Université de Montréal

# Le système de question-réponse QUANTUM

par
Luc Plamondon

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de M.Sc.
en informatique

Mars 2002



Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

# Le système de question-réponse QUANTUM

présenté par
Luc Plamondon

a été évalué par un jury composé des personnes suivantes :

Philippe Langlais
président-rapporteur

Guy Lapalme
directeur de recherche

Leila Kosseim
codirectrice

Gilles Brassard
membre du jury

Mémoire accepté le 29 avril 2002

**Résumé**

QUANTUM est un système de question-réponse qui prend en entrée une question posée en langage naturel, cherche sa réponse dans un corpus d'environ un million de documents et retourne 5 suggestions de réponse d'au plus 50 caractères. Pour ce faire, QUANTUM utilise à la fois des techniques de recherche d'information et d'analyse linguistique. L'analyse de la question se fait à l'aide d'expressions régulières portant sur les mots et sur leur étiquette grammaticale ; elle permet de choisir la fonction d'extraction à appliquer pour extraire les candidats-réponses. La fonction est appliquée non pas sur l'ensemble des documents mais plutôt sur de courts passages préalablement sélectionnés à l'aide du moteur de recherche Okapi. Pour extraire les candidats-réponses, les fonctions d'extraction font usage d'expressions régulières, de l'extracteur d'entités nommées Alembic et du réseau sémantique WordNet. De plus, QUANTUM a la capacité de détecter l'absence de réponse satisfaisante dans le corpus. Les standards d'évaluation de la conférence TREC-X ont permis de conclure que le module d'analyse des questions est efficace à 90 %, qu'Okapi est un outil approprié pour effectuer un premier filtrage du corpus de documents, qu'Alembic et WordNet s'avèrent utiles lors de l'extraction des réponses mais qu'ils ne permettent pas à eux seuls d'atteindre une performance globale satisfaisante, et que le module de détection d'absence de réponse échoue dans 90 % des cas. Nous présentons aussi $XR^3$, un système de question-réponse précédemment développé à l'Université de Montréal.

Mots-clés : question-réponse, informatique, linguistique, traitement des langues naturelles, recherche d'information, réponse automatique aux courriels, intelligence artificielle

**Abstract**

QUANTUM is a question-answering system that takes a natural language question as its input, looks for the answer in a 1-million document collection, and outputs five answer suggestions of up to 50 characters each. To achieve this goal, QUANTUM uses both information retrieval techniques and linguistic analysis. The analysis of a question is performed using regular expressions on words and on their grammatical tags so that the appropriate extraction function is chosen for the extraction of candidate answers. This extraction function is then applied to short passages retrieved by the Okapi search engine rather than on the entire document collection. Depending on the chosen function, candidates are extracted using regular expressions, the Alembic named entity extractor and/or the WordNet semantic network. QUANTUM also has the ability to detect that no suitable answer can be found in the document collection. We used TREC-X conference evaluation standards to evaluate the system. We found that the question analysis module is 90% effective, that Okapi should be retained as a pre-filtering tool, that Alembic and WordNet are useful for answer extraction but that they are not sufficient to achieve a satisfactory overall performance, and that the no-answer detection module has a 90% failure rate. We also describe $XR^3$, a question-answering system previously designed at the University of Montreal.

Keywords: question answering, computer science, linguistics, natural language processing, information retrieval, automatic e-mail reply, artificial intelligence

# TABLE DES MATIÈRES

# TABLE DES FIGURES

# LISTE DES TABLEAUX

# Remerciements

L'auteur désire remercier Guy Lapalme, sans qui ce projet n'aurait jamais vu le jour, et Leila Kosseim, dont l'expertise technique a été essentielle. Merci à Sylvain Laganière pour son aide concernant Okapi et à Massimo Fasciano pour son aide concernant Alembic.

Ce projet a été rendu possible par le concours financier des Laboratoires Universitaires Bell (LUB) et du Conseil de Recherches en Sciences Naturelles et en Génie du Canada (CRSNG). L'auteur bénéficie d'une bourse d'études du Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (FCAR) du Québec.

# CHAPITRE 1

# Introduction

La taille qu'a atteint le web à ce jour rend incontournables les moteurs de recherche pour qui veut trouver ce qu'il cherche. Des efforts considérables ont été investis dans les moteurs de sorte qu'ils ont non seulement réussi à gérer l'augmentation du volume de pages à traiter, ils ont aussi atteint une rapidité de réponse surprenante et les hyperliens qu'ils retournent sont de plus en plus pertinents. Cependant, l'efficacité technique des outils ne garantit pas à elle seule le succès d'une recherche car des facteurs humains entrent également en jeu, comme l'aptitude de l'internaute à composer une requête efficace et sa patience lorsque vient le temps d'explorer quelques-unes des pages proposées pour y dénicher la réponse à son interrogation initiale.

La qualité toujours plus grande des résultats proposés par les moteurs ravit la communauté des internautes mais lui fait oublier que toute cette mécanique n'est pas la plus intuitive. Pourquoi s'astreindre à un langage de requête alors que la question pourrait être formulée en toutes lettres, telle que l'internaute se la pose ? Pourquoi lire des pages en entier à la recherche d'une réponse que le moteur pourrait localiser pour lui ? C'est ici que le domaine de la recherche d'information devient celui de la question-réponse.

Ainsi peut-on imaginer un système de question-réponse qui permettrait d'interroger en langage naturel n'importe quelle base de données, qu'elle soit composée de textes médicaux ou d'archives de journaux, voire même de toutes les pages du web. Si la base de données était l'ensemble du site web d'une entreprise d'envergure, la question-réponse pourrait s'avérer être un constituant important d'un système de réponse automatique au courriel. Par exemple, il a été estimé par [Kosseim et Lapalme, 2001] qu'entre 20 et 40 % des courriels reçus par le service à la clientèle de Bell Canada Enterprises (BCE) sont des questions dont la réponse se trouve quelque part sur le site web de l'entreprise. En utilisant un système de question-réponse pour ce type de courriel, BCE diminuerait le volume de courriels à traiter manuellement et, par conséquent, le délai de réponse moyen au client.

C'est dans l'optique de pouvoir répondre au nombre grandissant de courriels reçus par BCE, et cela en des temps représentatifs de la nature quasi-instantanée du courrier électronique, que les Laboratoires Universitaires Bell et le Laboratoire de Recherche Appliquée en Linguistique Informatique (RALI) de l'Université de Montréal ont mis sur pied le projet MERKURE. Le RALI a opté pour une solution composite, étant donné la variété des courriels à traiter. Trois approches complémentaires ont été proposées : la classification pour rediriger certains courriels vers des filiales de BCE, la question-réponse pour répondre aux courriels courts dont la réponse se trouve sur le site web corporatif et le raisonnement par cas pour répondre aux courriels nécessitant une réponse complexe.

Le système de question-réponse QUANTUM décrit dans cet ouvrage trouve sa motivation première dans le projet MERKURE ; cependant, la version actuelle n'a pas été développée pour répondre spécifiquement aux courriels reçus par BCE. Elle correspond plutôt aux spécifications de la conférence TREC-X car cette conférence fournit un cadre d'expérimentation et d'évaluation rigoureux. Ce cadre comporte notamment des exigences précises pour les systèmes de question-réponse à développer, un corpus de recherche standard de près d'un million de documents, une méthode uniforme d'évaluation des systèmes, l'accès aux travaux de toute une communauté de chercheurs dédiée à cette problématique et la possibilité d'échanger avec eux.

Le système de question-réponse QUANTUM constitue le point principal du présent ouvrage. Son fonctionnement et une analyse de ses performances obtenues à TREC-X sont présentés au chapitre 4. Le chapitre 3 est quant à lui une description sommaire de $XR^3$, le premier système de question-réponse élaboré par le RALI ; $XR^3$ est indépendant de QUANTUM mais son examen a permis de tirer quelques leçons avant de concevoir QUANTUM. Mais d'abord, le chapitre 2 décrit le cadre applicatif dans lequel les deux systèmes ont été conçus, c'est-à-dire les conférences TREC.

# CHAPITRE 2

## Les conférences TREC

Le National Institute of Standards and Technology (NIST), un organisme gouvernemental américain, tient chaque année une conférence nommée TREC (pour Text REtrieval Conference) afin de stimuler la recherche dans le domaine de la recherche d'information. Cette large problématique est divisée en plusieurs pistes dont une, la *Question Answering Track*, s'intéresse aux systèmes de question-réponse. C'est à TREC-8 (1999), lors de l'introduction de cette piste, que la recherche sur la question-réponse prend véritablement son essor. En 2000, divers spécialistes de cette communauté nouvellement créée publient le *Vision Statement to Guide Research in Question & Answering and Text Summarization* [Carbonell *et al.*, 2000] : ce rapport fait le point sur les systèmes de question-réponse mais surtout, il explore la façon dont ces systèmes serviront l'*Intelligence Community* gouvernementale américaine (constituée entre autres du FBI, de la CIA et du Department of Defense). La suite de ce rapport, *Issues, Tasks and Program Structures to Roadmap Research in Question & Answering* [Burger *et al.*, 2001], est un cadre de travail plus précis qui oriente la recherche et pose des objectifs graduels pour les années 2001 à 2005.

La piste pourrait être qualifiée de compétition amicale. D'abord, les chercheurs du monde entier sont invités à concevoir un système de question-réponse. Quelques mois avant la tenue de la conférence, une liste de courtes questions et un corpus d'environ 1 million de documents sont distribués aux équipes de chercheurs inscrites. Ces dernières mettent alors à l'épreuve leur système et toute intervention humaine qui pourrait influencer les résultats est interdite. Les réponses trouvées par les systèmes sont soumises aux responsables de TREC pour qu'ils les évaluent d'une façon standard, de sorte que les systèmes puissent être comparés de façon juste. Lors de la tenue de la conférence annuelle, les participants mettent en commun leurs travaux par la publication d'un article décrivant leur système.

Le premier système décrit dans cet ouvrage, $XR^3$, a été développé dans le but de prendre part à la neuvième édition de la conférence TREC (TREC-9) ; le second système, QUANTUM, a pris part

| | |
|---|---|
| TREC-8 | *Who was the first American in Space ? (21)*<br>*When did Nixon visit China ? (24)*<br>*Why are electric cars less efficient in the north-east than in California ? (159)* |
| TREC-9 | *How many states have a "lemon law" for new automobiles ? (486)*<br>*Can you give me the name of a clock maker in London, England ? (577)*<br>*CPR is the abbreviation for what ? (783)* |
| TREC-X | *What is caffeine ? (920)*<br>*In Poland, where do most people live ? (1130)*<br>*What is the conversion rate between dollars and pounds ? (1221)* |

FIG. 2.1: Quelques questions de TREC-8, TREC-9 et TREC-X.

à la dixième édition (TREC-X).

## 2.1 Type de questions

Les questions sont courtes et demandent des réponses factuelles : la figure 2.1 en montre quelques exemples. Pour identifier les questions tirées des corpus de questions de TREC, nous indiquons leur numéro officiel entre parenthèses ; les questions de TREC-8 portent les numéros 1 à 200, celles de TREC-9 les numéros 201 à 893 et celles de TREC-X les numéros 894 à 1393.

Les responsables de TREC assurent que les questions admissibles ont une réponse "factuelle" sans toutefois définir ce concept. Nous nous contenterons de supposer qu'il s'agit de questions pour lesquelles il n'est pas nécessaire de faire appel à des capacités de jugement pour produire la réponse, ni de la déduire à partir de plusieurs faits énoncés dans le corpus. Ces hypothèses sont discutables mais nous croyons qu'il n'est pas nécessaire de chercher à définir de façon exacte quelles questions sont acceptables dans le cadre des conférences TREC puisque le but ultime est d'élaborer des systèmes de question-réponse qui puissent répondre à toute question.

## 2.2 Calcul du score

Pour chacune des questions, le système cherche dans le corpus de documents des extraits d'une longueur déterminée (50 ou 250 caractères, selon l'édition de TREC) susceptibles de répondre à la question (fig. 2.2). En fait, ces *extraits* peuvent être des chaînes générées par le système ; l'important est que la réponse exacte se trouve dans un des documents du corpus et qu'elle soit appuyée par le numéro du document source. Le système peut soumettre jusqu'à 5 suggestions de réponse et il doit les ordonner de la plus plausible (rang 1) à la moins plausible (rang 5).

```
Question :  Where are the British crown jewels kept?

Réponse :   1  FT921-14782    are kept in Edinburgh Castle - together with jewel
            2  AP901114-0171  kept in the Tower of London as part of the British
            3  AP900620-0160  treasures in Britain's crown jewels. He gave the K
            4  NIL
            5  AP900610-0018  the crown jewel settings were kept during the war.
```

FIG. 2.2: Question 1326 (TREC-X) et exemple de réponse donnée par QUANTUM. Chacune des 5 suggestions inclut un rang, le numéro du document source et une chaîne de 50 ou 250 caractères, selon l'édition de TREC. Ici, la réponse correcte est trouvée dans la suggestion au deuxième rang. La suggestion *NIL* signifie que QUANTUM suggère que le corpus de documents ne contient pas de réponse acceptable (cette éventualité ne s'applique pas à TREC-8 et TREC-9).

Lors de l'évaluation, le score obtenu pour chaque question est calculé comme suit : si la réponse correcte ne se trouve pas parmi les 5 suggestions, le score est 0 ; par contre, si une des 5 suggestions contient la réponse correcte, le score est égal à la réciproque du rang de cette suggestion. Ainsi, le score est de 1, $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$ ou $\frac{1}{5}$ selon que la réponse correcte est trouvée dans la suggestion au rang 1, 2, 3, 4 ou 5. Le corpus de documents peut contenir plusieurs variantes de la réponse correcte ; le verdict d'équivalence est subjectif et doit évidemment être posé par des évaluateurs humains. La moyenne des scores obtenus pour chaque question constitue le score final appelé *Mean Reciprocal Rank* (MRR).

Depuis TREC-9, une réponse doit être supportée par le document duquel elle provient, c'est-à-dire que le contexte dans lequel elle est citée doit être similaire au contexte visé par la question. Ainsi, à la question *What is Pittsburgh's baseball team called ? (844)*, la chaîne de caractères *Pirates* extraite d'un document traitant de baseball serait jugée supportée, alors que la même chaîne extraite d'un document traitant de sécurité informatique ne le serait pas. Les évaluateurs procèdent à deux types d'évaluation qui tiennent compte ou non de cette contrainte : une évaluation stricte (*strict*) qui n'accorde aucun point à une réponse correcte mais non supportée, et une évaluation tolérante (*lenient*) qui ne vérifie pas si la réponse est supportée.

## 2.3 Corpus de documents

Le corpus de documents dans lesquels les réponses doivent être puisées est constitué de près d'un million de documents totalisant environ 3 gigaoctets de données textuelles. Les documents sont des dépêches provenant de 6 agences de presse : Associated Press, Financial Times, Foreign Broadcast Information Service, Los Angeles Times, San Jose Mercury News et Wall Street Journal. Les documents sont en format SGML mais les annotations diffèrent selon l'agence.

*What tourist attractions are there in Reims ? (411)*
*What are the names of the tourist attractions in Reims ? (711)*
*What do most tourists visit in Reims ? (712)*
*What attracts tourists to Reims ? (713)*
*What are tourist attractions in Reims ? (714)*
*What could I see in Reims ? (715)*
*What is worth seeing in Reims ? (716)*
*What can one see in Reims ? (717)*

FIG. 2.3: Reformulations de la question 411 (TREC-9).

## 2.4 TREC-8 (1999)

Les 200 questions de TREC-8 sont créées par les responsables de TREC, ce qui n'en fait pas nécessairement des questions représentatives de celles posées par les utilisateurs potentiels. Chaque suggestion de réponse peut contenir jusqu'à 250 caractères. Il existe, dans le corpus de documents, au moins une réponse correcte pour chaque question. L'Université de Montréal n'a pas développé de système de question-réponse pour cette édition de TREC.

## 2.5 TREC-9 (2000)

Le nombre de questions passe à 693 : 500 sont des questions posées par le public sur des sites web et 153 sont des reformulations créées de toutes pièces. La figure 2.3 montre un exemple de reformulations d'une même question. Le nombre de questions sensiblement plus élevé qu'à TREC-8 permet de couvrir un plus large éventail de types de questions, dont les définitions. Comme pour TREC-8, il existe au moins une réponse correcte dans un des documents du corpus. Par contre, une contrainte s'ajoute : la réponse doit être supportée par le document duquel elle provient. Les équipes sont invitées à participer à deux sous-pistes : une dont les réponses sont d'au plus 250 caractères et une dont les réponses sont d'au plus 50 caractères (les deux sous-pistes utilisent la même série de questions). On trouvera une description plus complète de ces sous-pistes dans l'exposé de présentation de la conférence [Voorhees et Harman, 2000b]. L'Université de Montréal a participé à cette édition de TREC avec le système $XR^3$ décrit au chapitre 3.

## 2.6 TREC-X (2001)

Le corpus de questions est formé de 500 questions provenant de sites web. Il n'y a pas de reformulations. Les réponses doivent être supportées par le document dans lequel elles se trouvent, tout comme pour TREC-9. Certaines questions peuvent ne pas avoir de réponse dans le corpus de

documents et le système doit reconnaître cette éventualité. La sous-piste des réponses de 250 caractères est abandonnée pour ne conserver que celle de 50 caractères. Une autre sous-piste vient s'ajouter : les questions ayant pour réponse une liste d'éléments. Pour plus de détails, se référer à la présentation de la conférence [Voorhees et Harman, 2001b] et au chapitre 4.

