse. De ce fait la réponse n'est, en général, pas satisfaisante. Dans la mesure du possible, l'utilisateur doit reconstituer une réponse complète en s'appuyant sur ses connaissances

du monde. Mais dans certains cas, c'est impossible, notamment parce qu'il se peut que le sens de la réponse se trouve changé. Pour la question What city's newspaper is called "The Enquirer"?, quelqu'un qui n'a jamais entendu parler de Cincinnati ne sera pas capable de reconstituer une réponse complète à partir de la bribe suivante: *Blackwell also snagged the endorsement of Cincinn .*

Bien qu'aucun terme ne soit tronqué, dans l'exemple suivant Who was the 23rd president of the United States? la réponse est aussi incomplète *: and Benjamin was later to become the 23rd,* puisque la réponse attendue est *: Benjamin Harrisson*.

La troncature peut modifier le sens de la réponse sans que l'utilisateur ne s'en aperçoive. Pour la question What does a defibrillator do? les deux réponses suivantes, issues du même document et du même passsage ont été validées : *automatic defibrillator to stop irregular heart be ; defibrillator to stop irregular heart beats does*. Comme nous pouvons le constater ici, la limitation de la réponse à une chaîne de 50 caractères a un effet de bord non négligeable sur le sens de la réponse, si le dernier mot est coupé par exemple. Ici on a *be* au lieu de *beats*, ce qui modifie totalement la signification de la bribe de réponse et conduit à un non-sens.

- **bonnes réponses non validées** : la bribe répond bien à la question mais pour une raison inconnue (erreur? Inattention?), elle n'a pas été validée. Dans le cas suivant What is the effect of acid rain? on peut considérer que le degré de granularité des réponses ne convenait pas au juge, mais il pourrait peut-être satisfaire un utilisateur *: Acid Rain: How it Happens, Why it Kills* (trop vague?) *; in the ozone layer and destruction of rain forest.*

Concernant l'exemple What is sodium chloride?, de bonnes réponses, *salt, or sodium chloride, is one kind of sodium ; salt content salinity specific toxicity high* n'ont pas été validées, ce qui s'explique par un manque d'attention, puisqu'ailleurs ce type de réponse a été validé. Notons qu'il se peut qu'il y ait des différences d'évaluation au sein même de TREC : pour la question What year did the United States abolish the draft? la réponse, *1973*, est validée dans un cas, pas dans un autre, similaire.

5.4. Incidence du contexte sur la validité de la réponse

Dans de nombreux cas, la bribe contient la réponse attendue mais est invalidée soit parce que le document cité comme source n'est pas le bon, soit parce que la réponse est ambiguë. Dans l'exemple suivant What province is Montreal in? nous devons nous assurer que la bribe de réponse ne contient pas le nom d'une autre province : *in Quebec and Ontario*, ce qui pourrait rendre la réponse ambiguë. D'autre part, la réponse, *Quebec*, ne doit pas être suivie du mot *city*, parce que nous recherchons le nom d'une province – pas celui d'une ville. Néanmoins, lorsque *Quebec* apparaît seul dans la bribe, il peut référer à la ville sans que cela soit explicite. La réponse est considérée comme valide si

elle est supportée par le document, même si la chaîne de caractères extraite désignait en fait la ville. Autre point : même si le mot *Quebec* est présent et désigne bien la province, dans la mesure où il est composant d'un composé, peut-on vraiment considérer les réponses suivantes comme de bonnes réponses? *by Caisse de Depot et Placement du Quebec , the ; 's future within Canada, Hydro-Quebec said it does.*

Autre cas : What is the most frequently spoken language in the Netherlands? La réponse attendue est le nom ou l'adjectif *dutch* en relation avec le concept de langue. De nombreuse réponses contiennent *dutch* mais elles traitent des habitants : *Even the Dutch, renowned for their ability to spea; the struggle. While the Dutch take it for granted*. En revanche, nous accepterions sans aucune restriction *: home and continued to speak Dutch when alone ; the 20 million Dutch speakers in the Netherlands a.*

