Université de Neuchâtel
Faculté des Sciences

Cette thèse intitulée

Extraction automatique d'information : génération de résumé et question-réponse

a été présentée avec succès le 18 mars 2005 par

Laura Perret

pour l'obtention du grade de Docteur ès Sciences

Composition du jury

Prof. Jacques Savoy
Directeur de thèse
Université de Neuchâtel, Suisse

Prof. Kilian Stoffel
Université de Neuchâtel, Suisse

Prof. Eric Wehrli
Université de Genève, Suisse

Dr. Patrice Bellot
Université d'Avignon, France

IMPRIMATUR POUR LA THESE

EXTRACTION AUTOMATIQUE D'INFORMATION : GENERATION DE RESUME ET QUESTION-REPONSE

Laura PERRET

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel, sur le rapport des membres du jury

MM. J. Savoy (directeur de thèse), K. Stoffel, E. Wehrli (Genève) et P. Bellot (Avignon)

autorise l'impression de la présente thèse.

Neuchâtel, le 22 mars 2005

La doyenne:

Prof. M. Rahier

Thèse de doctorat

Extraction automatique d'information : génération de résumé et question-réponse

Laura Perret

Institut d'informatique
Université de Neuchâtel
Pierre-à-Mazel 7
CH-2000 Neuchâtel, Suisse
Laura.Perret@unine.ch

RÉSUMÉ. *Dans cette thèse, nous abordons diverses techniques d'extraction automatique d'information, à savoir la génération de résumé et la question-réponse. Dans la première partie, nous avons implémenté et évalué plusieurs méthodes de génération de résumé à partir d'articles médicaux. L'utilisation de la régression logistique s'est révélée produire les meilleurs résultats. Nous avons également combiné cette méthode avec une méthode de classification proposée par une autre équipe de recherche, obtenant une amélioration significative des performances observées. Dans la deuxième partie, nous avons développé un système de question-réponse pour le français. Dans ce but, nous avons combiné un modèle probabiliste de la recherche d'information classique avec une approche linguistique basée sur l'analyse syntaxique. Puis, nous avons exploité des ressources de traduction automatique afin de prendre en charge des questions exprimées dans d'autres langues que le français. Le système ainsi conçu a été soumis à deux campagnes d'évaluation et a obtenu des résultats encourageants.*

ABSTRACT. *In this dissertation, we tackle two aspects of the information retrieval field, namely text summarization and question answering. The first part is devoted to the implementation and the evaluation of several methods for summarization of medical articles. The logistic regression was found to produce the best results. We also combine this method with a classification approach provided by another research group obtaining a significant improvement of performance. In the second part, we developed a question answering system for the French language. To do so, we combined a probabilistic model from the classical information retrieval field with a linguistic approach based on syntactic analysis. Then, we took advantage of automatic translation resources in order to allow questions formulated in other languages than French. This system took part in two evaluation campaigns and produced encouraging results.*

MOTS-CLÉS : *extraction d'information, génération de résumé, question-réponse, recherche d'information.*

KEYWORDS:*information extraction, text summarization, question answering, information retrieval.*

Remerciements

Le travail ayant abouti à la présente thèse n'aurait pas été possible sans le concours de nombreuses personnes et institutions. J'aimerais ici toutes les remercier sans toutefois pouvoir être exhaustive.

Je voudrais commencer par remercier mes parents et amis qui m'ont encouragée et soutenue sans réserve pendant toute la durée de cette thèse.

Puis, j'aimerais particulièrement remercier mon directeur de thèse, le professeur Jacques Savoy, qui m'a donné l'opportunité d'effectuer cette thèse. Il m'a aiguillée pendant toute la durée de ce projet, m'a prodigué des conseils avisés et m'a encouragée régulièrement.

La présente thèse a été évaluée par le jury de thèse composé des professeurs Jacques Savoy (Institut d'informatique, Université de Neuchâtel), Kilian Stoffel (Institut d'informatique, Université de Neuchâtel) et Eric Wehrli (Laboratoire d'Analyse et de Technologie du Langage, Université de Genève) ainsi que du Dr. Patrice Bellot (maître de conférences, Laboratoire Informatique d'Avignon, Université d'Avignon). Je tiens à leur remercier pour le temps consacré à l'expertise de cette thèse. Leurs remarques et suggestions ont contribué à l'élaboration de la version révisée de ce rapport.

Je tiens à témoigner ma reconnaissance à mon collègue et ami Pierre-Yves Berger. Nos sujets de thèse étant liés, nous avons été amenés à collaborer étroitement. Outre la pertinence de ses analyses, il a également été une source d'inspiration et de soutien tout au long de ce travail.

Je remercie également C. Buckley de SabIR pour m'avoir donné l'opportunité d'utiliser le système SMART ainsi que E. Wehrli du LATL pour la disponibilité de l'outil FIPS.

Enfin, cette recherche a été partiellement soutenue par le Fonds national suisse de la recherche scientifique avec la bourse 21-66 742.01. L'Université de Neuchâtel a également contribué en mettant à ma disposition un poste partiel d'assistante.

TABLE DES MATIERES

1. INTRODUCTION 1

1.1 LA RECHERCHE D'INFORMATION 2
1.2 PROBLEMATIQUES 4
1.3 OBJECTIFS 6
1.4 CONTRIBUTION 7
1.5 PLAN DE LA THESE 8

2. GENERATION DE RESUME 9

2.1 INTRODUCTION 9
2.2 PERSPECTIVES DE LA GENERATION DE RESUME 10
2.3 EVALUATION DE LA GENERATION DE RESUME 11
2.4 ETAT DE L'ART 11
2.5 MODELE PROPOSE 12
2.6 EVALUATION 20
2.7 CONCLUSION ET PERSPECTIVES 24

3. QUESTION-REPONSE 25

3.1 INTRODUCTION 25
3.2 L'ARCHITECTURE DES SYSTEMES DE QUESTION-REPONSE 25
3.3 EVALUATION DES SYSTEMES DE QUESTION-REPONSE 26
3.4 ETAT DE L'ART 27
3.5 SYSTEME PROPOSE 30
3.6 EVALUATION 50
3.7 CONCLUSION ET PERSPECTIVES 72

4. CONCLUSION 73

4.1 CONTRIBUTIONS 73
4.2 LIMITES 74
4.3 PERSPECTIVES 74

BIBLIOGRAPHIE 77

ANNEXES 83

1. Introduction

Avec l'avènement d'Internet, nous disposons d'une quantité si gigantesque et variée de données qu'il devient difficile d'y trouver l'information souhaitée. Face à cette situation, de nombreux outils d'assistance à la recherche ont vu le jour. Néanmoins, beaucoup de difficultés interviennent dans la recherche d'information. Premièrement, le support de l'information peut se présenter dans des formats différents et parfois incompatibles (texte, image, audio, vidéo, multimédia). A cela vient s'ajouter la dimension linguistique dès qu'il s'agit de comprendre et interpréter des informations ou que celles-ci sont exprimées dans diverses langues.

Dans le présent rapport, nous commencerons par poser le cadre de la recherche d'information en évoquant quelques problématiques qui la caractérisent. Puis, nous nous intéresserons à l'extraction automatique d'information que nous analyserons selon deux thématiques particulières que sont la génération de résumé et la question-réponse.

La génération de résumé sera restreinte au contexte médical. Il s'agit-là de produire des résumés d'articles médicaux relativement courts qui pourraient servir de description dans une banque de données spécialisée, voire à remplir des attributs dans une banque de données. A cet effet, nous avons proposé plusieurs méthodes qui ont été évaluées en comparant les résultats obtenus avec les résumés créés manuellement par des experts du domaine. Nous avons également combiné notre approche avec celle d'une autre équipe de recherche pour voir si l'on pouvait améliorer la performance, ce qui s'est avéré être le cas.

Pour la question-réponse, nous nous sommes intéressés à la langue française. Dans un premier temps, nous avons développé un système permettant de répondre à des questions formulées en français en effectuant la recherche dans une collection de documents rédigés en français. Puis, nous avons étendu notre système pour lui permettre de traiter des questions exprimées dans d'autres langues européennes tout en conservant la collection française comme cible de la recherche. Pour concevoir notre système, nous avons fait intervenir un modèle probabiliste pour le dépistage des documents intéressants. Puis, nous avons eu recours à une approche linguistique basée sur l'analyse syntaxique pour l'appariement. Enfin, nous avons fait usage de ressources libres de traduction automatique pour franchir la barrière des langues.

Dans la suite de ce chapitre introductif, nous décrirons la recherche d'information ainsi que quelques-unes des problématiques qui lui sont propres. Puis, nous évoquerons les objectifs que nous nous étions fixés ainsi que les contributions que nous avons apportées. Enfin, nous terminerons ce chapitre par la présentation de la structure de cette thèse.

1.1 La recherche d'information

1.1.1 Introduction

La recherche d'information (*information retrieval*) est un domaine qui a vu le jour après la Deuxième Guerre Mondiale au début des années 1950. L'objectif était d'automatiser la recherche d'informations grâce aux ordinateurs. Le premier aspect traité fut l'indexation des documents pour permettre de les retrouver. La recherche d'information (RI) s'est affinée à travers une forte tradition expérimentale.

Parmi les événements marquants de l'histoire de la RI, on peut citer le projet Cranfield (1957-1967) [CLE 67] dont l'objectif était de comparer l'efficacité des divers modèles d'indexation et de recherche de documents en termes de précision et de rappel, mesures encore utilisées aujourd'hui (voir annexe A).

Le projet SMART (1961-1965) [SAL 71] a proposé une série d'expérimentations portant notamment sur l'architecture des systèmes de RI, l'indexation et la rétroaction de pertinence (*relevance feedback*). SMART fut réécrit et amélioré par la suite et reste un système encore utilisé de nos jours pour les expérimentations.

La série de conférences TREC[1] (*Text REtrieval Conference*) (1992-) [HAR 92] propose annuellement diverses évaluations des méthodes et systèmes de RI. Les aspects évalués évoluent d'année en année au gré des intérêts de la communauté scientifique et des organes de subvention. Parmi les plus caractéristiques figurent la RI classique, le filtrage d'information, la RI multi-lingue, les systèmes de question-réponse, la RI multimédia, etc.

Au fil du temps, la granularité des informations à produire a augmenté. Si au début de son histoire la RI se proposait de retrouver de l'information pertinente au sens large, par la suite, elle s'est intéressée au dépistage de documents comme l'illustre la création des moteurs de recherche sur le Web. Puis, constatant un manque de précision dans la satisfaction des besoins des usagers et un plafonnement de performance en recherche documentaire, la RI a évolué pour rechercher des faits précis tels que des chiffres, des références d'articles scientifiques ou des définitions. Cela a conduit à l'élaboration de systèmes de question-réponse.

[1] Voir le site http://trec.nist.gov/

1.1.2 Concepts élémentaires en RI

Pour bien comprendre la RI, il convient de décrire les concepts fondamentaux qui s'y rapportent.

Un document

Un document est l'entité atomique qui contient l'information. Il peut s'agir d'un livre, d'un article scientifique, d'un paragraphe, d'un fragment de texte, etc. Un document contient un ou plusieurs éléments d'information sous formes variées (texte, son, image, vidéo, multimédia).

Une requête

Une requête est l'expression du besoin d'information de l'usager. Elle peut être exprimée en langue naturelle ou par un ensemble de mots-clés articulés éventuellement par des opérateurs logiques.

La pertinence

La pertinence d'un document mesure sa faculté à répondre au besoin d'information exprimé par l'usager sous forme de requête. Les mesures de performances utilisées sont décrites dans l'annexe A.

Un index

Un index est une structure qui permet essentiellement d'associer à chaque terme d'indexation la liste des documents contenant ce terme. Un index peut contenir des informations supplémentaires telles que le poids d'un terme dans un document, la position d'un terme dans un document ou son occurrence dans telle partie logique.

L'appariement

L'appariement est l'opération qui met en relation une requête et un document qui doit y répondre. Selon les critères déterminés, un document sera jugé pertinent ou non pertinent pour une requête donnée. Suivant la fonction d'appariement choisie, cette pertinence peut être absolue (un document pouvant être pertinent vs. non pertinent) ou échelonnée (un ordonnancement des documents par probabilité de pertinence ou en fonction de leur similarité avec la requête).

Un modèle de RI

Un modèle de RI décrit d'une part l'indexation des requêtes et documents et d'autre part l'appariement entre requêtes et documents. Plus précisément, il s'agit de définir la représentation des documents et requêtes dans l'index ainsi que la fonction

d'appariement qui mesurera la similarité des représentations des requêtes et documents. Il existe quatre classes de modèles de RI, à savoir le modèle booléen, le modèle vectoriel, le modèle probabiliste et le modèle logique. Une description de ces modèles est proposée dans [IHA 04].

1.2 Problématiques

De nombreuses problématiques occupent la communauté de RI au sens large. Voici une brève description des principales thématiques.

La recherche de documents

La recherche documentaire vise à retrouver de l'information sous forme de documents à partir des requêtes formulées par l'utilisateur [SAL 71]. L'information recherchée peut être dans divers formats (texte, image, vidéo) comme dans diverses langues. Les collections de documents deviennent de plus en plus grandes, particulièrement si l'on travaille avec le Web.

La classification et catégorisation de documents

La classification vise à organiser un ensemble de textes dans des classes homogènes. Elle est généralement non supervisée par un usager. La catégorisation vise à associer à chaque texte une catégorie existante. Elle est généralement supervisée par un usager [YAN 99].

Le filtrage d'information

Le filtrage d'information sélectionne des documents provenant de corpus inconnus *a priori*. Il traite des documents provenant de sources dynamiques et décide pour chaque document s'il répond au besoin d'information de l'utilisateur. Il s'agit donc d'un assistant qui permet à l'utilisateur de recevoir des documents pertinents de diverses sources dynamiques sur la base du profil utilisateur qu'il a défini [BEL 92].

Le *data mining* (*text mining* ou fouille de textes)

Le *data mining* représente l'ensemble des techniques d'exploration de données afin d'en tirer des connaissances présentées à l'utilisateur pour l'aide à la prise de décision [MEN 96].

Le Web sémantique

L'objectif du Web sémantique est de modifier la répartition du travail entre l'homme et la machine en favorisant l'accès par le sens aux ressources du Web. Cela se concrétise par l'ajout aux données de métadonnées qui sont des connaissances formalisées utilisables dans des traitements automatiques [BER 01], [DUM 00].

L'extraction d'information

L'extraction d'information vise à permettre l'identification automatique d'entités, relations ou événements de types prédéfinis à partir de texte libre. Elle fait entre autre intervenir la reconnaissance d'entités nommées [COW 96], MUC[2] (*Message Understanding Conference*).

La génération de résumé

La génération de résumé permet de produire un résumé à partir d'un ou plusieurs documents. La couverture peut être globale, c'est-à-dire traitant de tout le contenu, ou partielle, ne s'intéressant qu'à un aspect du contenu [MAN 01].

La question-réponse

Les systèmes de question-réponse vont au-delà de la recherche de documents en ce sens qu'ils visent à fournir la réponse précise à la question posée plutôt qu'une liste de documents susceptibles de contenir cette réponse [VOO 04]. La réponse doit être extraite ou recomposée à partie de documents.

Le multilinguisme

Le multilinguisme implique des systèmes utilisant le traitement linguistique pour reconnaître et normaliser les termes apparaissant dans les textes. Les requêtes formulées dans différentes langues s'adressent à un corpus constitué de documents exprimés dans plusieurs langues. Cela nécessite des ressources linguistiques pour chaque langue avec un fort intérêt pour les techniques génériques, indépendantes des langues [GRE 00], [CHE 04] [3].

La visualisation d'information

L'objectif de la visualisation d'information est d'exploiter les caractéristiques du dispositif visuel humain afin de faciliter la manipulation et l'interprétation des données. Cela s'applique par exemple à l'exploration rapide d'un ensemble d'informations ou à la classification interactive des informations [CAR 99].

[2] Voir le site http://www.muc.saic.com/

[3] Voir le site de CLEF http://clef.iei.pi.cnr.it/
et celui de NTCIR http://research.nii.ac.jp/ntcir/

1.3 Objectifs

Parmi les diverses problématiques de la recherche d'information présentées ci-dessus, nous avons choisi de nous intéresser à deux d'entre-elles. Il s'agit de la génération de résumé et de la question-réponse.

1.3.1 Génération de résumé

Dans le domaine médical ou juridique par exemple, la quantité de publications est telle que les experts ont de plus en plus de mal à s'y retrouver. Afin d'aider les experts à dépister des informations telles que la description des implications des gènes dans certaines maladies pour les uns ou des éléments de jurisprudence pour les autres, des banques de données spécialisées ont été créées. Pour faciliter la diffusion du savoir et de la technologie, les articles sont d'abord indexés en prenant en considération des mots-clés ou des courts résumés qui permettent en un coup d'oeil de déterminer si l'article correspondant est susceptible de répondre au besoin de l'usager.

C'est dans cette optique que l'objectif de la première partie de la thèse était de proposer un modèle de génération automatique de résumé. Les limitations formulées étaient les suivantes :

- la collection de documents serait constituée d'articles médicaux ;
- le résumé serait global, c'est-à-dire portant sur tout le document ;
- le résumé serait court pour permettre son utilisation comme description dans une banque de données médicale.

Afin de vérifier que l'objectif serait atteint, le modèle proposé devait être soumis à une évaluation dans une ou plusieurs conférences du domaine et présenter une bonne performance. Cette performance serait naturellement dépendante des mesures adoptées pour les évaluations considérées.

1.3.2 Question-réponse

L'anglais était jadis considéré comme la langue universelle, particulièrement sur la Toile. Par conséquent, c'est en priorité pour cette langue que la question-réponse a connu ses premiers développements à la fin des années 1990. Par la suite, des études ont montré que la proportion de pages Web exprimées en anglais tendait à décroître au profit d'autres langues telles que le chinois [GLO 01]. De plus, la nécessité de considérer plusieurs langues simultanément s'est fait sentir dans des institutions européennes ou dans les gouvernements de pays multiculturels comme la l'Union européenne, l'Inde ou la Suisse. Dès lors, la question-réponse devait évoluer pour traiter de nouvelles langues. Alors que la conférence TREC propose depuis 1999 l'évaluation de systèmes de question-réponse pour l'anglais, des conférences telles que CLEF[4] (*Cross-Language Evaluation Forum*) pour les langues

[4] Voir le site http://clef.iei.pi.cnr.it/

européennes, NTCIR[5] (*NII-NACSIS Test Collection for IR Systems*) pour les langues asiatiques et EQueR (*Evaluation des systèmes de Question-réponse*) pour la langue française ont proposé de nouvelles évaluations pour différentes combinaisons de langues.

Dans ce contexte, l'objectif de la deuxième partie de la thèse était de développer un système automatique de question-réponse pour la langue française. Puis, il s'agissait d'étendre son champ d'application pour prendre en charge des questions exprimées dans différentes langues. Toutefois, trois limitations ont été formulées :

- le système devrait être capable de traiter des questions factuelles fermées et précises excluant donc les questions ouvertes ou de définition ;

- la recherche se ferait dans des collections de documents textuels écartant les autres supports multimédia ;

- seules des langues européennes seraient considérées.

Afin de vérifier que l'objectif serait atteint, le système développé devait être soumis à une évaluation dans une ou plusieurs conférences du domaine et présenter une bonne performance. Cette performance serait naturellement dépendante des mesures adoptées pour les évaluations considérées.

1.4 Contribution

1.4.1 Génération de résumé

Dans la première partie de cette thèse, nous nous sommes intéressés à la génération de résumé d'articles médicaux. Les résumés, de petite taille, pourraient servir de description dans une banque de données médicale. Pour ce faire, nous avons implémenté divers modèles exploitant des informations statistiques sur les occurrences de termes importants. Le modèle le plus performant s'est avéré être celui qui fait usage de la régression logistique afin de prédire quelles parties de l'article en constitueraient un bon résumé. Nos modèles ont été évalués dans la piste génomique de la campagne d'évaluation TREC-2003 [HER 04] et ont montré des résultats intéressants. Ils ont fait l'objet de trois publications [SAV 04a], [PER 04b] et [RUC 05].

[5] Voir le site http://research.nii.ac.jp/ntcir/

1.4.2 Question-réponse

La deuxième partie de cette thèse a été consacrée au développement d'un système de question-réponse pour le français. A cette fin, nous avons combiné un modèle probabiliste issu de la recherche d'information classique avec une approche linguistique basée sur l'analyse syntaxique des textes. Le système ainsi obtenu a participé à la campagne EQueR[6] (Evaluation en Question-Réponse) organisée en 2004 par EVALDA[7] (EVALuation à ELDA (*Evaluation and Language resources Distribution Agency*)), projet financé par le Ministère français en charge de la Recherche dans le cadre du programme Technolangue[8], ayant pour objectif la mise en oeuvre d'une infrastructure dédiée à l'évaluation des technologies de la langue en France, pour la langue française. Le système a présenté des performances intéressantes pour certains types de questions. Par la suite, nous l'avons adapté afin de permettre la prise en compte de questions formulées dans d'autres langues que le français alors que la collection de documents restait en français. Notre système ainsi modifié a concouru lors de la campagne d'évaluation CLEF 2004 et nous a permis d'estimer la perte de performance due à la phase de traduction. Ce travail a fait l'objet d'une publication [PER 04a].

1.5 Plan de la thèse

La présente thèse traite de deux problématiques de la recherche d'information, à savoir la génération de résumé et les systèmes de question-réponse. La suite de ce rapport est organisée comme suit.

Le deuxième chapitre est dédié à la génération de résumé. Nous commençons par décrire la problématique puis nous évoquons les approches existantes. Ensuite, nous fixons le cadre en énonçant les limites que nous nous sommes fixées. Nous continuons par la présentation des modèles proposés ainsi que leur évaluation. Enfin, nous tirons des conclusions et élaborons quelques perspectives futures.

Le troisième chapitre est consacré à la question-réponse. Nous introduisons d'abord la problématique considérée et passons en revue les principales approches existantes. Puis, nous présentons le système proposé ainsi que les limites que nous nous sommes fixées. Nous poursuivons par l'évaluation du système avant de conclure par quelques perspectives futures.

Le quatrième chapitre est l'occasion de tirer des conclusions globales sur la présente contribution.

---

[6] Voir le site d'EQueR http://www.elda.org/article118.html
[7] Voir le site d'EVALDA http://www.elda.org/rubrique69.html
[8] Voir le site de Technolangue http://www.technolangue.net/article20.html

2. Génération de résumé

2.1 Introduction

Deux disciplines scientifiques en particulier posent continuellement des défis aux chercheurs en recherche d'information : il s'agit du droit et de la médecine. Dans le premier cas, il s'agit de permettre un accès non seulement aux lois ou règlements mais à un ensemble représentatif de la jurisprudence, voire de la doctrine, autorisant une lecture éclairée des décisions des tribunaux de diverses instances. La gestion d'un volume considérable d'informations constitue un défi majeur pour les systèmes documentaires juridiques. Par exemple, le système Thomson West[9] portant essentiellement sur le *common law* doit permettre un accès précis à un volume dépassant les 7 Tb pour ses quelques 700 000 usagers. Plus près de nous, le multilinguisme inhérent au droit européen[10] soulève une difficulté supplémentaire. En médecine, terme couvrant dans ce contexte également de nombreux domaines connexes, on a désiré très tôt offrir un accès à un nombre important d'articles scientifiques sélectionnés. Le système sous-jacent nommé Medline[11] couvre actuellement plus de 4 600 revues scientifiques publiées dans plus de 70 pays. Dans ce second cas, le volume impressionnant des sources documentaires à disposition soulève de réels défis.

Certes, le thème central de la recherche d'information concerne le dépistage efficace de documents correspondant aux souhaits d'un usager. Ainsi, plusieurs campagnes d'évaluation comme TREC, CLEF ou NTCIR ont été lancées ces dernières années afin de faciliter l'évaluation comparative des systèmes documentaires et le transfert technologique des centres de recherche vers l'industrie. La question générale à laquelle nous souhaitons répondre est la suivante : « Ayant dépisté un article particulier, l'ordinateur est-il capable d'en fournir un résumé ? »

Cette problématique [MAN 99] soulève tout de suite plusieurs interrogations liées au degré de couverture désiré, à la taille du résumé à générer, aux choix lexicaux les plus appropriés ainsi qu'à la mesure de qualité à laquelle on pourrait recourir. Afin de limiter quelque peu l'éventail de ces possibilités, la campagne d'évaluation TREC-2003 a proposé une piste d'extraction d'information liée à la

[9] Voir le site http://west.thomson.com/
[10] Voir le site http://europa.eu.int/eur-lex/
[11] Medline est géré par la National Library of Medecine (NLM) et est accessible sur http://www.nlm.nih.gov/

génétique [HER 04]. En effet, on connaît depuis quelques années un fort accroissement du nombre d'articles scientifiques publiés dans cette discipline comme en médecine en général. Par exemple, en douze mois, plus de 500 000 articles ont été répertoriés dans le système Medline. Afin de gérer un tel volume, la mise au point d'outils automatiques ou semi-automatiques s'avère nécessaire.