## 2.7 Les conférences futures

Les exigences seront haussées d'année en année en accord avec les échelons posés par [Burger *et al.*, 2001]. Les objectifs sont de produire d'ici 2005 un système capable de répondre à des questions telles que *What are the opinions of the Danes on the Euro ?* ou *How likely is it that the Fed will raise the interest rates at their next meeting ?*, questions qui exigent de résoudre des ambiguïtés lexicales et sémantiques, d'établir des relations entre des faits, de résumer des faits provenant de domaines hétérogènes, de compléter ce résumé avec des informations précises et de générer une réponse articulée.

# CHAPITRE 3

# XR$^3$

XR$^3$ (pour *eXtraction de Réponses Rapide et Robuste*) est le premier système de question-réponse développé à l'Université de Montréal. Ce système lui a permis de participer à la conférence TREC-9 et, par conséquent, d'être évalué de façon comparable à d'autres systèmes. XR$^3$ se distingue de beaucoup de ces systèmes par l'extraction de passages intermédiaires de longueur fixe et par l'utilisation presque exclusive d'expressions régulières. La conception de QUANTUM s'inspire en partie de XR$^3$, c'est pourquoi nous allons en décrire les principaux modules et analyser ses performances obtenues avec les questions de TREC-9.

## 3.1 Architecture

On retrouve dans [Laszlo, 2000] une explication détaillée du fonctionnement de XR$^3$ et dans [Laszlo *et al.*, 2000], une description du système une fois adapté pour TREC-9. Les questions et le corpus de recherche sont les seules entrées du système ; le reste de la procédure est entièrement automatisé et ne requiert pas d'intervention manuelle. La question est d'abord analysée pour en extraire des termes-clés et d'autres informations utiles. Les termes-clés sont ensuite utilisés pour chercher dans le corpus les passages de 250 caractères les plus prometteurs. Puis, des candidats-réponses sont extraits de ces passages. Enfin, les candidats sont étendus à 50 caractères et ils sont filtrés afin de ne conserver que les 5 meilleurs tout en évitant les redondances.

### 3.1.1 Analyse de la question

#### Sélection de termes-clés

Des termes-clés sont extraits de la question en vue de former la requête qui sera utilisée pour la recherche de passages dans le corpus de documents. Pour ce faire, la question est analysée par

| But | TREC-8 (%) | TREC-9 (%) | Total (%) |
|---|---|---|---|
| PNOUN | 60 | 56 | 57 |
| UNKNOWN | 7 | 22 | 18 |
| CARDINAL | 12 | 9 | 10 |
| TIMEPOINT | 12 | 8 | 9 |
| TIMESPAN | 2 | 3 | 2 |
| LINEAR | 3 | 1 | 1 |
| MONEY | 4 | 1 | 1 |
| REASON | 1 | 1 | 1 |
| AREA | 1 | 0 | 0 |
| PERCENTAGE | 1 | 0 | 0 |
| MEANS | 0 | 0 | 0 |
| VOLUME | 0 | 0 | 0 |
| MASS | 0 | 0 | 0 |

TAB. 3.1: Proportion des questions de chaque but, tel qu'analysé par $XR^3$ (le taux d'erreur n'est pas disponible).

un étiqueteur grammatical développé au RALI. Chaque nom commun ou nom propre constitue un terme-clé. Les groupes nominaux composés de la plus longue suite possible d'adjectifs, de participes et de noms précédant un nom (tels *world energy output* et *managing director*) forment une deuxième série de termes-clés. Les verbes, adverbes, prépositions et autres ne sont pas retenus. La requête est formée des termes-clés tels quels, c'est-à-dire qu'aucun lemme ni variante morpho-syntaxique ou sémantique n'est inclus ; cette décision s'appuie sur les travaux de [Clarke *et al.*, 2000] qui ont démontré que l'ajout de variantes dans une requête à un stade aussi précoce est nuisible.

### Identification du but

Le but d'une question est défini comme étant le type de réponse attendue, c'est-à-dire un nom propre, une date, une quantité, etc. Pour être un candidat intéressant, un passage doit contenir au moins une expression du même type que le but de la question. Un ensemble d'expressions régulières sert à identifier le but de la question, alors qu'un deuxième ensemble permet d'identifier dans les documents les expressions correspondant à ce but. Le tableau 3.1 énumère les buts reconnus par $XR^3$ ; ces buts ont été choisis en fonction de leur facilité à être identifiés à l'aide d'expressions régulières et non en fonction de critères sémantiques. En cas d'incertitude lors de l'analyse de la question, le but PNOUN (nom propre) est retenu par défaut car il est le plus fréquent du corpus de test (les questions de TREC-8).

### Identification de la date

Lorsque la question contient une date, les passages contenant cette date ou publiés à cette date sont jugés plus pertinents que les autres. À cet effet, $XR^3$ extrait la ou les année(s) présente(s) dans

la question.

### 3.1.2 Sélection de passages de 250 caractères

Un premier filtrage du corpus de documents est effectué à l'aide d'un moteur de recherche conventionnel avec pour requête la question dans son intégralité. Cette opération n'est pas effectuée par $XR^3$ ; dans le cadre de la conférence TREC-9, une liste des 1000 meilleurs documents retournés par un moteur d'AT&T pour chacune des 893 questions a été distribuée aux participants. Malheureusement, nous ne disposons pas de mesure de l'efficacité de ce moteur.

Pour une question donnée, $XR^3$ parcourt les documents retournés par le moteur de recherche, à la recherche des termes-clés identifiés lors de l'analyse de la question. Chaque fenêtre de texte longue de 250 caractères ayant en son centre un terme-clé constitue un passage susceptible de contenir la réponse puisqu'il est supposé que cette dernière apparaît dans le voisinage de termes présents aussi dans la question. Un passage qui en superpose un précédent par plus de 125 caractères (c'est-à-dire 0,5 fois la taille des passages, où 0,5 est appelé *facteur de recouvrement*) est considéré redondant et est de ce fait éliminé.

Afin d'ordonner les passages restants, un score leur est attribué. Ce score, appelé *score IR*, est le produit de 3 scores :

$$\textit{score IR} = \textit{score de termes-clés} \times \textit{score de but} \times \textit{score de date} \tag{3.1}$$

Le *score de termes-clés* dépend du nombre et du poids des termes de la requête qui sont présents dans le passage. La présence d'un terme donne 1 point ; le poids des termes débutant par une majuscule est surpondéré par un facteur *capBonus* et le poids des termes complexes (les groupes nominaux) est surpondéré par un facteur *groupBonus*. Les occurrences multiples d'un terme-clé dans un même passage sont ignorées. De façon formelle, étant donné une suite $K_1 \ldots K_n$ de termes-clés :

$$\textit{score de termes-clés} = \sum_{i=1}^{n} f(K_i) c_i g_i \tag{3.2}$$

où $c_i = \textit{capBonus}$ si $K_i$ débute par une majuscule ($c_i = 1$ sinon), $g_i = \textit{groupBonus}$ si le terme-clé $K_i$ est un groupe nominal ($g_i = 1$ sinon) et $f(K_i) = 1$ si $K_i$ est présent dans le passage (0 sinon).

Le *score de but* est le score alloué lorsque le passage contient au moins une expression qui concorde avec le but de la question. Le *score de date* est alloué lorsqu'une question comporte une année et que le passage la contient aussi, ou lorsque le document a été publié cette année-là.

Les passages sont réordonnés selon leur score IR. Lors de la conférence TREC-9, les 5 meilleurs passages pour chacune des questions ont été soumis à la sous-piste des réponses de 250 caractères. Cependant, afin de participer aussi à la sous-piste des réponses de 50 caractères, des traitements supplémentaires sont nécessaires.

### 3.1.3 Extraction des réponses exactes

Pour réduire la taille d'un passage de 250 à 50 caractères, $XR^3$ doit identifier les *réponses exactes* qu'il contient, c'est-à-dire les chaînes de caractères les plus courtes possibles qui sont susceptibles de constituer la réponse à la question (à titre d'indication, les réponses exactes ont en moyenne une dizaine de caractères). Ensuite, $XR^3$ doit évaluer la qualité de ces réponses exactes, les ordonner et les augmenter de sorte qu'elles aient 50 caractères.

Pour expliquer la technique employée par $XR^3$, ses concepteurs définissent le *focus* comme étant une portion de la question qui doit obligatoirement figurer près du candidat-réponse, que ce soit verbatim ou bien sous une forme alternative. Par exemple, le focus de la question *What was the monetary value of the Nobel Peace Prize in 1989 ? (2)* serait *Nobel Peace Prize* car l'hypothèse est faite que la réponse correcte devrait se trouver à proximité de l'expression *Nobel Peace Prize* ou d'une expression sémantiquement apparentée.

Ainsi, chaque passage de 250 caractères obtenu précédemment est scruté pour y cerner les occurrences du focus de la question et les candidats-réponses (c'est-à-dire les expressions de même nature que le but de la question). Toutes les paires focus-candidat sont analysées une à une. Pour chacune des combinaisons, le passage est schématisé de la façon 1 si le candidat est à gauche du focus et de la façon 2 si le candidat est à droite :

1. [*pre, answer, post (inter), focus, right*]
2. [*left, focus, pre (inter), answer, post*]

Le terme *answer* désigne un candidat-réponse ; les termes *pre* et *post* font référence à la portion du passage apparaissant avant et après le candidat ; et le terme *inter* désigne la portion située entre le candidat et le focus, qui est en fait la portion *pre* ou *post* selon l'ordre d'apparition du candidat et du focus. Le tableau 3.2 illustre cette schématisation avec des exemples.

Le score assigné à chacune des combinaisons focus-candidat, en d'autres mots le score final d'un candidat, est une combinaison linéaire de 3 scores :

$$score\ final = score\ IR + score\ de\ lien + score\ extra \tag{3.3}$$

Le *score IR* (éq. 3.1) est le score obtenu par le passage de 250 caractères duquel provient la combinaison.

Le *score de lien* implique un nouveau concept : la relation entre la réponse et le focus. Lors de l'analyse de la question, $XR^3$ détermine, à l'aide d'expressions régulières, quelle relation existe entre le focus de la question et la réponse cherchée (voir le tableau 3.3 pour la liste des 6 types de liens reconnus). Si la portion du passage située entre un candidat et le focus (la portion *inter* de la schématisation) exprime la relation attendue, le score de lien sera élevé.

Le *score extra* regroupe des critères *ad hoc* tels la présence de certaines unités de mesure lorsque la réponse doit être une quantité et la présence de mots de la question autres que ceux formant le focus (pour amoindrir les effets d'une mauvaise identification du focus de la question).

| Question | Élément | Contenu |
|---|---|---|
| *Which team won the Super Bowl in 1968 ? (116)* | *pre* | terback, Namath, one of the first since Babe Ruth to make a fortune playing a game. With Namath as their leader, the AFL's 1968 |
| | *answer* | New York Jets |
| | *post* | went into |
| | *focus* | Super Bowl |
| | *right* | III as an 18-point underdog and won, 16-7, against the NFL champion Baltimore Colts, wh |
| *How many people live in the Falklands ? (101)* | *left* | everything. If they bought someone out, where would they go ? Mr Di Tella denied that payments would be made to encourage people to leave the |
| | *focus* | Falkland Islands |
| | *pre* | and settle elsewhere. He said, 'We want to be very respectful of these' |
| | *answer* | 2,000 |
| | *post* | people. They have liv |

TAB. 3.2: Schématisation d'un passage. Dans le premier cas, le candidat se trouve avant le focus ; dans le deuxième cas, il est après.

| Lien | Exemple |
|---|---|
| EXISTENCE | *Who is the Queen of Holland ? (136)* |
| ATTRIBUTE | *How tall is the Matterhorn ? (161)* |
| TIME | *When was Yemen reunified ? (130)* |
| LOCATION | *Where is Inoco based ? (20)* |
| REASON | *Why are electric cars less efficient in the north-east than in California ? (159)* |
| MEANS | *How did Socrates die ? (198)* |

TAB. 3.3: Liens reconnus par $XR^3$. C'est lors de l'analyse de la question que $XR^3$ tente d'établir quel lien existe entre le focus de la question et la réponse cherchée.

### 3.1.4 Expansion des réponses à 50 caractères et filtrage

Les réponses exactes extraites par $XR^3$ ont en moyenne 10 caractères et elles peuvent donc être soumises à la sous-piste des réponses de 50 caractères de TREC-9. Cependant, les chaînes de caractères sont augmentées jusqu'à ce qu'elles atteignent la limite permise de 50 caractères, ceci dans le but de maximiser les chances de succès. Pour ce faire, $XR^3$ prélève autant de caractères à la gauche qu'à la droite de la réponse exacte dans le texte source.

Suivant la même méthode utilisée pour soumettre une série de réponses de 250 caractères, les réponses de 50 caractères sont ordonnées selon leur score et les redondances sont éliminées. Les 5 meilleures réponses pour chacune des questions forment la série soumise à la sous-piste des réponses de 50 caractères de TREC-9.

| Série | Longueur moy. des rép. (car.) | Évaluation stricte | | Évaluation tolérante | |
|---|---|---|---|---|---|
| | | MRR | Taux d'échec (%) | MRR | Taux d'échec (%) |
| UdeMlng1 | 249,86 | 0,352 | 50,1 | 0,359 | 48,8 |
| UdeMlng2 | 249,96 | 0,366 | 47,7 | 0,380 | 46,2 |
| UdeMshrt | 49,81 | 0,179 | 71,3 | 0,187 | 69,8 |
| UdeMexct | 9,46 | 0,149 | 77,7 | 0,159 | 75,8 |

TAB. 3.4: Résultats de $XR^3$ à TREC-9. *MRR* est le score (sect. 2.2) et le *taux d'échec* est la proportion des questions auxquelles $XR^3$ n'a pu répondre correctement, sur un total de 682 questions (le corpus en contenait initialement 693 mais 11 ont été retirées officiellement).

## 3.2 Analyse des performances de $XR^3$ à TREC-9

Quatre séries de réponses produites par $XR^3$ ont été soumises à TREC-9 pour évaluation : deux séries de réponses longues de 250 caractères (`UdeMlng1` et `UdeMlng2`), une série de réponses courtes de 50 caractères (`UdeMshrt`) et une série de réponses exactes (`UdeMexct`) d'une longueur moyenne de 9,46 caractères. Les résultats sont illustrés au tableau 3.4 et ils sont comparés aux performances des autres systèmes aux figures 3.1 et 3.2.

### 3.2.1 Construction des 4 séries

La figure 3.3 permet de visualiser quelles opérations ont conduit à la production des séries.

L'utilisation du moteur de recherche d'AT&T (avec pour requête la question intégrale) réduit le corpus de près d'un million de documents à seulement 1000 documents. Les 10 meilleurs passages de 250 caractères sont extraits en utilisant leur score IR (éq. 3.1), c'est-à-dire une combinaison du score de termes-clés, du score de but et du score de date. Les 5 meilleurs passages non redondants pour chacune des questions forment la série `UdeMlng1`, appelée *série longue 1.*

$XR^3$ procède ensuite à l'extraction des paires focus-candidat contenues dans les 10 meilleurs passages de 250 caractères. Les candidats sont ordonnés selon un score composé du score IR, du score de lien et du score extra (éq. 3.3). Les 5 meilleurs candidats pour chacune des questions forment la série `UdeMexct`, appelée *série exacte* puisque les réponses ont une taille minimale, soit environ 10 caractères.