5.5. Variantes au sein des réponses

Les systèmes de QR ne sont pas des systèmes de dialogue. La réponse trouvée dans les documents ne fait pas écho à la question, comme dans les conversations, où il y a souvent reprise partielle des termes de la question et où la réponse comble littéralement le vide pointé par la question (dans la mesure, bien sûr, où l'interlocuteur connaît la réponse). On parle ainsi du couple Question/Réponse pour exprimer cette complétude. Dans le cadre de la tâche QR de TREC, les données textuelles contenues dans la collection de documents n'ont pas pour vocation d'être des réponses. La distance que doivent s'efforcer de pallier les mécanismes d'appariement est donc plus ou moins grande. Cette distance se manifeste notamment par les variantes. Pour la question What is the average body temperature?, la réponse attendue est *98.6 degrees Fahrenheit*. Curieusement, on obtient la réponse suivante : *82-degree water, 16.6 degrees below normal body t*, qui nous amène effectivement à la réponse attendue (82 + 16.6 = 98.6) sans toutefois préciser l'unité de mesure. Vraisemblablement cette réponse a été trouvée par hasard…

La question suivante What does a defibrillator do? a donné lieu au plus de variations de réponses au sein de ce petit corpus. En fait, cette question est ambiguë : veut-on connaître les effets immédiats d'un défibrillateur, ou le but de son utilisation à plus long terme? Variantes et validité des réponses sont liées, puisque la validité dépend en partie de l'expression de la réponse (forme) et du degré de granularité (fond) plus ou moins satisfaisant selon l'utilisateur. Comme réponses validées, on a :

- l'effet : *administering an electric shock ; administers a sharp electric shock to control…; delivers electric shocks to the heart to get it… ; high-voltage shock that should return the heart…*

- l'effet ET le but : *heart's normal rhythm through electric shock ; a shock powerful enough to restore normal heart ; to shock the failing heart back into normal rhythm*

- le but à court terme : *defibrillators used to stimulate the heart ; used to restore normal heartbeat ; restores the heart's normal rhythm ; to correct abnormal rhythms ; defibrillator can restore normal beating ; to help maintain a consistent heartbeat ; a device used to restart a stopped heart ; defibrillator to stop irregular heart beats ; electrically manage the rhythm of the heart*

- le but à long terme : *heart beats does, indeed, save and prolong lives*. Cette bribe n'a pas été validée. Peut-être parce que cette réponse a été considérée comme trop vague, trop générale.

Finalement, nous nous sommes basés sur l'évaluation de TREC et avons apporté des modifications lorsque nous l'avons jugé utile et pertinent, comme indiqué dans la figure 4. Pour un corpus de 4108 bribes de réponses, notre jugement diverge de celui de TREC pour seulement 60 réponses, ce qui représente près de 1,5 %. Ce taux est tout à fait négligeable, et l'on peut donc se fier aux résultats de l'évaluation proposée par TREC. Cependant, d'un point de vue épistémologique, l'observation des erreurs ou des divergences de notation est riche en enseignements ; elle nous a fait prendre conscience de la difficulté à trouver et à évaluer une réponse.

rg	%RV	No	évaluationTREC	1					-1			divergences
			Notre évaluation		-	IX	IC	AM		2	+	total
1	37%	1110	196	73					119		2	**2**
2	26%	1112	204	53			4		151			**4**
3	23%	1113	219	50			3		169		2	**5**
4	19%	1100	207	40			1		167			**1**
5	17%	1101	198	33			23		164			**23**
6	16%	1095	214	35			3		176		3	**6**
7	15%	1094	207	32	1				175			**1**
8	14%	1102	192	27	1		1		165			**2**
9	14%	1106	203	28				1	149		1	**2**
10	12%	1105	213	25			1		187	1	1	**3**
11	7%	1107	203	14			1		188			**1**
12	5%	1097	221	10					211			**0**
13	4%	1109	196	7	4				175			**4**
14	3%	1096	195	5				1	190			**1**
15	2%	1104	211	4					207			**0**
16	2%	1103	217	4	2				213			**2**
17	1%	1099	191	2		1			188			**1**
18	1%	1108	224	2					222			**0**
19	0%	1098	192	0					190	2		**2**
20	0%	1111	205	0					205			**0**
			4108		**8**	**1**	**37**	**2**		**3**	**9**	**60**

Figure 4. *Comparaison entre notre évaluation et celle de TREC. Les chiffres indiqués dans la quatrième colonne correspondent au nombre de réponses données pour chaque question ; ceux indiqués dans la cinquième colonne correspondent au nombre de réponses validées par les juges de TREC, ceux indiqués dans la dixième colonne correspondent au nombre de réponses non validées.* ***1-*** *mauvaise réponse validée ;* ***1 IX*** *réponse validée inexacte ;* ***1 IC*** *réponse validée incomplète ;* ***1 AM*** *réponse validée ambiguë ;* ***-1 2*** *la réponse est considérée comme mauvaise mais nous semble acceptable ;* ***-1 +*** *la réponse est bonne mais n'a pas été validée.*