2.2 Perspectives de la génération de résumé

Mitkov [MIT 03a] a proposé une définition d'un résumé. On peut traduire cette définition ainsi : « un résumé est un texte produit à partir d'un ou plusieurs textes, contenant une partie significative de l'information présente dans le(s) texte(s) original(aux), dont la taille ne dépasse pas la moitié de l'original (des originaux). ». Il est possible de classer les divers types de résumés en quatre catégories. Premièrement, les résumés indicatifs fournissent une idée du thème abordé sans en énoncer le contenu. Deuxièmement, les résumés informatifs qui présentent une version raccourcie du contenu. Puis viennent les extraits qui reprennent des passages. Enfin, les résumés (*abstracts*) sont le résultat de la reformulation du contenu.

Il est également possible de considérer le résumé du point de vue de l'usager qui le requiert. En effet, celui-ci peut être intéressé à recevoir un résumé du document dans son ensemble. Par exemple, un écolier sera intéressé par tous les aspects de la photosynthèse décrits dans un article traitant de ce thème. Alors que d'autres usagers ne seront intéressés que par le résumé d'une partie du document considéré. Par exemple, un médecin pourrait ne pas s'intéresser à toutes les caractéristiques d'un gène mais seulement aux maladies dans lesquelles celui-ci est impliqué.

Enfin, la manière de produire le résumé pose divers problèmes. En effet, une façon d'obtenir un résumé est de sélectionner les parties jugées importantes dans les documents originaux, par exemple des phrases ou des paragraphes, et de les juxtaposer ensuite. Toutefois, un problème de cohérence intervient car il est relativement complexe d'articuler divers extraits pour former un tout cohérent. Si l'on descend en-dessous du niveau de la phrase, en considérant par exemple les groupes syntaxiques, nous sommes confrontés au problème de génération du discours propre aux langues naturelles.

2.3 Evaluation de la génération de résumé

La génération de résumés appliquée au domaine médical a été évaluée lors de l'introduction de la tâche génomique dans la campagne d'évaluation TREC-2003 aux Etats-Unis [HER 04]. Le principe d'évaluation consiste à proposer un ensemble d'articles scientifiques du domaine médical et de demander aux participants de produire les résumés correspondants. L'adéquation des résumés produits (fonction d'une protéine, incidence d'un gène sur une maladie, etc.) est ensuite évaluée automatiquement par l'application d'une mesure appropriée.

2.4 Etat de l'art

Plusieurs indicateurs permettent de repérer les parties importantes d'un texte. Premièrement, il est possible de tenir compte de la position des phrases dans le document. En effet, il s'avère que les titres, les introductions et les conclusions contiennent des informations importantes [BRA 95]. Un autre indicateur est la présence d'expressions-type dans le texte [TEU 99]. Par exemple, une expression telle que « *in this paper, we show that* » semble annoncer que le paragraphe considéré revêt une certaine importance. Diverses statistiques sur les fréquences d'apparition de termes [HOV 99] ainsi que la prédiction de la probabilité d'apparition de termes dans le résumé [WIT 99b] peuvent également contribuer à identifier les parties importantes d'un document. Enfin, il est possible de considérer l'enchevêtrement entre la requête et le titre du document examiné [HOV 99]. Naturellement, ces divers indicateurs peuvent être combinés afin d'évaluer la performance résultante [LIN 99]. Les sections ci-dessous décrivent quelques approches proposées pour la génération de résumé lors de la campagne d'évaluation TREC-2003.

2.4.1 Apprentissage automatique

Bhalotia *et al.* [BHA 04] suggère de choisir entre le titre et la dernière phrase du résumé. Ce choix s'effectue selon l'approche Naive Bayes [MIT 97], une approche probabiliste classique en apprentissage automatique. Les variables de décision retenues sont, pour l'essentiel, les verbes, les MeSH (*Medical Subject Headings*)[12] et les gènes, toutes trois pondérées par *tf-idf* (modèle vectoriel avec pondération locale et globale), ainsi que la présence du gène cible, représentée par une valeur booléenne.

[12] Thésaurus controlé disponible à l'adresse http://www.nlm.nih.gov/mesh/

2.4.2 Classification

Pour [RUC 04] cette génération doit se faire selon des choix linguistiques ou plus précisément stylistiques. Cette équipe propose de classifier les phrases du résumé et le titre selon quatre catégories, à savoir la conclusion, le sujet, les résultats et les méthodes. Un second classement se base sur une mesure de similarité avec le titre. Au besoin, une approche de traitement de la langue naturelle permet l'élimination de séquences débutant ou finissant une phrase (comme « *in this paper, we show that* »).

2.4.3 Combinaison de caractéristiques

Une autre approche intéressante de Kayaalp *et al.* [KAY 04] propose de décomposer les articles en phrases puis de combiner leurs diverses caractéristiques selon deux méthodes différentes. Les caractéristiques considérées sont par exemple la présence de la phrase dans le résumé, le nombre de mots, le nombre de chiffres ou le nombre de lettres majuscules contenus dans la phrase. Une première méthode de sélection s'appuie sur une combinaison linéaire d'un sous-ensemble de ces caractéristiques afin de sélectionner la meilleure phrase. Comme alternative, une deuxième méthode se base sur le calcul de prédicats à partir d'un autre jeu de caractéristiques.

2.5 Modèle proposé

2.5.1 Introduction

Le modèle que nous avons proposé s'inscrivait dans le contexte de la piste génomique de la campagne d'évaluation TREC-2003. Plus précisément, il s'agissait de produire des résumés de la taille d'une phrase pour des articles médicaux (voir tableau 1). Nous avons travaillé sur l'échantillon de collection mis à notre disposition par la campagne. Notre modèle se basait essentiellement sur la décomposition des textes en phrases puis sur la sélection de la meilleure phrase. Pour cette étape, nous avons implémenté et évalué différentes méthodes parmi lesquelles figure la régression logistique. Les sections ci-dessous décrivent la collection-test ainsi que les méthodes proposées.

2.5.2 Collection-test

Une collection-test en recherche d'information est constituée de trois éléments :

- un ensemble de documents;
- un ensemble de requêtes;
- un ensemble de jugements de pertinence.

Dans notre contexte, un document représente un article médical. Une requête est l'expression du besoin d'information d'un utilisateur, en l'occurrence l'obtention d'un résumé de l'article médical. Afin de déterminer si un document répond ou non à une requête, nous disposons pour chaque requête du titre et du résumé de l'article médical considéré.

2.5.2.1 Ensemble de documents

Pour la préparation de notre système, nous avons choisi d'utiliser un corpus d'articles médicaux mis à disposition par la campagne d'évaluation TREC. Dans le système Medline, pour chaque article répertorié, on dispose essentiellement[13] d'un titre, du nom du ou des auteur(s) et leur affiliation, de la référence complète (nom du journal, langue de l'article, pages, année de publication), des descripteurs manuellement extraits à partir du thésaurus contrôlé MeSH (comprenant plus de 21 000 rubriques vedette-matière) ainsi que d'un résumé. Les articles médicaux proviennent de cinq revues, à savoir *Journal of Biological Chemistry, Journal of Cell Biology, Science, Nucleic Acids Research* et *Proceedings of the National Academy of Sciences*, articles parus durant le deuxième semestre de l'année 2002. Un exemple d'article médical est présenté dans l'annexe B.

2.5.2.2 Requêtes et jugements de pertinence

Pour préparer notre système, nous disposions d'un ensemble de 139 articles médicaux pour lesquels il s'agissait de produire les GeneRIFs (ou *Gene Reference Into Function* utilisés dans la banque de données *LocusLink*[14]). De plus, nous disposions des GeneRIFs générés manuellement par des êtres humains. Afin de produire les jugements de pertinence, nous avons mesuré la similarité (mesure Dice qui sera expliquée dans le chapitre traitant de l'évaluation) entre les GeneRIFs obtenus automatiquement et ceux produits manuellement. Le tableau 1 présente quelques exemples de GeneRIFs à produire.

role of PIN1 in transactivation
keratinocyte growth factor (KGF), a key stimulator of epithelial cell proliferation during wound healing, preferentially binds to collagens I, III, and VI.
regulation of CYP7A1 and CYP27A1 in human liver

Tableau 1. *Exemples de GeneRIFs à produire*

[13] L'ensemble des informations potentiellement disponibles dans le système Medline est décrit à l'adresse http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#MEDLINEDisplayFormat

[14] LocusLink est une des banques de données sur les gènes accessible sur www.ncbi.nlm.nih.gov/LocusLink/. Une autre est Swiss-Prot accessible sur us.expasy.org/sprot/.

2.5.3 Méthodes proposées

Générer un GeneRIF à partir d'un article semble *a priori* une tâche ardue. Cependant, nous avons émis l'hypothèse que le titre, le résumé et la conclusion de l'article devraient contenir l'essentiel des éléments à inclure dans un GeneRIF. Une étude préliminaire réalisée par Mitchell *et al.* [MIT 03b] de la *National Library of Medicine* a montré que 95 % des GeneRIFs contiennent du texte provenant du titre ou du résumé de l'article. Parmi ceux-ci, 42 % sont extraits tels quels du titre ou du résumé alors qu'environ 25 % sont des séquences significatives extraites du titre ou du résumé. Dès lors, nous avons choisi de ne considérer que le titre et le résumé de l'article associé au GeneRIF à produire, en écartant les autres éléments de l'article tels que les titres de section ou la conclusion.

Nous nous sommes également intéressés à la localisation de la phrase qui a permis de générer le GeneRIF, tout en limitant nos investigations au titre et au résumé de l'article scientifique. En considérant le titre ainsi que les phrases des résumés, la figure 1 donne la distribution des phrases ayant généré les 139 GeneRIFs. Dans ce graphique, on présente les phrases dans l'ordre d'apparition, en commençant par le titre puis en poursuivant par les autres phrases numérotées de 1 à *n*, *n* représentant la dernière phrase du résumé (en moyenne *n* se situe à 9,22). La distribution nous montre que le titre (55 observations) et la dernière phrase du résumé (36 observations) sont, a priori, de bons candidats pour produire les GeneRIFs.

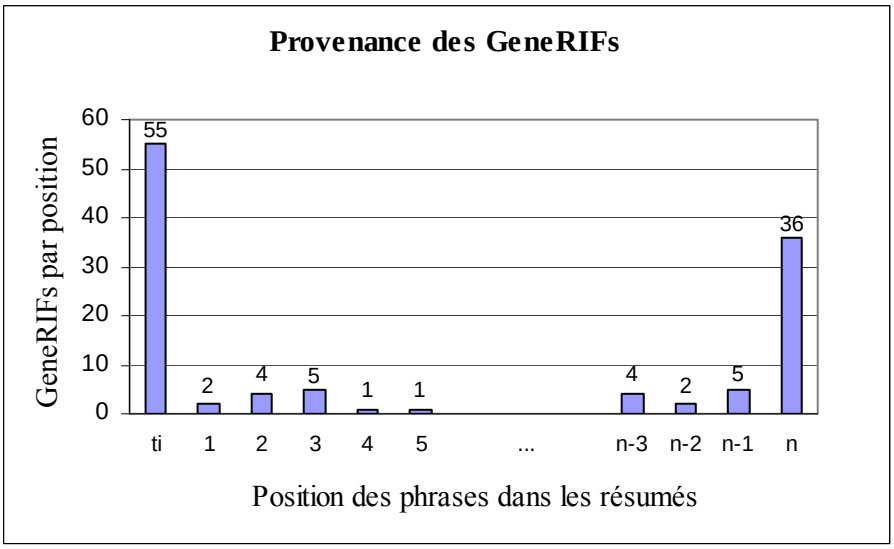


Figure 1. *Distribution de la localisation de la phrase source des GeneRIFs*

2.5.3.1 Limites de performance

Comme nos investigations se limitent au titre et au résumé de l'article scientifique servant de référence au GeneRIF, nous pouvons nous interroger sur la performance minimale et maximale que nous pourrions atteindre. Afin de définir une performance minimale, nous pouvons admettre que le titre est un bon candidat comme le confirme la distribution présentée dans la figure 1. En sélectionnant systématiquement ce titre, nous avons une performance minimale qui nous servira de référence pour d'autres approches d'extraction d'information.

Afin de connaître la performance maximale, et connaissant les bonnes réponses à nos 139 requêtes, nous pouvons déterminer la phrase à sélectionner dans le résumé ou le titre de l'article scientifique pour obtenir la réponse idéale. Or le bon GeneRIF contient aussi des termes n'apparaissant pas dans le titre ou le résumé de l'article scientifique. En limitant nos recherches au titre et résumé, il nous est donc impossible d'obtenir toujours le GeneRIF correct et donc une performance parfaite de 100 %. Afin de connaître la performance maximale que nous pourrions atteindre, nous avons créé la meilleure réponse possible en sélectionnant la meilleure phrase du résumé ou le titre de l'article.

2.5.3.2 Choix constant et naïf

Sur la base des 139 bonnes réponses, nous avons calculé la fréquence d'occurrences de tous les termes puis nous les avons classés par ordre décroissant. Comme les mots les plus fréquents ne sont pas ou peu porteurs d'information, nous avons supprimé tous les termes apparaissant dans la liste des 571 mots-outils du système SMART [SAL 71]. Ensuite, nous avons sélectionné les termes dont la fréquence était supérieure ou égale à 9. Dans cette liste, le terme le plus fréquent « *cell* » apparaît 36 fois, suivi de « *role* », 25 fois, et de « *protein* », 21 fois. Les termes ainsi obtenus, classés selon leur fréquence d'occurrences décroissante, sont les suivants :

cell role protein expression gene receptor activation regulate human apoptosis alpha sp1 signaling domain regulation kinase suggest pathway

Cette suite de mots constitue la réponse constante que cette stratégie retourne comme GeneRIF pour chaque résumé. L'une des lacunes importantes de cette approche est l'absence presque totale de sémantique. Cette liste de termes reste dépourvue de sens mais nous pouvons imaginer une deuxième stratégie, certes simpliste, mais qui fournit une réponse compréhensible directement par un être humain et qui est présentée dans la section suivante.

2.5.3.3 Tirage aléatoire

Comme deuxième stratégie automatique d'extraction d'information, nous avons subdivisé automatiquement le résumé en phrases. A cet ensemble de phrases, nous avons ajouté le titre. Dans cet ensemble, chaque phrase possède la même probabilité d'être tirée pour former la réponse. Si l'on étudie de plus près le nombre de phrases par résumé, on constate que le nombre moyen s'élève à 9,22 (écart-type 2,42; maximum 20; minimum 4) et ainsi on peut dire que chaque phrase a une chance sur dix d'être sélectionnée comme réponse.

Cependant, comme le titre est souvent un bon candidat, ce dernier a été introduit quatre fois dans la liste des réponses candidates. De plus, une petite étude statistique nous indique que la longueur moyenne d'un GeneRIF, longueur mesurée en nombre de mots significatifs, est de 11,96 (écart-type 4,2 ; maximum 28 ; minimum 3). L'expression « mot significatif » indique le fait que l'on ne compte pas les termes apparaissant dans la liste des mots-outils. Ainsi, la phrase « *Role of PIN1 in transactivation* » comporte trois mots significatifs « *Role* », « *PIN1* » et « *transactivation* ». Chaque phrase possédant un nombre de mots significatifs supérieur ou égal à 8 et inférieur ou égal à 16 est également introduite une seconde fois dans l'ensemble des réponses possibles. Ces valeurs limites correspondent à la moyenne (12) ± une fois l'écart-type (4). Finalement le système choisit aléatoirement une phrase de cette liste dans laquelle chaque élément possède la même chance d'être sélectionné. Ainsi, notre système retourne une réponse compréhensible comme GeneRIF, tout en accordant une légère préférence aux titres et aux phrases de taille moyenne.

2.5.3.4 Fréquence des termes significatifs

Les deux autres stratégies d'extraction que nous avons conçues se basent sur des principes similaires et sont présentées ci-après. Dans tous les cas, nous considérons le titre de l'article scientifique cible d'une part et d'autre part, l'ensemble des phrases que l'on peut extraire du résumé. Nos approches retourneront soit le titre soit l'une des phrases, garantissant ainsi que la réponse proposée par l'ordinateur possède un sens. Notre problème d'extraction se résume alors à définir quelle phrase du résumé ou le titre présente la plus grande similarité avec un GeneRIF typique.

Pour chaque phrase du résumé ainsi que pour le titre, nous avons supprimé les mots-outils puis nous avons éliminé la marque du pluriel des mots retenus au moyen de l'enracineur S (*S stemmer* [HAR 01]). Ce dernier suit les trois règles suivantes :

1. Si un mot se termine par « -ies », mais pas par « -eies » ou « -aies »alors remplacer « -ies » par « -y »;

2. Si un mot se termine par « -es », mais pas par « -aes », « -ees » ou « -oes » alors remplacer « -es » par « -e »;

3. Si un mot se termine par « -s », mais pas par « -us » ou « -ss » alors éliminer le dernier « -s ».

Pour chaque phrase et pour le titre, nous avons calculé un score selon la formule suivante :

$$\text{score} = \sum_{j=1}^{len} w(tf_j) \Big/ len \qquad (1)$$

dans laquelle tf_j indique la fréquence de ce terme dans l'ensemble des GeneRIFs, *len* la longueur de la phrase mesurée en nombre de mots et *w( )* une fonction définie dans le tableau 2 retournant un entier (choisi empiriquement) en fonction de la fréquence tf_j.

tf_j	$w(tf_j)$
$9 < tf_j$	4
$4 < tf_j \leq 9$	3
$2 < tf_j \leq 4$	2
$1 < tf_j \leq 2$	1
$tf_j \leq 1$	0

Tableau 2. *Poids attribué en fonction de la fréquence*

Finalement, le système sélectionne la phrase possédant le score le plus élevé comme GeneRIF. Nous favorisons ainsi la phrase ayant le plus de termes en commun avec le vocabulaire apparaissant dans les GeneRIFs. De plus, si ces termes communs sont aussi des mots fréquents, le score sera augmenté.

2.5.3.4 Régression logistique

Notre stratégie précédente attribuait un score à chacune des phrases du résumé ainsi qu'au titre. Or, nous savons que le titre forme souvent un bon candidat pour produire un GeneRIF. Afin de tenir compte de cette information, nous avons construit un modèle d'extraction basé sur la régression logistique devant sélectionner entre le titre d'une part et, d'autre part, la phrase ayant obtenu le meilleur score selon notre pondération vue à la section précédente.

Un exemple va illustrer de manière plus aisée le fonctionnement de notre modèle. En considérant la requête n° 30, nous devons choisir entre le titre et la phrase candidate présentés dans le tableau 3. Le tableau 4 présente les mêmes phrases après suppression des mots courants et de la marque du pluriel en anglais.

Titre	Comparative surface accessibility of a pore-lining threonine residue (T6') in the glycine and GABA(A) receptors.
Candidate	This action was not induced by oxidizing agents in either receptor.

Tableau 3. *Titre et phrase candidate pour la requête n° 30*

Titre	Comparative surface accessibility pore-lining threonine residue (T6') glycine GABA(A) receptor
Candidate	action induced oxidizing agent either receptor

Tableau 4. *Titre et phrase candidate comprenant les mots significatifs*

Pour chaque phrase candidate, nous pouvons calculer quelques statistiques, comme la longueur (notée *Len*), le nombre d'acronymes (noté *Abrv*), le nombre de termes indexés (c'est-à-dire apparaissant dans le vocabulaire des GeneRIF, variable notée *Terms*). A ces éléments, nous avons ajouté quelques variables liées à l'*idf*, ou logarithme de l'inverse de la fréquence documentaire. Ce choix s'appuie sur les travaux de Cronen-Townsend *et al.* [CRO 02] qui ont démontré que l'*idf* pouvait être, sous certaines conditions, un bon prédicteur de la performance d'une requête.

Le tableau 5 présente quelques variables explicatives. Nous avons mentionné les valeurs de ces six variables pour la phrase candidate et le titre. La dernière colonne indique la différence de ces valeurs entre la phrase candidate et le titre.

Variable	Signification	Candidate	Titre	Différence
Len	longueur	6	10	-4
Abrv	nombre d'acronymes	0	1	-1
Terms	nombre de mots indexés	5	10	-5
Max2Idf	2ème max idf	3,44	9,01	-5,57
MinIdf	min idf	2,25	2,35	-0,11
Min2Idf	2ème min idf	2,65	2,65	0,0

Tableau 5. *Variables utilisées pour notre modèle de prédiction*

La régression logistique [HOS 00] retourne une estimation de la probabilité de réalisation d'un événement en fonction d'une ou de plusieurs variables explicatives. Un des attraits majeurs de cette approche statistique réside dans le fait que les variables explicatives ne doivent pas être toutes des variables réelles ou entières mais peuvent, pour une partie d'entre elles, être des variables binaires, voire catégorielles.

Notre modèle de prédiction basé sur la régression logistique doit sélectionner entre le titre et une phrase candidate en fonction des variables explicatives décrites dans le tableau 6. Bien que nous ayons disposé au départ d'un ensemble plus conséquent de variables, seules celles présentées dans le tableau 6 ont finalement été retenues parce qu'elles étaient corrélées avec la variable d'apprentissage. Les variables commençant par la lettre 'd' représentent des différences. Par exemple, *d.Len* représente la différence de longueur entre la phrase candidate et le titre. La valeur retournée par cette régression logistique indique la probabilité que la phrase candidate soit un bon GeneRIF. Dans la dernière colonne nous avons indiqué les estimations obtenues pour chacune des variables. Par exemple, l'estimation pour la variable *d.Len* est négative, indiquant que si la longueur de la phrase candidate est supérieure à celle du titre, la probabilité que cette phrase candidate soit un bon GeneRIF diminue. De même, si la phrase candidate possède de nombreux mots en commun (*nb.Com*) avec le titre, la probabilité qu'elle soit un bon GeneRIF augmente.

Variable (x_i)	**Signification**	**Estimation (β_i)**
Terms	nombre de mots indexés	-19,867
Min2Idf	$2^{ème}$ min idf de la phrase candidate	-36,733
nb.Com	nombre de termes significatifs communs entre la phrase candidate et le titre	18,999
d.Len	différence de longueur (candidate – titre)	-57,029
d.Abrv	différence du nombre d'acronymes (candidate – titre)	17,141
d.Terms	différence des mots indexés (candidate – titre)	46,910
d.Max2Idf	différence $2^{ème}$ max idf (candidate – titre)	30,926
d.MinIdf	différence min idf (candidate – titre)	22,121

Tableau 6. *Ensemble des variables et leurs estimations*

Nous avons regroupé les variables dans un vecteur $X = [x_1, x_2, ..., x_k]$ que nous avons utilisé pour estimer la probabilité que la phrase candidate soit meilleure que le titre selon la formule 2 dans laquelle α, β_1, β_2, ..., β_k, sont les estimations calculées sur la base d'un ensemble d'observations à l'aide du logiciel R [VEN 99].

$$p[\text{candidate meilleure que titre} \mid X] = \frac{e^{\alpha + \sum_{i=1}^{8} \beta_i x_i}}{1 + e^{\alpha + \sum_{i=1}^{8} \beta_i x_i}} \qquad (2)$$

Si la probabilité calculée par la formule 2 pour une phrase candidate est supérieure à 0,5, cette phrase est sélectionnée comme GeneRIF. Dans le cas contraire, c'est le titre de l'article qui est retourné en guise de GeneRIF. En nous basant sur les résultats de la régression logistique, nous avons retourné le titre 129 fois et la phrase candidate 10 fois.

2.6 Evaluation

2.6.1 Introduction

Notre système de génération de résumé a été évalué dans le cadre de la tâche génomique de la campagne d'évaluation TREC-2003 [HER 04], [SAV 04a]. La tâche définie dans la piste génomique de TREC consistait à fournir le GeneRIF correspondant à un article scientifique répertorié dans Medline. Le GeneRIF correspond à un texte décrivant les caractéristiques particulières d'un gène, sa fonction et/ou ses implications dans telle ou telle maladie. Naturellement, un gène peut posséder plusieurs GeneRIFs car un article scientifique ne met souvent en lumière que l'une ou l'autre des fonctions d'un gène donné. Le GeneRIF équivaut très souvent à une phrase extraite ou construite à l'aide de segments choisis provenant de l'article associé. Quelques exemples sont repris dans le tableau 7.

Référence à l'article	**GeneRIF**
J Biol Chem 2002 Sep 13; 277(37): 34343-8	the death effector domain of FADD is involved in interaction with Fas.
J Biol Chem 2002 Dec 27;277(52):50834-41	Apocytochrome c blocks caspase-9 activation and Bax induced apoptosis
J Biol Chem 2002 Dec 13;277(50):47976-9	role of PIN1 in transactivation
Nucleic Acids Res 2002 Aug 15; 30(16):3609-14	In the case of Fas-mediated apoptosis, when we transiently introduced these hybrid-ribozyme libraries into Fas-expressing HeLa cells, we were able to isolate surviving clones that were resistant to or exhibited a delay in Fas-mediated apoptosis

Tableau 7. *Quatre exemples de GeneRIFs ainsi que les articles associés*

2.6.2 Mesure d'évaluation

L'évaluation de tout système générant des résumés ou permettant d'extraire de l'information s'avère difficile. Ainsi, plusieurs problèmes d'évaluation sont récurrents dans les campagnes d'évaluation des systèmes de question-réponse [VOO 04]. Dans le cas présent, nous rencontrons des problèmes similaires, mais

avec l'avantage que la bonne réponse est connue de façon précise. Si bien qu'il n'y pas besoin de discuter si la réponse à la question « Où est situé le Tahaj Mal ? » est bien « Aux Indes » ou « Atlantic City ».