Étant donné qu'il est permis à TREC-9 de soumettre des réponses d'au plus 50 caractères, les réponses exactes sont augmentées autant vers la gauche que vers la droite pour atteindre la taille maximale permise (sect. 3.1.4). Par conséquent, les 5 meilleures réponses de 50 caractères peuvent contenir plus de candidats que les 5 de la série exacte. Ces réponses de 50 caractères forment la série `UdeMshrt`, ou *série courte.*

La sélection des réponses longues de 250 caractères n'a pas bénéficié de l'analyse poussée faite lors de l'extraction des réponses exactes et courtes, notamment en ce qui a trait au score de lien et au score extra. Afin de mettre à profit ces précisions supplémentaires lors de la soumission d'une deuxième série de réponses longues, les 10 meilleurs passages ayant conduit à la série `UdeMlng1`

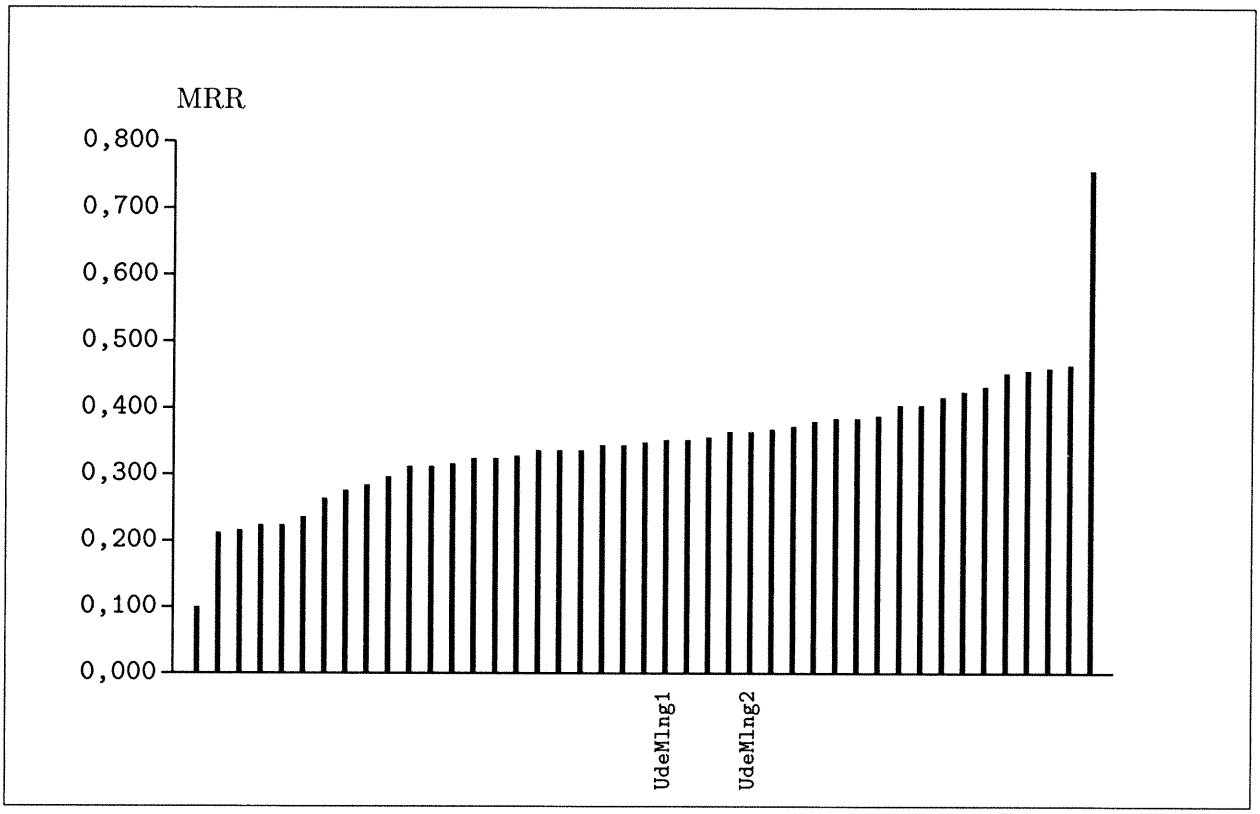


FIG. 3.1: MRR des séries de réponses de 250 caractères soumises à TREC-9 par l'ensemble des systèmes (évaluation stricte). Chaque barre représente une série de réponses.

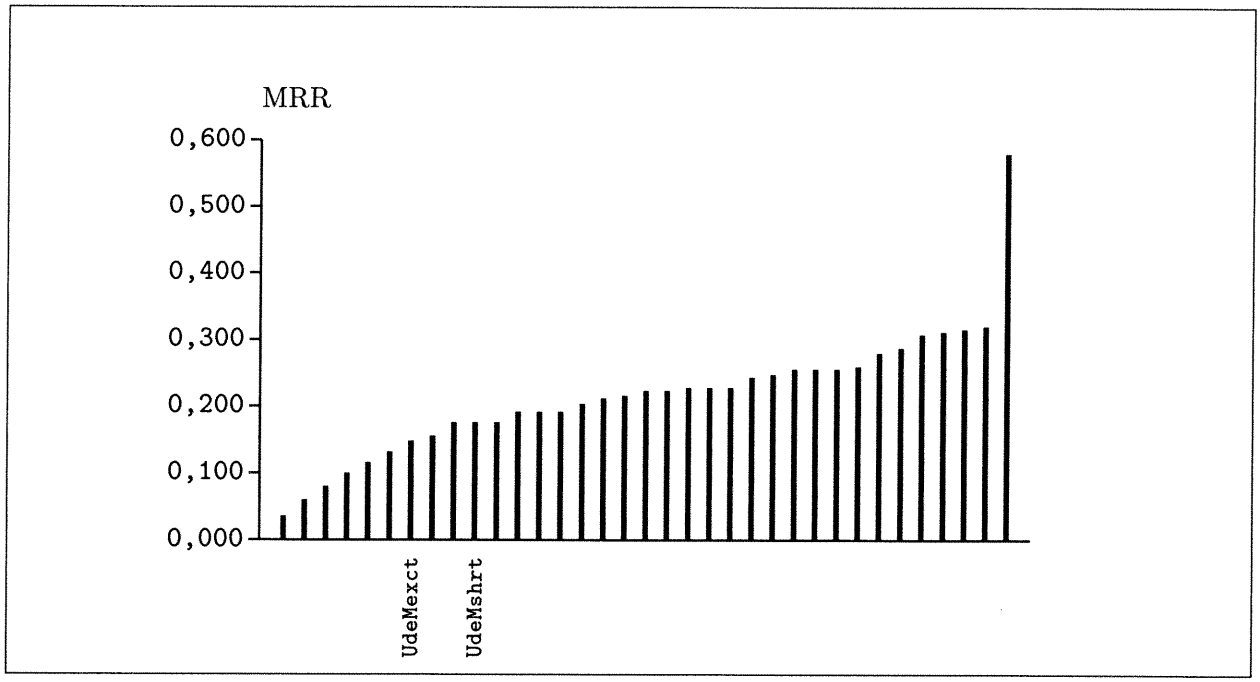


FIG. 3.2: MRR des séries de réponses de 50 caractères soumises à TREC-9 par l'ensemble des systèmes (évaluation stricte). Chaque barre représente une série de réponses.

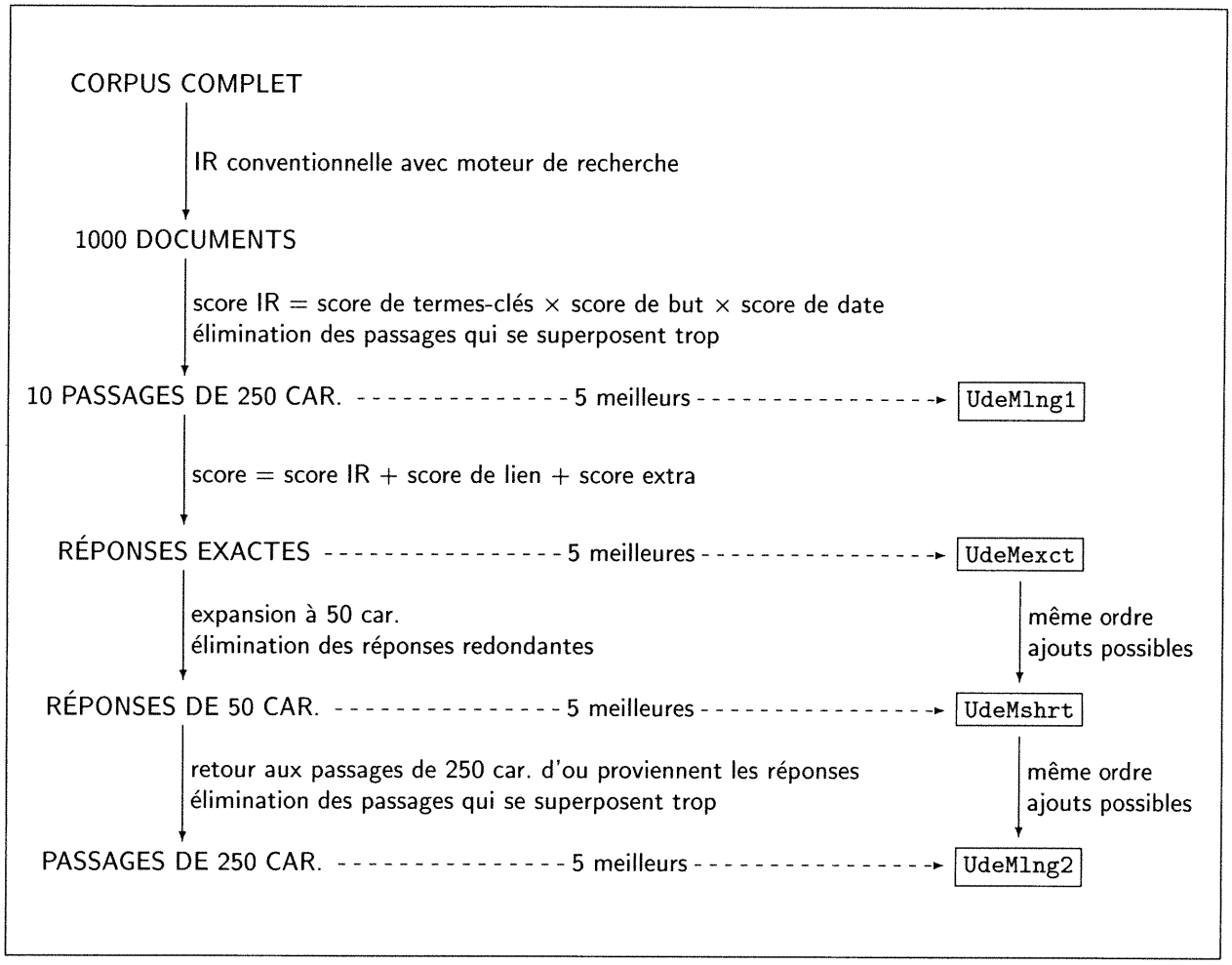


FIG. 3.3: Construction des 4 séries de réponses soumises à TREC-9 par XR$^3$.

sont réordonnés selon le score des réponses exactes qu'ils contiennent ; les 5 meilleurs forment la série `UdeMlng2`, ou *série longue 2*. Notons que les 5 meilleurs passages ainsi choisis contiennent forcément les 5 meilleures réponses exactes et qu'il est possible que les passages soumis aux séries `UdeMlng1` et `UdeMlng2` ne soient pas les mêmes.

### 3.2.2 Analyse

La série longue 1 (MRR = 0,352, évaluation stricte) utilise le score de termes-clés, le score de but et le score de date. La série longue 2 (MRR = 0,366) utilise ces 3 scores, en plus du score de lien et du score extra. Nous constatons que cet ajout améliore le MRR de 0,014. Il est cependant important de noter que seuls les 10 meilleurs passages de la série longue 1 sont utilisés pour produire la série longue 2 ; des essais avec un nombre plus élevé de passages ont entraîné une diminution des performances, montrant ainsi que le module d'identification des réponses exactes est sensible à la qualité des passages qui lui sont soumis en entrée [Laszlo *et al.*, 2000].

La comparaison de la série exacte (MRR = 0,149) avec la série courte (MRR = 0,179) montre que la chance a apporté une amélioration du MRR de 0,03 lors de l'expansion des réponses exactes en réponses de 50 caractères. Nous pouvons aussi en tirer que dans seulement 3 % des cas (14 questions sur 530), la réponse correcte se trouvait dans un rayon de 25 caractères autour d'une des réponses exactes suggérées par $XR^3$.

Nous observons une dégradation marquée de la performance entre les réponses de 250 caractères et les réponses de 50 caractères (MRR = 0,352 pour la série longue 1, contre 0,179 pour la série courte), ce qui met en évidence la difficulté avec laquelle $XR^3$ extrait les réponses exactes. Un examen plus détaillé des résultats par question montre que dans 24 % des cas (166 questions sur 682), la réponse correcte a été "perdue" lors du passage de 250 à 50 caractères ; dans 11 % des cas (76 questions), la réponse correcte a été placée à un rang supérieur (moins bon) ; dans 9 % des cas (63 questions), la réponse exacte a été soumise au même rang ; dans 6 % des cas (37 questions), la réponse exacte a été placée à un rang inférieur (meilleur) ; dans 3 % des cas (22 questions), une réponse correcte a été "gagnée" par le passage de 250 à 50 caractères (rappelons que les chaînes de 50 caractères jugées les meilleures ne proviennent pas nécessairement des passages de 250 caractères jugés les meilleurs) ; et dans 47 % des cas (320 questions), la réponse exacte est demeurée introuvable.

L'extraction des réponses exactes ou courtes pose un dilemme : d'une part, le nombre et la taille des passages à examiner doivent être petits car bien que $XR^3$ puisse identifier des candidats, il lui est difficile de choisir les meilleurs ; et d'autre part, le taux de précision (nombre de questions auxquelles $XR^3$ trouve la réponse correcte divisé par le nombre total de questions) avec ces petits passages est faible (de l'ordre de 50 %). À compter de TREC-X, la sous-piste des réponses de 250 caractères est abandonnée au profit de la sous-piste des réponses de 50 caractères. Il est tentant de ne plus passer par l'extraction de passages intermédiaires de 250 caractères et d'extraire les réponses exactes directement des textes complets, ceci afin de viser une précision de 100 %. Mais puisqu'il est si difficile de séparer les candidats intéressants des candidats farfelus, l'amélioration des techniques d'extraction ne sera pas suffisante et il faudra dans un futur proche composer avec les passages intermédiaires, quitte à mettre des efforts pour en améliorer le taux de précision tout en diminuant leur taille et leur nombre.

# CHAPITRE 4

# QUANTUM

QUANTUM (pour *QUestion ANswering Technology of the University of Montréal*) est le système de question-réponse que nous avons développé en vue de participer à la conférence TREC-X. Il a été conçu pour trouver des réponses de 50 caractères à des questions courtes, factuelles et syntaxiquement bien formées. Il procède en 4 étapes : analyse de la question, recherche de passages, extraction de candidats et expansion à 50 caractères. Ces étapes ne sont guère différentes de celles suivies par $XR^3$ (sect. 3.1) ; par contre, la façon dont elles sont accomplies est tout autre.

D'abord, nous avons entièrement revu la classification des questions et nous avons intégré des outils d'analyse morpho-syntaxique au module d'analyse des questions afin de tirer parti des étiquettes grammaticales et de l'identification des groupes nominaux (sect. 4.1). Ensuite, pour l'étape de recherche de passages, nous fournissons deux possibilités à l'utilisateur : un mécanisme de recherche de passages de longueur fixe similaire à celui de $XR^3$ ou le moteur de recherche Okapi (sect. 4.2). Puis, l'extraction des candidats-réponses qui se faisait uniquement à l'aide d'expressions régulières dans $XR^3$ est accomplie dans QUANTUM avec le concours des outils externes WordNet et Alembic (sect. 4.3). Pour les besoins spécifiques de TREC-X, un module de traitement des questions sans réponse a été ajouté (sect. 4.5). Nous faisons suivre la description des 5 modules de QUANTUM par une évaluation de chacun d'eux (sect. 4.6) et nous concluons en situant notre système par rapport à d'autres systèmes de question-réponse (sect. 4.7).

## 4.1 Analyse de la question

L'analyse d'une question vise d'abord à la classer, ceci afin de déterminer quelles fonctions d'extraction présentées au tableau 4.2 appliquer pour extraire les candidats-réponses. Une fois la fonction choisie, il est parfois nécessaire de poursuivre l'analyse pour identifier le *focus* de la question, c'est-à-dire le groupe nominal servant à paramétrer la fonction d'extraction ; nous précisons

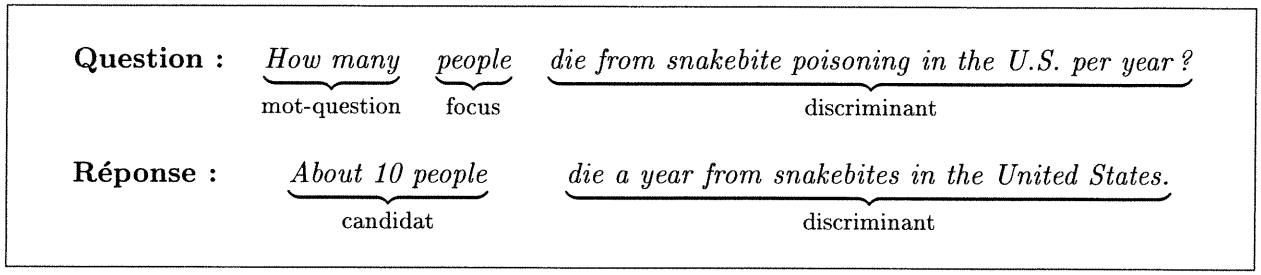


FIG. 4.1: Décomposition de la question 302 (TREC-9) et de sa réponse.

cette notion de focus à la section 4.1.1. C'est à l'aide d'environ 40 expressions régulières portant non seulement sur les mots, mais aussi sur leur étiquette grammaticale et sur la présence de groupes nominaux que QUANTUM effectue son analyse ; nous présentons à la section 4.1.2 les outils d'analyse morpho-syntaxique nécessaires à ce traitement. Enfin, à la section 4.1.3, nous justifions la classification que nous avons retenue.

### 4.1.1 Composantes des questions et des réponses

Pour décrire le fonctionnement de QUANTUM, nous avons besoin de désigner les différentes composantes d'une question et d'une réponse de façon non équivoque. Cependant, à notre connaissance, le domaine de la question-réponse ne dispose pas encore d'un vocabulaire standard ; c'est pourquoi nous définissons ici certains des termes que nous employons. Nous ne prétendons pas établir de définition formelle : nous désirons plutôt orienter l'intuition du lecteur pour assurer la compréhension de la suite de cet ouvrage.

Soit le couple de question-réponse de la figure 4.1. La question se divise en trois parties : le mot-question (*how many*), le focus (*people*) et le discriminant (*die from snakebite poisoning in the U.S. per year*). La réponse a deux parties : le candidat (*about 10 people*) et le discriminant (*die a year from snakebites in the United States*).

#### Mot-question

Le mot-question est le plus souvent *what, when, where, when, who, why* ou *how*, sans toutefois s'y limiter. Des locutions peuvent aussi être considérées comme mots-questions : *how many* et *how much* sont les plus courantes. Le mot-question détermine en partie quel type de réponse est attendu (en partie, car des questions formulées à l'aide d'un même mot-question peuvent requérir des types de réponse différents ; voir à ce sujet l'annexe B). Dans l'exemple de la figure 4.1, *how many* indique que la réponse cherchée est un nombre. Toutes les phrases du corpus de recherche qui ne contiennent pas de nombre peuvent d'ores et déjà être éliminées (au risque de perdre certains candidats qui seraient de la forme *most of, as many as*, etc.).