6. Au cœur du système : l'appariement

6.1. Fonctionnement global des systèmes de QR

Ce n'est pas un hasard si la tâche QR a été promue par la TREC, originairement axée sur l'évaluation en recherche d'information. Le « *Question Answering* » (QA) s'appuie effectivement sur la recherche d'information. La plupart des systèmes concourrant à la TREC observent globalement la stratégie suivante : ils classent la question selon le type de réponse attendue, puis, après avoir transformé la question en requête, ils utilisent un moteur de recherche afin de sélectionner quelques documents parmi la collection fournie. Ces documents sont étiquetés afin de repérer les entités correspondant au type de la réponse attendue. Si l'entité repérée se trouve à proximité de mots-clefs, elle est retournée en guise de réponse. Sinon, le système s'appuie sur des techniques permettant de trouver le passage qui correspond le mieux à la requête. Un des problèmes soulevés par cette technique réside dans le fait que la réponse peut se trouver dans un document qui n'a rien à voir avec le sujet (*topic*) de la question posée. En outre, de nombreux systèmes utilisent WordNet pour étendre la requête, ou pour vérifier que l'entité extraite correspond bien au type de la réponse attendue. Depuis la TREC-10, certains systèmes vont même chercher des extraits hors collection, dans des encyclopédies ou sur le web, afin de générer la réponse. Cette réponse doit tout de même être appuyée par un document issu de la collection de TREC. Cette pratique n'est pas interdite mais elle a pour effet de bord la génération de réponses constituées de mots mis bout à bout afin d'augmenter les chances de fournir la « bonne » réponse. On ne peut alors parler de génération de réponse au sens propre et habituel du terme. Quoi qu'il en soit, globalement, les mots-clefs jouent un rôle capital : ils permettent de déterminer des extraits sur lesquels on va ensuite glisser une fenêtre s'arrêtant sur la portion susceptible de contenir la réponse.

6.2. Constat

Effectivement, la plupart du temps, même si la bribe de réponse ne contient pas de mot(s)-clef(s), le voisinage semble en contenir. Il semblerait que la sélection de la réponse se fasse par tâtonnements, en glissant une fenêtre de gauche à droite du mot-clef. Ce constat peut être fait en comparant différentes bribes de texte issues du même document. Le problème n'est-il pas, au fond, de localiser la réponse dans le document? A partir des bribes présentées dans la figure 3 p. 16, on peut reconstituer une bribe plus importante qui serait : *is part of sodium chloride, ordinary table salt, which manufacturers add to food to enhance*.

On constate aussi qu'un certain nombre de réponses non validées sont extraites du même document, mais - par malchance? - la précision du zoom n'a pas été aussi efficace : *Sodium is part of sodium chloride, ordinary tab* ; *Sodium is part of sodium chloride, ordinary table*.

Si on lance des requêtes sur Internet avec le moteur de recherche Google, on s'aperçoit que la plupart des questions posées telles quelles trouvent leur réponse dans les 3 premiers documents fournis par le moteur de recherche. Pour les questions 1098 et 1111 (auxquelles aucun candidat n'a su répondre), la réponse est respectivement dans le premier et troisième document fournis par le moteur de recherche. On peut alors soit supposer que la réponse n'était pas dans la collection de TREC, soit que la formulation de la réponse noyée dans la collection de documents fournie par TREC était très distante de celle de la question.

6.3. Hypothèse

Partons de l'hypothèse suivante : si pour chaque question, le rapport mots-clefs/réponse validée ET mots-clefs/réponse non validée est quasiment le même, alors quelle est la part de hasard dans le fait de trouver la « bonne » réponse, et quelle serait la part résultant au contraire d'efforts concertés (technique plus fine d'ajustement du zoom nécessitant un minimum de « compréhension ») ? Peut-on établir un lien entre les réponses auxquelles on a su répondre et la présence de mots-clefs dans les réponses? Et y-a-t-il un lien avec le nombre de documents contenant les mots-clefs? Bref, dans quelle mesure la présence de mots-clefs permet-elle d'obtenir de bons scores?