Une première mesure pour savoir si une réponse correspond au GeneRIF souhaité serait de vérifier l'égalité stricte des deux chaînes de caractères. Cela ne nous avancerait pas beaucoup car cette mesure d'évaluation apporte une réponse booléenne {vrai, faux}. De ce fait, si le GeneRIF attendu est « *Role of PIN1 in transactivation* », cette fonction d'évaluation répondrait faux aux trois réponses suivantes :

1. *The role of PIN1 in transactivation*
2. *role transactivation PIN1*
3. *This action was not induced by oxidizing agents in either receptor.*

En effet, la première réponse débute par le déterminant « the » et la deuxième ne possède pas les mots-outils « of » et « in ». Mais ces deux premières réponses sont beaucoup plus proches du GeneRIF attendu que la dernière. Pour tenir compte des éléments en commun entre la réponse proposée et le GeneRIF adéquat, les responsables de la campagne d'évaluation de TREC-2003 Hersh & Bhupatiraju [HER 04] ont retenu le coefficient de Dice qui s'explique de la manière suivante.

Etant donné deux phrases A et B, on définit |A| comme la cardinalité de l'ensemble A ou, dans notre cas, le nombre de mots différents dans A, |B| comme le nombre de mots différents dans B, et $|A \cap B|$ comme le nombre de mots différents communs à A et à B. La similarité entre les phrases A et B se mesure alors par :

$$\text{Similarité Dice}(A, B) = 2 * |A \cap B| \Big/ |A| + |B| \qquad (3)$$

Cette mesure de similarité présente quelques difficultés. D'abord, on ne peut raisonnablement pas attribuer la même importance à un terme significatif comme « *cell* » ou « *protein* » et à un mot-outil tel que « *the* », « *in* » ou « *of* ». De plus, si la différence entre la réponse et le GeneRIF tient à la présence de la marque du nombre (par exemple « *protein* » ou « *proteins* ») ou à une dérivation suffixale (« *signal* » ou « *signaling* »), la valeur de la similarité ne devrait pas être affectée de manière importante.

Pour remédier à ces lacunes évidentes, le degré de similarité de Dice sera calculé après élimination des mots-outils (dans la campagne TREC-2003, 321 mots composent cette liste) et suppression de certaines séquences finales au moyen de l'algorithme de Porter [POR 80].

Avec cette mesure, les deux premières réponses de notre exemple précédent obtiennent une similarité de 1 (ou 100 %), tandis que la troisième réponse possède une similarité de 0 avec le bon GeneRIF « *Role of PIN1 in transactivation* ».

Cette première mesure n'est toutefois pas parfaite. En effet, l'ordre des mots n'a pas d'importance dans cette évaluation si bien que ce coefficient de Dice donne le même degré de similarité pour les deux premières réponses. Or l'anglais, tout comme le français d'ailleurs, est une langue dans laquelle l'ordre des mots revêt une grande importance. Ainsi, les phrases « *the dog bites the postman* » ou « *the postman bites the dog* » ne possède pas le même sens. Certes, après l'élimination des mots-outils, l'ordre des mots n'a pas toujours une grande importance comme dans l'exemple « *the information retrieval* » (« *information retrieval* ») ou « *the retrieval of information* » (« *retrieval information* »).

Afin de tenir compte de l'ordre des mots, une mesure de Dice modifiée pour s'appliquer à des paires de mots a été utilisée dans la campagne génomique de TREC-2003. Ce n'est donc plus sur les mots que l'on évalue la similarité mais sur les doublets (paire ordonnée de mots).

Avec cette deuxième mesure, la première réponse de notre exemple précédent obtient une similarité de 1 (ou 100 %), tandis que la deuxième réponse, qui contient les termes souhaités mais dans un ordre différent, possède une similarité de 0 avec le bon GeneRIF « *Role of PIN1 in transactivation* ». Cette dernière valeur surprend car la réponse « *Role transactivation PIN1* » est à nos yeux relativement proche de la bonne réponse. Cet exemple illustre le fait que nous devons garder un certain recul vis-à-vis de différences faibles dans nos mesures de performance.

2.6.3 Résultats

En utilisant les deux mesures d'évaluation présentées dans la section précédente pour nos différents modèles d'extraction, nous obtenons les résultats indiqués dans le tableau 8. La stratégie simple qui retourne toujours le titre possède une performance relativement élevée (ligne « Titre » dans le tableau 8). Elle dépasse de loin le choix naïf et constant retournant toujours les mêmes mots (ligne « Choix constant » dans le tableau 8), même si ceux-ci sont fréquents dans les GeneRIFs. Dans ce dernier cas, on remarque que la performance obtenue lorsque l'on tient compte de l'ordre des mots (colonne notée « Dice modifié ») est très faible. La sélection aléatoire, même en favorisant les phrases de longueur moyenne ou le titre, n'apporte pas une performance acceptable. Notre première stratégie d'extraction raisonnable, basée sur la fréquence d'occurrence des termes significatifs, possède une évaluation légèrement inférieure à la stratégie « Titre ». Finalement, notre stratégie de sélection entre le titre et la phrase candidate dispose d'une performance la situant au-dessus de la stratégie de référence « Titre ».

Stratégie	Dice	Dice modifié
Seuil minimal - titre (section 2.5.3.1)	50,47 %	34,82 %
Choix constant (section 2.5.3.2)	9,42 %	0,15 %
Tirage aléatoire (section 2.5.3.3)	29,93 %	14,84 %
Fréquence (section 2.5.3.4)	46.44 %	32,37 %
Régression logistique (section 2.5.3.5)	**52,28 %**	**37,43 %**
Seuil maximal (section 2.5.3.1)	71,17 %	64,08 %

Tableau 8. *Evaluation de nos stratégies d'extraction*

Notre système est arrivé en quatrième position lors de cette évaluation. Le meilleur système proposé a obtenu un score de 57,83 % (Dice) et 46,75 % (Dice modifié). Ce qui signifie que notre système arrive à 90 % de la meilleure performance selon la mesure Dice et 80 % selon la mesure Dice modifiée.

Dans notre meilleure stratégie, si l'on compare les choix induits par la régression logistique avec les meilleurs choix possibles, on constate que dans 44,60 % des cas, notre système a choisi correctement entre le titre et la phrase candidate considérée. Pour les 55,40 % restants, plusieurs cas de figure se présentent. Tout d'abord, lorsque le titre comprenait plusieurs phrases, notre système a privilégié le titre complet au détriment d'une des phrases le composant. D'autre part, après analyse des GeneRIFs, il s'avère que 26,62 % d'entre eux sont des paraphrases du titre (16,55 %) ou d'une phrase issue du résumé (10,07 %). Notre système n'étant pas conçu pour traiter des paraphrases, il a retourné le titre ou une phrase présente dans le résumé. De plus, la bonne réponse peut être construite en concaténant des séquences provenant de diverses phrases du résumé ou du titre, opérations de sélection inter-phrases que notre système n'est actuellement pas capable d'effectuer. Enfin, le vocabulaire des GeneRIFs n'apparaît pas toujours dans le résumé ou le titre de l'article mais peut provenir d'autres sources telles que l'article complet.

2.6.4 Combinaison de stratégies

Suite à notre participation à l'évaluation de la campagne TREC-2003, nous avons souhaité combiner notre approche avec celle d'une autre équipe de recherche de l'Hôpital Universitaire de Genève (HUG) [RUC 04] qui avait également participé à la piste génomique de TREC-2003 et avait obtenu le troisième rang. Pour notre modèle, nous avons choisi la méthode ayant produit la meilleure performance, à savoir l'application de la régression logistique. Le HUG a quant à lui proposé une stratégie de classification qui utilise des caractéristiques argumentatives, positionnelles et structurelles afin de sélectionner la phrase candidate à retourner comme résumé. Puis, nous avons combiné les deux stratégies de la manière suivante. Chaque système a sélectionné les phrases candidates. Si les deux systèmes avaient sélectionné la même phrase, celle-ci était retenue. En cas de divergence, si l'estimation de la probabilité calculée par la régression logistique était supérieure à un certain seuil (0,5), alors la phrase sélectionnée par notre modèle était retenue. Dans le cas contraire, la phrase sélectionnée par le HUG était choisie. Enfin, après

avoir sélectionné une candidate, un post-traitement permettait d'éliminer les fragments jugés inutiles dans la phrase. Du point de vue de la performance, une augmentation significative est apparue. En effet, la performance individuelle de notre méthode se situait à 52,28 % selon la mesure Dice. Pour la méthode du HUG, cette performance atteignait 52,78 %. La combinaison des stratégies a été évaluée dans les mêmes conditions et a obtenu une performance supérieure à 55 % (+ 5,2 % par rapport à notre méthode) [RUC 05].

2.7 Conclusion et perspectives

La génération automatique de GeneRIFs décrivant les caractéristiques d'un gène sur la base d'un article scientifique donné demeure une tâche complexe. Notre approche, basée sur la régression logistique, correspond à une sélection entre le titre et une phrase candidate choisie dans le résumé en fonction de son vocabulaire. Notre démarche relativement simple possède une évaluation qui s'avère supérieure à la stratégie visant à retourner systématiquement le titre de l'article.

Nous espérons améliorer notre stratégie selon deux axes. En premier lieu, nous devons y incorporer plus de considérations linguistiques. Notre sélection actuelle se fonde, pour l'essentiel, sur le vocabulaire ou sur le fait que la phrase descriptive est souvent le titre de l'article. Deuxièmement, nous devrions pouvoir décomposer une phrase dans ses différents éléments constitutifs afin d'écarter certaines parties d'une part et, d'autre part, de sélectionner les groupes nominaux jugés très importants dans la description des caractéristiques d'un gène donné.

3. Question-réponse

3.1 Introduction

Le besoin d'information se répand de plus en plus parmi les usagers. Ainsi des usagers occasionnels recherchent une simple information factuelle comme la date d'un événement ou le nom du président de tel ou tel pays. D'autres sont des clients intéressés par les caractéristiques d'un produit. Viennent ensuite les analystes qui collectent des données par exemple sur les marchés financiers et enfin, les spécialistes tels que les criminologues qui recherchent des informations très précises et spécifiques, requérant une grande expérience.

Les systèmes de recherche de documents retournent pour une requête donnée la liste des documents jugés pertinents par la machine et susceptibles de contenir la réponse souhaitée. C'est à l'usager qu'incombe ensuite la responsabilité d'extraire, à partir des documents proposés, la réponse précise à sa question. Les systèmes de question-réponse franchissent un pas supplémentaire en direction de la recherche d'information puisqu'ils permettent de retrouver la réponse précise à une question à partir d'une importante collection de documents. Les systèmes de question-réponse utilisent des modèles de recherche de documents pour identifier les documents et les passages pertinents puis appliquent des techniques de traitement de la langue naturelle pour extraire la réponse précise souhaitée [MIT 03a].

Les systèmes de question-réponse peuvent être ouverts vers tous les domaines ou spécifiques à un domaine particulier comme la médecine par exemple. Les questions prises en charge peuvent être factuelles ou plus complexes traitant d'événements ou de situations. Plus le degré de complexité de la question augmente, plus il est nécessaire de faire intervenir un raisonnement et des connaissances élaborés.

3.2 L'architecture des systèmes de question-réponse

Un système de question-réponse se compose généralement de trois parties principales, à savoir un module d'analyse de la question, un module de traitement des documents et un module d'extraction de la réponse [MIT 03a], [GRA 04]. L'analyse de la question permet son interprétation et l'identification du type de réponse attendue. Par exemple, pour la question « Qui est le président des Etats-Unis ? », le type de réponse attendue est une personne qui pourra être repérée dans un fragment tel que « George Bush ». L'analyse de la question permet aussi la sélection des termes importants, voire d'y inclure d'autres, qui serviront à construire la requête adressée au module de traitement des documents. Celui-ci va dépister les documents susceptibles de contenir la réponse à la requête. Puis, le module d'extraction de la réponse va identifier, dans un document, le fragment qui contient

la réponse à la question posée. Le système peut également exploiter des ressources externes telles que la Toile, thésaurus ou ressources diverses. La figure 2 illustre les grandes fonctions incluses dans l'architecture d'un système de question-réponse.

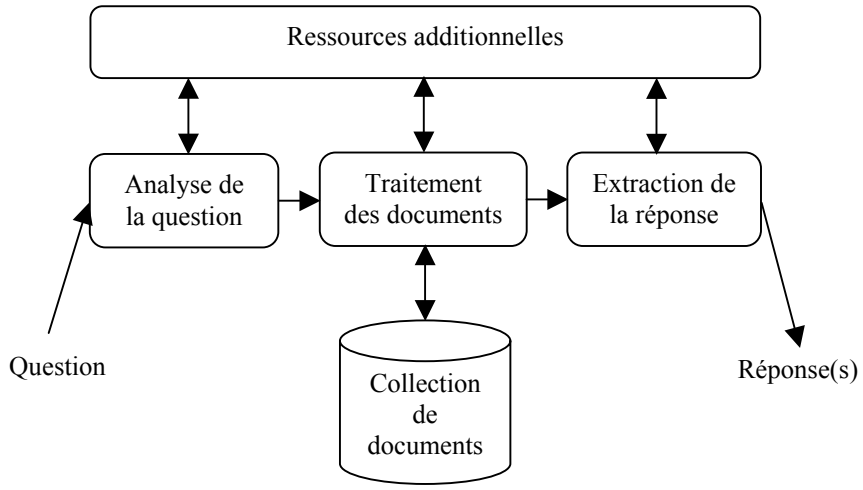


Figure 2. *Architecture d'un système de question-réponse*

3.3 Evaluation des systèmes de question-réponse

La première évaluation de systèmes de question-réponse pour l'anglais remonte à 1999 lors de l'introduction de la tâche question-réponse dans la campagne d'évaluation TREC-8 aux Etats-Unis [VOO 00a]. Le principe d'évaluation consiste à proposer un jeu de questions aux participants qui soumettront en retour les réponses accompagnées de l'identificateur du document de justification. Puis des juges humains évaluent les réponses selon des directives communes fournies. Depuis cette première édition, TREC a poursuivi cette tâche chaque année en modifiant divers paramètres tels que la collection de documents, la complexité des questions, la longueur maximale des réponses à fournir ou l'aptitude des systèmes à estimer leur confiance dans la réponse fournie.

Alors que la tâche question-réponse de TREC se concentre exclusivement sur l'anglais, un autre cadre d'évaluation européen a vu le jour. Il s'agit de CLEF qui est aussi une série de campagnes d'évaluation annuelles qui existe depuis l'année 2000 et qui est issue de TREC. CLEF [BRA 04] s'est spécialisé dans le traitement des langues européennes en mettant à disposition l'infrastructure permettant de tester différents aspects du développement de systèmes de recherche d'information monolingues et multilingues [MAG 03b]. A travers les campagnes d'évaluation,

CLEF constitue et met à disposition des collections-test formées de documents, requêtes et jugements de pertinence. Chaque année, plusieurs tâches sont proposées aux participants qui reçoivent un corpus de documents et un ensemble de requêtes. Les résultats retournés par les participants sont jugés par des experts humains qui décident de leur pertinence. C'est ainsi que les jugements de pertinence sont construits. Il faut toutefois remarquer que pour la question-réponse, on ne dispose pas vraiment de collection-test parce qu'un système peut proposer une nouvelle réponse qui n'a jamais été jugée et ne figure donc pas dans les jugements de pertinence. CLEF est soutenu dans ses efforts en partie par DELOS Network of Excellence[15] spécialisé dans les bibliothèques numériques.

Enfin, plus récemment, une évaluation des systèmes de question-réponse pour le français a été créée en France en 2004 (EQueR). Ce projet a été coordonné par l'ELDA (*Evaluation and Language resources Distribution Agency*) une agence de distribution de ressources linguistiques et d'évaluation.

3.4 Etat de l'art

Comme nous l'avons vu précédemment, l'architecture générale d'un système de question-réponse se compose de trois parties. Premièrement, un composant d'analyse de la question permettant d'identifier le type de réponse attendue. Deuxièmement, un composant de recherche des documents pertinents relativement à la question. Enfin, un composant d'extraction de la réponse précise à partir des documents précédemment identifiés. Lorsqu'aucune réponse n'a pu être trouvée, divers comportements sont envisagés. Sur la base de cette architecture générale, il existe de nombreuses approches qui varient entre deux courants majeurs, l'un basé sur l'appariement de représentations syntaxiques et sémantiques, l'autre sur l'application de patrons d'extraction.

Il est important de relever qu'au-delà des qualités intrinsèques des systèmes, les ressources additionnelles telles que la Toile, WordNet [FEL 98], [LEA 98], dictionnaires ou schémas de recherche jouent un grand rôle dans la performance observée. D'autre part, il faut garder une certaine prudence lorsqu'on compare des systèmes agissant sur des langues différentes. En effet, les diverses langues renferment des caractéristiques spécifiques et souvent peu ou pas étudiées ayant pour conséquence que certaines affirmations peuvent s'avérer fausses lorsqu'on les transpose vers une autre langue. Les sections ci-dessous décrivent quelques approches caractéristiques qui ont obtenu les meilleurs résultats dans les tâches de question-réponse lors des récentes campagnes d'évaluation TREC et CLEF.

[15] Voir le site http://www.delos.info/

3.4.1 Architecture à flux multiples

L'Université d'Amsterdam [JIJ 03], [JIJ 04b] et [JIJ 04a] propose un système de question-réponse pour le néerlandais destiné à améliorer le problème du rappel (rapport entre le nombre de documents pertinents retrouvés et le nombre total de documents pertinents). En effet, il y a souvent une différence de formulation entre la question et les documents contenant potentiellement la réponse, impliquant qu'une partie des documents pertinents ne sont pas retrouvés. On peut exprimer le même concept avec des termes ou tournures de phrases différents (l'étoile du Soir, la planète Vénus, l'étoile du matin, etc.). Le système proposé se base sur une architecture à flux multiples, chacun des flux procédant d'une manière différente pour identifier les réponses candidates. On peut classer ces flux en quatre groupes :

- « *lookup* » : flux qui exploite une base de connaissances produite préalablement à partir de la collection ;

- « *DutchTequesta* » : système de question-réponse pour le nérlandais [MON 02] incluant un étiqueteur POS (part-of-speech), un reconnaisseur d'entités nommées et une sélection de réponses candidates basée sur la notion de proximité ;

- « *pattern matching* » : flux effectuant l'appariement grâce à la génération d'expressions régulières à partir des questions ;

- « *ngram* » : flux qui applique la technique des n-grammes pour rechercher les réponses candidates.

Chacun de ces flux produit un ensemble de réponses candidates qui sont d'abord filtrées pour éliminer les réponses jugées incorrectes [SCH 04] avant d'être fusionnées dans une seule liste. Les réponses provenant de la Toile et n'ayant aucune justification dans le corpus sont également éliminées. Enfin, un système de vote permet de départager les réponses restantes afin de sélectionner la meilleure réponse que le système retournera.

Ce système a montré les meilleures performances lors de la campagne d'évaluation QA@CLEF-2004 [MAG 04].

3.4.2 Réseaux sémantiques

L'Université de Hagen (Allemagne) [HAR 04b] a développé un système de question-réponse pour l'allemand nommé *InSicht* basé sur la construction de réseaux sémantiques. Dans un premier temps, les questions et les documents sont analysés individuellement par un analyseur syntactico-sémantique et une représentation sous forme de réseau sémantique est construite [HEL 02]. L'extension de requête permet de générer plusieurs représentations de la même question. Puis, chaque représentation de la question est comparée aux représentations des documents candidats. Si la correspondance est établie, alors une réponse est générée à partir du réseau sémantique du document. Disposant de plusieurs réponses candidates possibles, il s'agit ensuite de sélectionner la meilleure en favorisant deux caractéristiques, à savoir la longueur et la fréquence d'apparition

des réponses dans la collection. Il est important de relever que ce système n'utilise aucune autre source que la collection fournie garantissant ainsi que la réponse provient bien de la collection considérée [NEU 03].

Ce système a participé à la campagne d'évaluation QA@CLEF-2004 [MAG 04].

3.4.3 Diogene

Le centre de recherche scientifique et technologique italien ITC-irst a conçu un système de question-réponse multilingue appelé Diogene [MAG 03a], [NEG 03] et [TAN 04a]. Dans une première phase, les questions sont analysées à l'aide d'un analyseur POS et le type de réponse attendue est identifié. Puis, un système de recherche d'information classique nommé *Managing Gigabytes* (MG) [WIT 99a] et implémentant le modèle booléen est utilisé pour dépister les documents candidats. Ensuite, la réponse est extraite à l'aide de la reconnaissance d'entités nommées [MAG 02a] et de schémas linguistiques [TAN 04b] utilisant des expressions régulières. Enfin, la réponse est validée par une recherche sur la Toile [MAG 02b] avec l'outil *AllTheWeb*[16].

Ce système a participé à la campagne d'évaluation QA@CLEF-2004 [MAG 04].

3.4.4 Raisonnement logique

L'équipe de recherche du LCC (*Language Computer Corporation*) a réalisé un système de question-réponse basé sur le raisonnement logique [HAR 04a] et [MOL 03]. Ce système se compose de trois modules, à savoir le module d'analyse de la question, celui de traitement des documents et celui d'extraction de réponse. Pour l'analyse de la question, le type de réponse attendue est identifié soit à partir d'un reconnaisseur d'entités nommées soit sur la base d'une hiérarchie de concepts sémantiques dérivée de WordNet [LEA 98]. Puis, les passages intéressants sont dépistés dans la collection de documents en utilisant les mots-clés de la question. Enfin, la réponse est extraite de différentes manières. Premièrement, le reconnaisseur d'entités nommées est appliqué. Si celui-ci ne parvient pas à trouver une réponse, un théorème est prouvé logiquement à l'aide d'idiomes provenant de WordNet ainsi que d'axiomes approximant les relations sémantiques ou pragmatiques.

Ce système a participé à la campagne d'évaluation TREC-2003 [VOO 04] et a obtenu les meilleurs résultats dans la tâche QA monolingue anglais.

3.4.5 QUALIFIER

QUALIFIER (*QUestion Answering by LexIcal FabrIc and External Resources*) est un système de question-réponse développé par l'équipe de la *School of Computing* à l'Université de Singapour [YAN 04]. Dans un premier temps, la question est analysée afin d'en déterminer la classe et le type de réponse attendue. Une ontologie de classes de questions est construite sur la base des entités

[16] Voir le site http://www.alltheweb.com/

nommées. Puis, la question est étendue [YAN 03a] à l'aide de la Toile et de ressources lexicales telle que WordNet [LEA 98]. Les réponses candidates sont ensuite dépistées à partir des termes de la question étendue à l'aide du système de recherche d'information *Managing Gigabytes* (MG) [WIT 99a]. Les phrases des documents retournés sont classées selon des règles associatives. Enfin, la réponse est extraite en appliquant la reconnaissance d'entités nommées, la résolution de co-références, la résolution d'anaphores et la justification de réponse [YAN 03b].

Ce système a participé à la campagne d'évaluation TREC-2003 [VOO 04].

3.4.6 Patrons d'extraction

L'équipe d'*InsightSoft-M* (Moscou) a développé un système de question-réponse basé sur l'application de patrons d'extraction et la maintenance d'une base de connaissances [SOU 03] et [SOU 02]. Ils font usage de patrons décrivant les structures de surface de chaînes de caractères susceptibles de contenir certaines informations sémantiques [RAV 02]. Les questions sont analysées pour en déterminer le type et sélectionner les patrons correspondants. Puis, chaque patron, caractérisé par une sémantique générale, est appliqué aux passages susceptibles de contenir la réponse. L'appariement est fait sur la base du patron ainsi que des éléments de la base de connaissance tels que pays, monnaies ou mesures.

Ce système a participé à la campagne d'évaluation TREC-2002 [VOO 03].

3.5 Système proposé

3.5.1 Introduction

Dans un premier temps, nous nous sommes concentrés sur le développement d'un système de question-réponse monolingue pour la langue française. Cela supposait de disposer de requêtes et de documents formulés en français. S'agissant des documents, nous avons utilisé une partie de la collection mise à disposition par la campagne d'évaluation CLEF. Pour les requêtes, nous avons constaté à l'époque qu'il n'existait encore aucune collection de questions utilisable pour nos besoins si bien que nous avons créé, à partir de la sous-collection de documents, un jeu de questions-test avec les jugements de pertinence correspondants. La sous-collection de documents et le jeu de questions-test nous ont permis de constituer une collection-test qui sera utilisée pour évaluer notre système de question-réponse monolingue. Ce système a participé aux campagnes d'évaluation CLEF 2004 et EQueR.

Après avoir achevé le développement et l'évaluation du système monolingue, nous avons adapté notre modèle pour qu'il puisse répondre à des requêtes exprimées dans diverses langues européennes en recherchant les réponses dans des documents formulés en français. Le système bilingue ainsi obtenu a participé à la campagne d'évaluation CLEF 2004.

Les sections ci-dessous décrivent plus précisément les étapes de constitution de notre système de question-réponse monolingue et bilingue.