**Focus**

Le focus est le nom ou le groupe nominal qui permet de préciser les mécanismes génériques de recherche de réponse utilisés par QUANTUM. L'identification du focus se fait après l'analyse de la structure de la question et après le choix d'un mécanisme de recherche. Le choix du focus dépend des besoins du mécanisme de recherche ; parfois, il n'est pas nécessaire d'identifier un focus si le mécanisme de recherche ne le requiert pas. Dans notre exemple, la question comporte les mots *how many* suivis d'un groupe nominal, ce qui conduit QUANTUM à choisir un mécanisme de recherche de cardinalité et à désigner ensuite le groupe nominal *people* comme étant le focus ; le mécanisme générique choisi consiste à rechercher les candidats formés d'un nombre suivi du focus, ce qui s'instancie par NOMBRE *people* dans ce cas précis. Bien souvent, la réponse correcte ne contient pas le focus tel qu'il apparaît dans la question, mais plutôt une variante sémantique (*humans* et *Americans* seraient des variantes vraisemblables).

**Discriminant**

Le discriminant est la portion restante de la question une fois le mot-question et le focus enlevés. Alors que le focus sert à filtrer les candidats-réponses qui n'ont pas la forme voulue, le discriminant apporte les précisions nécessaires de sorte qu'un seul des candidats retenus puisse être la réponse correcte (en supposant que le corpus de documents ne contienne pas d'énoncés contradictoires). Toujours selon notre exemple, parmi tous les candidats de la forme NOMBRE *people*, seuls ceux qui proviennent de phrases contenant aussi le discriminant *die from snakebite poisoning in the U.S. per year* sont retenus. Cet exemple est évidemment naïf : il est rare que le discriminant ait la même forme dans la question et dans le corpus de documents. Les informations du discriminant peuvent être déplacées (*In the U.S., 10 people die from snakebite poisoning per year*), formulées différemment (*10 Americans die from snakebite poisoning per year*), ou encore être réparties dans des phrases différentes, voire même dans tout le document.

**Candidat**

Le candidat est un groupe de mots présent dans l'un des documents du corpus et susceptible d'être la réponse correcte. Les questions de TREC étant factuelles, les réponses correctes sont rarement plus longues qu'un groupe nominal ou un groupe prépositionnel. La forme du candidat dépend du type de question, d'où la nécessité de classifier les questions lors de leur analyse. Dans notre exemple, toutes les séquences NOMBRE FOCUS sont des candidats car ils sont susceptibles de répondre à une question de type *How many* FOCUS.

### 4.1.2 Outils d'analyse morpho-syntaxique

Pour les besoins de QUANTUM, l'analyse syntaxique des questions et des documents ne va pas au-delà de l'identification des groupes nominaux. Les outils utilisés pour y parvenir ont tous été

développés au RALI sans relation avec QUANTUM.

Le tokeniseur est utilisé pour segmenter les questions et les documents du corpus en *tokens*. Ensuite, l'étiqueteur attribue une catégorie grammaticale à chacun des *tokens*. Cet étiqueteur s'appuie sur un modèle statistique et il est à noter qu'il a été entraîné sur un corpus contenant très peu de questions (en l'occurrence les débats de la Chambre des Communes du Canada). Or, en anglais, les phrases interrogatives présentent des particularités syntaxiques, dont la plus courante est l'inversion du sujet et de l'auxiliaire, ce qui peut dérouter l'étiqueteur. Il n'a pas été possible de le réentraîner à temps pour TREC-X. Nous examinons ses performances à la section 4.6.1.

L'étiquetage des questions est requis par QUANTUM car leur analyse se fait à l'aide d'expressions régulières utilisant parfois les étiquettes des mots. De plus, ces étiquettes sont les données entrantes de l'extracteur de groupes nominaux. L'extraction des groupes nominaux constitue un premier filtrage des passages car les mots ou groupes de mots admissibles pour être des candidats-réponses sont presque toujours des groupes nominaux. Pour les besoins de QUANTUM, un groupe nominal peut être aussi complexe que *these three fourth-generation, extensively tested and efficient pain relievers*. La tête du groupe est généralement un nom commun ou un nom propre (ici, *relievers*). Dans de plus rares cas, la tête peut être :

- un quantificateur (*the Fifteen* [*members of the European Union*]),
- une lettre (*5 g* pour *5 grams*),
- un postfixe (*am* et *o'clock* dans *2 am, 8 o'clock*).

La tête peut être précédée d'une combinaison des éléments suivants :

- des déterminants (*these* dans l'exemple ci-haut),
- des quantificateurs (*three*),
- des adjectifs ordinaux (*fourth*),
- des noms (*generation, pain*),
- des adverbes suivis de participes passés ou suivis d'adjectifs (*extensively tested*),
- des adjectifs qualificatifs (*efficient*).

L'extracteur de groupes nominaux utilisé se distingue par la facilité avec laquelle il est possible de modifier les règles qui définissent un groupe nominal. Cependant, nous ne disposons pas de mesure de son efficacité.

### 4.1.3 Classification des questions

Classer une question fragmente le problème complexe de la recherche d'une réponse. En effet, le choix de la classe détermine le mécanisme de recherche à appliquer pour identifier la réponse dans le corpus de documents. Nous examinons d'abord quelques classifications nous ayant orienté vers celle que nous proposons ensuite.

| Classe | Exemple | TREC |
|---|---|---|
| Comparaison | *What is the difference between AM radio stations and FM radio stations ? (1165)* | oui |
| Complétion de concept | *Who invented the instant Polaroid camera ? (1284)* | |
| Définition | *What is dianetics ? (1160)* | |
| Exemple | *Name a food high in zinc. (1268)* | |
| Caractérisation | *What color is a poison arrow frog ? (1004)* | |
| Quantification | *How far is it from Denver to Aspen ? (894)* | |
| Cause | *Why is the sun yellow ? (1220)* | |
| Conséquence | *What is the effect of acid rain ? (1103)* | |
| But | *What was the purpose of the Manhattan project ? (541)* | |
| *Enablement* | *How do you measure earthquakes ? (996)* | |
| Instrument/procédure | *How did Socrates die ? (198)* | |
| Expectative | *Why can't ostriches fly ? (315)* | |
| Requête/Directive | *Tell me what city the Kentucky Horse Park is near ? (403)* | |
| Vérification | *Dit it rain yesterday ?* | non |
| Disjonction | *Did he order chicken, beef, lamb or fish ?* | |
| Interprétation | *Does the graph show a main effect for "A" ?* | |
| Jugement | *What do you think about the new taxes ?* | |
| Assertion | *I need to know how to get to the Newark airport.* | |

TAB. 4.1: Classification des questions selon Graesser *et al.* avec exemples tirés des conférences TREC. Nous avons isolé les classes de questions qui, à notre avis, n'ont pas été et ne seront pas, dans un avenir proche, visées par les conférences TREC. L'inclusion d'une classe dans le groupe des classes pertinentes à TREC n'implique pas que toutes les formes de questions de cette classe soient appropriées à TREC. Voir l'annexe A pour la description des classes et les exemples originaux donnés par Graesser *et al.*

### Quelques classifications

Dès 1978, [Lehnert, 1978] analyse le phénomène du questionnement, ce qui lui permet d'établir une liste finie de classes conceptuelles décrivant toutes les formes de questions. Plus tard, [Graesser *et al.*, 1992] confirment cette classification à l'aide de principes psycholinguistiques et l'augmentent. Le résultat final est reproduit au tableau 4.1. Nous y regroupons les classes qui nous semblent pertinentes dans le cadre de TREC.

Le contexte applicatif de TREC conduit les participants à se forger leur propre classification selon les techniques de traitement informatique de la langue dont ils disposent. Les classifications proposées sont presque aussi nombreuses que les systèmes. À titre d'exemple, nous exposons à la figure 4.2 l'ontologie des *réponses attendues* de [Harabagiu *et al.*, 2000], les concepteurs du système qui a le mieux performé à TREC-9.

Bien qu'elle soit intéressante, nous n'avons pas utilisé la classification de Harabagiu *et al.*, et ce pour plusieurs raisons. D'abord, elle s'appuie sur la capacité de reconnaître les entités nommées listées à la figure 4.3 alors que l'extracteur d'entités nommées dont nous disposons ne permet pas d'atteindre ce niveau de détails (sect. 4.3.4). De plus, la classification nous semble trop dépendante des questions provenant des conférences TREC précédentes. En effet, la classe RACE DE CHIEN a

**Catégories racines**

| | | | |
|---|---|---|---|
| DATE | TEMPS | ORGANISATION | RAISON |
| MANIÈRE | NATIONALITÉ | PRODUIT | ARGENT |
| LANGUE | MAMMIFÈRE | JEU | RACE DE CHIEN |
| LIEU | REPTILE | VALEUR NUMÉRIQUE | CITATION |
| ALPHABET | POURCENTAGE | | |

**Catégories feuilles**

| | | | |
|---|---|---|---|
| VILLE | PAYS | POURCENTAGE | COMPTE |
| MONTANT | TEMPÉRATURE | VITESSE | TAUX |
| DURÉE | DEGRÉ | PROVINCE | DIMENSION |
| UNIVERSITÉ | CONTINENT | AUTRE LIEU | |

**Exemple d'une hiérarchie racine-feuilles**

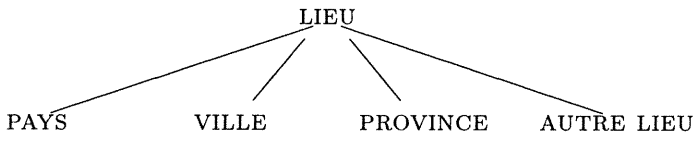


FIG. 4.2: Ontologie présentée par Harabagiu *et al.* à TREC-9.

| | | | |
|---|---|---|---|
| DATE | TEMPS | ORGANISATION | VILLE |
| PRODUIT | PRIX | PAYS | ARGENT |
| HUMAIN | MALADIE | NO DE TÉLÉPHONE | CONTINENT |
| POURCENT | PROVINCE | AUTRE LIEU | PLANTE |
| MAMMIFÈRE | ALPHABET | CODE D'AÉROPORT | JEU |
| OISEAU | REPTILE | UNIVERSITÉ | RACE DE CHIEN |
| NOMBRE | QUANTITÉ | ATTRACTION | |

FIG. 4.3: Entités nommées reconnues par l'extracteur de Harabagiu *et al.*

| Fonction | Exemple |
|---|---|
| *définition*($\rho$, $\varphi$) | *What is an* ***atom*** *? (897)* |
| *spécialisation*($\rho$, $\varphi$) | *What* ***metal*** *has the highest melting point ? (910)* |
| *cardinalité*($\rho$, $\varphi$) | *How many* ***Great Lakes*** *are there ? (933)* |
| *mesure*($\rho$, $\varphi$) | *How much* ***fiber*** *should you have per day ? (932)* |
| *attribut*($\rho$, $\varphi$) | *How* ***far*** *is it from Denver to Aspen ? (894)* |
| *personne*($\rho$) | *Who was the first woman to fly across the Pacific Ocean ? (907)* |
| *temps*($\rho$) | *When did Hawaii become a state ? (898)* |
| *lieu*($\rho$) | *Where is John Wayne airport ? (922)* |
| *manière*($\rho$) | *How do you measure earthquakes ? (996)* |
| *raison*($\rho$) | *Why does the moon turn orange ? (902)* |
| *objet*($\rho$) | Fonction par défaut lorsque aucune des précédentes ne s'applique. |

TAB. 4.2: Classification des questions selon 11 fonctions d'extraction. Lorsque la fonction requiert l'identification d'un focus $\varphi$, ce dernier est en gras dans la question fournie en exemple.

vraisemblablement été introduite à cause de la question *What breed of dog was the "Little Rascals" dog ? (532)* et elle nous apparaît hyper-spécialisée en comparaison avec les autres classes (REPTILE, JEU, PRODUIT, etc.).

Plutôt que de concentrer nos efforts sur l'écriture des expressions régulières nécessaires à l'implantation de la classification de Harabagiu *et al.* dans QUANTUM, ou encore de continuer dans la voie de $XR^3$ et augmenter la liste des expressions déjà existantes, nous avons privilégié les mécanismes génériques exposés à la section suivante.

### Classification proposée

Nous proposons de classer les questions de façon à pouvoir leur attribuer une ou plusieurs fonctions d'extraction. Une fonction d'extraction est un mécanisme générique de recherche de réponse qui peut être paramétré selon un élément provenant de la question. De façon formelle,

$$\mathcal{C} = f(\rho,\ \varphi) \tag{4.1}$$

où $f$ est une fonction d'extraction, $\rho$ est un passage sur lequel s'opère la recherche de réponse, $\varphi$ est le focus de la question (section 4.1.1) et $\mathcal{C}$ est l'ensemble des candidats trouvés dans $\rho$. Chaque élément de $\mathcal{C}$ est un triplet $(c_i, d_i, s_i)$ où $c_i$ est le candidat, $d_i$ est le document source et $s_i$ est un score attribué au candidat par la fonction d'extraction.

C'est à l'aide d'expressions régulières portant sur les mots et leur étiquette grammaticale que les questions sont analysées et que la fonction d'extraction appropriée est choisie. La liste des fonctions que nous proposons est présentée au tableau 4.2 ; ces fonctions sont décrites à la section 4.3.

Dans la plupart des systèmes de question-réponse qui procèdent à une classification des questions ([Harabagiu *et al.*, 2000] et [Ferret *et al.*, 2000], par exemple), chaque classe correspond à un type d'entité que le système est en mesure de discerner dans le texte : un lieu géographique, un nom

de personne, une race d'animal, un poids, une longueur, etc. Afin de déduire le type d'entité sur lequel focalise une question, il faut avoir prévu toutes les formes possibles de questions portant sur ce type d'entité. Ceci introduit une difficulté supplémentaire vu le grand nombre de reformulations possibles d'une question (la figure 2.3 en donne un échantillon). De plus, cette difficulté est d'autant multipliée qu'il y a de types d'entité dans la classification.

Par contre, l'analyse lexicale et syntaxique de questions factuelles en anglais nous a montré que les mécanismes de recherche de réponse sont peu nombreux, contrairement aux types d'entité à chercher. C'est pourquoi les classes composant la classification que nous proposons correspondent à des *mécanismes de recherche* à appliquer plutôt qu'à des *types d'entité* à chercher. Les fonctions d'extraction du tableau 4.2 forment les 11 classes de notre classification dite *fonctionnelle.*

Les questions sont plus faciles à classer de cette façon car le nombre de classes est petit et les classes sont fortement liées à la syntaxe des questions (c'est ce que nous avons observé lors de l'élaboration des expressions régulières pour la classification). Même si le nombre de classes est réduit par rapport à une classification fondée sur les types d'entité, il est possible d'atteindre le même degré de précision en paramétrant les fonctions avec le focus de la question lorsque utile. Le paramétrage automatisé d'un mécanisme générique permet en théorie de traiter des questions à propos de tout, alors qu'une classification à base de types d'entité est limitée aux questions portant sur les types d'entité prévus dans la classification. Dans les pires cas, la fonction $f$ choisie et son paramètre $\varphi$ peuvent conduire à une recherche trop générique et non optimale mais l'espoir de trouver la réponse correcte n'est pas nul.

## 4.2 Recherche de passages

Après avoir analysé la question et choisi une fonction d'extraction, QUANTUM procède à l'extraction des candidats-réponses. Cette tâche demande beaucoup de temps ; par conséquent, il est préférable d'appliquer la fonction d'extraction à quelques courts passages sélectionnés parmi tout le corpus de documents, tout en s'assurant qu'il s'agit des passages les plus pertinents. En plus d'améliorer la rapidité du traitement, le fait d'avoir des passages courts et peu nombreux diminue le nombre de candidats extraits, et donc le bruit (lors de l'analyse des performances de $XR^3$ à la section 3.2.2, nous avons conclu qu'il est difficile de distinguer les candidats intéressants des candidats farfelus). Nous avons expérimenté deux techniques de recherche de passages : la première est similaire à celle utilisée par $XR^3$ et produit des passages de longueur fixe, tandis que la deuxième tire parti d'Okapi pour produire des passages de longueur variable.

### 4.2.1 Passages de longueur fixe

Il s'agit d'une variante de la technique utilisée par $XR^3$ pour sélectionner des passages de 250 caractères (sect. 3.1.2). Nous rappelons ici brièvement les étapes que cette méthode comporte et de quelle façon elles ont été adaptées pour QUANTUM.

| Catégorie de termes-clés | Poids des termes-clés |
|---|---|
| Syntagmes nominaux contenant plus d'un mot | 10 |
| Noms propres et noms communs | 1 |
| Chaînes entres guillemets | 20 |
| Entités nommées | 10 |
| Années | 10 |

TAB. 4.3: Catégories de termes-clés extraits d'une question pour la recherche de passages de longueur fixe et poids de chacun des termes-clés de la catégorie.