Dans le cadre d'une évaluation des systèmes de QR, nous avons besoin de vérifier si l'on peut *facilement* retrouver des réponses en utilisant les méthodes traditionnelles de la recherche d'information basée sur les mots-clefs. Si en utilisant les mots-clefs on a 50% de chances de trouver la bonne réponse, il faudra revoir l'évaluation en proposant plutôt des questions qui possèdent des réponses dans la collection de documents, mais qui ne sont pas formulées en reprenant ne serait-ce que partiellement les mots-clefs issus de la question. Notons qu'on ne sait pas quels sont les différents participants à l'origine des réponses proposées. Il est donc difficile d'associer à un type de réponse fournie un type de fonctionnement précis de système.

6.4. Méthodologie

Afin de confirmer ou d'infirmer les hypothèses précédentes, nous avons décidé de comptabiliser les mots-clefs dans les bribes de réponses valides et non valides. En

conclusion, nous verrons si l'appariement QR basé essentiellement sur la recherche de mots clefs est plutôt bruyant ou payant au niveau des résultats.

Pour chaque question nous n'avons sélectionné que les mots *non*-vides, et qui ont *effectivement* une correspondance dans les réponses valides. Nous avons fait une exception toutefois pour la question What is the ***most frequently*** spoken language in the Netherlands? car most et frequently ont toute leur importance ici. N'oublions pas que le problème essentiel du passage de la question à la requête, c'est qu'on perd le lien entre les mots, on se retrouve avec un sac de mots qui peuvent être mélangés dans n'importe quel ordre[11]. Le sens général a disparu. L'exemple précédent nous montre qu'il peut être particulièrement judicieux de ne pas supprimer certains mots-vides qui agissent comme des modalités, établissant par exemple ici une hiérarchie : ce qui nous intéresse, ce n'est pas de savoir quelles langues on parle aux Pays-Bas, mais quelle est la langue la plus parlée parmi toutes ces langues. Nous avons éliminé les « termes étiquettes » (en gras : What is the ***name*** of the satellite that the Soviet Union sent into space in 1957?) qui semblent superflus dans la mesure où *il valent pour* ce que nous cherchons. Nous avons aussi éliminé un des termes redondants le cas échéant (en gras : How ***fast*** is the speed of light ?). En ce qui concerne les termes composés, nous n'avons pas recherché les composants séparément mais bien ensemble. La méthodologie que nous avons suivie ici *manuellement* lors du choix et du traitement des mots-clefs a l'avantage de soulever certains points développés ultérieurement et de nous faire prendre conscience des écueils auxquels les systèmes *automatiques* peuvent être confrontés.

6.5. Résultats

Nous avons cherché à savoir quel était le pourcentage de mots-clefs au sein des réponses validées et non validées par les juges de TREC. En général on a pu retenir un, deux voire trois mots-clefs pour effectuer nos calculs. Nous n'avons pas traité la question 1094 compte tenu du nombre trop élevé de mots-clefs, et de la présence d'un mot-clef composé.

Il est difficile de dire à partir de ces données s'il existe une corrélation entre le nombre de bonnes réponses trouvées et la correspondance de mots clefs. On notera que le rang des questions n'est pas systématiquement lié au total des correspondances de mots-clefs : les réponses valides de la question 1110, 1er rang, contiennent 62% de correspondances au total, tandis que les réponses valides de la question 1112, 2ème rang,

[11]. Certains systèmes, tel celui d'IBM, utilisent néanmoins un analyseur de dépendance syntaxique [ITT 01].

contiennent 100% de correspondances au total, comme la question 1108, classée 18ème![12]. Le contenu du corpus n'est pas étranger aux résultats : comment expliquer le fait que deux questions semblables (What is fungus? 18ème rang ; What is influenza? 3ème rang) se positionnent à des rangs très différents, si ce n'est à cause du contenu du corpus ? Il faut donc aussi tenir compte des documents fournis (collection entière de documents mise à la disposition des concurrents par TREC) ; des documents exploités pour fournir une réponse, quelle que soit la nature de la réponse (validée ou non) ; des documents desquels on a puisé les réponses valides. La présence ou l'absence de mots-clefs dans tous ces documents rend nos résultats relatifs. Afin de mesurer le degré de difficultés, il nous faudrait savoir dans quelle mesure on retrouve les mots-clefs dans les documents (pour une question donnée, quelle est la proportion de documents qui contiennent les mots-clefs, quelle est la proportion de documents qui contiennent la réponse).