3.5.2 Collection-test

3.5.2.1 Ensemble de documents

Pour la préparation de notre système, nous avons choisi d'utiliser des corpus de documents mis à disposition par la campagne d'évaluation CLEF. Parmi les divers corpus mis à disposition, nous en avons choisi deux contenant des documents formulés en français. Le corpus *Le Monde* contient divers articles publiés en 1994 alors que l'*ATS* (Agence Télégraphique Suisse) comprend des dépêches parues également en 1994. Le tableau 9 donne quelques précisions sur ces corpus. Des exemples de documents qui en sont issus sont présentés dans les annexes C et D.

Corpus	Année	Taille	Nb documents
Le Monde	1994	157 Mo	44 013
ATS français	1994	86 Mo	43 178

Tableau 9. *Collection de documents*

3.5.2.1 Requêtes et jugements de pertinence

Il n'existait à notre connaissance aucun jeu de requêtes et jugements de pertinence pour la langue française. Si bien que nous avons décidé de construire à partir des corpus mis à notre disposition un ensemble de questions avec leurs jugements de pertinence. Pour cela, nous avons pris des documents au hasard dans les deux corpus (*Le Monde* et *ATS*) à partir desquels nous avons formulé des questions factuelles et relevé les réponses correspondantes. Une question factuelle est supposée recevoir une réponse précise telle qu'une date, une localité ou une personne. Il n'était pas prévu de traiter des questions plus complexes attendant des réponses sous forme de définition. Après avoir épuré l'ensemble des questions et réponses associées, nous avons conservé 57 questions factuelles. Le tableau 10 contient quelques exemples issus de nos questions-test alors que l'annexe J contient l'ensemble des questions.

Question	Réponse attendue	Document source
Où se trouve le siège de l'OCDE ?	Paris	LEMONDE94-000001-19941201
Qui est le premier ministre canadien ?	Jean Chrétien	LEMONDE94-000034-19941201
Combien de collaborateurs emploie ABB ?	206 000	ATS.941214.0105

Tableau 10. *Exemples de questions-test*

A ce stade, les jugements de pertinence dont nous disposions ne contenaient qu'un seul document pertinent par question, le document source à partir duquel la question avait été formulée. Nous souhaitions compléter cette liste par d'autres documents pertinents. Pour ce faire, nous avons utilisé le système de recherche d'information SMART[17] [SAL 71] auquel nous avons soumis les 57 requêtes construites à partir des questions correspondantes. Puis, nous avons évalué les documents retournés et avons ajouté ceux jugés pertinents aux jugements de pertinence. Le tableau 11 illustre les jugements de pertinence pour une question-test, le premier document étant le document source de la question.

Question	Réponse	Documents pertinents
Qui est secrétaire général de l'OCDE ?	Jean-Claude Paye	LEMONDE94-000001-19941201 LEMONDE94-002606-19940922 ATS.941129.0021 ATS.941130.0072 ATS.940405.0079 ATS.940921.0142 ATS.941130.0140

Tableau 11. *Exemple de jugements de pertinence*

[17] SMART version 11.0 disponible à l'adresse ftp://cs.cornell.edu/pub/smart/

3.5.3 Système de question-réponse monolingue

3.5.3.1 Introduction

Dans ce chapitre seront présentées les diverses étapes ayant abouti à la réalisation de notre système de question-réponse monolingue pour le français. Nous exposerons les traitements préliminaires tels que la décomposition de la collection de documents en paragraphes, l'élimination de mots-outils et l'application d'un enracineur. Puis, nous décrirons le système de recherche de documents et les mesures de performance utilisées. Nous poursuivrons par la présentation de l'analyse syntaxique et la reconnaissance d'entités nommées. Enfin, nous expliquerons notre méthode d'extraction de réponse.

3.5.3.2 Découpage des documents en paragraphes

Les deux corpus que nous avons choisis, à savoir *Le Monde* et *ATS* de 1994, contenaient des documents relativement différents, bien qu'extraits du monde journalistique. En effet, *Le Monde* offre des articles sur des sujets variés alors qu'*ATS* contient des dépêches. Si l'on considère les questions factuelles, les besoins de l'utilisateur sont précis. Nous avons supposé que la réponse à une question factuelle se trouverait généralement au sein d'un paragraphe, si bien que nous avons choisi d'augmenter la granularité des documents considérés en prenant en considération non pas les documents dans leur intégralité mais les paragraphes qu'ils contiennent. Pour cela, nous avons découpé tous les documents en paragraphes en nous basant sur la structure SGML des documents tout en limitant la longueur à 50 lignes au maximum (environ 400 mots). A partir de 87 191 documents de départ (44 013 pour *Le Monde* et 43 178 pour *ATS*), nous avons obtenu 730 098 paragraphes. Cela signifie que nos documents d'origine ont été décomposés en 8,37 paragraphes en moyenne. Après cette étape, chaque paragraphe était considéré comme un document pour la recherche.

Notre collection-test était finalement constituée de :

- 730 098 documents (chaque paragraphe étant considéré comme un document)

- 57 requêtes (dérivées à partir des questions)

- 202 jugements de pertinence (3,5 jugements par requête en moyenne)

Les figures 3, 4 et 5 montrent des exemples de requête, document, et jugements de pertinence de notre collection-test. Pour la figure 5, le paragraphe du document est ajouté à l'identificateur du document. Par exemple, ATS.940921.0142-1 renvoie au premier paragraphe du document ATS.940921.0142.

```
<top>
<num> 7 </num>
<title>
Qui est secrétaire général de l'OCDE ?
</title>
</top>
```

Figure 3. *Exemple de requête*

```
<DOC>
<DOCNO>LEMONDE94-000001-19940101-2</DOCNO>
<TEXT>Il y a toujours eu des publications destinées à un public averti, où la liberté
de création avait le loisir de s'exprimer et de se
développer. Ce qui me semble nouveau, c'est le fait que ces
publications sont désormais exhibées à tous les regards, sans aucune
pudeur ni respect pour les sensibilités des plus jeunes. Ainsi, en
emmenant ses enfants voir un film " tous publics ", on s'expose à
l'agression des bandes-annonces racoleuses. En flânant sur les
Champs-Elysées on ne peut plus promener ses yeux sur un kiosque ou
une vitrine sans rencontrer la couverture d'une revue ou une affiche
de publicité particulièrement suggestives. En faisant ses achats dans
un grand magasin à vocation culturelle, on découvre que les bandes
dessinées pour enfants sont exposées dans le même présentoir que
celles destinées aux adultes, fussent-elles pornographiques.
</TEXT>
</DOC>
```

Figure 4. *Exemple de document*

```
LEMONDE94-000046-19941201-2
LEMONDE94-002606-19940922-1
LEMONDE94-000001-19941201-2
ATS.940921.0142-1
ATS.941130.0140-1
ATS.941129.0021-1
ATS.941130.0072-1
ATS.940405.0079-1
```

Figure 5. *Exemples de jugements de pertinence pour la requête #2*

3.5.3.3 Traitements préparatoires

Avant d'effectuer la recherche effective dans nos documents, nous leur avons appliqué deux traitements préparatoires. Premièrement, nous avons éliminé les mots-outils. Puis, nous avons utilisé un enracineur afin d'éliminer certains suffixes (voir annexe E).

Lors de la procédure d'indexation, pour ne pas encombrer le système de mots qui ne sont pas significatifs, nous avons eu recours à une liste de mots-outils comportant des termes tels que prépositions, conjonctions, articles ou auxiliaires. Nous sommes partis d'une liste de mots-outils existante pour le français[18]. Cette liste ne nous convenait pas tout à fait parce qu'elle comportait des termes qui pouvaient s'avérer utiles dans le contexte de la question-réponse. En effet, des informations telles que des valeurs numériques ne devaient pas être éliminées parce qu'une question factuelle pouvait parfaitement attendre une valeur numérique comme réponse. Forts de ce constat, nous avons retranché de notre liste les adjectifs numériques tels que « premier », « dix-huit » ou « soixante ». Notre liste ainsi modifiée contiendra finalement 421 entrées. La figure 6 illustre un extrait de cette liste.

ça	ceci	celle-ci	celles-ci	celui-ci
car	cela	celle-là	celles-là	celui-là
ce	celle	celles	celui	cependant

Figure 6. *Extrait de notre liste de mots-outils*

Dans un deuxième temps, nous avons appliqué une méthode de racinisation aux documents. La racinisation est un procédé qui permet de regrouper dans une même classe des mots issus de la même racine à travers une ou plusieurs relations morphologiques flexionnelles ou dérivationnelles. La morphologie flexionnelle traite les affixes grammaticaux qui ne modifient ni la catégorie grammaticale ni le sens des mots. Elle concerne des traits tels que le genre, le nombre, la personne, le cas ou le temps. La morphologie dérivationnelle traite de la formation de nouveaux mots, c'est-à-dire d'affixes susceptibles de modifier aussi bien la catégorie grammaticale que le sens d'un mot. Les tableaux 12 et 13 illustrent des exemples de relations morphologiques.

[18] Voir le site http://www.unine.ch/info/clef/

racine	affixes	mot
aim	-er	aimer
aim	-ait	aimait
chien	-s	chiens
petit	-e	petite
parl	-er-i-ons	perlerions

Tableau 12. *Exemples de relations morphologiques flexionnelles*

mot de base	mot dérivé	commentaire
relatif	relativiser	verbe dérivé d'adjectif
centre	central	adjectif dérivé de substantif
préparer	préparation	substantif dérivé de verbe
utile	inutile	adjectif dérivé d'adjectif
raisonnable	raisonnablement	adverbe dérivé d'adjectif

Tableau 13. *Exemples de relations morphologiques dérivationnelles*

Dans notre contexte de recherche, nous avons choisi d'utiliser un enracineur léger afin de ne pas réduire les chances de trouver des réponses précises. En effet, nous ne souhaitions pas de modification de catégorie grammaticale. Par conséquent, notre enracineur s'est restreint à la morphologie flexionnelle en supprimant les suffixes du pluriel, du féminin ainsi que l'infinitif. L'algorithme de racinisation utilisé dans nos expériences est présenté dans l'annexe E.

3.5.3.4 Le modèle de recherche probabiliste Okapi

Parmi les différents modèles de RI disponibles (booléen, vectoriel, probabiliste, etc.), notre choix s'est porté sur le modèle probabiliste. En effet, contrairement au modèle booléen, le modèle probabiliste n'exige pas du document qu'il contienne tous des mots-clés de la requête pour être sélectionné. Il est capable de dépister un document qui contient une partie des termes de la requête en donnant plus de poids aux termes rares dans la collection. Alors que le modèle booléen se révèle efficace sur de grandes collections comme la Toile, le modèle probabiliste offre de bonnes performances même sur de plus petites collections comme celle dont nous disposons.

Un modèle probabiliste estime donc aussi précisément que possible la probabilité qu'un document soit pertinent pour une requête en fonction des informations disponibles. Etant donné une requête Q, le système doit estimer $P_Q(\text{rel} \mid D_i)$ la probabilité que le document D_i soit pertinent (exprimé par *rel*) pour la requête Q.

Pour nos expériences, nous avons utilisé le modèle de recherche probabiliste Okapi [ROB 94], [ROB 95], [ROB 00]. Ce modèle repose sur la pondération des termes de la requête avec prise en compte de la fréquence d'apparition des termes dans le document et la requête ainsi que d'un facteur de correction tenant compte de la longueur du document. La formulation simplifiée de ce modèle est la suivante :

$$RSV(D_i, Q) = \sum_{j=1}^{t} \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}} \cdot w_{ij} \quad (4)$$

$$\text{avec } K = k_1 \left[(1 - b) + b \cdot \frac{l_i}{avdl} \right] \quad (5)$$

dans laquelle k_1 et b sont des constantes, *avdl* indique la taille moyenne d'un document, l_i la longueur du document D_i, tf_{ij} la fréquence d'occurrence du terme t_j dans le document D_i et w_{ij} correspond à la valeur calculée par la formule 6.

$$w_{ij} = tf_{ij} \cdot \ln \frac{n - df_j}{df_j} \quad (6)$$

Dans cette formule, tf_{ij} représente la fréquence d'occurrence du terme t_j dans le document D_i, n le nombre de documents D_i dans la collection et df_j la fréquence documentaire, i.e. le nombre de documents dans lesquels le terme t_j apparaît.

Dans le contexte de système de la question-réponse, notre objectif est de trouver la réponse à une question. Il est dès lors important d'avoir une grande pertinence sur les cinq à dix premiers documents retrouvés. Par conséquent, nous avons retenu la précision après dix documents comme critère d'évaluation et de réglage de notre système. Après une série d'évaluations, nous avons appliqué le modèle Okapi BM25 en fixant aux constantes les valeurs suivantes : $b = 0,8$, $k_1 = 2$ et $avdl = 400$. Pour chaque requête nous avons extrait les 10 meilleurs documents, ce qui nous a donné une précision moyenne de 24,91 % après cinq documents et de 15,26 % après dix documents.

3.5.3.5 Analyse syntaxique

A l'étape précédente, nous avions extrait les dix meilleurs documents pour chaque requête. Il s'agit ensuite de trouver dans ces dix meilleurs documents la réponse à la question posée. Nous avons supposé à ce stade qu'une analyse et compréhension de la question ainsi que des documents pouvaient s'avérer bénéfiques. Nous nous sommes alors orientés vers la linguistique informatique appliquée à la langue française et avons exploité l'analyseur syntaxique FIPS

(*French Interactive Parsing System*) développé au LATL[19] (Laboratoire d'analyse et de Technologie du Langage) de l'Université de Genève [LAE 91], [WEH 97], [WEH 04].

FIPS est un analyseur syntaxique capable d'associer à chaque phrase d'un texte une structure syntaxique reflétant les principales propriétés syntaxiques de cette phrase. Il se base sur des grammaires inspirées des théories chomskyennes [CHO 95], [HAE 94] et attribue à chaque phrase considérée une structure de constituants qui correspond au schéma général suivant :

$$[_{XP}\ L\ X\ R\ ] \qquad (7)$$

dans lequel *XP* est une projection maximale de la tête *X*, *L* et *R* sont des listes, éventuellement vides, de projections maximales des sous-constituants gauche et droit de *X*. La tête *X* est une variable qui peut prendre pour valeur l'ensemble des catégories lexicales suivantes :

- *Adv* adverbe (p.ex. « certainement ») ;
- *A* adjectif (p.ex. « intéressant ») ;
- *N* nom (p.ex « livre ») ;
- *D* déterminant (p.ex « le ») ;
- *V* verbe (p.ex. « lire ») ;
- *P* préposition (p.ex. « sur ») ;
- *C* conjonction (p.ex. « que ») ;
- *Inter* interjection (p.ex. « Eh ! ») ;
- *T* morphème de temps/inflexion.

Pour chaque tête de la projection *X*, on considère sa projection maximale *XP* signifiant « *X Phrase* ». Afin d'illustrer ces catégories, voici un exemple simple avec l'analyse de la phrase « Un joli chat persan étalé. ».

$$[_{DP}\ un\ [_{NP}\ [_{AP}\ joli\ ]\ chat\ [_{AP}\ persan]\ [_{AP}\ étalé]]] \qquad (8)$$

Pour ses analyses, FIPS applique la stratégie suivante. Il commence par lire la phrase mot par mot de gauche à droite. Pour chaque mot, une analyse lexicale permet d'identifier l'élément lexical associé à l'aide d'une base de données lexicale et d'un analyseur morphologique. Puis, cet élément lexical est converti en une projection syntaxique. Lors du traitement du mot suivant, le système tente de combiner le nouvel élément avec la structure immédiatement sur sa gauche. La combinaison des deux constituants produit un nouveau constituant. Ce procédé est

[19] Voir le site http://www.latl.unige.ch/

répété jusqu'à la fin de la phrase. En considérant la phrase un peu plus complexe « Le joli chat persan se promène dans les champs à la recherche d'une souris. » la figure 7 présente l'analyse syntaxique obtenue. La représentation arborescente de cette analyse est illustrée dans l'annexe F.

[$_{TP}$ [$_{DP}$ le [$_{NP}$ [$_{AP}$ joli][$_{N}$ chat [$_{AP}$ [$_{DP}$ ej][$_{A}$ persan]]]]j]i[$_{T}$ se_i promène [$_{VP}$ [$_{AdvP}$ [$_{PP}$ dans [$_{DP}$ les [$_{NP}$ champs]]]][$_{AdvP}$ [$_{PP}$ à [$_{DP}$ la [$_{NP}$ recherche [$_{PP}$ d' [$_{DP}$ une [$_{NP}$ souris]]]]]]]]]

Figure 7. *Exemple d'analyse syntaxique*

3.5.3.6 Reconnaissance des entités nommées

Les entités nommées se réfèrent à des concepts uniques et partagés. Elles comprennent des éléments tels que :

- organisations (entreprises, administrations, musées, etc.)
- lieux (villes, régions, pays, etc.)
- personnes (hommes politiques, artistes, chefs d'entreprises, etc.)
- valeurs numériques (poids, longueurs, monnaies, etc.)

La reconnaissance d'entités nommées sera utilisée pour la classification des questions ainsi que l'identification des types de réponses attendues dans les documents.

Comme nous disposions de FIPS pour l'analyse syntaxique des phrases, nous nous sommes interrogés sur la capacité de cet outil à reconnaître les entités nommées. Etant donné la grande quantité de données générées lors de l'analyse, l'équipe du LATL a procédé à une adaptation afin de rendre exploitables certaines informations pour identifier diverses entités. Il était dès lors possible de reconnaître deux types d'entités nommées, les valeurs numériques et les valeurs nominales. Le tableau 14 présente la liste des entités numériques disponibles alors que le tableau 15 montre les entités nommées nominales reconnues.

Entité nommée	Exemple
numérique	premier
pourcent	23 %
ordinal	1er
nombre spécial	751.04.09
cardinal	1291
nombre	12, douze

Tableau 14. *Entités nommées numériques reconnues par FIPS*

Entité nommée	Exemple	Entité nommée	Exemple
humain	homme	action	grève
animé	chat	collectif	équipe
quantité	kilo	pays	Suisse
temps	heure	ville	Paris
jour	lundi	rivière	Gange
mois	mai	mountain	Everest
poids	gramme	personne	John
longueur	mètre	nom propre	Yangze
lieu	bureau	entreprise	IBM
abstraction	liberté	titre	Monsieur
objet physique	livre	fonction	président

Tableau 15. *Entités nommées nominales reconnues par FIPS*

Finalement, en utilisant toutes les possibilités mises à notre disposition par FIPS, nous obtenons davantage d'informations que nous illustrons pour la phrase « Le joli chat persan se promène dans les champs à la recherche d'une souris. » dont l'analyse syntaxique était présentée dans la figure 7. Le tableau 16 montre les informations disponibles pour chaque mot de la phrase. La première colonne contient le terme rencontré dans la phrase, la deuxième colonne en donne les traits morphologiques, la troisième annonce le numéro de concept associé qui est lié à la représentation interne, la quatrième donne la liste des entités nommées reconnues. L'avant-dernière colonne annonce la position du terme dans la phrase et la dernière colonne présente le lemme utilisé comme entrée de dictionnaire. Par exemple, si nous nous attachons au dernier mot « souris », ses traits morphologiques sont « *NOM-SIN-FEM* » ce qui signifie qu'il s'agit d'un nom singulier féminin. La reconnaissance d'entités nommées « {1} » nous indique par ce code qu'il s'agit d'une créature animée. Enfin, l'entrée du dictionnaire est identique au mot, à savoir « souris ».

Terme	Traits morphologiques	N° concept	Entités nommées	Position	Lemme
le	DET-SIN-MAS	211045001		0	le
joli	ADJ-SIN-MAS	211013949		2	joli
chat	NOM-SIN-MAS	211000112	{1, 13}	7	chat
persan	ADJ-SIN-MAS	211017577		12	persan
se	PRO-CLI-SIN-ING	211000020		19	se
promène	VER-IND-PRE-3-SIN	211015582		22	promener
dans	PRE	211045077		30	dans
les	DET-PLU-MAS	211045001		35	le
champs	NOM-PLU-MAS	211001730	{13}	39	champ
à	PRE	211045070		46	à
la	DET-SIN-FEM	211045001		48	le
recherche	NOM-SIN-FEM	211006909	{9}	51	recherche
d'	PRE	211047305		61	de
une	DET-SIN-FEM	211045921 211045922	{5}	64	un
souris	NOM-SIN-FEM	211023936	{1}	67	souris

Tableau 16. *Informations retournées pour les mots de la phrase*

3.5.3.7 Analyse des questions

A l'issue de l'étape précédente, nous disposions pour chaque question de son analyse syntaxique ainsi que diverses autres données. Il s'agit ensuite d'exploiter ces informations afin de définir les types de réponses attendues. Pour ce faire, nous avons commencé par sélectionner dans la question les termes que nous jugeons pertinents. Un terme est considéré pertinent lorsque son *idf* (*inverted document frequency* ou fréquence documentaire inverse) est supérieur à 3,5. L'*idf* mesure la fréquence d'apparition d'un terme dans la collection. Par exemple, le terme « *computer* » aura un *idf* grand dans une collection généraliste, ce qui signifie qu'il est rare donc important dans ce contexte. Par contre, le même terme aura un *idf* petit dans une collection spécialisée dans l'informatique, ce qui signifie qu'il est trop fréquent pour être discriminant. L'*idf* s'exprime de la manière suivante :

$$\mathrm{idf} = \ln\left(n \diagup \mathrm{df}\right) \tag{9}$$

Dans cette expression, *n* dénote le nombre de documents dans la collection et *df* le nombre de documents qui contiennent le terme. Le seuil a été fixé empiriquement à la valeur 3,5, ce qui signifie pour notre collection de 730 098 documents un *df* situé autour de 20 000. Il faut préciser que *df* ne peut pas être nul car le calcul de l'*idf* n'a de sens que si le terme existe dans la collection, i.e. $df >= 1$.

Après avoir sélectionné les termes pertinents dans la requête, nous nous sommes intéressés au mot interrogatif qui débute généralement une question, à l'exception de questions déclaratives telles que « Citer le nom d'une capitale européenne. ». Nous avons recensé les mots interrogatifs les plus fréquents et n'en avons retenu que la formulation canonique servant d'entrée dans le dictionnaire puisque FIPS est capable de retrouver le lemme d'un terme. Cette liste est présentée dans la figure 8.

combien, comment, pourquoi, quand, que, quel, qui, quoi, où

Figure 8. *Mots interrogatifs considérés*

Sur la base des mots interrogatifs retenus, nous avons défini une classification des types de réponses attendues. Pour ce faire, nous avons d'abord identifié la cible de la question en prenant le premier terme suivant le mot interrogatif qui a comme trait morphologique la valeur « NOM », ce qui correspond à un substantif. Lorsque aucun mot interrogatif n'est trouvé dans une requête, la cible est cherchée de la même manière mais depuis le début de la question. Le tableau 17 illustre les classes de questions avec leurs cibles respectives.

Classe	Mot interrogatif	Cible spécifique	Exemple
Classe 1	quel, quoi, comment, pourquoi, que, qu'est-ce que	-	Comment appelle-t-on l'intérieur d'un bateau ? Qu'a inventé le baron Marcel Bich ?
Classe 2	où	-	Où se trouve le siège de l'OCDE ?
Classe 3	combien quel + cible numérique aucun + cible numérique	cible numérique: pourcentage, nombre, quantité, distance, poids, longueur, hauteur, largeur, âge, grandeur, dimension, superficie	Combien de membres compte l'OCDE ? A quel âge est mort Massimo Troisi ?
Classe 4	quand, quel + cible temporelle aucun + cible temporelle	cible temporelle: date, jour, mois, année, an, époque, période	Quand est né Albert Einstein ? En quelle année est né Alberto Giacometti ?
Classe 5	qui, quel + cible fonctionnelle aucun + cible fonctionnelle	cible fonctionnelle: président, directeur, ministre, juge, sénateur, acteur, chanteur, artiste, présentateur, réalisateur	Qui est Jacques Chirac ? Quel est le président du parti socialiste suisse ?
Classe 6	-	-	Donnez le nom d'un liquide inodore et insipide.

Tableau 17. *Classification des questions*

Nous avons toutefois éliminé des cibles potentielles les termes peu intéressants parce que trop généraux. La liste des cibles exclues est présentée dans la figure 9.

nombre, quantité, grandeur, dimension, date, jour, mois, an, année, époque, période, nom, surnom, titre, lieu

Figure 9. *Cibles exclues*

Après avoir assigné des classes aux questions, nous avons identifié pour chaque classe les types de réponses attendues comme le montre le tableau 18.