D'abord, la question intégrale est soumise à un moteur de recherche conventionnel (AT&T lors de TREC-9, PRISE lors de TREC-X) afin de réduire le corpus de recherche aux 200 meilleurs documents. Cette opération n'est pas effectuée par QUANTUM puisque cette liste de documents est disponible aux participants de TREC. La première opération véritablement effectuée par QUANTUM est l'extraction des termes-clés de la question. Le tableau 4.3 montre quels sont les termes-clés extraits et quel poids est assigné à chacun d'eux (les poids ont été inspirés de ceux utilisés par $XR^3$, eux-mêmes obtenus par expérimentation avec les données de TREC-8). Les documents sont ensuite balayés et pour chaque occurrence d'un terme-clé rencontrée, le passage de 250 caractères centré sur ce terme est retenu. Ensuite, chacun des passages reçoit un score égal à la somme des poids des termes-clés qu'il contient. Si deux passages se chevauchent par plus de 125 caractères et que leur score est égal, un seul passage est conservé ; nous avons arbitrairement choisi de conserver celui qui apparaît en premier dans le texte.

### 4.2.2 Passages de longueur variable avec Okapi

Okapi [Robertson et Walker, 1999] est un moteur de recherche qui permet de sélectionner des passages dont la longueur est un nombre entier de paragraphes. QUANTUM soumet une requête formée de la question intégrale et Okapi se charge de sélectionner et de tronquer les termes-clés qu'il utilisera pour la recherche. Afin d'obtenir les passages les plus courts possible, leur longueur est limitée à 1 paragraphe. Les passages retournés par Okapi ont en moyenne 350 caractères ; bien qu'ils soient plus longs que ceux obtenus par la méthode des passages de longueur fixe (250 caractères), ils sont de meilleure qualité, de sorte qu'un bon taux de précision (nombre de questions dont la réponse se trouve dans les passages sélectionnés divisé par le nombre total de questions) puisse être atteint avec un petit nombre de passages (tab. 4.4).

## 4.3 Extraction des candidats

Une fois les passages les plus pertinents obtenus, que ce soit par la méthode des passages de longueur fixe ou à l'aide d'Okapi, la fonction d'extraction qui a été choisie après l'analyse de la question est appliquée sur ces passages pour en extraire des candidats. Les fonctions sont présentées

| Nombre de passages par question | Taux de précision (%) selon la taille des passages | | |
|---|---|---|---|
| | Taille = 1 par. | Taille = 2 par. | Taille = 3 par. |
| 5 | 63 | 72 | 74 |
| 10 | 72 | 80 | 81 |
| 30 | 83 | 88 | 89 |
| 50 | 86 | 90 | 91 |

TAB. 4.4: Taux de précision (nombre de questions dont la réponse se trouve dans les passages sélectionnés divisé par le nombre total de questions) obtenu par Okapi sur les 682 questions de TREC-9, selon le nombre de passages retenus par question et leur taille (mesurée en paragraphes).

au tableau 4.2 et nous les décrivons sommairement ici. Rappelons qu'une fonction d'extraction $f(\rho, \varphi)$, où $\rho$ est un passage et $\varphi$ est le focus de la question, extrait des candidats-réponses $c_i$ avec leur numéro de document $d_i$. Elle leur attribue aussi un score d'extraction $s_i$ ; ce score est combiné à d'autres scores décrits à la section 4.3.6 pour former le score final du candidat sur la base duquel les meilleurs candidats sont sélectionnés.

Avant de passer aux mécanismes particuliers des fonctions, mentionnons que les passages sont tous tokenisés, étiquetés et soumis à un extracteur de groupes nominaux (sect. 4.1.2). Les réponses aux questions de TREC sont généralement des groupes nominaux et il est donc inutile d'extraire des candidats qui n'en sont pas. Les étiquettes grammaticales sont d'autant plus utiles que certaines fonctions les utilisent dans des expressions régulières.

### 4.3.1 Fonctions de hiérarchie : *définition*($\rho$, $\varphi$) et *spécialisation*($\rho$, $\varphi$)

La *définition* et la *spécialisation* sont des opérations inverses portant sur une hiérarchisation hyponyme/hyperonyme des termes. L'*hyperonyme* d'un terme est un terme plus général : par exemple, *liquor* est un hyperonyme de *brandy* (fig. 4.4). L'*hyponyme* est quant à lui un terme plus spécifique : *brandy* est un des hyponymes de *liquor*. QUANTUM utilise WordNet (version 1.6[1]) pour obtenir automatiquement ces relations ; le lecteur peut trouver une description détaillée de l'organisation de WordNet dans [Fellbaum, 1998]. Ces relations sont transitives mais il est à noter que dans ce texte, les termes *hyponyme* et *hyperonyme* ne désignent pas nécessairement l'*hyponyme immédiat* et l'*hyperonyme immédiat*. Ainsi, nous considérons *drink* comme un hyperonyme de *brandy* alors que seul *liquor* peut être l'hyperonyme immédiat de *brandy*.

L'opération de *définition* consiste à décrire un terme spécifique à l'aide d'un terme générique (son hyperonyme) auquel des adjectifs, adverbes, noms et propositions relatives sont accessoirement greffés dans le but de distinguer l'hyponyme en question des autres hyponymes possibles. À l'inverse, nous nommons *spécialisation* l'opération consistant à trouver un des hyponymes d'un terme, c'est-à-dire une instance plus spécifique.

Par exemple, la question *What is ouzo ? (644)* demande de définir *ouzo*. Parmi les réponses

[1]WordNet : `www.cogsci.princeton.edu/~wn/index.shtml`

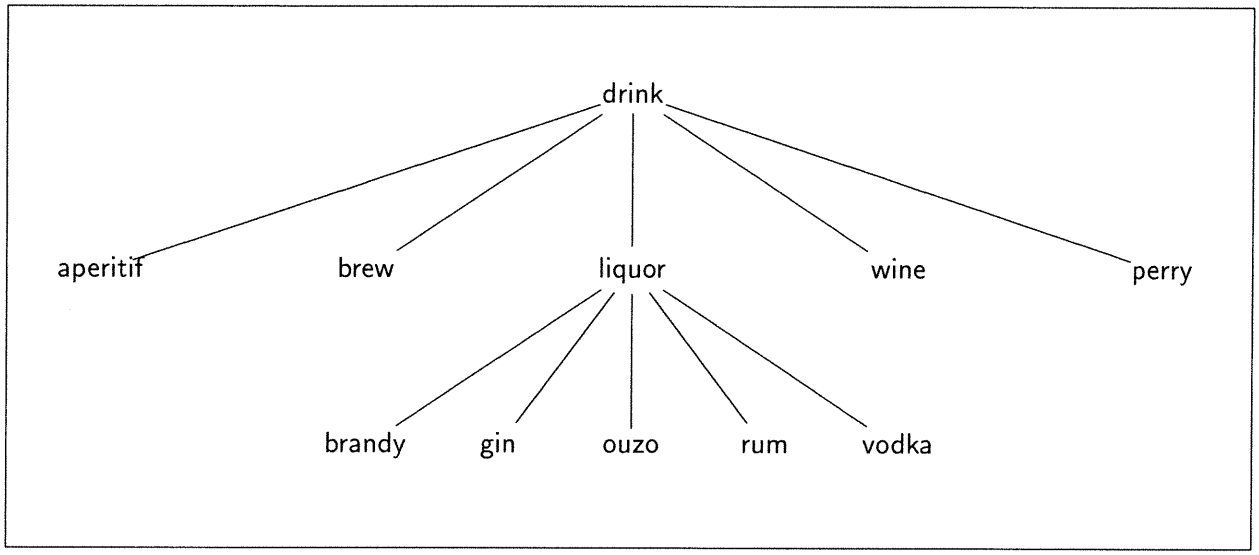


FIG. 4.4: Portion de la hiérarchie de WordNet pour le terme *ouzo*.

jugées correctes, nous retrouvons : *anise flavored drink, anise flavored liquor, fiery liquor* et *Greek traditional drink*. Les têtes de ces syntagmes nominaux (*drink, liquor*) sont effectivement des hyperonymes de *ouzo* dans la hiérarchie de WordNet (fig. 4.4). À l'inverse, une question de spécialisation aurait la forme : *What drink is traditional to Greeks ?* ou *Name a liquor flavored with anise.* La réponse *ouzo* est comme prévu un hyponyme de *drink* et de *liquor*.

La fonction d'extraction *définition*($\rho$, $\varphi$) extrait d'un passage $\rho$ les groupes nominaux dont le dernier terme est un hyperonyme du terme $\varphi$ à définir et elle leur attribue un certain score. Si, de plus, un candidat apparaît dans un contexte qui s'apparie à une des expressions régulières du tableau 4.5, son score est augmenté de façon à privilégier ce candidat. Par exemple, si la question était *What is leukemia ? (1081)*, la fonction appliquée au passage $\rho$ serait *définition*($\rho$, *leukemia*) et les candidats en gras du tableau 4.5 seraient privilégiés.

La fonction *spécialisation*($\rho$, $\varphi$) retient tous les groupes nominaux dont la tête est un hyponyme de $\varphi$. Lorsque $\varphi$ (en gras ici) est un lieu comme dans *What* ***city*** *had a world fair in 1900 ? (904)*, une personne comme dans *What* ***person***'*s head is on a dime ? (905)* ou une expression de temps comme dans *Mercury, what* ***year*** *was it discovered ? (990)*, il est préférable de faire appel aux mécanismes plus fins des fonctions *lieu*($\rho$), *personne*($\rho$) ou *temps*($\rho$). Pour déterminer si une telle redirection est nécessaire, QUANTUM vérifie si $\varphi$ compte respectivement *location*, *person* ou *time* parmi ses hyperonymes (nous avons simplifié ici car en réalité, la liste des hyperonymes à vérifier est plus complexe et les différents sens d'un même terme doivent être distingués).

| Expression régulière | Exemple |
|---|---|
| $\varphi$ **(cand)** | People suffering from *leukemia* **(a blood disease)**... |
| $\varphi$, **cand** | People suffering from *leukemia*, **a blood disease**... |
| $\varphi$ is **cand** | *Leukemia* is **a blood disease**... |
| **cand**, $\varphi$ | People suffering from **the blood cancer**, *leukemia*... |
| the definition of $\varphi$ is **cand** | The definition of *leukemia* is **a cancer** that... |
| $\varphi$ could be defined as **cand** | *Leukemia* could be defined as **a cancer** that... |

TAB. 4.5: Expressions régulières utilisées par la fonction *définition*($\rho$, $\varphi$) pour l'identification des candidats. Le focus est en italique et le candidat est en gras.

### 4.3.2 Fonctions de quantification : *cardinalité*($\rho$, $\varphi$) et *mesure*($\rho$, $\varphi$)

La fonction *cardinalité*($\rho$, $\varphi$) est utilisée pour répondre aux questions similaires à *How many **Great Lakes** are there ? (933)*, c'est-à-dire lorsque le candidat type est un nombre suivi du focus (que nous avons identifié en gras dans la question et qui pourrait être présent dans la réponse de façon partielle ou sous la forme d'un synonyme ou d'un hyperonyme). Le candidat *5 lakes*, par exemple, serait accepté.

La fonction *mesure*($\rho$, $\varphi$), quant à elle, est plus appropriée pour des questions comme *How much **folic acid** should a pregnant woman get each day ? (729)*. Le candidat type est un nombre suivi d'une unité de mesure et possiblement du focus : *400 micrograms of folic acid.* QUANTUM détermine qu'un terme est une unité de mesure en vérifiant s'il est un hyponyme de *unit of measurement.*

### 4.3.3 Fonction de caractérisation : *attribut*($\rho$, $\varphi$)

Les questions de type *How **far** is it from Denver to Aspen ? (894)* ou encore *How **fast** is the speed of light ? (1105)* sont complexes à traiter du fait de la difficulté d'automatiser le lien entre le focus (*far*, *fast*) et la réponse (*200 miles*, *299,792 km/sec*, *670 million mph*). Le focus est ici un adjectif alors que la réponse est une mesure : cette relation n'est pas disponible avec WordNet. Par contre, WordNet offre la relation d'attribut qui associe certains adjectifs avec l'attribut qu'ils qualifient, comme *far*↔*distance* et *fast*↔*speed.* La fonction d'extraction *attribut*($\rho$, $\varphi$) utilise WordNet afin d'obtenir l'attribut auquel fait référence le focus de la question, puis elle consulte une table de correspondance que nous avons dressée manuellement afin de chercher les unités de mesure appropriées.

### 4.3.4 Fonctions de complétion de concept : *personne*($\rho$), *temps*($\rho$), *lieu*($\rho$) et *objet*($\rho$)

Le terme *complétion de concept* est employé par [Lehnert, 1978] et [Lauer *et al.*, 1992] pour désigner les questions qui demandent de trouver l'entité référencée par un des pronoms interrogatifs *who/when/where/what* : *Who is the tallest man in the world ? (1144)*, *When was Ulysses S.*

| Type | Description |
|---|---|
| PERSON | Noms propres de personnes (*G. Washington*), titres (*Mr. President*) |
| ORGANIZATION | Noms d'organisations (*NATO*, *Congress*) |
| LOCATION | Toponymes (*Lake Ontario*, *North Africa*) |
| DATE | Années (*1983*), dates (*Sep. 12, 1943*), mois (*May*), jours (*Friday*) |
| TIME | Heures (*23 :03 :12*, *4 a.m.*, *8 o'clock*) |

TAB. 4.6: Entités nommées reconnues par Alembic et utilisées par QUANTUM. Se référer à [Aberdeen *et al.*, 1995] pour des descriptions plus détaillées.

*Grant born ? (1279)*, *Where is Perth ? (1013)*, *What do penguins eat ? (257)*. Les réponses sont généralement des entités nommées : noms propres de personnes, toponymes, dates, etc.

QUANTUM fait appel à l'extracteur d'entités nommées Alembic de MITRE Corporation [Aberdeen *et al.*, 1995] lorsque la réponse attendue est l'une des entités nommées du tableau 4.6. La fonction *personne*($\rho$) requiert les entités nommées de type PERSON et ORGANIZATION, la fonction *lieu*($\rho$) requiert les entités de type LOCATION et la fonction *temps*($\rho$) requiert les entités DATE et TIME. L'utilisation de l'extracteur nous évite la conception d'expressions régulières pour reconnaître ces entités nommées. Alembic répond aux spécifications de la conférence MUC-7. Notre choix s'est porté sur lui en raison de son efficacité et du fait qu'il est disponible gratuitement pour des fins de recherche[2].

Afin d'obtenir un taux de rappel plus grand, les candidats qui ne sont pas retenus par Alembic mais qui sont des hyponymes de *person*, *location* et *time unit* sont quand même considérés ; ils reçoivent un score moindre, cependant.

Les fonctions de complétion de concept ne nécessitent pas d'isoler un focus $\varphi$ dans la question. En effet, le mécanisme de recherche n'a pas besoin d'être paramétré. Cela ne signifie pas pour autant que les autres éléments de la question soient inutiles : ils ont une influence par le biais des différents scores composant le score final d'un candidat (sect. 4.3.6).

Aucune des entités extraites par Alembic n'est appropriée à la fonction *objet*($\rho$). En fait, chercher un *objet* est trop général pour que des critères d'extraction puissent être appliqués. Dans ce cas, tous les groupes nominaux du passage $\rho$ sont des candidats et seules les autres composantes du score final peuvent les départager. Cette fonction constitue la fonction par défaut que QUANTUM utilise lorsque aucune autre fonction plus précise ne peut s'appliquer.

### 4.3.5 Autres fonctions : *manière*($\rho$) et *raison*($\rho$)

Les questions qui expriment une manière, telle *How did Janice Joplin die ? (1163)*, ou encore une raison, telle *Why is the sun yellow ? (1220)*, demandent en général une réponse plus complexe qu'un groupe nominal. De plus, ces questions représentent une proportion minime des corpus de questions de TREC : ces deux catégories confondues comptent pour moins de 2 % du corpus de

[2]Alembic Workbench Project : `www.mitre.org/resources/centers/it/g063/workbench.html`

TREC-X. Nos ressources étant limitées, nous n'avons pas implémenté les fonctions *manière*($\rho$) et *raison*($\rho$).

### 4.3.6 Score des candidats

Les fonctions d'extraction assignent un score aux candidats en même temps qu'elles les extraient. Ce score d'extraction n'est qu'un des 3 scores partiels composant le score final d'un candidat :

$$\textit{score final} = \textit{score d'extraction} + \textit{score du passage} + \textit{score de proximité} \tag{4.2}$$

Le *score d'extraction* reflète notre niveau de confiance en la méthode utilisée pour extraire le candidat. Il est attribué par la fonction d'extraction elle-même selon qu'elle a extrait le candidat à l'aide de l'extracteur d'entités nommées, des expressions régulières, de WordNet ou d'une combinaison de ces outils. De cette façon, il est possible de favoriser les candidats extraits par expressions régulières ou avec Alembic au détriment de ceux extraits parce qu'ils satisfont à une relation hyperonyme/hyponyme de WordNet (la consultation de WordNet peut introduire du bruit, notamment à cause de la polysémie).

Alors que le score d'extraction ne concerne que la forme et le type d'un candidat, le *score de passage* vise à prendre en compte l'information supplémentaire apportée par le discriminant de la question. Il mesure la similitude entre le contexte dans lequel est posée la question et le contexte dans lequel est cité le candidat. En général, les éléments du discriminant de la question apparaissent dans le texte sous forme modifiée et ils sont dispersés dans les quelques phrases entourant la réponse correcte : c'est pourquoi nous croyons qu'un moteur de recherche est le meilleur outil pour mesurer la concentration d'éléments du discriminant dans un passage. Nous utilisons donc le score attribué à un passage lors de sa sélection par la méthode des passages de longueur fixe ou, encore mieux, par Okapi.