L'analyse met en évidence le fait que le choix de réponses qui s'avèrent non valides peut s'expliquer par la sélection de mots-clefs non pertinents. On ne retrouve pas ces termes dans les réponses valides. Par exemple, pour la question How fast is the speed of light?, *fast* est retrouvé dans 15% des réponses non valides (et jamais dans les réponses valides). Ce « mauvais » choix peut aussi s'expliquer par la recherche séparée de composants de certains mots composés. Par exemple, pour la question 1098, *melting point* a été retrouvé dans 10% des réponses non valides, *melting* seul dans 13% des réponses non valides, *point* seul dans 22% des réponses non valides. Pour la question 1109, dans les réponses non valides, *language* apparaît 45 fois, soit dans 25% des cas, tandis que *spoken* apparaît 17 fois, *most* 13 fois, et *frequently* 3 fois. Peut-on en déduire que dans la plupart des cas, même si les réponses ne sont pas valides, l'analyse syntaxique a été efficace puisqu'elle a repéré language comme tête du syntagme nominal the most frequently spoken language ? ou est-ce plutôt parce que le terme *language* était plus présent dans les textes?

Aucun mot-clef n'a été trouvé dans les réponses validées des questions 1096 (5 réponses valides, 14ème rang), 1103 (4 réponses valides, 16ème rang) et 1099 (2 réponses valides, 17ème rang). Ce type de réponses doit néanmoins avoir été trouvé grâce à une recherche basée sur des mots-clefs, comme nous l'avons vu en 6.1. En se penchant sur les documents (y compris leurs balises HTML) dont sont extraites ces réponses, on remarque en effet qu'il y a occurrence ou cooccurrence de mots-clefs, et que dans certains cas la bribe extraite se trouve à proximité d'un ou plusieurs mots-clefs. Dans le document FBIS4-67349, de nombreuses bribes auraient pu être relevées et acceptées comme réponses à la question What is the effect of acid rain ? en fonction du degré

[12]. Les réponses non valides des questions 1110, 1112 et 1108 contiennent respectivement 64, 72 et 51 % de mots-clefs.

d'exhaustivité et de granularité souhaité (en fait, l'article porte justement sur les effets des pluies acides, et détaille ces effets points par points).

Comme nous l'avons vu (6.1.), la plupart des systèmes s'appuient sur les résultats des moteurs de recherche et donc sur les mots-clefs pour sélectionner des documents et trouver la réponse. L'étude que nous avons menée sur la fréquence des mots-clefs présents dans les réponses valides ET non valides tend à prouver que la présence de mots-clefs n'est pas un facteur garantissant la validité de la réponse. Les résultats se répartissent en effet plus ou moins équitablement : le nombre de réponses valides contenant des mots-clefs n'est pas nettement différent voire supérieur au nombre de réponses non valides contenant aussi des mots-clefs. De plus, certaines réponses valides ne contiennent pas de mots-clefs. On pourrait donc en déduire que l'ajustement du zoom sur la réponse à partir de la localisation des mots-clefs n'est pas au point dans de nombreux cas ; les réponses proposées relèvent de techniques d'ajustement tantôt fines, tantôt approximatives, qui nous font penser que le hasard n'est pas étranger à certains résultats. D'autre part, les systèmes ne possèdent pas suffisamment d'indicateurs leur permettant d'évaluer les réponses qu'ils trouvent avant de les soumettre.

La présence de mots-clefs ne permet donc pas nécessairement d'obtenir de bons scores. En fait, le mécanisme d'appariement basé sur la recherche de mots-clefs est critiquable. En effet, nombre de réponses peuvent se trouver dans des documents ne contenant pas de mots-clefs, ou contenant un nombre très restreint de mots-clefs ce qui fait que ces documents ne seront pas retenus (en fonction par exemple des calculs *tf-idf* pour le modèle vectoriel[13]). Les réponses peuvent aussi se trouver dans des documents qui contiennent des mots-clefs, mais pas nécessairement à proximité de la réponse. De toute façon, au cours de l'appariement - et c'est là essentiellement que réside la difficulté - on ne devrait pas chercher à apparier des éléments (question et réponse) en fonction de leur degré de similitude, puisque question et réponse ne se ressemblent pas mais se complètent[14]. Finalement, les concurrents devraient essayer de perfectionner leur système en vue de proposer une meilleure analyse et compréhension de la question, afin de mieux cibler le focus, essentiel lors de la phase d'appariement. On pourrait envisager aussi de minimiser le rôle des mots-clefs durant cette phase, en préférant des patrons illustrant des formulations de réponses possibles.