Classe	**Types de réponses attendues**
Classe 1	toutes les entités nommées nominales
Classe 2	lieu, pays, ville, rivière, montagne, nom propre
Classe 3	quantité, poids, longueur et toutes les entités nommées numériques
Classe 4	temps, jour, mois, numérique, ordinal, nombre spécial, cardinal, chiffre
Classe 5	humain, animé, collectif, personne, entreprise, titre, fonction, nom propre
Classe 6	toutes les entités nommées nominales

Tableau 18. *Types de réponses attendues par classe*

3.5.3.8 Ordonnancement des phrases

Après avoir analysé les questions et avoir identifié les types de réponses attendues, il s'agissait de rechercher dans les documents sélectionnés les phrases susceptibles de contenir la réponse. Pour cela, nous avons également appliqué l'analyse syntaxique sur les documents. FIPS ayant la capacité de décomposer des paragraphes en phrases avant de procéder à l'analyse syntaxique de chacune d'entre elles, nous disposions donc pour chaque document de l'ensemble de ses phrases dûment analysées. Il fallait ensuite classer les phrases par ordre de pertinence relativement à la question. A cette fin, nous avons calculé pour chaque phrase des dix meilleurs documents retenus un score selon la formule suivante :

$$\text{scoreS} = \frac{\text{srt} \cdot \text{sl}}{\text{sl} - \text{qrt}} \tag{10}$$

dans laquelle *srt* (*sentence relevant terms*) désigne le nombre de termes jugés pertinents dans la phrase, *sl* (*sentence length*) le nombre de termes de la phrase et *qrt* (*query relevant terms*) le nombre de termes jugés pertinents dans la requête. Nous avons ensuite classé les phrases dans l'ordre décroissant de leurs scores et avons retenu pour la suite du processus les dix phrases ayant les scores les plus élevés. Ces phrases deviennent les meilleures phrases candidates susceptibles de contenir la réponse à la question. Le tableau 19 donne comme exemple les meilleures phrases pour la question « Où se trouve la mosquée Al Aqsa ? ».

Rang	Scores	Document	Phrase
1	2,148	ATS.950417.0033-9	la police interdit aux juifs de prier sur l'esplanade où se trouve la mosquée al-Aqsa , troisième lieu saint de l'islam après la Mecque et Médine.
2	2,103	ATS.940304.0093-2	la police a expliqué qu' elle bouclait le site le plus sacré du judaïsme jusqu' à la fin de la prière du vendredi à la mosquée Al -- Aqsa , laquelle se trouve sur l' Esplanade du Temple qui domine le Mur des Lamentations.
3	1,4	ATS.940405.0112-1	la mosquée al Aqsa rouverte aux touristes.
4	1,118	ATS.940606.0081-4	cette phrase laisse ouverte la possibilité pour M. Arafat d'aller prier à la mosquée al-Aqsa à Jérusalem.
5	1,095	LEMONDE94-001632-19940514-7	ce camp, dénommé Hanan par Tsahal, a été immédiatement rebaptisé Al Aqsa, du nom de la grande mosquée de Jérusalem.
6	1,091	ATS.941107.0105-3	la mosquée Al Aqsa, à Jérusalem, est le troisième lieu saint de l'Islam après la Mecque et Médine en Arabie Saoudite.
7	1,067	ATS.940304.0093-3	des centaines de policiers ont été déployés à l'intérieur et à l'extérieur de la Vieille Ville vendredi, limitant l'accès à la mosquée Al-Aqsa, l'un des lieux saints de l'islam.
8	1,063	LEMONDE94-001740-19940820-9	" si nous conservons l'ensemble de la ville, il faudra maintenir sous notre protection la mosquée d'al-Aqsa et le Saint-Sépulcre, ainsi que 180 000 Arabes qui nous haïssent à mort.
9	1,063	ATS.940405.0112-1	le Waqf avait interdit l'accès de la mosquée aux non - musulmans en mars pour protester contre l'attitude de la police israélienne qui limitait l'entrée des musulmans au lieu saint.
10	1,059	ATS.951223.0020-3	" demain nous allons prier ici à Bethléem à la Mosquée d' Al Aqsa à Jérusalem et bientôt nous irons prier au St Sépulcre ainsi qu' ", a lancé M. Arafat sous un tonnerre d'applaudissements.

Tableau 19. *Phrases retenues pour la question considérée*

3.5.3.9 Extraction de fragments de texte

L'étape précédente nous a permis de disposer des dix phrases les plus prometteuses pour chaque requête avec leurs analyses syntaxiques. Mais une phrase est une entité trop grande pour fournir une réponse précise, si bien qu'il fallait descendre en-dessous du niveau de la phrase pour considérer des fragments de phrase. Comme la cible de la question est connue à ce stade, nous avons recherché cette cible dans les phrases candidates. Si aucune phrase ne contient la cible, alors la première phrase est retenue pour la suite du processus.

Pour chaque phrase candidate, si la cible a été trouvée, nous avons considéré une fenêtre d'au maximum neuf termes centrée sur la cible (quatre au plus de chaque côté). Dans cette fenêtre, nous avons recherché les termes pertinents correspondant aux types de réponses attendues. Pour chacun de ces termes, nous avons extrait un groupe syntaxique de type DP (*determiner phrase* ou groupe déterminant) ou NP (*noun phrase* ou groupe nom) à partir de la représentation arborescente de l'analyse syntaxique. Ainsi, chaque phrase peut générer un ou plusieurs groupes syntaxiques intéressants.

Le procédé décrit ci-dessus était susceptible de générer un sous-groupe à partir d'un groupe de base. Afin d'éviter la production inutile de partitions de groupes, nous avons éliminé les fragments contenus dans un groupe alors que la différence de niveau entre le groupe et le sous-groupe est inférieure à sept. Le niveau correspond à la profondeur du groupe considéré dans l'arbre syntaxique de la phrase. Le seuil a été choisi empiriquement.

D'autre part, nous ne souhaitions pas voir apparaître des termes de la requête dans les fragments retenus, si bien que nous avons élagué les parties des fragments contenant des termes de la requête. Nous avons ensuite constaté que les diverses étapes de « nettoyage » des fragments ont pu provoquer la disparition des termes pertinents associés aux types de réponses attendues. Par conséquent, nous avons encore éliminé des fragments restants tous ceux qui ne contenaient plus de termes pertinents.

Parallèlement, nous avons calculé un score relatif pour chacune des phrases candidates selon la formule suivante :

$$\text{scoreR} = \text{scoreS} \Big/ \max \text{ScoreS} \qquad (11)$$

dans laquelle *scoreS* désigne le score initial de la phrase considérée et *maxScoreS* le score de la meilleure phrase. Si le *maxScoreS* vaut 0, alors le score relatif est aussi mis à 0. Il est à remarquer que ce score ne tient pas compte de l'ordre des mots. Le tableau 20 montre la liste des fragments restants pour la requête « Où se trouve la mosquée Al Aqsa ? » avec les scores associés.

Document	ScoreR	Fragment
ATS.940304.0093	0,978	Al -- Aqsa
ATS.940606.0081	0,520	M. Arafat
ATS.940606.0081	0,520	Jérusalem
LEMONDE94-001632-19940514	0,509	Jérusalem
ATS.941107.0105	0,507	Jérusalem
ATS.940304.0093	0,496	Ville
ATS.940304.0093	0,496	Al-Aqsa l'un des lieux saints de l'islam
LEMONDE94-001740-19940820	0,494	le Saint-Sépulcre
ATS.940405.0112	0,494	le Waqf
ATS.951223.0020	0,492	à Jérusalem
ATS.951223.0020	0,492	Bethléem

Tableau 20. *Fragments restants pour la requête considérée*

3.5.3.10 Choix de la réponse

L'étape précédente nous a permis d'extraite un ensemble de réponses possibles. Il reste à choisir parmi ces réponses celle qui nous paraîtra la plus adéquate. Dans un premier temps, nous avons sélectionné la réponse dont le score relatif était le plus haut. Mais cette technique ne s'est pas révélée efficace. En effet, une candidate moins bien classée peut néanmoins constituer une bonne réponse si elle est supportée par plusieurs passages. Partant de ce constat, nous avons imaginé une procédure de vote afin de sélectionner la réponse que notre système va retourner.

Pour cela, nous avons décomposé les fragments en mots puis nous avons compté la fréquence d'apparition de chaque mot qui n'est pas un mot-outil dans tous les autres fragments. Cela nous a permis de calculer un score basé sur la redondance pour chaque réponse selon la formule suivante :

$$\text{scoreF} = \frac{\sum_{j} \text{tf}_j}{\text{len}} \qquad (12)$$

dans laquelle tf_j désigne la fréquence d'apparition du terme j dans les autres fragments et *len* le nombre de mots du fragment ou 1 lorsqu'il s'agissait d'une question de définition. En effet, comme les définitions sont potentiellement plus longues que les réponses factuelles, nous ne voulions pas pénaliser les longues réponses lorsqu'une définition était attendue.

Nous avons ensuite classé les fragments par ordre décroissant selon leur *scoreF* puis *scoreR*. Lorsque le plus haut score était nul, nous avons choisi le premier fragment mais avons baissé son score relatif (division par deux). Le tableau 21 illustre les scores obtenus sur la base du vote alors que le tableau 22 montre la réponse finalement sélectionnée par notre système.

Document	ScoreF	ScoreR	Fragment
ATS.940606.0081	3	0,520	Jérusalem
LEMONDE94-001632-19940514	3	0,509	Jérusalem
ATS.941107.0105	3	0,507	Jérusalem
ATS.940304.0093	1	0,978	Al -- Aqsa
ATS.940606.0081	0	0,520	M. Arafat
ATS.951223.0020	0.5	0,492	à Jérusalem
ATS.940304.0093	0.2	0,496	Al-Aqsa l'un des lieux saints de l'islam
ATS.940304.0093	0	0,496	Ville
LEMONDE94-001740-19940820	0	0,494	le Saint-Sépulcre
ATS.940405.0112	0	0,494	le Waqf
ATS.951223.0020	0	0,492	Bethléem

Tableau 21. *Scores obtenus après le vote*

Document	ScoreR	ScoreF	Fragment
ATS.940606.0081	0,520	3	Jérusalem

Tableau 22. *Réponse retournée par notre système pour la question considérée*

3.5.4 Système de question-réponse bilingue

3.5.4.1 Introduction

Dans le chapitre précédent, nous avons présenté notre système de question-réponse monolingue pour le français. Nous avons souhaité étendre le champ d'application de notre système en autorisant les questions formulées dans d'autres langues européennes que le français. Pour ce faire, nous avons choisi de traduire automatiquement les requêtes en français puis d'utiliser notre système monolingue pour rechercher les réponses dans la collection française.

3.5.4.2 Traduction automatique des questions

Nous souhaitions pouvoir traiter des questions formulées dans diverses langues européennes, à savoir le bulgare (BG), l'allemand (DE), l'anglais (EN), l'espagnol (ES), l'italien (IT), le néerlandais (NL) et le portugais (PT). Pour dépasser les

frontières linguistiques, notre approche s'est basée sur l'exploitation de ressources de traduction automatique disponibles gratuitement sur la Toile. Les ressources utilisées sont les suivantes :

1. Reverso (www.reverso.fr)
2. TranslationExperts.com (intertran.tranexp.com)
3. Free2Professional Translation (www.freetranslation.com)
4. AltaVista (babelfish.altavista.com)
5. Systran (www.systranlinks.com)
6. Google.com (www.google.com/language_tools)
7. WorldLingo (www.worldlingo.com)

Le tableau 23 présente les langues sources disponibles alors que le français est choisi comme langue cible. La traduction choisie pour chaque langue source est identifiée par une étoile (*).

	Langue source						
Ressource de traduction	**BG**	**DE**	**EN**	**ES**	**IT**	**NL**	**PT**
Reverso		√ *	√ *	√ *			
TranslationExperts.com	√ *	√	√	√	√	√	√
Free2Professional Translation			√				
AltaVista		√	√	√	√	√	√
Systran		√	√	√	√	√	√
Google.com		√	√				
WorldLingo		√	√	√	√ *	√ *	√ *

Tableau 23. *Ressources de traduction disponible avec le français comme cible*

Le bulgare étant une langue à alphabet cyrillique, les mots pour lesquels la ressource ne disposait pas de traduction restent en caractères cyrilliques. Nous avons par conséquent ajouté une étape de traitement spécifique à cette langue. Il s'agit d'appliquer une translittération après la traduction afin d'obtenir les termes non traduits en alphabet latin. Pour cela, nous avons utilisé une table de conversion disponible sur la Toile à l'adresse www.world-gazetter.com/pronun.htm#cyr. Le tableau 24 illustre les traductions obtenues à partir de questions équivalentes à la question française « Quel est le directeur général de FIAT ? »

Langue source	Question originale	Question traduite
Bulgare	Кой е управителният директор на ФИАТ?	Qui å upravitelniiat direktor na FIAT?
Allemand	Wer ist der Geschäftsführer von FIAT?	Qui est le directeur de FIAT ?
Anglais	Who is the managing director of FIAT?	Qui est le directeur général de DéCRET ?
Espagnol	¿Quién es el director gerente de FIAT?	Qui est-ce qui est le directeur gérant de CONSENTEMENT ?
Italien	Chi è l'amministratore delegato della Fiat?	Qui est le directeur exécutif général de Fiat ?
Néerlandais	Wie is de bestuursvoorzitter van Fiat?	Qui est-il le président d'administration de fiat ?
Portugais	Quem é o administrador-delegado da Fiat?	Qui est l'agent d'administrateur-commission de Fiat ?

Tableau 24. *Exemples des traductions obtenues*

3.5.4.3 Application du système monolingue

Après avoir obtenu la traduction française des questions formulées dans d'autres langues européennes, nous avons appliqué notre système monolingue pour répondre aux questions posées. Naturellement, la phase de traduction détériore considérablement la performance du système. En effet, l'analyseur syntaxique FIPS fait l'hypothèse forte que le texte à analyser est correct. Or, les traducteurs n'offrent pas cette garantie. Lors de la campagne d'évaluation CLEF 2004 nous avons mesuré cette détérioration qui sera discutée dans la section suivante dédiée à l'évaluation.

3.6 Evaluation

3.6.1 Introduction

Notre système de question-réponse a été proposé et évalué dans le cadre de deux campagnes d'évaluation en recherche d'information. Premièrement, nous avons participé à QA@CLEF-2004, la tâche de question-réponse multilingue de CLEF. Cette évaluation est décrite dans la section 3.6.2. Puis, nous avons soumis notre système à EQueR qui est une évaluation de systèmes de question-réponse pour la langue française. Cette évaluation est décrite dans la section 3.6.3.

3.6.2 Campagne d'évaluation CLEF

3.6.2.1 Introduction

Notre système de question-réponse monolingue et bilingue a été évalué dans le cadre de la campagne d'évaluation CLEF 2004 [MAG 04], [PER 04a]. Parmi les diverses tâches offertes, nous avons choisi celle dédiée aux systèmes de question-réponse « CLEF 2004 Multilingual Question Answering Track » aussi appelée QA@CLEF-2004. En 2003 avait lieu la première évaluation de systèmes de question-réponse multilingues. Trois langues avaient alors été traitées dans la tâche monolingue (néerlandais, italien et espagnol). Dans la tâche bilingue, seul l'anglais était offert comme langue cible alors que les requêtes étaient formulées dans cinq langues sources (néerlandais, français, allemand, italien et espagnol). Pour l'édition 2004, la piste QA@CLEF-2004 a considérablement augmenté l'offre des langues disponibles. En effet, neuf langues sources (bulgare, néerlandais, anglais, finnois, français, allemand, italien, portugais et espagnol) et sept langues cibles (néerlandais, anglais, français, allemand, italien, portugais et espagnol) étaient mises à disposition. Comme il n'existait pas de collections de documents pour le bulgare et le finnois, ces deux langues n'étaient offertes que comme langues sources. D'autre part, l'anglais étant traditionnellement évalué dans la piste question-réponse de la campagne TREC, il n'était pas disponible pour la tâche monolingue. Outre ces deux restrictions, toutes les combinaisons entre langues sources et cibles étaient autorisées. Cela représentait un total de 56 tâches possibles, 6 monolingues et 50 bilingues. Puisque le français était disponible à la fois comme source et cible, il nous était alors possible de présenter notre système de question-réponse pour les deux types de tâches.

3.6.2.2 Collection de documents

Pour les sous-tâches ayant le français comme langue cible, trois corpus de documents ont été sélectionnés. Il s'agit de *Le Monde 1994*, *ATS 1994* et *ATS 1995*. Le tableau 25 donne des détails sur ces collections. Un exemple de document est présenté dans l'annexe G.

Corpus	Année	Taille	Nb documents
Le Monde	1994	157 Mo	44 013
ATS français	1994	86 Mo	43 178
ATS français	1995	88 Mo	42 615
Total		**331 Mo**	**129 806**

Tableau 25. *Corpus français de QA@CLEF-2004*

3.6.2.3 Questions

Pour chaque sous-tâche, un ensemble de 200 questions était proposé, 90 % d'entre elles étant de type factuel et 10 % de type définition. De plus, 10 % des questions n'avaient pas de réponses connues dans la collection. Le tableau 26 présente un extrait des questions formulées en français alors que le tableau 27 illustre les mêmes questions exprimées en allemand. Dans ces deux tableaux, la première colonne indique le type de la question, à savoir F pour factuelle et D pour définition. La deuxième colonne contient la langue source et la troisième colonne la langue cible. Par exemple, le tableau 27 présente des questions écrites en allemand (langue source) pour lesquelles il s'agit de trouver des réponses en français (langue cible). Puis, la quatrième colonne indique le numéro de la question et enfin la dernière colonne contient la question.

Type	Source	Cible	Nb	Question
F	FR	FR	109	Quand est-ce que l'Autriche a adhéré à Schengen ?
F	FR	FR	110	Comment s'appelle le parti de Silvio Berlusconi ?
F	FR	FR	111	Combien de catholiques compte l'Afrique ?
D	FR	FR	112	Qui est Philippe Biberson ?
D	FR	FR	113	Qu'est-ce qu'Eurotunnel ?
F	FR	FR	114	Quand a eu lieu l'attaque de Pearl Harbor ?
F	FR	FR	115	Quel est le réalisateur de Nikita ?
F	FR	FR	116	Comment s'appelle la reine des Pays Bas ?
D	FR	FR	117	Qui est Montxo Armendariz ?
F	FR	FR	118	Qui a remporté la palme d'or a Cannes en 1995 ?

Tableau 26. *Quelques exemples de questions en français*

Type	Source	Cible	Nb	Question
F	DE	FR	109	Wann trat Österreich dem Schengener Abkommen bei?
F	DE	FR	110	Wie heißt die Partei von Silvio Berlusconi?
F	DE	FR	111	Wie viele Katholiken gibt es in Afrika?
D	DE	FR	112	Wer ist Philippe Biberson?
D	DE	FR	113	Was ist Eurotunnel?
F	DE	FR	114	Wann fand der Angriff auf Pearl Harbor statt?
F	DE	FR	115	Wer führte Regie in "Nikita"?
F	DE	FR	116	Wie heißt die Königin der Niederlande?
D	DE	FR	117	Wer ist Montxo Armendariz?
F	DE	FR	118	Wer gewann 1995 die Goldene Palme in Cannes?

Tableau 27. *Quelques exemples de questions en allemand*

3.6.2.4 Réponses

Pour chaque question posée, il s'agissait de retourner une seule réponse, la plus précise possible. Les questions factuelles étaient supposées recevoir des réponses très courtes alors que les questions de définition étaient susceptibles de nécessiter des réponses plus longues comme des phrases entières ou même des passages.

3.6.2.5 Tâches effectuées

Parmi les 56 tâches disponibles, nous en avons effectué huit, à savoir une tâche monolingue en français et sept tâches bilingues ayant pour cible le français et pour source l'allemand, l'anglais, l'espagnol, l'italien, le néerlandais et le portugais. Le tableau 28 illustre ce choix, en gris clair sont indiquées les tâches bilingues et en gris foncé la tâche monolingue. Le *X* désigne les autres tâches effectuées par d'autres participants.

		Langues cibles						
		DE	**EN**	**ES**	**FR**	**IT**	**NL**	**PT**
Langues sources	**BG**		X					
	DE	X	X					
	EN						X	
	ES			X				
	FI		X					
	FR		X					
	IT		X			X		
	NL						X	
	PT							X

Tableau 28. *Tâches effectuées*

3.6.2.6 Mesure d'évaluation

Des juges humains ont évalué les réponses proposées sur la base de deux critères, la correction et l'exactitude de la réponse. La correction signifie que la réponse est correcte pour la question considérée. L'exactitude signifie que la réponse est exacte, c'est-à-dire correcte et ni trop longue ni trop courte. Pour chaque réponse, une des quatre appréciations suivantes est attribuée :

- *R* pour une réponse correcte ;
- *X* pour une réponse inexacte ;
- *U* pour une réponse non supportée ;
- *W* pour une réponse incorrecte.

Une réponse est jugée correcte lorsqu'elle répond précisément à la question et qu'elle est supportée par le document spécifié. Par exemple, la paire ["Cesare Romiti", ATS.940531.0063] a été jugée correcte pour la question #1, « Quel est le directeur général de FIAT ? », parce que le document référencé contient la phrase « directeur général de Fiat Cesare Romiti ». D'autre part, une réponse est jugée inexacte lorsque celle-ci est trop longue ou trop courte par rapport à la réponse correcte bien qu'elle soit supportée par le document retourné. Par exemple, la paire ["premier ministre irlandais", ATS.940918.0057] a été jugée inexacte pour la question #177, « Quelle est la fonction d'Albert Reynolds en Irlande ? », parce que l'adjectif « irlandais » était redondant. Puis, une réponse est jugée non supportée lorsque le document proposé ne contient pas la réponse ou ne la justifie pas. Enfin, une réponse est jugée incorrecte lorsqu'elle ne répond pas à la question posée. Par exemple, la paire ["Underground", ATS.950528.0053] a été jugée incorrecte pour la question #118, « Qui a remporté la palme d'or à Cannes en 1995 ? ». En effet, « Underground » est le titre du film alors que « Emir Kusturica », son réalisateur, aurait constitué la bonne réponse. Puisque chaque question ne pouvait recevoir qu'une seule réponse, la mesure d'évaluation choisie est la justesse (*accuracy*) qui s'exprime comme suit :

$$\text{accuracy} = \frac{\sum_{j=1}^{200} \text{correct}(j)}{200} \tag{13}$$

$$\text{correct}(j) = \begin{cases} 1 \text{ si réponse correcte à la question } j \\ 0 \text{ sinon} \end{cases} \tag{14}$$

3.6.2.7 Résultats

Lors de cette édition, 18 groupes ont participé aux diverses tâches de question-réponse. Sur les 20 soumissions monolingues, la justesse moyenne s'élève à 23,7 % alors qu'elle descend à 14,7 % sur les 28 soumissions bilingues. Nous avons donc une première estimation de la perte de qualité entraînée par la traduction automatique. Dans le tableau 29, la deuxième ligne présente les résultats obtenus pour la tâche monolingue en français [MAG 04]. Les lignes suivantes montrent les résultats pour les tâches bilingues ayant le français pour langue cible et la langue spécifiée pour langue source. L'avant-dernière colonne montre la justesse calculée selon la formule 13 et la dernière colonne illustre la perte de justesse en pourcentage engendrée par l'étape de traduction de la langue source vers le français.

	Langue source	Correct	Inexact	Non supporté	Incorrect	Justesse	Perte traduction
	FR	49	6	0	145	24,5 %	-
Bilingue	**DE**	34	12	0	154	17,0 %	-30,6 %
	ES	34	4	0	162	17,0 %	-30,6 %
	NL	29	15	0	156	14,5 %	-40,8 %
	IT	29	7	0	164	14,5 %	-40,8 %
	PT	29	7	0	164	14,5 %	-40,8 %
	EN	27	9	0	164	13,5 %	-44,9 %
	BG	13	7	0	180	6,5 %	-73,5 %

Tableau 29. *Résultats obtenus*

Il était prévisible que la phase de traduction provoque une détérioration des résultats par rapport à la tâche monolingue. Cette détérioration est fonction directe de la langue source ainsi que de la qualité des ressources de traduction automatique utilisées. Il n'est pas surprenant de trouver le bulgare en dernière position (-73,5 %) car il existe peu de traducteurs pour cette langue, qui sont de faible qualité. Par contre, il est étonnant de trouver l'anglais en avant-dernière position avec une perte de 44,9 %. Cela provient probablement du fait que la traduction obtenue à partir de l'anglais contient tellement d'erreurs que l'analyseur FIPS peine à la décortiquer, affectant ainsi la suite du processus.

Lorsqu'on analyse plus précisément ces résultats, on peut trouver que pour 7,5 % des questions (15 / 200) la majorité des traductions (plus de quatre différentes langues sources) génèrent des réponses correctes alors que pour 2,5 % des questions (5 / 200) une majorité de traductions génèrent des réponses inexactes. Cela suggère que la traduction n'affecte pas considérablement la faculté du système à produire des réponses correctes ou inexactes dans 10 % des questions (20 / 200). Cela signifie donc que la traduction affecte la faculté du système à produire des réponses correctes ou inexactes dans 90 % des questions.

Il était aussi intéressant de dissocier les questions factuelles de celle de définition pour voir les performances relatives obtenues. Le tableau 30 montre pour chaque tâche la justesse globale calculée selon la formule 13 puis la justesse respectivement des questions factuelles et de définition.