La combinaison du score d'extraction et du score de passage favorise les candidats qui sont du type cherché et qui apparaissent dans un contexte apparenté au discriminant de la question. Afin d'établir un lien supplémentaire entre un candidat donné et la question, QUANTUM ajoute un *score de proximité* aux candidats adjacents à un groupe nominal contenant un terme-clé de la question. Un groupe nominal est considéré adjacent à un candidat si les deux ne sont pas séparés par un autre groupe nominal. Nous avons choisi une valeur faible pour ce score de proximité afin de minimiser son influence car cette mesure de la proximité est, dans sa forme actuelle, grossière. De plus, le bien-fondé du principe est encore à démontrer. À tout le moins, le score de proximité peut servir à départager des candidats *ex aequo*.

## 4.4 Expansion des candidats et élimination des redondances

Les réponses soumises à TREC-X peuvent contenir jusqu'à 50 caractères. QUANTUM procède donc à l'expansion des candidats-réponses en prélevant dans le document source autant de caractères à gauche qu'à droite du candidat, jusqu'à l'obtention d'une réponse de 50 caractères. De cette façon, QUANTUM augmente les chances que la réponse correcte y figure dans le cas malheureux où le candidat identifié serait erroné. Ces chances ne sont pas négligeables : le MRR obtenu par $XR^3$ à TREC-9 est passé de 0,149 (série `UdeMexct`, tab. 3.4) à 0,179 (série `UdeMshrt`) par la seule expansion des candidats à 50 caractères.

Toujours dans le but de favoriser le hasard, QUANTUM maximise le nombre de mots complets figurant dans la chaîne de 50 caractères car la réponse correcte ne peut être un mot incomplet. À cet effet, QUANTUM élimine les mots tronqués aux extrémités de la chaîne. La chaîne résultante ayant moins de 50 caractères, QUANTUM a le loisir de l'augmenter dans la direction qui permet d'inclure le plus grand nombre de mots complets.

L'expansion systématique de tous les candidats est inutile puisque 5 suggestions par question sont nécessaires. Afin de ne pas avoir de suggestions redondantes contenant les mêmes candidats, QUANTUM procède comme suit : d'abord, le meilleur candidat est étendu à 50 caractères ; ensuite, le deuxième candidat est étendu seulement s'il n'apparaît pas dans la première suggestion ; puis, le troisième candidat est étendu seulement s'il n'apparaît dans aucune des suggestions précédentes, et ainsi de suite jusqu'à ce que le nombre désiré de suggestions soit atteint. Pour inclure le plus grand nombre de candidats différents, les candidats en double sont éliminés même s'ils ne proviennent pas du même document, et cela au risque de perdre un candidat supporté par son document source au profit d'un candidat non supporté. Cependant, le risque qu'un candidat ne soit pas supporté est faible : à TREC-X, sur les 175 réponses trouvées par QUANTUM et jugées correctes lors de l'évaluation tolérante, seulement 5 (2,9 %) ont été jugées non supportées lors de l'évaluation stricte.

## 4.5 Traitement des questions sans réponse

Jusqu'à maintenant, nous avons pris pour acquis que toute question traitée par QUANTUM a une réponse dans le corpus de documents. Or, pour TREC-X, cela peut ne pas être le cas : la réponse à une question est parfois absente du corpus. Le système doit reconnaître cette éventualité et l'indiquer par une suggestion formée de la chaîne *NIL* et d'un rang compris entre 1 et 5 (fig. 2.2). Le système peut quand même suggérer d'autres réponses ; il ajuste le rang de la réponse *NIL* selon l'importance relative qu'il lui accorde.

Étant donné que l'ordre de grandeur du score des réponses diffère d'une question à l'autre (particulièrement lorsque les fonctions d'extraction appelées sont différentes), il n'est pas possible de fixer un score-seuil en dessous duquel une réponse *NIL* est plus probable qu'une réponse à faible score. Nous choisissons plutôt d'appliquer un seuil sur la baisse de score entre deux candidats de

| Intervalle de normalisation | $\Delta_r$ | $\Delta_i$ |
|---|---|---|
| $\delta_i^{i+4}$ | 33 % | 29 % |
| $\delta_i^{i+3}$ | 40 % | 35 % |
| $\delta_i^{i+2}$ | 56 % | 50 % |

TAB. 4.7: Valeur de $\Delta$ selon l'intervalle de normalisation. La baisse de score $\Delta_r$ entre une réponse correcte et la suivante (incorrecte) est plus élevée que la baisse de score moyenne $\Delta_i$ entre deux réponses incorrectes de rangs consécutifs, peu importe l'intervalle de normalisation. Les résultats ont été obtenus par QUANTUM avec les questions de TREC-9.

rangs consécutifs : une fois les candidats ordonnés du meilleur au pire, une réponse *NIL* est insérée à la première baisse de score importante mesurée. L'écart de score est normalisé afin que l'écart-seuil que nous allons fixer soit le même pour toutes les questions.

Soit $a_i$ la réponse au rang $i$, et $\delta_i^{i+j}$ la différence de score entre $a_i$ et son $j^e$ successeur $a_{i+j}$. L'écart de score normalisé $\Delta_i$ entre $a_i$ et $a_{i+1}$ est calculé de la façon suivante :

$$\Delta_i = \frac{\delta_i^{i+1}}{\delta_i^{i+4}} = \frac{s_i - s_{i+1}}{s_i - s_{i+4}} \tag{4.3}$$

où $s_i$ est le score de $a_i$. La décision de normaliser sur $\delta_i^{i+4}$, c'est-à-dire sur l'écart de score entre des candidats distants de 5 rangs, est arbitraire ; cependant, les observations qui suivent sont valables même si l'intervalle de normalisation est différent.

Nous avons exécuté QUANTUM sur les questions de TREC-9 et nous avons conservé toutes les suggestions (pas seulement les 5 meilleures). Nous avons ensuite appliqué le script de correction automatique fourni par NIST et nous avons noté à quel rang $r$ apparaît la réponse correcte pour chaque question lorsque QUANTUM la trouve. Nous avons calculé $\Delta_r$ pour mesurer la baisse de score entre la réponse correcte et la réponse suivante, laquelle est présumée incorrecte. Nous avons ensuite calculé la baisse moyenne $\Delta_i$ pour une paire quelconque de réponses de rangs consécutifs. Nous avons ainsi pu observer que la baisse de score entre une réponse correcte et une réponse incorrecte est légèrement plus élevée que la baisse moyenne entre deux réponses incorrectes. Le tableau 4.7 montre que cette observation est vérifiée pour différents intervalles de normalisation.

Par conséquent, QUANTUM applique l'algorithme suivant pour déterminer si une réponse *NIL* doit figurer parmi les 5 suggestions finales. D'abord, la suggestion au rang 1 est considérée comme correcte puisqu'elle a obtenu le score le plus élevé. La question est alors de déterminer si la suggestion au rang 2 est plus probable qu'une réponse *NIL* insérée au rang 2. Pour ce faire, QUANTUM mesure $\Delta_1$ entre $a_1$ et $a_2$. Si cet écart est élevé, cela constitue un indice supplémentaire que $a_1$ est correcte et que $a_2$ est incorrecte, puisque nous avons observé qu'une réponse correcte était généralement suivie d'une baisse importante de score ; par conséquent, QUANTUM a une confiance suffisante en $a_1$ pour affirmer que si $a_1$ s'avère incorrecte, alors il n'existe pas de réponse satisfaisante dans le corpus. Concrètement, si $\Delta_1$ est plus élevé qu'un écart-seuil $\Delta_t$, QUANTUM insère une réponse *NIL* au rang 2 et décale d'un rang la suggestion $a_2$ et les suivantes. Si, par contre, $\Delta_1$ est faible,

QUANTUM considère que $a_2$ est un bon deuxième choix ; il reprend alors l'algorithme en considérant cette fois que $a_2$ est correcte et il examine la pertinence d'insérer une réponse *NIL* au $3^e$ rang. La procédure s'arrête aussitôt qu'une réponse *NIL* est insérée parmi les 5 suggestions initiales ou après que l'insertion au rang 5 ait été considérée.

Si QUANTUM a réussi à formuler moins de 5 suggestions et qu'aucune baisse de score entre ces suggestions ne justifie l'insertion d'une réponse *NIL*, il ajoute une réponse *NIL* après la dernière suggestion.

Les écarts $\Delta_r$ entre une réponse correcte et une réponse incorrecte calculés au tableau 4.7 sont des bornes inférieures pour un écart-seuil $\Delta_t$ au-dessus duquel une réponse *NIL* doit être insérée. Nous avons fixé ce seuil $\Delta_t$ de façon expérimentale en créant un corpus de 400 questions dont 5 % n'ont pas de réponse dans le corpus de documents et dont les autres 95 % sont des questions de TREC-9. Nous avons ensuite choisi la valeur de $\Delta_t$ qui maximise le MRR obtenu avec ce corpus de questions. Nous avons obtenu un MRR maximal de 0,257 avec $\Delta_t = 80$ %. Cependant, ce seuil peut ne pas être optimal si la proportion des questions sans réponse n'est pas de 5 % ; notamment, le corpus de questions de TREC-X est composé à 10 % de questions sans réponse.

La technique de l'écart-seuil souffre d'un handicap : elle ne permet pas l'insertion d'une réponse *NIL* au rang 1 car $\Delta_0$ ne peut être calculé. Le seul cas où QUANTUM peut insérer une réponse *NIL* au premier rang est lorsqu'il n'a pu extraire de candidat. Selon nous, il est très rare qu'une telle situation se produise car les fonctions d'extraction sont très permissives au sujet de la qualité des candidats dans le but d'obtenir un taux de rappel élevé.

## 4.6 Analyse des performances de QUANTUM à TREC-X

QUANTUM a produit 3 séries de réponses pour la sous-piste des réponses de 50 caractères de TREC-X. Ces séries diffèrent par le choix de la méthode de recherche de passages et par le choix de l'écart-seuil pour l'insertion de réponses *NIL*. Les combinaisons formant les 3 séries et leur score officiel sont montrés au tableau 4.8 et comparés aux autres systèmes de TREC-X à la figure 4.5. QUANTUM a aussi produit 2 séries de réponses pour la sous-piste des listes d'éléments mais cette sous-piste dépasse le propos du présent ouvrage.

La meilleure série produite par QUANTUM a été obtenue en utilisant Okapi comme moteur de recherche de passages et en fixant un écart-seuil $\Delta_t$ de 80 % pour l'insertion de réponses *NIL*. En comparant les séries `UdeMmainOk80` (MRR = 0,191) et `UdeMmainQt80` (MRR = 0,137) qui ne diffèrent que par la méthode de recherche de passages, nous concluons qu'Okapi est beaucoup plus performant que notre propre méthode de recherche de passages de longueur fixe. En comparant les séries `UdeMmainOk80` et `UdeMmainOk60` (MRR = 0,183) qui ne diffèrent que par l'écart-seuil pour l'insertion d'une réponse *NIL*, nous concluons qu'un écart-seuil de 80 % est plus approprié pour le corpus de questions de TREC-X.

Dans les sections suivantes, nous évaluons trois modules-clés de QUANTUM : le module d'analyse

| Série | Recherche de passages | Seuil $\Delta_t$ | Évaluation stricte | | Évaluation tolérante | |
|---|---|---|---|---|---|---|
| | | | MRR | Taux d'échec (%) | MRR | Taux d'échec (%) |
| UdeMmainOk80 | Okapi | 80 % | 0,191 | 65 % | 0,197 | 64 % |
| UdeMmainOk60 | Okapi | 60 % | 0,183 | 67 % | 0,189 | 66 % |
| UdeMmainQt80 | QUANTUM | 80 % | 0,137 | 76 % | 0,145 | 75 % |

TAB. 4.8: Construction des 3 séries produites par QUANTUM pour TREC-X et résultats. Lorsque l'outil de recherche de passages utilisé est Okapi, les passages intermédiaires ont une longueur de 1 paragraphe (350 caractères en moyenne) ; lorsque QUANTUM utilise sa propre méthode de recherche, les passages ont tous 250 caractères (sect. 4.2). Le seuil $\Delta_t$ est l'écart-seuil normalisé au-dessus duquel une réponse *NIL* est considérée plus probable que toute autre réponse (sect. 4.5). Le taux d'échec (proportion des questions auxquelles QUANTUM n'a pas trouvé de réponse correcte) est calculé par rapport à 492 questions (500 étaient initialement prévues mais 8 ont été retirées officiellement).

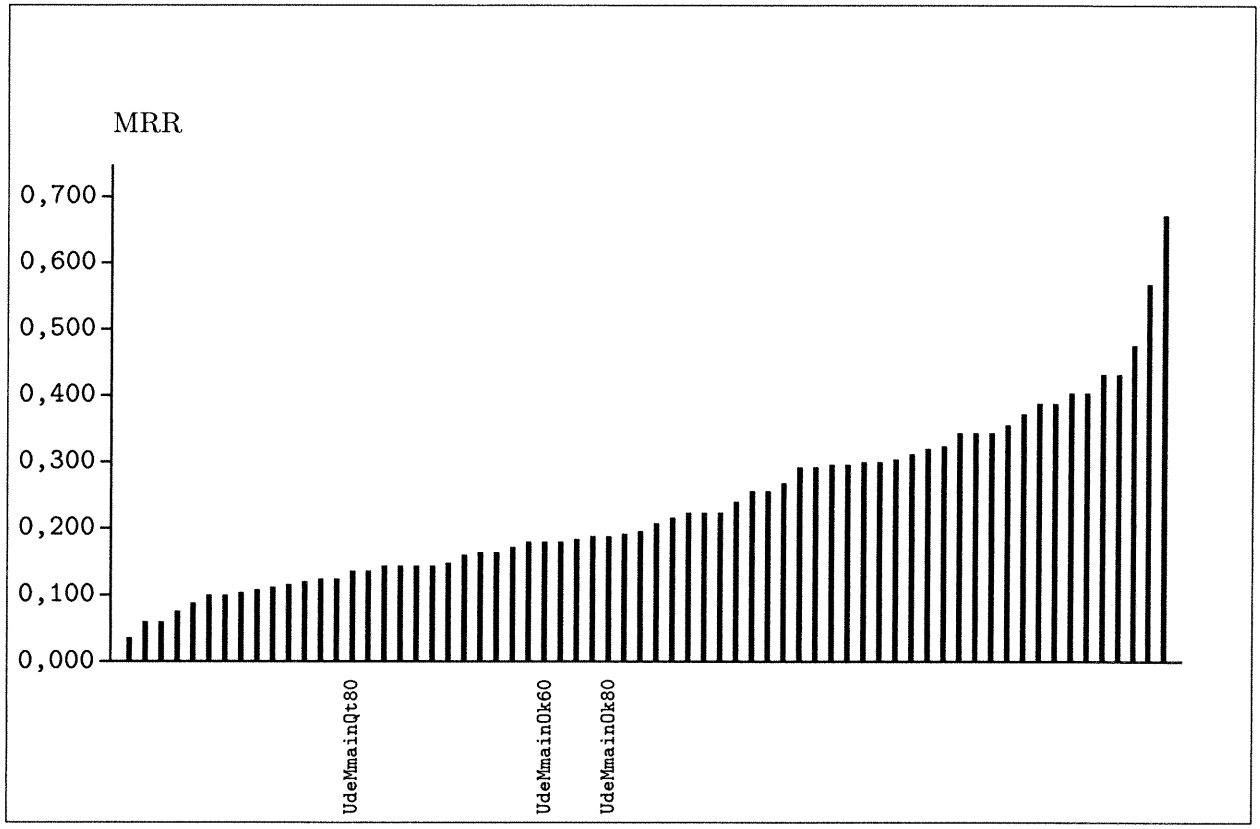


FIG. 4.5: MRR des séries de réponses soumises à TREC-X par l'ensemble des systèmes (évaluation stricte). Chaque barre représente une série de réponses. Les différences majeures entre QUANTUM et les systèmes les plus performants sont présentées à la section 4.7.

| Fonction | Nombre de questions | | Analyse par QUANTUM | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Fonct. correcte Focus correct | | Fonct. correcte Focus incorrect | | Fonct. incorrecte | |
| *définition*($\rho$, $\varphi$) | 140 | (29 %) | 107 | (76 %) | 11 | (8 %) | 22 | (16 %) |
| *spécialisation*($\rho$, $\varphi$) | 194 | (40 %) | 177 | (91 %) | 7 | (4 %) | 10 | (5 %) |
| *cardinalité*($\rho$, $\varphi$) | 13 | (3 %) | 12 | (92 %) | 0 | (0 %) | 1 | (8 %) |
| *mesure*($\rho$, $\varphi$) | 1 | (0 %) | 1 | (100 %) | 0 | (0 %) | 0 | (0 %) |
| *attribut*($\rho$, $\varphi$) | 22 | (4 %) | 20 | (91 %) | 0 | (0 %) | 2 | (9 %) |
| *personne*($\rho$) | 43 | (9 %) | 40 | (93 %) | — | — | 3 | (7 %) |
| *temps*($\rho$) | 26 | (5 %) | 26 | (100 %) | — | — | 0 | (0 %) |
| *lieu*($\rho$) | 27 | (5 %) | 27 | (100 %) | — | — | 0 | (0 %) |
| *inconnu*($\rho$) | 26 | (5 %) | 23 | (88 %) | — | — | 3 | (12 %) |
| **Total** | **492** | **(100 %)** | **433** | **(88 %)** | **18** | **(4 %)** | **41** | **(8 %)** |

TAB. 4.9: Erreurs de classification par fonction d'extraction, pour les 492 questions de TREC-X. Les fonctions *raison*($\rho$), *manière*($\rho$) et *objet*($\rho$) sont traitées comme *inconnu*($\rho$) dans la version courante de QUANTUM.

des questions, le module d'extraction des réponses et le module d'insertion de réponses *NIL*.