---

[13]. Les documents sont représentés par des vecteurs correspondant aux mots. Les mesures utilisées sont : *tf* (*term frequency*) : calcule le nombre d'occurrences d'un mot dans un document ; *df* (*document frequency*) : calcule le nombre de documents quoi contiennent un terme donné ; *idf* (*inverse document frequency*).

[14]. Les systèmes de QR utilisant des bases de FAQs essaient de déjouer cette difficulté en appariant des éléments plus ou moins semblables : les questions d'utilisateurs et les questions préenregistrées des FAQs.

7. Conclusion

En conclusion, si l'on veut proposer une compétition qui stimule réellement les participants, fasse avancer la recherche dans le domaine du traitement automatique de la langue et aboutisse à des applications susceptibles de satisfaire la plupart des utilisateurs, on a tout intérêt à augmenter et cibler ses exigences. La question de fond est en fait la suivante : doit-on évaluer les systèmes en fonction de « vraies » questions d'utilisateurs (dont le focus est difficile à déterminer) ou bien en fonction de questions sur mesure, correspondant *grosso-modo* à l'état de l'art des systèmes (basés essentiellement sur la recherche et l'extraction d'information de type « entité nommée »)? En tout cas, le passage d'une réponse de 50 caractères à *la* réponse exacte lors de la TREC-11 empêchera les concurrents de générer la réponse en concaténant une suite de mots puisés çà et là aboutissant généralement à une bribe de réponse ubuesque. Cette mesure a pour avantage de laisser de côté cette technique qui peut fournir de bons résultats mais n'aide pas à une meilleure compréhension des phénomènes et ne fait pas avancer la recherche.

Il nous semble que TREC pourrait évaluer les systèmes en fonction d'autres critères que les réponses proposées. D'une part parce qu'évaluer un système à partir de ses sorties paraît trop restrictif, et d'autre part parce qu'évaluer un seul des éléments alors qu'ils sont tous plus ou moins interdépendants est difficile[15]. La compétition pourrait être basée sur un ensemble de questions plus variées, plus proches des questions d'utilisateurs et comprenant des difficultés linguistiques graduelles pour le TAL. Cependant, nous avons vu que le degré de difficulté des questions posées n'est pas absolu, il dépend aussi de la manière dont les systèmes fonctionnent et du contenu de la base de données. Comme il semble que le degré de difficulté soit essentiellement lié à la distance QR lors de l'appariement, on pourrait attribuer des pondérations à chaque question en fonction de la présence ou non d'une réponse dans la collection de documents, de la formulation de la réponse et du nombre de réponses recensées dans les documents. Ces dispositions pourraient permettre de mieux cerner et donc évaluer la distance à parcourir pour réaliser l'appariement. Elles permettraient aussi d'établir une première échelle de difficultés entre les questions. Puisque certains systèmes font déjà appel aux données du Web pour trouver une réponse valide dans la collection de documents fournie par TREC (entre autres, celui de Microsoft, [BRI 01]), pourquoi ne pas directement se servir du Web comme corpus? On se rapprocherait ainsi des vraies conditions d'utilisation, avec une base plus importante et plus hétérogène. La nature et la taille du corpus doivent en effet être pris en compte lors de l'évaluation. Cependant, il deviendrait alors difficile de savoir si la réponse est présente ou non dans le corpus, et sous quelle forme. On pourrait aussi inciter vivement les participants à utiliser le même

[15]. Évaluer le système en tenant compte de tous les éléments est néanmoins aussi difficile, et soulève d'autres problèmes…

moteur de recherche. En effet, nous avons vu que la piste du QR est directement issue de la recherche d'information et s'appuie sur les résultats fournis par les moteurs de recherche pour trouver la réponse, sans que le moteur de recherche, qui joue pourtant un rôle important, soit lui-même évalué.