	Langue source	Justesse globale [200]		Justesse factuelle [180]		Justesse définition [20]	
	FR	49	24,5 %	43	23,89 %	6	30,0 %
Bilingue	**DE**	34	17,0 %	28	15,56 %	6	30,0 %
	ES	34	17,0 %	31	17,22 %	3	15,0 %
	NL	29	14,5 %	24	13,33 %	5	25,0 %
	IT	29	14,5 %	28	15,56 %	1	5,0 %
	PT	29	14,5 %	24	13,33 %	5	25,0 %
	EN	27	13,5 %	22	12,22 %	5	25,0 %
	BG	13	6,5 %	11	6,11 %	2	10,0 %

Tableau 30. *Justesse par type de question*

Un autre point de vue sur la performance de notre système consiste à analyser les réponses selon les types de réponse attendue. On peut classer les questions selon les dix types suivants (classification proposée par CLEF) :

- définition - organisation (dorg) pour une organisation
- définition - personne (dper) pour une personne
- factuelle - lieu (floc) pour un lieu
- factuelle - manière (fman) pour une manière
- factuelle - mesure (fmea) pour une mesure
- factuelle - objet (fobj) pour un objet
- factuelle - organisation (forg) pour une organisation
- factuelle - autre (foth) pour d'autres types
- factuelle - personne (fper) pour une personne
- factuelle - temps (ftim) pour une référence temporelle

Le tableau 31 montre le nombre de réponses correctes obtenues par type de réponse et le pourcentage correspondant. Dans la deuxième colonne, le nombre correspond au nombre de questions de ce type parmi les 200 questions. L'avant-dernière ligne présente le nombre de réponses correctes tous types confondus alors que la dernière ligne en donne la justesse moyenne.

	Type	FR	DE	ES	NL	IT	PT	EN	BG
Déf.	**dorg [8]**	0	0	0	0	0	0	0	0
		0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
	dper [12]	6	6	3	5	1	5	5	2
		50,0 %	50,0 %	25,0 %	41,7 %	8,3 %	41,7 %	41,7 %	16,7 %
Factuelle	**floc [29]**	10	7	8	5	5	5	7	2
		34,5 %	24,1 %	27,6 %	17,2 %	17,2 %	17,2 %	24,1 %	6,9 %
	fman [14]	0	0	0	0	1	0	0	1
		0,0 %	0,0 %	0,0 %	0,0 %	7,1 %	0,0 %	0,0 %	7,1 %
	fmea [28]	9	5	4	7	4	5	5	2
		32,1 %	17,9 %	14,3 %	25,0 %	14,3 %	17,9 %	17,9 %	7,1 %
	fobj [15]	1	3	2	2	3	2	1	2
		6,7 %	20,0 %	13,3 %	13,3 %	20,0 %	13,3 %	6,7 %	13,3 %
	forg [20]	6	3	5	2	4	2	2	0
		30,0 %	15,0 %	25,0 %	10,0 %	20,0 %	10,0 %	10,0 %	0,0 %
	foth [21]	6	2	3	4	3	3	1	2
		28,6 %	9,5 %	14,3 %	19,0 %	14,3 %	14,3 %	4,8 %	9,5 %
	fper [32]	4	6	4	3	4	3	4	2
		12,5 %	18,8 %	12,5 %	9,4 %	12,5 %	9,4 %	12,5 %	6,3 %
	ftim [21]	7	2	5	1	4	4	2	0
		33,3 %	9,5 %	23,8 %	4,8 %	19,0 %	19,0 %	9,5 %	0,0 %
Total	**Nb**	49	34	34	29	29	29	27	13
	%	24,5 %	17,0 %	17,0 %	14,5 %	14,5 %	14,5 %	13,5 %	6,5 %

Tableau 31. *Réponses correctes selon le type de réponse*

3.6.2.8 Analyse des résultats

Nous pouvons déduire que notre système produit de bonnes réponses lorsqu'il s'agit de retrouver des personnes, des mesures ou des lieux. En revanche, il s'avère moins efficace lorsque la question porte sur des définitions d'organisations ou sur la manière dont quelque chose se passe (par exemple « Comment est mort Jimi Hendrix ? »).

Nous nous sommes également intéressés aux causes ayant conduit à la production de réponses incorrectes outre la phase de traduction abordée précédemment. Tout d'abord, il s'avère que pour certaines questions, notre système n'a pas réussi à trouver de document pertinent dans la collection, condition *sine qua non* pour la suite du processus. Dans d'autres cas, nous avons choisi la mauvaise cible lors de l'analyse de la question, ce qui pouvait déboucher sur un mauvais choix du type de réponse attendue. Une autre cause est la difficulté engendrée par les références temporelles. Par exemple, pour la question #22 « Combien a coûté la construction du Tunnel sous la Manche ? » nous avons répondu avec la paire ["28,4 milliards de francs", LEMONDE94-002679-19940621] supportée par la phrase "à l'origine, la construction du tunnel devait coûter 28,4 milliards de francs". Dans ce cas, notre réponse correspond à la prévision initiale mais pas au coût final. Enfin, les questions de définition étaient nettement plus complexes que les questions factuelles de par leur nature plus étendue et le fait qu'il était difficile de définir un type de réponse attendue basé sur la reconnaissance d'entités nommées Il s'agit-là des causes principales identifiées expliquant la majorité des réponses incorrectes soumises.

3.6.2.9 Conclusion

Pour la tâche monolingue, la performance obtenue par notre système (français) est de 24,5 % alors que la performance moyenne des tâches monolingues toutes langues confondues est de 23,7 %. Pour la tâche bilingue, les performances obtenues par notre système varient de 6,5 % (Bulgare-Français) à 17 % (Allemand-Français et Espagnol-Français) alors que la performance moyenne des tâches bilingues toutes langues confondues est de 14,7 %. Comme nous avons utilisé le même moteur pour les deux tâches, nous constatons comme prévu que la phase de traduction produit une perte considérable de performance (de 30,6 % à 73,5 % suivant les langues). Une amélioration importante de notre système pourrait s'articuler autour de la recherche de ressources de traduction plus performantes. Une autre optique serait de demander aux traducteurs l'analyse du texte plutôt que le texte traduit. Cela éviterait de refaire l'analyse syntaxique d'un texte qui n'est pas garanti correct. Enfin, il faudrait aussi envisager de faire une sélection intelligente des meilleures traductions disponibles [SAV 04b].

3.6.3 Campagne d'évaluation EQueR

3.6.3.1 Introduction

Notre système de question-réponse monolingue a été évalué lors de la première édition de la campagne EQueR, financée par le Ministère de la Recherche français, qui offre un cadre d'évaluation aux systèmes de question-réponse pour la langue française. EQueR a proposé deux tâches de recherche automatique de réponses, à savoir une tâche générique sur une collection hétérogène de textes et une tâche spécifique liée au domaine médical sur une collection de textes du domaine.

3.6.3.2 Collection de documents

Pour la tâche générale, six corpus de documents ont été sélectionnés. Il s'agit de *Le Monde 1992-2000*, *Le Monde Diplomatique 1992-2000, ATS 1994-1995, Rapports d'Informations (RI) du Sénat 1996-2001, Rapports Interparlementaires d'Amitiés (RA) du Sénat 1996-2001, Lois, Rapports Législatifs (L) du Sénat 1996-2001*. Le tableau 32 donne des détails sur ces collections. Un exemple de document est présenté dans l'annexe H.

Corpus	**Années**	**Taille**	**Nb documents**
Le Monde	1992 - 2000	1,19 Go	463 009
Le Monde Diplomatique	1992 - 2000	44,8 Mo	7 840
ATS français	1994 - 1995	145 Mo	85 793
Sénat (RI)	1996 - 2001	67 Mo	230
Sénat (RA)	1996 - 2001	2,96 Mo	23
Sénat (L)	1996 - 2001	68,6 Mo	417
Total		**1,51 Go**	**557 312**

Tableau 32. *Corpus de la tâche générale*

Pour la tâche médicale, un corpus de documents a été constitué à partir de pages qui se trouvaient sur la Toile en 2004. Il s'agit d'articles scientifiques et de recommandations de bonne pratique médicale sélectionnés par le CISMeF (*Catalogue et Index des Sites Médicaux Francophones*) du Centre Hospitalier Universitaire de Rouen. Le tableau 33 donne des détails sur cette collection. Un exemple de document est présenté dans l'annexe I.

Corpus	Année	Taille	Nb documents
CISMeF	jusqu'à 2004	132 Mo	5 584

Tableau 33. *Corpus de la tâche médicale*

3.6.3.3 Questions

Pour la tâche générale, un ensemble de 500 questions était proposé, 410 d'entre elles étant de type factuel, 30 de type définition, 30 de type liste et 30 de type booléen. De plus, 100 des 500 questions étaient des reformulations d'autres questions. Pour la tâche médicale, un ensemble de 200 questions était proposé, 125 d'entre elles étant de type factuel, 25 de type définition, 25 de type liste et 25 de type booléen. De plus, 50 des 200 questions étaient des reformulations d'autres questions. Le tableau 34 présente un extrait des questions de la tâche générale alors que le tableau 35 illustre un extrait des questions de la tâche médicale. Dans l'identificateur de la question situé dans la première colonne, la première lettre désigne la tâche, G pour générale et M pour médicale. Puis, le type de la question est identifié par F pour une question factuelle, D pour une question de définition, B pour une question booléenne et L pour une question liste. Enfin, la présence d'un R indique que la question est une reformulation. Par exemple, dans le tableau 34, la cinquième question (GRD401) est une reformulation de la deuxième question (GD141), toutes deux attendant comme réponse « Confédération générale du travail (CGT) ».

Identificateur	Question
GF17	Combien y a-t-il d'habitants en Lettonie ?
GD141	Que signifie CGT ?
GB341	Fidel Castro est-il déjà allé au Vietnam ?
GL382	Quels sont les trois critères qui caractérisent le miel ?
GRD401	À quoi correspond le sigle CGT ?

Tableau 34. *Quelques questions de la tâche générale*

Identificateur	Question
MF7	Où doit se dérouler une consultation diététique ?
MD16	Que veut dire "noyade sublétale" ?
MB102	Le cancer peut-il être transmissible par voie sexuelle ?
ML129	Citez 10 symptômes de l'aniridie.
MRD155	Comment l'IPS peut-il être défini ?

Tableau 35. *Quelques questions de la tâche médicale*

3.6.3.4 Réponses

Pour les deux tâches, les mêmes contraintes sur les réponses étaient imposées. Premièrement, il était possible de retourner soit le passage contenant la réponse soit la réponse précise assortie du passage qui la contient. Pour les questions de type factuel ou définition, cinq réponses ordonnées étaient attendues au maximum. Pour les questions de type booléen, une seule réponse était attendue. Enfin, pour les questions de type liste, le nombre de réponses était plafonné à 20. Dans tous les cas, la longueur maximale des passages était fixée à 250 caractères.

3.6.3.5 Mesures d'évaluation

Des juges humains ont évalué les réponses proposées sur la base de deux critères, la correction et l'exactitude de la réponse. La correction signifie que la réponse est correcte pour la question considérée. L'exactitude signifie que la réponse est exacte, c'est-à-dire ni trop longue ni trop courte. Pour chaque question, la réponse précise et le passage étaient évalués, obtenant une des cinq appréciations suivantes :

-1 pour une réponse non jugée ;

0 pour une réponse correcte ;

1 pour une réponse incorrecte ;

2 pour une réponse inexacte ;

3 pour une réponse non supportée.

Une partie des réponses n'ont pas été évaluées. Par exemple, lorsque la bonne réponse à une question était trouvée, les réponses suivantes n'étaient pas évaluées. Pour les autres appréciations, elles ont la même signification que dans l'évaluation de QA@CLEF-2004 (voir section 3.6.2.5).

Puisqu'il y avait différents types de questions, deux mesures d'évaluation différentes ont été utilisées. Pour les questions de type factuel, définition et booléen, la mesure adoptée est la moyenne des réciproques du rang de la première bonne réponse (*Mean Reciprocal Rank - MRR*), qui s'exprime comme suit :

$$\mathrm{mrr} = \frac{\sum_{i=1}^{\mathrm{nbQ}} \frac{1}{\mathrm{rc_i}}}{\mathrm{nbQ}} \tag{15}$$

où *nbQ* désigne le nombre de questions et rc_i désigne le rang de la première réponse correcte pour la question *i*. Une réponse NIL est acceptée uniquement lorsqu'elle apparaît en première position.

Pour les questions de type liste, la mesure adoptée est la moyenne de la précision moyenne non-interpolée qui, s'exprime comme suit :

$$\mathrm{MAvgP} = \frac{\sum_{i=1}^{\mathrm{nbQ}} \mathrm{AvgP_i}}{\mathrm{nbQ}} \tag{16}$$

où *nbQ* désigne le nombre de questions considérées et $AvgP_i$ désigne la précision moyenne non-interpolée pour la question *i*, qui se calcule somme suit :

$$\mathrm{AvgP_i} = \frac{\sum_{j=1}^{\mathrm{nbP_i}} \frac{j}{\mathrm{r_j}}}{\mathrm{nbP_i}} \tag{17}$$

où nbP_i désigne le nombre de réponses pertinentes pour la question *i* et r_j désigne le rang de la $j^{ème}$ réponse pertinente retrouvée pour cette question.

3.6.3.6 Résultats

Pour cette première édition, sept groupes ont participé à la tâches générale en fournissant douze séries de réponses à évaluer. Pour la tâche médicale, cinq groupes ont participé en proposant sept séries de résultats. Lors de l'évaluation, il s'est avéré que certaines questions contenaient des erreurs orthographiques. Par conséquent, le nombre de questions de la tâche générale prises en compte a été réduit à 495. Nous avons soumis deux séries de résultats par tâche.

Tâche générale

Le tableau 36 présente l'évaluation des passages de la tâche générale. Puis, le tableau 37 répartit les passages corrects par type de réponse attendue. De même, le tableau 38 présente l'évaluation des réponses courtes de la tâche générale alors que le tableau 39 répartit les réponses courtes correctes par type de réponse attendue.

	neuc04g1	**neuc04g2**
Nombre passages corrects [464] Moyenne à 44,3 %	186 (40,08 %)	183 (39,65 %)
MRR - GF, GD, GRF, GRD, GB	0,31	0,29
MRR - GF, GRF	0,30	0,28
MRR - GD, GRD	0,43	0,49
MRR - GB	0,35	0,22
MAVGP - GL	0,08	0,02

Tableau 36. *Evaluation des passages de la tâche générale*

Type de réponse	**#**	**neuc04g1**	**neuc04g2**
définition-organisation	19	9 (47,37%)	10 (52,63 %)
définition-personne	14	10 (71,43%)	8 (57,14 %)
factuelle-lieu	65	30 (46,15 %)	26 (40,00 %)
factuelle-manière	26	6 (23,08 %)	5 (19,23 %)
factuelle-mesure	77	26 (33,77 %)	27 (35,06 %)
factuelle-organisation	28	11 (39,29 %)	13 (46,43 %)
factuelle-objet	67	22 (32,84 %)	20 (29,85 %)
factuelle-personne	95	38 (40,00 %)	43 (45,26 %)
factuelle-date	42	23 (54,76 %)	24 (57,14 %)
oui/non	31	11 (35,48 %)	8 (25,81 %)
Total	464	186 (40,08 %)	184 (39,65 %)

Tableau 37. *Passages corrects de la tâche générale par type de réponse*

	neuc04g1	neuc04g2
Nombre réponses correctes [464] Moyenne à 27,1 %	104 (22,41 %)	101 (21,76 %)
MRR - GF, GD, GRF, GRD, GB	0,16	0,16
MRR - GF, GRF	0,16	0,16
MRR - GD, GRD	0,13	0,17
MRR - GB	0,29	0,12
MAVGP - GL	0	0

Tableau 38. *Evaluation des réponses courtes de la tâche générale*

Type de réponse	#	neuc04g1	neuc04g2
définition-organisation	19	3 (15,79%)	5 (26,32 %)
définition-personne	14	4 (28,57%)	4 (28,57 %)
factuelle-lieu	65	15 (23,08 %)	12 (18,46 %)
factuelle-manière	26	0 (0 %)	0 (0 %)
factuelle-mesure	77	19 (24,68 %)	20 (25,97 %)
factuelle-organisation	28	9 (32,14 %)	9 (32,14 %)
factuelle-objet	67	7 (10,45 %)	8 (11,94 %)
factuelle-personne	95	26 (27,37 %)	27 (28,42 %)
factuelle-date	42	12 (28,57 %)	12 (28,57 %)
oui/non	31	9 (29,03 %)	4 (12,90 %)
Total	464	104 (22,41 %)	101 (21,76 %)

Tableau 39. *Réponses courtes correctes de la tâche générale par type de réponse*

Tâche médicale

Le tableau 40 présente l'évaluation des passages de la tâche médicale alors que le tableau 41 présente l'évaluation des réponses courtes de cette même tâche.

	neuc04m1	neuc04m2
Nombre passages corrects [175]	26 (14,86 %)	27 (15,43 %)
MRR - MF, MD, MRF, MRD, GB	0,11	0,13
MRR - MF, MRF	0,19	0,23
MRR - MD, MRD	0,05	0,02
MRR - MB	0,04	0,08
MAVGP - ML	0,02	0,02

Tableau 40. *Evaluation des passages de la tâche médicale*

	neuc04m1	neuc04m2
Nombre réponses correctes [175]	12 (6,86 %)	17 (9,71 %)
MRR - MF, MD, MRF, MRD, GB	0,05	0,09
MRR - MF, MRF	0,05	0,11
MRR - MD, MRD	0	0
MRR - MB	0,2	0,25
MAVGP - ML	0,01	0

Tableau 41. *Evaluation des réponses courtes de la tâche médicale*

3.6.3.7 Analyse des résultats

En proposant une tâche générale et une tâche médicale, EQueR souhaitait évaluer le bénéfice qu'apporterait l'exploitation de connaissances médicales, telles que les ontologies, dans la tâche spécifique. Toutefois, nous ne disposions pas de telles connaissances, de surcroît difficiles à obtenir pour le français. Par conséquent, nous nous sommes concentrés sur la tâche générale dont l'analyse des résultats a permis de formuler les considérations qui vont suivre.

Nous avons commencé par faire une analyse de la répartition des réponses de la tâche générale dans la collection. Nous avons éliminé de cette analyse les cinq questions supprimées parce qu'elles comportaient des erreurs ainsi que les cinq questions sans réponse connue dans la collection. De ce fait, il restait 490 questions à considérer. Nous avons d'une part reporté l'observation de la répartition des réponses dans la collection et d'autre part nous avons effectué un test de *Chi-carré* afin de prédire cette répartition. Les résultats obtenus tendent à démontrer que la répartition est un peu biaisée et que le nombre de documents dans une collection n'est pas un bon prédicteur pour sa contribution au nombre de réponses. Le tableau 42 illustre ces résultats.

Corpus	Observé	Taux observé	Attendu	Taux attendu
LEMONDE	340	69,39 %	407,09	83,08 %
ATS	85	17,35 %	75,43	15,39 %
Sénat + MD	65	13,27 %	7,48	1,53 %
Total	490	100 %	490	100 %

Tableau 42. *Répartition des réponses dans la collection*

D'autre part, notre système n'était pas prévu pour générer des listes ni pour répondre à des questions booléennes. De ce fait, les résultats qui nous intéressent sont ceux obtenus pour les questions factuelles ou de définition. En observant les performances par type de réponse attendue, nous pouvons déduire que notre système produit de bonnes réponses lorsqu'il s'agit de retrouver des définitions portant sur des personnes ou des organisations ainsi que des réponses factuelles portant sur des personnes, des dates et des lieux. En revanche, il s'avère moins efficace pour des questions de manière, de mesure, d'organisation et d'objet, ce qui s'explique par le fait que les entités nommées n'ont pas été prévues pour ces catégories.

Pour les questions de manière (p.ex. « Comment est mort Claude François ? »), cela s'explique par le fait que la réponse est souvent composée d'un groupe adverbial (p.ex. « En prenant un bain »). Or notre système se concentre sur l'extraction de groupes nominaux. Pour les questions portant sur des objets, leur domaine était trop vague si bien que nous avons parfois manqué la sélection de la cible ou du type de réponse attendue. Si l'on considère des causes génériques pouvant s'appliquer à tous les types de questions, il y a tout d'abord le fait que notre système n'a pas toujours réussi à trouver de document pertinent dans la collection, condition *sine qua non* pour la suite du processus. Puis, la décomposition des paragraphes en phrases a pu produire deux effets. Premièrement, la découpe pouvait échouer, c'est-à-dire couper une seule phrase en plusieurs parties. Deuxièmement, lorsqu'une réponse est à cheval sur deux ou plusieurs phrases, la prise en compte des phrases individuelles ne permet pas de rendre la réponse attendue. Une autre cause enfin concerne les références temporelles, c'est-à-dire le choix d'une réponse selon des contraintes temporelles, ce que notre système ne savait pas faire. Il s'agit-là des causes principales identifiées expliquant la majorité des réponses incorrectes soumises.

Nous nous sommes également intéressés aux réponses jugées inexactes et avons souhaité découvrir les raisons de l'inexactitude. Pour ce faire, nous avons considéré les 51 questions jugées inexactes lors de l'évaluation. En fait, nous avons découvert deux questions faisant l'objet d'erreurs d'évaluation, nos réponses faisant partie des réponses correctes acceptées. Pour les 49 questions restantes, nous avons distingué trois cas de figure :

- notre réponse est inclue dans une bonne réponse mais est trop courte ;
- notre réponse contient une bonne réponse mais est trop longue ;
- les autres cas.

Le premier cas de figure s'est présenté neuf fois. Le problème commun réside dans le fait qu'en présence d'un fragment contenant la bonne réponse, nous avons trop coupé lors de l'extraction de la réponse. Par exemple, pour la question # GF70 « Quel est le nom du meurtrier du gitan assassiné en mars 1996 à Ingwiller ? », la réponse attendue était « Alfred Henninger » alors que nous avons retourné « Alfred ». De même, pour la question # GD332 « Qui est Alexander Popov ? », une réponse correcte était « nageur russe » alors que nous avons répondu par « Russe ».

Le deuxième cas de figure s'est produit 27 fois. En analysant les questions et réponses concernées, nous n'avons pas pu identifier une cause commune mais plutôt trois causes différentes. Premièrement, il y a des réponses dans lesquelles une partie de la question était présente et de fait redondante. Par exemple, pour la question # GF29 « Quel grade occupe Juan-Carlos Rolon dans la marine ? », la réponse attendue était « capitaines de frégate » alors que nous avons retourné « capitaines de frégate Juan Carlos ». Deuxièmement, pour des questions de type liste, nous avons retourné deux réponses en une seule. Par exemple, pour la question # GL390 « Citez les 5 Français membres de l'équipage du Galathée. », une première réponse correcte était « Philippe Ellé » et une deuxième « François Clavel ». Or notre réponse « Philippe Ellé et François Clavel » contenait deux réponses correctes. Enfin, dans la majorité des cas, nous n'avons pas identifié assez précisément la réponse en rendant une chaîne de caractères trop longue. Par exemple, pour la question # GF238 « Où est-ce que les pièces de la nouvelle monnaie argentine sont fabriquées ? », la réponse attendue était « Chili » alors que nous avons répondu par « frappées au Chili ». Ou encore, pour la question # GF272 « Quel a été le montant du prêt accordé au Mexique par la Banque mondiale ? », une réponse correcte était « 1,5 milliard de dollars » alors que nous avons retourné « total de 1,5 milliard de dollars ».

Le troisième cas de figure s'est manifesté treize fois. Nous avons déterminé deux causes majeures ayant conduit à la production de réponses inexactes. Premièrement, nous avons parfois rendu une réponse trop longue comme pour la question # GRF490 « Lors de quel meeting Frankie Fredericks a-t-il battu Michael Johnson en 1996 ? ». Une réponse correcte était « à Oslo » alors que notre réponse était « OSLO VENDREDI 5 juillet ». Deuxièmement, nos réponses étaient parfois imprécises. Par exemple, pour la question # GD340 « Qui est Jean-Marie Spaeth ? », une réponse possible était « président du conseil d'administration de la Caisse nationale de l'assurance maladie des travailleurs salariés » alors que nous avons rendu « Caisse nationale d'assurance maladie des travailleurs salariés CNAMTS ». Ou bien, pour la question # GF242 « Par quoi le Taj Mahal est-il menacé ? », la réponse attendue était « pollution » alors que nous avons retourné « fermeture de onze usines les rejets polluants ».

3.6.3.8 Comparaison avec la performance humaine

Nous avons souhaité comparer la performance de notre système avec la performance humaine. Pour cela, nous avons considéré un échantillon formé des 50 premières questions de la tâche générale duquel nous avons éliminé la question n° 44 parce qu'elle contenait une erreur. Cela nous laissait donc 49 questions pour notre expérience. Nous avons demandé à une personne de notre équipe d'effectuer deux tâches manuellement. Premièrement, à partir des dix paragraphes retournés par notre système de recherche d'information pour chaque question, retrouver le paragraphe contenant la bonne réponse. Deuxièmement, à partir des dix meilleures phrases sélectionnées par notre système pour chaque question, retrouver la phrase contenant

la réponse. Le tableau 43 présente les résultats obtenus. Dans la deuxième colonne figure le maximum théorique, c'est-à-dire le nombre de questions pour lesquelles il était possible de trouver la réponse dans les éléments considérés.