### 4.6.1 Évaluation du module d'analyse des questions

Le tableau 4.9 détaille les erreurs commises lors de l'analyse des questions, regroupées selon la fonction d'extraction qui aurait dû être attribuée. Elle montre que 88 % des 492 questions de TREC-X sont correctement analysées par QUANTUM, de sorte que la fonction d'extraction appropriée est utilisée pour trouver des candidats-réponses ; 4 % des questions sont partiellement bien analysées, c'est-à-dire que la fonction est correctement choisie mais le focus est mal identifié ; enfin, QUANTUM échoue son analyse dans 8 % des cas.

La fonction d'extraction souffrant le plus des erreurs d'analyse est *définition*($\rho$, $\varphi$) : 24 % des questions de définition se voient assigner une fonction d'extraction ou un focus incorrects. Ce taux d'erreur a d'autant plus d'impact que ces questions comptent pour presque 30 % du corpus de TREC-X.

La catégorie la plus commune du corpus, *spécialisation*($\rho$, $\varphi$), représente 40 % du corpus et est plutôt bien analysée avec un taux de succès de 91 %.

Les catégories *temps*($\rho$) et *lieu*($\rho$) obtiennent un taux de succès de 100 % ; l'absence de focus à identifier pour ces fonctions diminue les risques d'erreur d'analyse.

Les autres catégories, dont le nombre de représentants varie de 3 à 9 % du corpus total, sont analysées avec succès dans 88 à 93 % des cas. La catégorie *mesure*($\rho$, $\varphi$) a un taux de succès de 100 % mais il n'est pas possible d'en tirer de conclusions puisqu'elle n'a qu'un seul représentant.

Le tableau 4.10 montre que près de la moitié (47 %) des erreurs d'analyse sont dues à des formes syntaxiques de question auxquelles les filtres d'analyse de QUANTUM ne peuvent s'appliquer. La conception de filtres supplémentaires permettrait de réduire les erreurs de ce type. Environ un tiers (29 %) des erreurs d'analyse sont causées par l'étiqueteur grammatical : l'étiqueteur s'appuie

| Type d'erreurs | Nombre de questions | Proportion |
|---|---|---|
| Formes syntaxiques non prévues | 28 | 47 % |
| Étiquettes grammaticales incorrectes | 17 | 29 % |
| Groupes nominaux mal identifiés | 9 | 15 % |
| Autres | 5 | 8 % |

TAB. 4.10: Types d'erreurs d'analyse pour les 59 questions mal analysées de TREC-X.

sur des techniques probabilistes mais il n'a pas été entraîné sur un corpus de questions, d'où son incapacité à saisir les particularités de la syntaxe des phrases interrogatives.

### 4.6.2 Évaluation du module d'extraction des candidats

Pour évaluer l'utilité des outils utilisés par le module d'extraction des candidats, nous effectuons des simulations avec et sans ces outils. Nous comparons ensuite le résultat de ces simulations avec la performance du module original.

Le MRR d'une simulation est obtenu par correction automatique. À la suite de la conférence, les participants de TREC-X reçoivent une liste d'expressions régulières qui correspondent aux réponses jugées correctes par les évaluateurs. Ces expressions sont construites d'après les réponses soumises par l'ensemble des systèmes ayant participé à TREC-X. Un script utilise les expressions pour déterminer si les résultats d'une simulation contiennent les réponses correctes. Bien que beaucoup plus rapide qu'une évaluation humaine, l'évaluation automatique ne donne qu'une estimation du MRR qu'aurait obtenu une série de réponses si elle avait été évaluée selon les directives officielles de TREC. En effet, dans certains cas, les expressions régulières sont trop restrictives : c'est le cas lorsqu'une réponse acceptable est trouvée dans le corpus de documents mais n'a jamais été rencontrée par les évaluateurs et ne figure donc pas dans la liste des expressions. Dans la majorité des cas cependant, l'évaluation automatique est trop généreuse : c'est le cas lorsqu'une expression régulière s'apparie avec une chaîne de caractères citée dans un contexte différent de celui de la question.

Le corpus de questions utilisé pour l'évaluation du module d'extraction des candidats est composé des questions de TREC-X que QUANTUM analyse correctement, ceci afin que les données d'entrée du module d'analyse soient exemptes d'erreur. Pour mesurer les données de sortie telles qu'elles sont produites par le module, le module d'insertion de réponses *NIL* est désactivé et les questions sans réponse sont éliminées du corpus de questions.

Dans ces conditions, QUANTUM obtient un MRR de 0,223 (correction automatique). L'outil le plus utile à QUANTUM est l'extracteur d'entités nommées Alembic : sans lui, la performance de QUANTUM passe à 0,175, soit une baisse de 22 %. L'examen du MRR par fonction d'extraction (tab. 4.11) montre que les fonctions s'appuyant principalement sur l'extracteur d'entités nommées, telles *lieu*($\rho$), *personne*($\rho$) et *temps*($\rho$), voient leur MRR diminuer de moitié en l'absence d'Alembic.

| Fonction | Nb. de questions | MRR | | | |
|---|---|---|---|---|---|
| | | QUANTUM complet | QUANTUM sans WordNet | QUANTUM sans Alembic | QUANTUM sans expr. rég. |
| *définition*($\rho$, $\varphi$) | 113 | 0.179 | 0.159 | 0.181 | 0.158 |
| *spécialisation*($\rho$, $\varphi$) | 153 | 0.205 | 0.170 | 0.175 | 0.205 |
| *cardinalité*($\rho$, $\varphi$) | 12 | 0.096 | 0.086 | 0.096 | 0.096 |
| *mesure*($\rho$, $\varphi$) | 1 | 0.000 | 0.000 | 0.000 | 0.000 |
| *attribut*($\rho$, $\varphi$) | 18 | 0.019 | 0.056 | 0.074 | 0.074 |
| *personne*($\rho$) | 38 | 0.348 | 0.375 | 0.205 | 0.346 |
| *temps*($\rho$) | 24 | 0.411 | 0.418 | 0.206 | 0.411 |
| *lieu*($\rho$) | 25 | 0.451 | 0.415 | 0.188 | 0.418 |
| *inconnu*($\rho$) | 21 | 0.129 | 0.129 | 0.129 | 0.148 |
| **Total** | **405** | **0.223** | **0.207** | **0.175** | **0.218** |

TAB. 4.11: MRR global et MRR par fonction d'extraction pour différentes versions de QUANTUM. Les 405 questions sont les questions de TREC-X que QUANTUM analyse correctement et qui ont une réponse dans le corpus de documents. Les différentes versions de QUANTUM testées ici ne font pas d'insertion de réponses *NIL*.

Sans WordNet, la performance de QUANTUM est de 7 % moindre que sa performance optimale. Il est intéressant de noter que les fonctions *personne*($\rho$) et *temps*($\rho$) bénéficient du retrait de WordNet : nous croyons que pour ces fonctions, WordNet est une source de bruit qui brouille les bons résultats donnés par Alembic. En dernier lieu, le retrait des expressions régulières utilisées par certaines fonctions pour l'extraction de candidats engendre une baisse de 2 % par rapport à la performance du système complet. Le fait que cette baisse soit légère s'explique par le petit nombre d'expressions régulières que nous avons décidé d'inclure dans ; la plupart d'entres elles servent à l'extraction des définitions où elles semblent par ailleurs être utiles.

### 4.6.3 Évaluation du module d'insertion de réponses *NIL*

De 0,223, le MRR diminue à 0,199 avec l'inclusion de questions sans réponse dans le corpus de questions décrit à la section précédente et avec l'activation du module d'insertion de réponses *NIL*. Nous croyons que cette détérioration de la performance est due aux difficultés considérables que pose la détection d'absence de réponse et à la piètre efficacité du module d'insertion : QUANTUM répond correctement à seulement 5 des 49 questions sans réponse du corpus de TREC-X. Notons que la performance du module d'insertion de réponses *NIL* est tributaire de celle du module d'extraction des candidats par le biais du score que ce dernier assigne aux candidats.

## 4.7 Discussion, comparaison et travaux connexes

Trente-six organisations ont présenté un système à la piste question-réponse de la conférence TREC-X. Parmi les meilleurs de ces systèmes, quelques-uns font appel presque exclusivement à

des techniques de recherche d'information alors que d'autres font de l'analyse linguistique évoluée. Certains, comme c'est le cas aussi de QUANTUM, combinent les deux approches. Il semble encore être trop tôt pour conclure à la supériorité d'une approche par rapport à une autre.

Les prochaines sections sont consacrées à la description de quelques systèmes présentés à TREC-9 et à TREC-X qui suivent des approches intéressantes et comment QUANTUM se situe par rapport à eux. Malheureusement, il est difficile de comparer la performance des modules internes de QUANTUM avec celle des modules des autres systèmes ayant participé à TREC-X, cela pour deux raisons : d'abord parce que peu d'informations quantitatives sont disponibles en dehors des MRR globaux, ensuite parce que les systèmes ont des architectures très variées. Nous avons cependant pu distinguer deux grands types d'architecture : les systèmes à base de recherche d'information et ceux à base d'analyse linguistique.

### Systèmes à base de recherche d'information

Certains systèmes de question-réponse peuvent être considérés comme faisant principalement de la recherche d'information vu le peu de connaissances linguistiques qu'ils intègrent. C'est le cas du système de [Ittycheriah *et al.*, 2000] dont l'approche est essentiellement statistique : un modèle d'entropie maximale est construit par entraînement sur un corpus pour classifier les questions et reconnaître les entités nommées ; dans sa version TREC-X [Ittycheriah *et al.*, 2001], le système utilise aussi des arbres syntaxiques partiels. Le système de [Clarke *et al.*, 2000] fait aussi intervenir une quantité minimale de connaissances syntaxiques et sémantiques, et fonde sa technique surtout sur un modèle mathématique de rareté des termes [Prager *et al.*, 2001] ; la version TREC-X [Clarke *et al.*, 2001] se distingue par l'utilisation du web pour privilégier les réponses fréquentes. À l'extrême, le système de [Brill *et al.*, 2001] cherche d'abord la réponse sur le web et tente ensuite de la retrouver dans le corpus de documents duquel la réponse aurait dû être extraite.

Pour QUANTUM, la recherche d'information se limite à un filtrage du corpus de documents à l'aide du moteur de recherche Okapi (sect. 4.2). À notre avis, la performance de ce moteur surpasse ce que nous pourrions accomplir de nous-mêmes dans un futur proche, c'est pourquoi nous préférons continuer à l'utiliser et consacrer nos efforts sur les modules les moins performants.

Certains systèmes se distinguent par une indexation préalable du corpus de documents avec des concepts explicites. C'est le cas de Nova [Woods *et al.*, 2000], de PISAB [Attardi et Burrini, 2000] et d'un système de chez IBM [Prager *et al.*, 2000] [Prager *et al.*, 2001]. La nécessité de réindexer la base de données (en des temps non négligeables) lorsqu'elle est changée et les modifications qu'il faudrait apporter à l'algorithme d'indexation d'Okapi rendent cette approche plus ou moins attrayante ; cela dépend de la facilité avec laquelle l'indexation de concepts peut être intégrée dans le fonctionnement d'Okapi et de l'utilisation ultime de QUANTUM, c'est-à-dire si la base de données sera volumineuse et si elle sera modifiée souvent.

### Systèmes à base d'analyse linguistique

D'autres systèmes requièrent davantage de connaissances syntaxiques et sémantiques. Par exemple, le système *QA-Lassie* [Scott et Gaizauskas, 2000] utilise une représentation quasi-logique pour unifier la question et les réponses potentielles. Un des systèmes développés chez Microsoft [Elworthy, 2000] produit des arbres de dépendances sémantiques à l'aide de l'outil NLPWin. Le système de [Harabagiu *et al.*, 2000] utilise des représentations sémantiques et logiques avec règles d'unification évoluées ainsi qu'une approche itérative, c'est-à-dire que le système peut modifier sa requête initiale lorsqu'il juge les résultats insatisfaisants. Quant à Webclopedia [Hovy *et al.*, 2000][Hovy *et al.*, 2001b], il utilise un segmenteur sémantique de textes, une ontologie hiérarchique très fine des questions [Hovy *et al.*, 2001a] et un parseur grammatical CONTEX avec auto-apprentissage.

L'analyse linguistique effectuée par QUANTUM est faite principalement par le module d'extraction de candidats car cette étape nécessite une compréhension plus fine du texte, ce que la recherche d'information ne permet pas d'atteindre. L'analyse linguistique se limite toutefois à l'étiquetage grammatical des mots, à l'identification des groupes nominaux et à la consultation de WordNet : les ressources dont nous disposions n'ont pas permis de mettre en place des techniques d'analyse aussi évoluées que des arbres syntaxiques et des représentations en forme logique. Le module d'extraction de candidats étant le point faible de QUANTUM, il serait temps d'explorer des méthodes d'analyse plus ambitieuses.

### Systèmes ayant le mieux performé à TREC-X

Bien que la méthodologie d'évaluation mise en place par les responsables de TREC soit la plus uniforme et la plus objective possible, ces derniers signalent que les MRR officiels ne doivent pas être interprétés comme une mesure incontestable de la performance des systèmes. Conséquemment, les conférences TREC ne constituent pas une compétition mais un cadre de recherche. Ceci étant dit, voici les points saillants des techniques utilisées par les systèmes ayant obtenu les plus hauts MRR à TREC-X.

Le système de la compagnie InsightSoft-M [Soubbotin, 2001] a obtenu le MRR le plus élevé, soit 0,676. Il utilise seulement des filtres (expressions régulières) pour analyser les questions et extraire les réponses. Le deuxième meilleur MRR, soit 0,570, a été obtenu par le système QAS de Language Computer Corporation [Harabagiu *et al.*, 2001]. Ce système est la continuation de celui de [Harabagiu *et al.*, 2000] décrit brièvement plus haut : il construit des représentations sémantiques et logiques des questions et du texte, il fait une utilisation accrue des relations recensées par WordNet (hyponymie et hyperonymie, mais aussi méronymie et glossaire pour répondre aux définitions), il apparie les candidats suivant des règles d'unification et il modifie les critères de recherche automatiquement s'il juge que les candidats trouvés ne sont pas de qualité satisfaisante. La troisième meilleure performance (0,477) a été livrée par le moteur de recherche Oracle9i Text [Alpha *et al.*, 2001], au sujet duquel nous savons seulement qu'il mesure la fréquence des candidats

et leur proximité avec les termes-clés de la question présents dans les extraits les plus pertinents. Presque *ex aequo*, nous retrouvons le système de l'Information Science Institute (MRR de 0,435) et celui de l'University of Waterloo (MRR de 0,434). Le premier tire sa force d'une ontologie très détaillée des types d'entités à chercher [Hovy *et al.*, 2001b]. Quant au deuxième, il utilise le web pour privilégier les candidats les plus fréquents [Clarke *et al.*, 2001].

Il ressort de ces descriptions sommaires qu'il n'y a pas de technique qui se démarque des autres ni qui garantisse des résultats de loin supérieurs à la moyenne. Ou, si l'on considère que l'écriture de davantage de filtres et une utilisation accrue du web donneraient un avantage certain, cela n'en fait pas nécessairement des solutions prometteuses. Nous ne pouvons pas encore nous prononcer sur la supériorité des techniques comportant un certain degré d'automatisation sur la technique des filtres écrits à la main, mais nous préférons investir du temps dans la recherche de mécanismes généraux plutôt que dans une liste de filtres qui ne sera jamais complète.

Bien que l'utilisation du web pour privilégier les réponses redondantes ait fait bondir l'efficacité de certains systèmes [Clarke *et al.*, 2001][Brill *et al.*, 2001], nous ne prévoyons pas l'intégrer à QUANTUM. L'information sur le web ne provient pas nécessairement de sources crédibles. De plus, le web en tant que corpus de documents est peut-être approprié pour les questions à caractère général de TREC mais il ne convient pas lorsque le domaine d'application est particulier. Par exemple, un utilisateur désirant interroger une base de données médicales s'attend à ce que les sources d'information consultées par le système de question-réponse fassent autorité dans le domaine. D'autre part, dans un contexte de réponse automatique aux courriels reçus par le service à la clientèle d'une entreprise, le client désire obtenir le prix d'un article offert par l'entreprise et non le prix d'un article similaire affiché sur le site web d'un concurrent. Hors du cadre de TREC, le système *ExtrAns* [Aliod *et al.*, 1998] constitue un exemple d'application à un domaine particulier auquel QUANTUM pourrait aussi être adapté et pour lequel la nécessité de consulter le web est discutable : trouver dans des *man pages* la réponse à des questions portant sur l'utilisation de logiciels.

# CHAPITRE 5

# Conclusion

Nous avons décrit QUANTUM, un système de question-réponse dont le but est de trouver, dans un grand corpus de documents, la réponse à une question posée en langage naturel. Nous avons montré comment une combinaison de techniques de recherche d'information et d'analyse linguistique pouvait conduire à ce résultat. Puis, nous avons utilisé les standards de TREC-X pour évaluer les modules composant QUANTUM. Il s'est avéré que

- le module d'analyse de questions est efficace à 90 %, ce qui ne nécessite que des corrections mineures ;
- il est préférable d'utiliser Okapi pour l'étape de recherche d'information brute plutôt que d'élaborer notre propre algorithme de recherche ;
- des outils comme WordNet et Alembic peuvent servir le module d'extraction s'ils sont utilisés pour répondre à certains types de questions, mais ils sont nettement insuffisants à eux seuls pour atteindre les performances souhaitées ;
- le module de détection d'absence de réponse dans le corpus est efficace à 10 %, d'où la nécessité d'approfondir les tests si le module est réutilisé.