Nous pouvons d'ores et déjà répondre au moins partiellement à un certain nombre de questions présentes de façon latente durant toute cette étude. Qu'est-ce qui permet dans le contexte immédiat de la réponse de la valider ou de l'invalider? *A priori*, et de manière triviale, on peut dire que les réponses contenant des termes tronqués au début ou à la fin de la bribe de 50 caractères peuvent être éliminées. Le zoom doit être réajusté de façon adéquate. Cette réponse laisse néanmoins en suspens le problème de la sélection de la bribe. En l'absence de mots-clefs, comment les réponses valides ont-elles été trouvées? Bien souvent, des mots-clefs se trouvent à proximité de la bribe sélectionnée, ou à tout le moins dans le document. Comment expliquer les erreurs, les réponses non-valides? Quelques-unes s'expliquent par un mauvais choix de mots-clefs ou par la recherche séparée de composants de composés. Il reste néanmoins un certain nombre de questions sans réponses, auxquelles nous tâcherons de répondre dans une étude ultérieure : Qu'est-ce que les mauvaises réponses avec ou sans mots-clefs contiennent? Qu'est-ce qui distingue efficacement les bonnes réponses avec mots-clefs des mauvaises réponses? Y-a-t-il un sens à comparer les réponses valides / non valides ne comprenant pas de mots-clefs? Comment améliorer l'appariement question/réponse de façon systématique (par exemple en utilisant des règles de réécriture ou des phrases à trous) afin de réduire le fossé entre des formulations trop différentes? Peut-on calculer la « distance » entre question et réponse, et en tirer des règles? Sous quelle forme se présentent les réponses qui n'ont pas été trouvées, notamment parce qu'elles ne contiennent pas ou ne sont pas à proximité de mots-clefs ? Quelles sont les bonnes réponses *non* reconnues, qui sont passées sous silence? Comment sont-elles formulées? Est-ce leur formulation qui explique le fait qu'on ne les a pas retrouvées? Nous avons vu que dans certains cas la transformation de la question en requête gomme le lien entre les mots et altère le sens global de la question. La typologie des questions peut-elle y remédier systématiquement, en proposant pour chaque type ou sous-type de question une représentation syntaxico-sémantique de la question? Un problème pratique subsiste néanmoins concernant ce type d'étude portant sur la distance dans l'appariement : comment retrouver les réponses qui sont restées dans l'ombre, dans la jungle de la collection de documents utilisée par TREC…

Remerciements

Ce travail a été réalisé lors d'une collaboration entre Karine Lavenus et le laboratoire RALI. Il a été rendu possible grâce à la participation financière de Bell Canada à travers son programme de support à la R-D des Laboratoires universitaires Bell, et grâce au soutien financier de l'A.N.R.T. (Association Nationale de la Recherche Technique) et de LexiQuest.

8. Bibliographie

[ALP 01] Alpha S. et alii, "Oracle at TREC-10 : Filtering and Question Answering", Gaithersburg, *TREC-10*, novembre 2001, p. 419-428.

[BRI 01] Brill E. et alii, "Data-Intensive Question Answering" , Gaithersburg, *TREC-10*, novembre 2001, p. 183-189.

[BUR 00] Burger J. et alii, "Issues, tasks ans program structures to roadmap research in question answering", Octobre 2000.

[HAR 01] Harabagiu S. et alii, "Answering complex, list and context questions with LCC's Question-Answering Server", Gaithersburg, *TREC-10*, novembre 2001.

[HER 01] Hermjakob U., "Parsing and question classification for question answering", Toulouse, *ACL*, juillet 2001.

[HOV 01] Hovy E. et alii, "The Use of external knowledge in factoid QA", Gaithersburg, *TREC-10*, novembre 2001, p. 166-174.

[ITT 01] Ittycheriah A. et alii, "IBM's Statistical Question Answering System", Gaithersburg, *TREC-10*, novembre 2001, p. 317-323.

[KEY 96] Keyes J.G., "Using conceptual categories of questions to measure differences in retrieval performance", Baltimore, *ASIS*, octobre 1996.

[LEH 78] Lehnert W., *The process of question answering : a computer simulation of cognition*, New York, Erlbaum Associates, 1978.

[SOU 01] Soubbotin M.M., "Patterns of potential expressions as clues to the right answers", Gaithersburg, *TREC-10*, novembre 2001, p. 175-182.

[VOO 00] Voorhees E., Tice D., "The TREC-8 question answering track evaluation", Gaithersburg, TREC-8, novembre 1999.

[VOO 01]Voorhees E., "Overview of the TREC-9 question answering track", Gaithersburg, *TREC-10*, novembre 2001, P.71-79.