	Maximum théorique	**Réponses trouvées**	**Taux de réussite**
Tâche paragraphes	38	36	94,74 %
Tâche phrases	30	28	93,33 %

Tableau 43. *Exécution manuelle sur un échantillon de 49 questions*

Sur notre échantillon, la performance humaine se situe vers 94,7 % alors que la meilleure performance obtenue lors de cette évaluation par un système automatique se situe à 81,5 %. Cela nous conduit à considérer dans ce contexte précis qu'un traitement automatique peut atteindre approximativement 86,1 % de la performance humaine.

3.6.3.9 Estimation des pertes à chaque étape

Comme le processus de traitement se décompose en plusieurs phases, nous souhaitions pouvoir mesurer la perte induite par chacune d'entre elles. Nous avons alors identifié quatre étapes principales :

- le dépistage des meilleurs paragraphes répondant à une question (10 paragraphes) ;
- la sélection des meilleures phrases répondant à une question (10 phrases) ;
- la sélection des meilleurs fragments répondant à une question (*n* fragments) ;
- le choix du fragment à retourner comme réponse.

Pour faire cette analyse, nous avons soumis les 464 questions de type factuel, définition et booléen de l'évaluation EQueR à des être humains en leur demandant d'effectuer les tâches suivantes :

1. Pour chaque question, chercher la réponse dans les dix paragraphes retournés par SMART ; Si une réponse figure dans un paragraphe, indiquer l'identificateur du paragraphe ; sinon, inscrire NIL ;

2. Pour chaque question, chercher la réponse dans les dix phrases sélectionnées selon l'*idf* des termes ; Si une phrase contient la réponse, noter la phrase et indiquer l'identificateur du paragraphe auquel elle appartient ; sinon, inscrire NIL ;

3. Pour chaque question, chercher la réponse dans les fragments sélectionnés après l'identification des syntagmes nominaux et l'appariement ; si un fragment représente la réponse, noter la réponse et indiquer l'identificateur du paragraphe auquel elle appartient ; sinon, inscrire NIL ;

4. Pour chaque question, vérifier que le fragment choisi est une bonne réponse ; Si c'est le cas, noter la réponse et indiquer l'identificateur du paragraphe auquel elle appartient ; sinon, inscrire NIL ;

Les résultats obtenus sont présentés dans le tableau 44. La première colonne indique le nombre des questions pour lesquelles aucune réponse n'était disponible à l'étape considérée. La deuxième colonne contient le complément par rapport à 464. Puis, le succès absolu indique ce qu'il était possible d'obtenir à chaque étape par rapport au maximum de 464. La colonne suivante indique la perte absolue par rapport à 464. Le succès relatif indique ce qu'il était possible d'obtenir à chaque étape par rapport à l'étape précédente. La dernière colonne indique la perte relative par rapport à l'étape précédente.

Etape	NIL [464]	Non NIL [464]	Succès absolu	Perte absolue	Succès relatif	Perte relative
1	94	370	79,74 %	20,26 %	79,74 %	20,26 %
2	183	281	60,56 %	39,44 %	75,95 %	24,05 %
3	331	133	28,66 %	71,34 %	47,33 %	52,67 %
4	360	104	22,41 %	77,59 %	78,20 %	21,80 %

Tableau 44. *Pertes induites par chaque étape*

A la lumière des résultats obtenus, nous pouvons énoncer les considérations suivantes. Premièrement, lors du dépistage des paragraphes pertinents, nous subissons une perte de l'ordre de 20 %, ce qui correspond aux prévisions et aux résultats obtenus par les autres équipes. Puis, lors de la sélection des meilleures phrases, nous perdons encore près de 25 % de la performance. Ceci tend à indiquer que la méthode de classement des phrases pourrait être améliorée. De plus, les pertes induites par cette phase varient fortement d'une équipe à l'autre. La perte la plus importante provient de l'identification des réponses précises et de leur classement puisque la perte s'élève à plus de 50 % par rapport à l'étape précédente. Il nous semble dès lors évident que cette étape est le noeud du problème. Comme elle fait intervenir des considérations linguistiques, leur exploitation mériterait d'être perfectionnée en tenant mieux compte du type de question pour identifier l'entité nommée recherchée. Nous pourrions également envisager de recourir à l'approche linguistique en amont du processus, c'est-à-dire dans la phase d'indexation pour laquelle une analyse morphologique pourrait être profitable. Enfin, la procédure de vote aboutissant à la sélection de la réponse à fournir est responsable d'une perte d'environ 20 % et mériterait elle aussi d'être révisée.

Nous avons aussi calculé la moyenne de l'inverse du rang (MRR - *Mean Reciprocal Rank*) pour la sélection des phrases candidates. Cette mesure estime la capacité du système de trouver des réponses correctes dans les premières positions. Dans ce contexte, seules les questions de type factuel pour lesquelles notre système était développé nous intéressaient, représentant 401 des 500 questions de l'évaluation. Le tableau 45 présente les résultats obtenus. On y lit par exemple que dans 102 cas, la bonne phrase a été trouvée en première position. Puisque nous devions rendre cinq réponses par question, nous avons donc évalué les rangs de un à cinq. La dernière ligne indique la moyenne de l'inverse du rang pour les 401 questions.

Rang	**# réponses par rang**
1	102
2	19
3	19
4	6
5	8
Total	**154**
MRR	**30,16 %**

Tableau 45. *MRR sur les phrases des questions factuelles*

La valeur globale MRR obtenue (30,16 %) est plus faible que ce que nous avions escompté en regard des autres tâches en recherche d'information telle que la recherche de page connue sur le Web où cette valeur avoisine les 80 % [SAV 03]. Mais il faut être prudent lors de telles comparaisons en raison précisément de la disparité et difficulté des tâches. Néanmoins, nous pouvons émettre l'hypothèse que cette sélection de phrases pourrait être significativement améliorée. Il ne s'agit pas seulement de viser un reclassement des phrases car celui-ci n'augmenterait pas significativement le MRR. Il s'agit surtout de faire en sorte de sélectionner des phrases correctes pour davantage de questions dans les cinq premières réponses. Cette voie reste à explorer dans une future contribution.

Cet exercice avait pour objectif d'évaluer les réponses courtes et les passages associés. En prenant un peu de distance, nous pouvons nous demander ce que l'usager souhaiterait obtenir comme réponse à une question factuelle. Par exemple, pour la question « Où se trouve la mosquée Al Aqsa ? », l'usager serait-il satisfait de la réponse « Jérusalem » ou préférerait-il une réponse soutenue par un contexte telle que « La mosquée d'Al Aqsa se trouve à Jérusalem » ? En effet, bien que la réponse courte soit exacte, aucun élément ne permet à l'usager d'en vérifier la pertinence. Il doit alors faire confiance au système qui lui a proposé cette réponse. Par contre, la deuxième réponse offre un contexte qui permet de justifier ou soutenir la réponse fournie. De ce fait, il nous semble que les réponses qui seraient appréciées par l'usager devraient mettre en évident la réponse précise dans un contexte suffisant pour en permettre la justification. Enfin, la portée du contexte

devrait probablement varier selon la nature de la question. Dans certains cas, une phrase pourrait suffire alors que dans d'autres cas, il faudrait vraisemblablement remonter au niveau du paragraphe.

3.6.3.10 Conclusion

Notre approche se décompose en deux phases, à savoir la recherche d'information classique puis l'analyse syntaxique. Pour la première phase, l'exécution manuelle nous a montré que la performance maximale qu'on pouvait atteindre en dépistant des paragraphes est d'environ 80 %. Le dépistage de phrase (unique) répondant à une question pourrait constituer une thématique de recherche intéressante pour laquelle la mise à disposition de collection-test serait envisageable. Pour la deuxième partie, nous n'avons utilisé qu'une seule source d'évidence, l'analyse syntaxique avec exploitation des groupes nominaux. Notre système pourrait être amélioré par une meilleure utilisation de l'analyse syntaxique ainsi que l'ajout de nouvelles sources d'évidence.

3.7 Conclusion et perspectives

Nous avons élaboré un système de question-réponse acceptant des questions formulées dans diverses langues européennes et effectuant la recherche dans une collection de documents en français. Lors des deux évaluations auxquelles notre système a participé, ses performances se sont révélées encourageantes. Malheureusement peu de groupes de recherche ont participé aux mêmes tâches restreignant de fait le *benchmarking*.

Dans le modèle proposé, nous avons commencé par la traduction automatique des questions en français. Pour le dépistage des documents potentiellement intéressants, nous avons appliqué un modèle probabiliste. Puis, nous avons exploité des connaissances linguistiques afin d'extraire le fragment de texte que nous allions retouner comme réponse. Naturellement, chacune de ces étapes est responsable d'une perte de qualité résultant dans une baisse de performance. Si l'on peut raisonnablement considérer que la phase de recherche d'information classique offre de bons résultats, il s'avère que la sélection de phrases puis l'extraction des fragments sont responsables d'une grande perte de performance.

Afin de réduire ce phénomène, nous pourrions commencer par rechercher de meilleures ressources de traduction automatique. Nous pourrions également effectuer une sélection des traductions ou une combinaison de plusieurs traductions [SAV 04b]. De plus, il serait intéressant de disposer des analyses générées par les traducteurs afin d'éviter la phase d'analyse et les difficultés qu'elle engendre lors du traitement de texte traduit contenant des erreurs. Cela nous permettrait peut-être d'augmenter la quantité et la qualité des documents dépistés. Concernant le dépistage de documents pertinents, nous pourrions implémenter de nouveaux modèles et comparer les performances obtenues. Il serait également intéressant de compléter les listes de mots-outils et d'expérimenter d'autres algorithmes de racinisation. Concernant l'appariement et l'extraction de la réponse, nous pourrions commencer par une meilleure sélection des phrases intéressantes accompagnée d'une exploitation plus profonde de l'analyse syntaxique et des informations linguistiques mises à notre disposition. Nous aurions avantage à y ajouter d'autres sources d'informations telles que EuroWordNet ou une base de connaissances (pays, capitales, monnaies, etc.). Il serait probablement également possible de sélectionner des ressources lexicales en fonction du domaine d'application (médical, juridique, etc.). La taxonomie des questions et des types de réponses attendues mériterait une extension afin de permettre la prise en compte de questions plus complexes telles que les définitions. Enfin, nous devrions incorporer les références temporelles qui nous aideraient à mieux cibler les informations pertinentes.

4. Conclusion

4.1 Contributions

Les contributions que nous avons apportées par cette thèse dans le cadre de l'extraction automatique d'information sont les suivantes.

La première partie de notre travail concernait la génération de résumé à partir d'articles médicaux. Nous avons implémenté diverses méthodes exploitant des informations statistiques afin de sélectionner la meilleure phrase représentant l'article. Les critères retenus sont assez différents des autres approches proposées et qui utilisent également des informations statistiques. L'application de la régression logistique a obtenu une bonne performance en arrivant au quatrième rang lors de l'évaluation de la tâche génomique de TREC-2003. De plus, la combinaison de notre méthode basée sur la régression logistique avec une méthode de classification proposée par une équipe du HUG a augmenté significativement la performance obtenue.

La deuxième partie de cette thèse traitait de la question-réponse. Nous avons développé un système de question-réponse pour le français. L'originalité de notre travail réside dans la combinaison d'un modèle classique en recherche d'information avec une approche linguistique. Le modèle probabiliste Okapi nous a permis de dépister les meilleurs paragraphes à partir de la collection. Puis, l'analyse syntaxique nous a permis de segmenter les paragraphes en phrases et de décomposer les phrases en fragments basés sur les syntagmes nominaux. Enfin, un procédé de vote nous permettait de choisir le meilleur fragment comme réponse à la question.

Nous avons ensuite étendu notre système pour prendre en compte des questions formulées dans d'autres langues que le français tout en effectuant la recherche dans des documents en français. Pour cela, nous avons fait usage de ressources de traductions automatiques disponibles librement sur la Toile pour traduire les questions en français avant d'utiliser notre système de question-réponse monolingue pour répondre à la question.

Notre système a participé à CLEF 2004, première conférence ayant abordé la problématique de la question-réponse dans d'autres langues que l'anglais. Malheureusement, la grande quantité de combinaisons de langues a eu pour conséquence de ventiler les résultats proposés par les diverses équipes et ne nous a pas permis la comparaison avec d'autres modèles. Puis, nous avons soumis notre système à EQueR, première évaluation spécialisée dans la question-réponse en français. Pour les deux évaluations, notre système a présenté des performances encourageantes laissant néanmoins des possibilités intéressantes d'amélioration.

4.2 Limites

Pour la partie traitant de la génération de résumé, une des limitations provient du fait que le résumé que nous proposons est une phrase décrivant le mieux le contenu de l'article. Notre système ne permet pas de construire un résumé combinant plusieurs parties de phrases de l'article. D'autre part, nous ne considérons qu'une source d'évidence, à savoir l'article.

Concernant la question-réponse, plusieurs limitations existent. Premièrement, la langue dans laquelle les questions et les documents sont formulés restreint les ressources linguistiques à disposition. En effet, s'il existe de nombreuses ressources pour l'anglais, ce n'est pas le cas pour le finnois, le russe ou le néerlandais. De plus, de nombreuses ambiguïtés surgissent au niveau syntaxique, sémantique et pragmatique. Par exemple, quel est le sens du terme « avocat », fruit ou métier ? Ou encore l'expression « Il a cassé sa pipe. » est-elle à prendre au sens propre ou au sens figuré en tant qu'expression idiomatique ? De plus, la robustesse des systèmes est mise à l'épreuve par les erreurs d'orthographe, de grammaire ou de conjugaison. Il ne faut pas oublier la qualité de la source d'information qui joue également un rôle non négligeable dans la pertinence des réponses fournies.

Du point de vue de l'extraction d'information, nous pouvons considérer que la recherche d'information classique offre une performance d'environ 80 % même s'il serait possible de l'augmenter de l'ordre de 5 % en tenant par exemple compte de la proximité des termes [RAS 03]. Par contre, lorsqu'il s'agit d'extraire un court fragment de texte, l'absence de ressources telles que dictionnaire ou ontologie est une sérieuse limitation. En effet, en présence d'un nom propre par exemple, il serait fortement utile de pouvoir distinguer un nom de personne, d'entreprise ou de pays.

4.3 Perspectives

Pour la génération automatique de résumé, nous pouvons envisager des perspectives d'amélioration sur les axes suivants. Premièrement, il s'agirait d'étendre le système à d'autres langues et domaines d'application. Par exemple, nous pourrions nous intéresser à la jurisprudence de la Communauté européenne. D'autre part, le fait d'intégrer des considérations linguistiques permettrait une meilleure analyse du texte, une segmentation en syntagmes nominaux et une identification des parties importantes. Il faudrait également prévoir l'articulation de plusieurs parties de phrases afin de constituer un résumé global. Enfin, il faudrait incorporer des informations relatives au domaine d'application telles que le vocabulaire spécifique.

Pour la question-réponse, nous avons identifié les axes de développement suivants. Premièrement, nous suggérons d'exploiter d'avantage les informations linguistiques en termes d'analyse. En effet, nous nous sommes restreints aux syntagmes nominaux alors que les syntagmes verbaux prennent tout leur sens dans des questions de manière. Il faudrait également étendre les types de questions pris en charge, particulièrement les questions de définition ou booléennes (oui/non) qui

nécessitent un traitement plus complexe. L'ajout de nouvelles sources d'évidence comme EuroWordNet, une base de connaissances (pays, capitales, monnaies, etc.) ou des ressources liées au domaine d'application (terminologie) permettrait de prendre en compte des relations de synonymie pour l'expansion de la question et aussi de modifier la cible de la recherche selon le type de question. Enfin, l'introduction des dimensions temporelle et spatiale permettrait de réduire les documents à prendre en considération selon les contraintes induites par la question.

Si l'on s'intéresse à l'aspect multilingue, il serait naturellement intéressant de prendre en considération des requêtes dans des langues supplémentaires nécessitant la disponibilité de ressources de traduction adéquates. De même, nous pourrions utiliser des collections de documents dans d'autres langues que le français. Ceci nécessiterait de disposer d'outils linguistiques pour chacune des langues considérées. Un autre axe serait de considérer des collections multilingues, contenant des documents dans différentes langues (chaque document est exprimé dans une langue) voire des documents contenant des informations dans plusieurs langues (par exemple des éléments de jurisprudence exprimés dans plusieurs langues nationales au sein du même document). Cette problématique pourrait par exemple conduire au développement de techniques génériques applicables à plusieurs langues ou alors à l'identification de la langue avant l'application des méthodes spécifiques à la langue.

Bibliographie

[BEL 92] Belkin N. J., Croft W. B., « Information retrieval and information filtering : Two sides of the same coin ? », In *CACM*, 1992, p. 29-38.

[BEL 03] Bellot P., Crestan E., El-Bèze M., Gillard L., de Loupy C., « Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC-11 Question-Answering Track », In *Proceedings of The Eleventh Text REtrieval Conference (TREC-2002)*, NIST Special Publication # 500-251, 2003, p. 398-406.

[BER 01] Berners-Lee T., Hendler J., Lassila O., « The Semantic Web », In *Scientific American*, 284(5), 2001, p. 34-43.

[BHA 04] Bhalotia G., Nakov P.I., Schwartz A.S., Hearst M.A., « BioText Team Report for the TREC 2003 Genomics Track », In *Proceedings of the Twelfth Text REtrieval Conference (TREC-2003),* NIST Special Publication # 500-255, 2004, p. 612-621.

[BRA 95] Brandow R., Mitze K., Rau L., « Automatic condensation of electronic publishing publications by sentence selection », In *Information Processing and Management*, 31(5), 1995, p. 675-685.

[BRA 04] Braschler M., Peters C., « Cross-Language Evaluation Forum: Objectives, Results, Achievements », In *Special Issue of Information Retrieval*, 7 (1-2), 2004, p. 7-31.

[CAR 99] Card S. K., Mackinlay J. D., Shneiderman B., *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, 1999.

[CHE 04] Chen A., Gey F. C., « Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding », In *Information Retrieval*, 7(1-2), 2004, p. 149-182.

[CHO 95] Chomsky N., *The Minimalist Program*, MIT Press, Cambridge, 1995.

[CHU 00] Chuang W.T., Yang J., « Extracting Sentence Segments for Text Summarization: A Machine Learning Approach », In *Proceedings of the ACM-SIGIR'2000*, 2000, p. 152-159.

[CLE 67] Cleverdon C. W., « The Cranfield tests on index language devices », In *Aslib Proceedings*, 19(6), 1967, p. 173-193.

[CRO 02] Cronen-Townsend S., Zhou Y., Croft W.B., « Predicting Query Performance », In *Proceedings of the ACM-SIGIR'2002*, 2002, p. 299-306.

[COW 96] Cowie J., Lehnert W., « Information Extraction », In *Communications of the ACM*, 39(1), 1996, p. 80-91.

[DUM 00] Dumbill E., *The Semantic Web: A Primer*, O'Reilly, 2000.

[FEL 98] Fellbaum C., *WordNet : An Electronic Lexical Database*, MIT Press, Cambridge, 1998.

[GLO 01] Global Reach, *Global Internet statistics*, http://www.euromktg.com/globstats, 2001.

[GOL 99] Goldstein J., Kantrowitz M., Mittal V., Carbonell J., « Summarizing Text Documents: Sentence Selection and Evaluation Metrics », In *Proceedings of the ACM-SIGIR'99*, 1999, p. 121-128.

[GRA 04] Grau B., « Les systèmes de question-réponse », In *Méthodes avancées pour les systèmes de recherche d'informations*, Ihadjadene M. (éd.), Hermès, Paris, 2004, p. 189-218.

[GRE 00] Grefenstette G., *Cross-language Information Retrieval*, Kluwer Academic Publisher, 2000.

[HAE 94] Haegeman L., *Introduction to government and binding theory*, Blackwell, 1994.

[HAR 91] Harman D., « How effective is suffixing? », In *Journal of the American Society for Information Science*, 42(1), 1991, p. 7-15.

[HAR 04a] Harabagiu S., Moldovan D., Clark C., Bowden M., Williams J., Bensley J., « Answer Mining by Combining Extraction Techniques with Abductive Reasoning », In *Proceedings of the Twelfth Text REtrieval Conference (TREC-2003)*, NIST Special Publication # 500-255, 2004, p. 375-382.

[HAR 04b] Hartrumpf S., « Question Answering using Sentence Parsing and Semantic Network Matching », In *Working Notes for the CLEF 2004 Workshop*, Peters C. and Borri F. (Eds), Pisa (Italy), 2004, p. 385-392.

[HEL 02] Helbig H., Carsten G., « Multilayered extended semantic networks as a language for meaning representation in NLP systems », In *Computational Linguistics and Intelligent Text Processing (CICLing 2003)*, Springer, vol. 2276, 2002, p. 69-85.

[HER 04] Hersh W., Bhupatiraju R.T., « TREC Genomics Track Overview », In *Proceedings of the Twelfth Text REtrieval Conference (TREC-2003)*, NIST Special Publication # 500-255, 2004, p. 14-23.

[HOS 00] Hosmer D.W., Lemeshow S., *Applied Logistic Regression.*, John Wiley (Eds), New York, 2000.

[HOV 99] Hovy E., Lin C.-Y., « Automatic text summarization in SUMMARIST », In *Advances in AutomaticText Summarization*, Mani I. and Maybury M. (éd.), MIT Press, Cambridge, 1999, p. 81-97.

[IHA 04] Ihadjadene M., *Les systèmes de recherche d'informations*, Hermès, Paris, 2004.

[JIJ 03] Jijkoun V., Mishne G., de Rijke M., « How frogs built the Berlin Wall », In *Proceedings CLEF 2003*, LNCS, Springer, 2004.

[JIJ 04a] Jijkoun V., Mishne G., de Rijke M., Scholbach S., Ahn D., Müller K., « The University of Amsterdam at QA@CLEF-2004 », In *Working Notes for the CLEF 2004 Workshop*, Peters C. and Borri F. (éd.), Pisa (Italy), 2004, p. 321-324.

[JIJ 04b] Jijkoun V., Mishne G., Monz C., de Rijke M., Scholbach S., Tsur O., « The University of Amsterdam at the TREC 2003 Question Answering Track », In *Proceedings of the Twelfth Text REtrieval Conference (TREC-2003)*, NIST Special Publication # 500-255, 2004, p. 586-593.

[KAY 04] Kayaalp M., Aronson A.R., Humphrey S.M., Ide N.C., Tanabe L.K., Smith L.H., Demner D., Loane R.R., Mork J.G., Bodenrieder O., « Methods for accurate retrieval of MEDLINE citations in functional genomics », In *Proceedings of the Twelfth Text*

REtrieval Conference (TREC-2003), NIST Special Publication # 500-255, 2004, p. 441-450.

[LAE 91] Laenzlinger C., Wehrli E., « FIPS : Un analyseur interactif pour le français », In *TA Informations*, 32(2), 1991, p. 35-49.

[LAV 04] Lavenus K., Grivolla J., Gillard L., Bellot P., « Question-answer matching: two complementary methods », In *Actes de la 7è conférence RIAO*, 2004, p. 244-259.

[LEA 98] Leacock C., Chodorow M., *WordNet : An Electronic Lexical Database and Some of its Applications*, MIT Press, Cambridge, 1998, p. 265-283.

[LEF 00] Lefèvre P., *La recherche d'informations*, Hermès, Paris, 2000.

[LIN 99] Lin C., « Training a selection function for extraction », In *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM)*, 1999, p. 1-8.

[MAG 02a] Magnini B., Negri M., Prevete R., Tanev H., « A WordNet-Based Approach to Named Entities Recognition », In *Proceedings of SemaNet02, COLING Workshop on Building and Using Semantic Networks*, 2002, p. 38-44.

[MAG 02b] Magnini B., Negri M., Prevete R., Tanev H., « Comparing Statistical and Content-Based Techniques for Answer Validation », In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, 2002, p. 425-432.

[MAG 03a] Magnini B., Negri M., Prevete R., Tanev H., « Mining Knowledge from repeated Co-occurrences : DIOGENE at TREC-2002 », In *Proceedings of The Eleventh Text REtrieval Conference (TREC-2002)*, NIST Special Publication # 500-251, 2003, p. 349-357.

[MAG 03b] Magnini B., Romagnoli S., Vallin A., Herrera J., Penas A., Peinado V., Verdejo F., de Rijke M., « The Multiple Language Question Answering Track at CLEF 2003 », In *Working Notes for the CLEF 2003 Workshop*, Peters C. and Borri F. (éd.), Pisa (Italy), 2003, p. 299-310.

[MAG 04] Magnini B., Vallin A., Ayache C., Erbach G., Penas A., de Rijke M., Rocha P., Simov K., Sutcliffe R, « Overview of the CLEF 2004 Multilingual Question Answering Track », In *Working Notes for the CLEF 2004 Workshop*, Peters C. and Borri F. (éd.), Pisa (Italy), 2004, p. 281-294.

[MAN 99] Mani I., Maybury M.T., *Advances in Automatic Text Summarization*, MIT Press, Cambridge (MA), 1999.

[MAN 01] Mani I., *Automatic Summarization*, Benjamin J. (éd.), 2001.

[MEN 96] Menlo Park C., Fayyad U. M., *Advances in knowledge discovery and data mining*, MIT Press, Cambridge (MA), 1996.

[MIT 97] Mitchell T., *Machine Learning*, McGraw Hill, 1997.

[MIT 03a] Mitkov R., *The Oxford Handbook of Computational Linguistics*, Oxford University Press, 2003.