QUANTUM est actuellement développé dans l'esprit des conférences TREC, dans le but d'en arriver à un système efficace qui pourrait être intégré à un système de réponse automatique aux courriels. Les questions sur lesquelles QUANTUM a été testé sont courtes, factuelles, syntaxiquement bien formées et exemptes d'erreurs d'orthographe... ce qui n'est pas le cas de la majorité des courriels adressés au service à la clientèle des entreprises. Les difficultés sont telles qu'un système de question-réponse pourrait ne pas être approprié à cette tâche. Néanmoins, comme le suggèrent [Kosseim et Lapalme, 2001], une interface de question-réponse accessible aux clients directement sur le site web d'une entreprise permettrait déjà de réduire le volume de courriels. L'avantage de la question-réponse sur d'autres techniques tel le raisonnement par cas est que le corpus de documents peut être modifié du tout au tout sans que le système ne requière de modification (hormis bien sûr

le décodage du format des documents). Les possibilités qui s'offrent alors sont sans limite, allant de la consultation de bases de données médicales, financières ou encyclopédiques à l'aide en ligne de logiciels. L'interrogation de la base de données prendrait la forme d'une conversation entre l'utilisateur et le système, et les réponses factuelles feraient place à des réponses élaborées combinant analyse et synthèse. Les systèmes de question-réponse deviendront dans quelques années la nouvelle génération de moteurs de recherche.

# BIBLIOGRAPHIE


[Aberdeen *et al.*, 1995] J. Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson et M. Vilain. MITRE : Description of the Alembic System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, San Francisco, Californie, 1995. Morgan Kaufman Publishers.

[Aliod *et al.*, 1998] Diego Mollá Aliod, Jawad Berri et Michael Hess. A Real World Implementation of Answer Extraction. In *Proceedings of the 9th International Workshop on Database and Expert Systems Applications*, Natural Language and Information Systems Workshop (NLIS 98), Vienne, Autriche, août 1998. IEEE Computer Society.

[Allan, 2001] James Allan, éditeur. *Proceedings of HLT 2001 (Human Language Technology Conference)*, San Diego, Californie, mars 2001. HLT.

[Alpha *et al.*, 2001] Shamim Alpha, Paul Dixon, Ciya Liao et Changwen Yang. Oracle at TREC 10 : Filtering and Question-Answering. In [Voorhees et Harman, 2001a].

[Attardi et Burrini, 2000] Giuseppe Attardi et Cristian Burrini. The PISAB Question Answering System. In [Voorhees et Harman, 2000a].

[Brill *et al.*, 2001] Eric Brill, Jimmy Lin, Michele Banko, Susan Dumais et Andrew Ng. Data-Intensive Question Answering. In [Voorhees et Harman, 2001a].

[Burger *et al.*, 2001] John Burger *et al.* Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). Rapport technique, NIST, 2001.

[Carbonell *et al.*, 2000] Jaime Carbonell, Donna Harman, Eduard Hovy, Steve Maiorano, John Prange et Karen Sparck-Jones. Vision Statement to Guide Research in Question & Answering (Q&A) and Text Summarization. Rapport technique, NIST, 2000.

[Clarke *et al.*, 2000] C. L. A. Clarke, G. V. Cormack, D. I. E. Kisman et T. R. Lynam. Question Answering by Passage Selection. In [Voorhees et Harman, 2000a].

[Clarke *et al.*, 2001] C. L. A. Clarke, G. V. Cormack, T. R. Lynam, C. M. Li et G. L. McLearn. Web Reinforced Question Answering (Multitext Experiments for TREC 2001). In [Voorhees et Harman, 2001a].

[Elworthy, 2000] David Elworthy. Question Answering Using a Large NLP System. In [Voorhees et Harman, 2000a].

[Fellbaum, 1998] Christiane Fellbaum, éditeur. *WordNet : An Electronic Lexical Database.* The MIT Press, 1998.

[Ferret *et al.*, 2000] Olivier Ferret, Brigitte Grau, Martine Hurault-Plantet, Gabriel Illouz, Christian Jacquemin et Nicolas Masson. QALC – The Question-Answering System of LIMSI-CNRS. In [Voorhees et Harman, 2000a].

[Graesser *et al.*, 1992] Arthur Graesser, Natalie Person et John Huber. *Mechanisms that Generate Questions.* In [Lauer *et al.*, 1992].

[Harabagiu *et al.*, 2000] Sanda Harabagiu, Dan Moldovan, Marius Paşca, Rada Mihalcea, Mihai Surdeanu, Răzvan Bunescu, Roxana Gîrju, Vasile Rus et Paul Morărescu. FALCON : Boosting Knowledge for Answer Engines. In [Voorhees et Harman, 2000a].

[Harabagiu *et al.*, 2001] Sanda Harabagiu, Dan Moldovan, Marius Paşca, Mihai Surdeanu, Rada Mihalcea, Roxana Gîrju, Vasile Rus, Finley Lăcătuşu, Paul Morărescu et Răzvan Bunescu. Answering Complex, List and Context Questions with LCC's Question-Answering Server. In [Voorhees et Harman, 2001a].

[Hovy *et al.*, 2000] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk et Chin-Yew Lin. Question Answering in Webclopedia. In [Voorhees et Harman, 2000a].

[Hovy *et al.*, 2001a] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin et Deepak Ravichandran. Toward Semantics-Based Answer Pinpointing. In [Allan, 2001].

[Hovy *et al.*, 2001b] Eduard Hovy, Ulf Hermjakob et Chin-Yew Lin. The Use of External Knowledge in Factoid QA. In [Voorhees et Harman, 2001a].

[Ittycheriah *et al.*, 2000] Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu et Adwait Ratnaparkhi. IBM's Statistical Question Answering System. In [Voorhees et Harman, 2000a].

[Ittycheriah *et al.*, 2001] Abraham Ittycheriah, Martin Franz et Salim Roukos. IBM's Statistical Question Answering System – TREC-10. In [Voorhees et Harman, 2001a].

[Kosseim et Lapalme, 2001] Leila Kosseim et Guy Lapalme. Analyse des dossiers de BCE et description du projet de réponse au courriel de BCE. Rapport technique, RALI/DIRO, Université de Montréal, 2001.

[Laszlo *et al.*, 2000] Michael Laszlo, Leila Kosseim et Guy Lapalme. Goal-Driven Answer Extraction. In [Voorhees et Harman, 2000a].

[Laszlo, 2000] Michael Laszlo. Extraction de réponses par méthodes de surface. Rapport technique, RALI/DIRO, Université de Montréal, 2000.

[Lauer *et al.*, 1992] Thomas W. Lauer, Eileen Peacock et Arthur C. Graesser, éditeurs. *Questions and Information Systems.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1992.

[Lehnert, 1978] Wendy G. Lehnert. *The Process of Question Answering : A Computer Simulation of Cognition.* Halsted Press, Hillsdale, New Jersey, 1978.

[Moldovan *et al.*, 1999] Dan Moldovan, Sanda Harabagiu, Marius Paşca, Rada Mihalcea, Richard Goodrum, Roxana Gîrju et Vasile Rus. LASSO : A Tool for Surfing the Answer Net. In [Voorhees et Harman, 1999].

[Prager *et al.*, 2000] John Prager, Eric Brown, Dragomir R. Radev et Krzysztof Czuba. One Search Engine or Two for Question-Answering. In [Voorhees et Harman, 2000a].

[Prager *et al.*, 2001] John Prager, Dragomir Radev et Krzysztof Czuba. Answering What-Is Questions by Virtual Annotation. In [Allan, 2001].

[Robertson et Walker, 1999] S. E. Robertson et S. Walker. Okapi/Keenbow at TREC-8. In [Voorhees et Harman, 1999].

[Scott et Gaizauskas, 2000] Sam Scott et Robert Gaizauskas. University of Sheffield TREC-9 Q & A System. In [Voorhees et Harman, 2000a].

[Soubbotin, 2001] M. M. Soubbotin. Patterns of Potential Answer Expressions as Clues to the Right Answers. In [Voorhees et Harman, 2001a].

[Voorhees et Harman, 1999] E. M. Voorhees et D. K. Harman, éditeurs. *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, Gaithersburg, Maryland, novembre 1999. NIST.

[Voorhees et Harman, 2000a] E. M. Voorhees et D. K. Harman, éditeurs. *Proceedings of the 9th Text REtrieval Conference (TREC-9)*, Gaithersburg, Maryland, novembre 2000. NIST.

[Voorhees et Harman, 2000b] Ellen M. Voorhees et Donna Harman. Overview of the Ninth Text REtrieval Conference (TREC-9). In [Voorhees et Harman, 2000a].

[Voorhees et Harman, 2001a] E. M. Voorhees et D. K. Harman, éditeurs. *Notebook Proceedings of the 10th Text REtrieval Conference (TREC-X)*, Gaithersburg, Maryland, novembre 2001. NIST.

[Voorhees et Harman, 2001b] Ellen M. Voorhees et Donna Harman. Overview of TREC 2001. In [Voorhees et Harman, 2001a].

[Woods *et al.*, 2000] W. A. Woods, Stephen Green, Paul Martin et Ann Houston. Halfway to Question Answering. In [Voorhees et Harman, 2000a].

# ANNEXE A

## Classification proposée par [Graesser *et al.*, 1992]

Le tableau de la page suivante reproduit la classification des questions proposées par [Graesser *et al.*, 1992]. L'intérêt de cette classification est qu'elle couvre toutes les formes de questionnement et que, de plus, elle n'a pas été élaborée en fonction d'une quelconque application informatique. Elle peut servir, en quelque sorte, de référence théorique.

Afin de se rapporter à cette référence, nous avons pris soin, à la section 4.3, de regrouper lorsque possible la description de nos fonctions d'extraction selon les concepts proposés ici. Par exemple, les fonctions *cardinalité*($\rho$, $\varphi$) et *mesure*($\rho$, $\varphi$) sont des fonctions de *quantification* ; la fonction *attribut*($\rho$, $\varphi$) relève du concept *feature specification* ; et les fonctions *personne*($\rho$), *temps*($\rho$), *lieu*($\rho$) et *objet*($\rho$) sont des fonctions de *concept completion*. La fonction *définition*($\rho$) correspond bien sûr au concept de *definition* alors que son contraire, *spécialisation*($\rho$, $\varphi$), correspond au concept *example*. Si elle avait été implémentée, la fonction *raison*($\rho$) engloberait *causal antecedent*, *causal consequence* et *goal orientation*. La fonction *manière*($\rho$) serait quant à elle associée à *enablement* et *instrumental/procedural*.

| Question | Spécification abstraite | Exemple |
|---|---|---|
| Verification | Is a fact true ?<br>Did an event occur ? | Is an F-test a type of statistic ?<br>Dit it rain yesterday ? |
| Comparison | How is X similar to Y ?<br>How is X different from Y ? | In what way is Florida similar to China ?<br>How is an F-test different from a t-test ? |
| Disjunctive | Is X or Y the case ?<br>Is X, Y, or Z the case ? | Do the mountains increase or decrease the rain in Oregon ?<br>Did he order chicken, beef, lamb or fish ? |
| Concept completion | Who ? What ? When ? Where ?<br>What is the referent of a noun argument slot ? | Where are the large population densities in North America ?<br>Who wrote the song ?<br>What did the child steal ? |
| Definition | What does X mean ? What is the superordinate category and some properties of X ? | What is a factorial design ?<br>What does interaction mean ? |
| Example | What is an example of X ? What is a particular instance of the category ? | What is an example of an ordinal scale ?<br>What experiment supports this claim ? |
| Interpretation | How is a particular event interpreted or summarized ?<br>How is a pattern of information interpreted or summarized ? | Does the graph show a main effect for "A" ?<br>What happened yesterday ? |
| Feature specification | What qualitative attributes does entity X have ?<br>What is the value of a qualitative variable ? | What is George like ?<br>What color is the dog ? |
| Quantification | What is the value of a quantitative variable ?<br>How much ? How many ? | How many rooms are in the house ?<br>How much profit was made last year ? |
| Causal antecedent | What caused some event to occur ?<br>What state or event causally led to an event or state ? | How does warm air get to Ireland ?<br>Why is the kite going backwards ? |
| Causal consequence | What are the consequences of an event or state ?<br>What causally unfolds from an event of state ? | What happens to the warm winds when they reach the mountains ?<br>What are the consequences of double-digit inflation ? |
| Goal orientation | What are the motives behind an agent's action ?<br>What goals inspired an agent to perform an action ? | Why did Roger move to Chicago ?<br>What was the purpose of the city's cutting taxes ? |
| Enablement | What object or resource enables an agent to perform an action ? | What device allows you to measure an earthquake ?<br>What do I need to bake this fish ? |
| Instrumental/ Procedural | How does an agent accomplish a goal ?<br>What instrument or body part is used when an agent performs an action ?<br>What plan of action accomplishes an agent's goal ? | How does a person perform long division ?<br>How do you move a mouse on a computer ? |
| Expectational | Why did some expected event not occur ? | Why wasn't there a war in Iraq ?<br>Why doesn't this doll have a mouth ? |
| Judgmental | The questioner wants the answerer to judge an idea or to give advice on what to do. | What do you think about the new taxes ?<br>What should I do to stop the fight ? |
| Assertion | The speaker expresses that he or she is missing some information. | I don't understand what this message on the computer means.<br>I need to know how to get to the Newark airport. |
| Request/ Directive | The speaker directly requests that the listener supply some information. | Please tell me how to get a printout of this file. |

# ANNEXE B

## Classification proposée par [Moldovan *et al.*, 1999]

Nous présentons à la page suivante la classification des questions proposée par l'équipe de la Southern Methodist University [Moldovan *et al.*, 1999] à TREC-8. Le tableau met en évidence qu'un mot-question peut désigner plusieurs types d'entités : par exemple, *which* peut référer à une personne (*which-who*), à un lieu (*which-where*), à un temps (*which-when*) ou à un objet (*which-what*). Nous avons établi, à la dernière colonne, de quelle façon les questions données en exemple auraient théoriquement été analysées par QUANTUM. Il est à noter que plusieurs questions sont d'abord analysées comme des questions de spécialisation mais avant de passer à l'extraction des candidats, QUANTUM prend soin de vérifier si la nature du focus justifie l'appel d'une fonction d'extraction plus précise, notamment $personne(\rho)$, $lieu(\rho)$ ou $temps(\rho)$.

| Classe | Sous-classe | Exemple | Analyse de QUANTUM |
|---|---|---|---|
| what | basic-what | What was the monetary value of the Nobel Peace Prize in 1989? | *spécialisation*($\rho$, *monetary value*) |
| | what-who | What costume designer decided that Michael Jackson should only wear one glove? | *spécialisation*($\rho$, *costume designer*) |
| | what-when | In what year did Ireland elect its first woman president? | *spécialisation*($\rho$, *year*) |
| | what-where | What is the capital of Uruguay? | *spécialisation*($\rho$, *capital*) |
| who | | Who is the author of the book "The Iron Lady : A Biography of Margaret Thatcher" ? | *personne*($\rho$) |
| how | basic-how | How did Socrates die? | *manière*($\rho$) |
| | how-many | How many people died when the Estonia sank in 1993? | *cardinalité*($\rho$, *people*) |
| | how-long | How long does it take to travel from Tokyo to Niigata? | *attribut*($\rho$, *long*) |
| | how-much | How much did Mercury spend on advertising in 1993? | *mesure*($\rho$, *money*) |
| | how-much-<modifier> | How much stronger is the new vitreous carbon material invented by the Tokyo Institute of Technology compared with the material made from cellulose? | *attribut*($\rho$, *stronger*) |
| | how-far | How far is Yaroslav from Moscow? | *attribut*($\rho$, *far*) |
| | how-tall | How tall is Mt. Everest? | *attribut*($\rho$, *tall*) |
| | how-rich | How rich is Bill Gates? | *attribut*($\rho$, *rich*) |
| | how-large | How large is the Arctic Refuge to preserve unique wildlife and wilderness value on Alaska's north coast? | *attribut*($\rho$, *large*) |
| where | | Where is Taj Mahal? | *lieu*($\rho$) |
| when | | When did the Jurassic Period end? | *temps*($\rho$) |
| which | which-who | Which former Klu Klux Klan member won an elected office in the U.S. ? | *spécialisation*($\rho$, *Klu Klux Klan member*) |
| | which-where | Which city has the oldest relationship as sister-city with Los Angeles? | *spécialisation*($\rho$, *city*) |
| | which-when | In which year was New Zealand excluded from the ANZUS alliance? | *spécialisation*($\rho$, *year*) |
| | which-what | Which Japanese car maker had its biggest percentage of sale in the domestic market? | *spécialisation*($\rho$, *Japanese car maker*) |
| name | name-who | Name the designer of the show that spawned millions of plastic imitations, known as "jellies" ? | *spécialisation*($\rho$, *designer*) |
| | name-where | Name a country that is developing a magnetic levitation railway system? | *spécialisation*($\rho$, *country*) |
| | name-what | Name a film that has won the Golden Bear in the Berlin Film Festival? | *spécialisation*($\rho$, *film*) |
| why | | Why did David Koresh ask for a word processor? | *raison*($\rho$) |
| whom | | Whom did the Chicago Bulls beat in the 1993 championship? | *personne*($\rho$) |