[MIT 03b] Mitchell J.A., Aronson A.R., Mork J.G., Folk L.C., Humphrey S.M., Ward J.M., « Gene indexing: characterization and analysis of NLM's GeneRIFs », In *Proceedings of the AMIA 2003 Annual Symposium*, 2003, p. 460-464.

[MOL 03] Moldovan D., Clark C., Harbagiu S., Maiorano S., « Cogex : A logic prover for question answering », In *Proceedings of the Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference (HLT-2003)*, 2003, p. 87-93.

[MON 02] Monz C., de Rijke M., « Tequesta: The University of Amsterdam's textual question answering system », In *Proceedings of The Tenth Text REtrieval Conference (TREC-2001)*, NIST Special Publication 500-250, 2002, p. 519-528.

[NEG 03] Negri M., Tanev H., Magnini B., « Bridging Languages for Question Answering: DIOGENE at CLEF 2003 », In *Working Notes for the CLEF 2003 Workshop*, Peters C. and Borri F. (éd.), Pisa (Italy), 2003, p. 321-329.

[NEU 03] Neumann G., Xu F., « Mining answers in German web pages », In *Proceedings of the International Conference on Web Intelligence (WI-2003)*, 2003, p. 125-131.

[PER 04a] Perret L., « Question-answering system for the French language », In *Working Notes for the CLEF 2004 Workshop*, Peters C. and Borri F. (Eds), Pisa (Italy), 2004, p. 295-303.

[PER 04b] Perret L., Berger P.-Y., « Extraction d'information à partir d'articles médicaux », In *Actes CORIA'04*, 2004, p. 197-211.

[POR 80] Porter M.F., « An algorithm for suffix stripping », In *Program*, 14, 1980, p. 130-137.

[RAS 03] Rasolofo Y., Savoy J., « Term Proximity Scoring for Keyword-based Retrieval Systems », In *Lecture Notes in Computer Science #2633*, Springer-Verlag, 2003, p. 207-218.

[RAV 02] Ravichandran D., Hovy E., « Learning surface text patterns for a Question Answering System », In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, 2002, p. 41-47.

[ROB 94] Robertson S.E., Walker S., « Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval », In *Proceedings of the 17th International Conference of the ACM-SIGIR'94*, 1994, p. 232-241.

[ROB 95] Robertson S.E., Walker S., Hancock-Beaulieu M.M., « Large test collection experiments on an operational, interactive system: Okapi at TREC », In *Information Processing & Management*, 31(3), 1995, p. 345-360.

[ROB 00] Robertson S.E., Walker S., Beaulieu M., « Experimentation as a way of life : OKAPI at TREC », In *Information Processing & Management*, 36(1), 2000, p. 95-108.

[RUC 04] Ruch P., Chichester C., Cohen G., Coray G., Ehrler F., Ghorbel H., Müller H., Pallotta V., « Report on the TREC 2003 Experiment: Genomic Track », In *Proceedings of the Twelfth Text REtrieval Conference (TREC-2003)*, NIST Special Publication # 500-255, 2004, p. 756-761.

[RUC 05] Ruch P., Perret L., Savoy J., « Features Combination for Extracting Gene Functions from MedLine », In *Proceedings of the 27th European Conference on IR Research (ECIR'05)*, to appear.

[SAL 71] Salton G., *The SMART Retrieval System : Experiments in Automatic Document Processing*, Prentice Hall, 1971.

[SAL 83] Salton G., McGill M. J., *Introduction to Modern Information Retrieval*, Mc Graw Hill, 1983.

[SAV 03] Savoy J., Rasolofo Y., « Report on the TREC-11 Experiment: Arabic, Named Page and Topic Distillation Searches. », In *Proceedings of The Eleventh Text REtrieval Conference (TREC-2002)*, NIST Special Publication # 500-251, 2003, p. 765-774.

[SAV 04a] Savoy J., Rasolofo Y., Perret L. « Report on the TREC-2003 Experiment: Genomic and Web Searches », In *Proceedings of the Twelfth Text REtrieval Conference (TREC-2003)*, NIST Special Publication # 500-255, 2004, p. 739-750.

[SAV 04b] Savoy J., Berger P.-Y., « Recherche bilingue et multilingue d'information », In *Actes CORIA'04*, 2004, p. 271-286.

[SCH 04] Scholbach S., Olsthoorn M., de Rijke M., « Type checking in open-domain question answering », In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*, 2004, p. 398-402.

[SOU 02] Soubbotin M.M., « Patterns of Potential Answer Expressions as Clues to the Right Answers », In *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)*, NIST Special Publication # 500-250, 2002, p. 293-302.

[SOU 03] Soubbotin M.M., Soubbotin S.M., « Use of Patterns for Detection of Likely Answer Strings: A Systematic Approach », In *Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)*, NIST Special Publication # 500-251, 2003, p. 325-331.

[TAN 04a] Tanev H., Negri M., Magnini B., Kouylekov M., « The DIOGENE Question Answering System at CLEF-2004 », In *Working Notes for the CLEF 2004 Workshop*, Peters C. and Borri F. (éd.), Pisa (Italy), 2004, p. 325-333.

[TAN 04b] Tanev H., Kouylekov M., Negri M., Coppola B., Magnini B., « Multilingual Pattern Libraries for Question Answering : a Case Study for Definition Questions », In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, to appear.

[TEU 99] Teufel S., Moens M., « Argumentative classification of extracted sentences as a first step toward flexible abstracting », In *Advances in AutomaticText Summarization*, Mani I. and Maybury M. (éd.), MIT Press, Cambridge, 1999, p. 155-175.

[VEN 99] Venaples W. N., *Modern Applied Statistics with S-PLUS*, Springer-Verlag, New York, 1999.

[VOO 00a] Voorhees E. M., Tice D., « The TREC-8 Question Answering Track Report », In *Proceedings of the Eight Text REtrieval Conference (TREC-8)*, NIST Special Publication # 500-246, 2000, p. 77-82.

[VOO 00b] Voorhees E. M., Tice D., « Building a question answering test collection », In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, 2000, p. 33-40.

[VOO 01a] Voorhees E. M., « Overview of the TREC-9 Question Answering Track », In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, NIST Special Publication # 500-249, 2001, p. 71-80.

[VOO 02] Voorhees E. M., « Overview of the TREC 2001 Question Answering Track », In *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)*, NIST Special Publication # 500-250, 2002, p. 42-51.

[VOO 03] Voorhees E. M., « Overview of the TREC 2002 Question Answering Track », In *Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)*, NIST Special Publication # 500-251, 2003, p. 57-68.

[VOO 04] Voorhees E. M., « Overview of the TREC 2003 Question Answering Track », In *Proceedings of the Twelfth Text REtrieval Conference (TREC-2003)*, NIST Special Publication # 500-255, 2004, p. 54-68.

[WEH 97] Wehrli E., *L'analyse syntaxique des langues naturelles : Problèmes et méthodes*, Masson, 1997.

[WEH 04] Wehrli E., « Un modèle multilingue d'analyse syntaxique », In *Structures et discours -- Mélanges offerts à Eddy Roulet*, 2004, p. 311-329.

[WIT 99a] Witten I. H., Moffat A., Bell T., *Managing Gigabytes: compressing and Indexing Documents and Images (second ed.)*, Morgan Kaufmann Publishers, New York, 1999.

[WIT 99b] Witbrock M., Mittal V., « Ultra-summarization : a statistical approach to generating highly condensed non-extractive summaries », In *Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'1999)*, 1999, p. 315-316.

[YAN 99] Yang Y., « An Evaluation of Statistical Aapproaches to Text Categorization », In *Information Retrieval*, 1(1-2), 1999, p. 69-90.

[YAN 03a] Yang H., Chua T.-S., « The Integration of Lexical Knowledge and External Resources for Question Answering », In *Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)*, NIST Special Publication # 500-251, 2003, p. 156-161.

[YAN 03b] Yang H., Chua T.-S., Wang S., Koh C., « Structured Use of External Knowledge for Event-based Open Domain Question Answering », In *Proceedings of the Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'2003)*, 2003, p. 33-40.

[YAN 04] Yang H., Cui H., Maslennikov M., Qiu L., Kan M.-Y., Chua T.-S., « QUALIFIER In TREC-12 QA Main Task », In *Proceedings of the Twelfth Text REtrieval Conference (TREC-2003)*, NIST Special Publication # 500-255, 2004, p. 480-488.

Annexes

Annexe A : Mesure de pertinence

Afin d'évaluer la pertinence des réponses et l'efficacité de la recherche d'information, nous avons considéré les mesures classiques décrites ci-après.

Soit P l'ensemble des documents pertinents pour une requête et N_P la cardinalité de cet ensemble. Soit R l'ensemble des documents retrouvés en réponse à une requête et N_R la cardinalité de cet ensemble. Soit $P \cap R$ l'ensemble des documents pertinents retrouvés pour une requête et N_{PR} la cardinalité de cet ensemble. La figure 10 illustre ces ensembles. Alors on peut définir les notions suivantes :

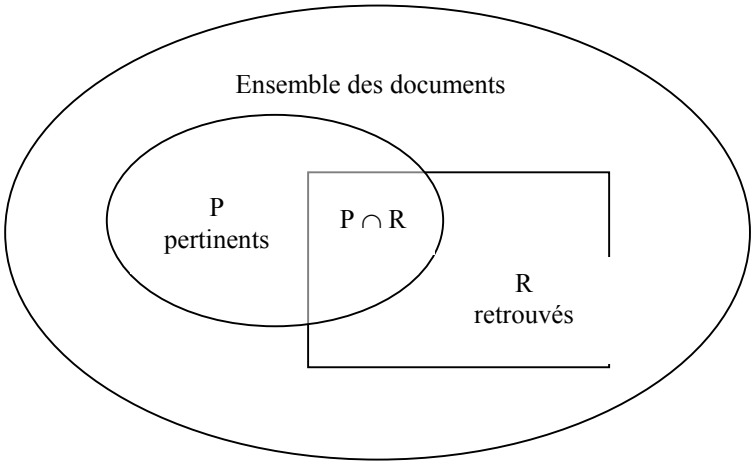


Figure 10. *Caractérisation des documents*

La précision représente le rapport entre le nombre de documents pertinents retrouvés et le nombre total de documents retrouvés. Elle s'exprime de la manière suivante :

$$\text{précision} = \frac{N_{PR}}{N_R} \tag{18}$$

Le bruit représente les documents retrouvés non pertinents. Il s'exprime de la manière suivante :

$$\text{bruit} = 1 - \text{précision} = \frac{N_R - N_{PR}}{N_R} \tag{19}$$

Le rappel représente le rapport entre le nombre de documents pertinents retrouvés et le nombre total de documents pertinents. Il s'exprime de la manière suivante :

$$\text{rappel} = \frac{N_{PR}}{N_P} \tag{20}$$

Le silence représente les documents pertinents non retrouvés. Il s'exprime de la manière suivante :

$$\text{silence} = 1 - \text{rappel} = \frac{N_P - N_{PR}}{N_P} \tag{21}$$

Il est possible d'établir des courbes de précision-rappel qui indiquent les variations conjointes de la précision et du rappel. La figure 11 illustre une courbe de précision-rappel.

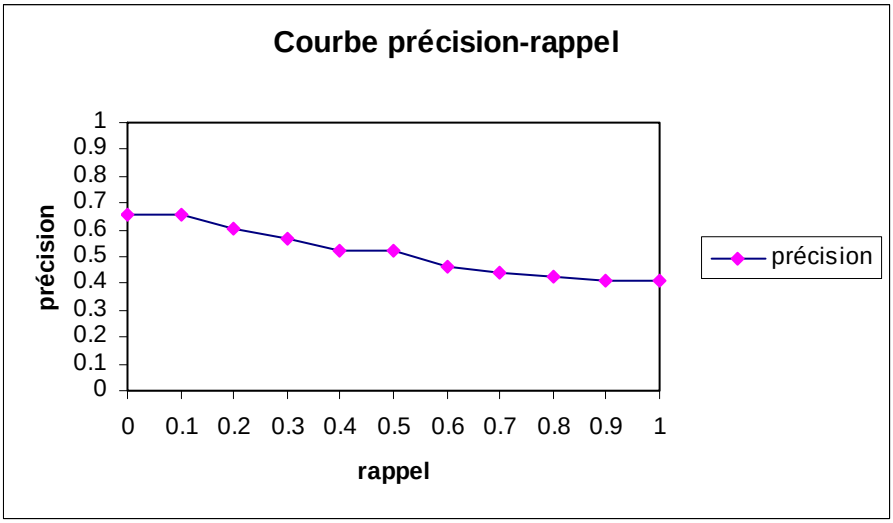


Figure 11. *Courbe de précision-rappel*

Annexe B : Exemple d'article issu de la collection médicale (TREC)


```
<!DOCTYPE art SYSTEM "keton.dtd">
<ART JID="SCI" AID="0787" VID="297" ISS="5585" DATE="08-23-2002"
PPF="1330" PPL="1333">
<FM>
<DOCHEAD>Reports</DOCHEAD>
<DOCSUBJ>BIOCHEM</DOCSUBJ>
<ATL>Structure of the Extracellular Region of HER3 Reveals an Interdomain
Tether
</ATL>
<AUG>
<AU><FNM>Hyun-Soo</FNM><SNM>Cho</SNM></AU>
<AU><FNM>Daniel J.</FNM><SNM>Leahy</SNM></AU><FNR
RID="FN150">
<AFF>Department of Biophysics and Biophysical Chemistry, Howard Hughes
Medical Institute, Johns Hopkins University School of Medicine, 725
North Wolfe Street, Baltimore, MD 21205, USA.</AFF>
</AUG>
<RE>3 June 2002</RE><ACC>18 July
2002</ACC>
<PUBFRONT>
<FPAGE>1330</FPAGE>
<LPAGE>1333</LPAGE>
<DOI>10.1126/science.1074611</DOI>
</PUBFRONT>
<FN ID="FN150">
<P>To whom correspondence should be addressed. E-mail:
<EMAIL>leahy&commat;groucho.med.jhmi.edu</EMAIL> </P></FN>
<ABS>
<P>We have determined the 2.6 angstrom crystal structure of
the entire extracellular region of human HER3 (ErbB3), a member of the
epidermal growth factor receptor (EGFR) family. The structure consists
of four domains with structural homology to domains found in the type I
insulin-like growth factor receptor. The HER3 structure reveals a
contact between domains II and IV that constrains the relative
orientations of ligand-binding domains and provides a structural basis
for understanding both multiple-affinity forms of EGFRs and
conformational changes induced in the receptor by ligand binding during
signaling. These results also suggest new therapeutic approaches to
modulating the behavior of members of the EGFR
family.</P></ABS></FM></ART>
```

Annexe C : Exemple d'article issu de la collection Le Monde (CLEF)

```
<DOC>
<DOCNO>LEMONDE94-000001-19940101</DOCNO>
<DOCID>LEMONDE94-000001-19940101</DOCID>
<ACCOUNT>318622</ACCOUNT>
<GENRE>CORRESPONDANCE</GENRE>
<DATE>19940101</DATE>
<LMDOC>JGB</LMDOC>
<FAB>12171006</FAB>
<SUBJECTS>ENFANCE,SEXUALITE,PUBLICITE,PORNOGRAPHIE</SUBJECTS
>
<NOTE>Courbevoie (Hauts-de-Seine)</NOTE>
<PUM1>QUO</PUM1>
<REFERENCE1>2-002-32</REFERENCE1>
<SEC1>IDE</SEC1>
<AUTHOR>ARACTINGI FARID</AUTHOR>
<PAGE>42</PAGE>
<TITLE>AU COURRIER DU MONDE&gt; SEXUALITE&gt; A tous les
regards</TITLE>
<TEXT>Il y a toujours eu des publications destinées à un public averti, où la liberté de
création avait le loisir de s'exprimer et de se
développer. Ce qui me semble nouveau, c'est le fait que ces
publications sont désormais exhibées à tous les regards, sans aucune
pudeur ni respect pour les sensibilités des plus jeunes. Ainsi, en
emmenant ses enfants voir un film " tous publics ", on s'expose à
l'agression des bandes-annonces racoleuses. En flânant sur les
Champs-Elysées on ne peut plus promener ses yeux sur un kiosque ou
une vitrine sans rencontrer la couverture d'une revue ou une affiche
de publicité particulièrement suggestives. En faisant ses achats dans
un grand magasin à vocation culturelle, on découvre que les bandes
dessinées pour enfants sont exposées dans le même présentoir que
celles destinées aux adultes, fussent-elles pornographiques.</TEXT>
<TEXT>
  On me répondra que cela n'aura qu'un effet limité sur la majorité
des enfants. Je n'en suis pas sûr. D'abord, parce que les effets
peuvent être souterrains et donc sournois. Et puis, l'image que nous
sommes en train de donner d'une sexualité débridée ne risque-t-elle
pas de donner comme modèle de normalité l'aventurisme sexuel pour le
plus grand nombre et la violence sexuelle pour une minorité marginale
?&lt;
</TEXT>
...
</DOC>
```

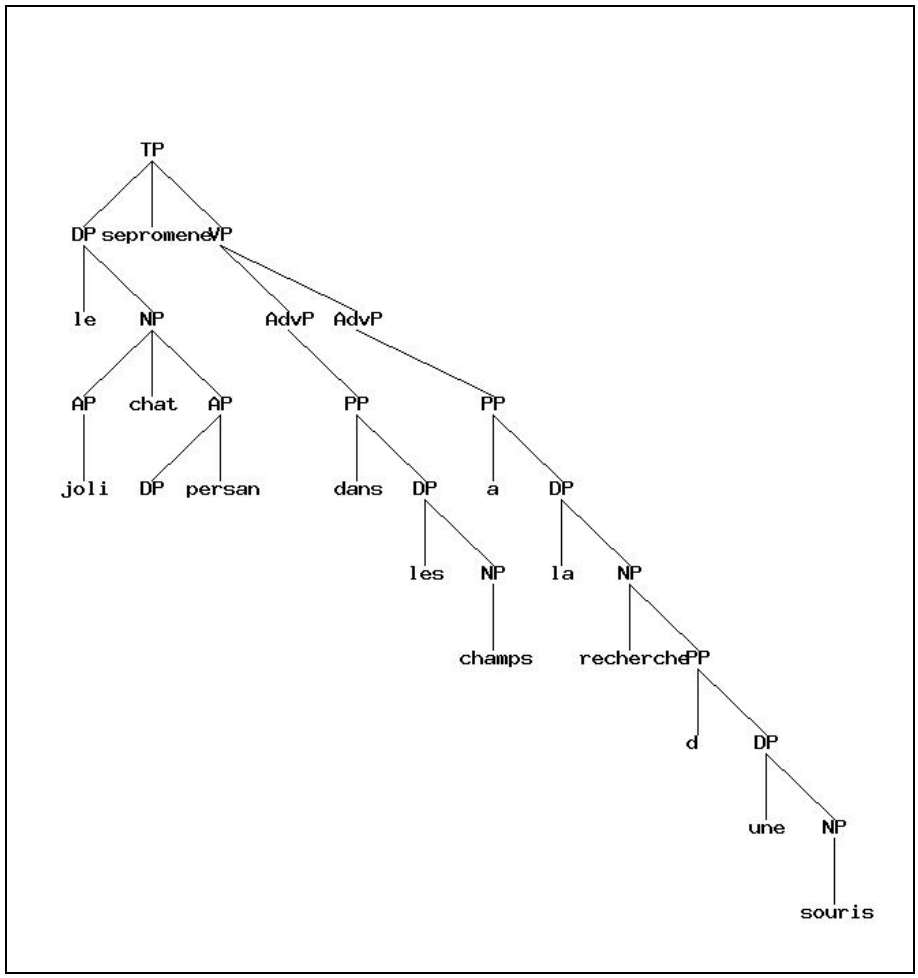
Annexe D : Exemple de dépêche issue de la collection ATS (CLEF)

```
<DOC>
<DOCID>ATS.940101.0005</DOCID>
<DOCNO>ATS.940101.0005</DOCNO>
<LC>
F
</LC>
<DT>
940101
</DT>
<PT>
202000
</PT>

...

<KW>
sarajevo combats bilan
</KW>
<TI>
Deux morts et cinq blessés samedi à Sarajevo.
</TI>
<LD>
Sarajevo, 1er jan (ats/afp) Deux personnes ont été tuées et cinq autres
blessées par des éclats d'obus samedi à Sarajevo, a indiqué un responsable de
l'hôpital de Kosevo. Un obus a atteint un appartement du centre-ville tuant
une jeune femme, a-t-on précisé de même source.
</LD>
<TX>
Un soldat de l'armée bosniaque, à majorité musulmane, a été mortellement
blessé sur l'une des lignes de front de Sarajevo. Les tirs d'artillerie et
d'armes automatiques ont été nettement moins nourris samedi que la veille,
lorsqu'ils ont fait cinq morts et une quarantaine de blessés.
</TX>
</DOC>
```

Annexe E : Algorithme de racinisation

si longueur de mot supérieure à 5
  si le mot se termine par « aux » alors remplacer « aux » par « al »
    p.ex. animaux -> animal
  sinon
    si le mot se termine par ‘s’ alors supprimer ‘s’
      p.ex. chattes -> chatte
    si le mot se termine par ‘r’ alors supprimer ‘r’
      p.ex. parler -> parle
    si le mot se termine par ‘e’ alors supprimer ‘e’
      p.ex. chatte -> chatt
    si le mot se termine par ‘é’ alors supprimer ‘é’
      p.ex. chanté -> chant
    si le mot se termine par double lettre alors supprimer la dernière lettre
      p.ex. chatt -> chat

Annexe F : Représentation arborescente de l'analyse syntaxique



Annexe G : Exemple d'article issu de la collection Le Monde 1994 (EQueR)

```
<DOC>
<DOCNO>LEMONDE94-001875-19940118</DOCNO>
<DOCID>LEMONDE94-001875-19940118</DOCID>
<ACCOUNT>320609</ACCOUNT>
<GENRE>BULLETIN</GENRE>
<DATE>19940118</DATE>
<LMDOC>DAR</LMDOC>
<DOS>GEN</DOS>
<SUBJECTS>RUSSIE,DEMISSION,MUTATION,MEMBRE DU
GOUVERNEMENT,CONSEQUENCE,POLITIQUE
ECONOMIQUE,REFORME</SUBJECTS>
<FAB>01170100</FAB>
<NUM1>940118-2-001-61</NUM1>
<NAMES>GAIDAR EGOR</NAMES>
<PUM1>QUO</PUM1>
<REFERENCE1>2-001-61</REFERENCE1>
<SEC1>ETR</SEC1>
<PAGE>83</PAGE>
<LEAD1>16 JANVIER 1994</LEAD1>
<TITLE>BULLETIN&gt; La Russie &gt; sans M. Gaïdar</TITLE>
<TEXT>EN promettant, alors qu'il prenait congé de Boris Eltsine, de " prier pour le
succès des réformes " en Russie, Bill Clinton faisait un pari
dangereux. Quelques heures plus tard, on apprenait en effet la
démission d'Egor Gaïdar, le symbole même de cette politique de
réformes.</TEXT>
<TEXT>
  Ce départ de l'ancien premier ministre peut paraître logique.
Principal perdant, avec Boris Eltsine, des élections législatives du
12 décembre, Egor Gaïdar ne fait que tirer les conséquences de
l'échec de son parti, le Choix de la Russie, qui, en dépit de moyens
matériels considérables et du soutien inconditionnel de la
télévision, s'était fait nettement distancer par Vladimir Jirinovski.
Ce scrutin avait mis en évidence l'un des handicaps majeurs d'Egor
Gaïdar et de ses amis : leur incapacité à montrer qu'ils comprennent
le sort et les préoccupations de leurs compatriotes.</TEXT>
</DOC>
```

Annexe H : Exemple d'article issu de la collection Rapports Interparlementaires d'Amitiés (RA) du Sénat (EQueR)

<DOC>
<DOCNO>Senat-RA1</DOCNO>
<P>Les Emirats Arabes Unis, un nouveau tigre au Moyen-Orient ? </P>
<P>Rapport GA 21 - Compte rendu de la visite aux Emirats Arabes Unis d'une délégation du Groupe sénatorial France-Pays du Golfe - du 9 au 16 mai 1998 - </P>
<P>Compte rendu de la visite aux Emirats Arabes Unis d'une délégation du Groupe sénatorial France-Pays du Golfe - du 9 au 16 mai 1998 - </P>
<P>COMPOSITION DE LA DÉLÉGATION </P>
<P>MM. Daniel GOULET </P>
<P>Président du Groupe sénatorial, Sénateur (RPR) de l'Orne </P>
<P>Jean BIZET </P>
<P>Vice-président délégué du Groupe sénatorial, Sénateur (RPR) de la Manche </P>
<P>François TRUCY </P>
<P>Vice-président délégué du Groupe sénatorial, Sénateur (RI) du Var </P>
<P>Mme Danielle BIDARD-REYDET </P>
<P>Sénateur (CRC) de Seine-Saint-Denis </P>
<P>La délégation était accompagnée de M. Marc LE DORH, Administrateur principal des Services du Sénat, Secrétaire exécutif du Groupe sénatorial. </P>
<P>CARTE </P>
<P>Mesdames, </P>
<P>Mesdemoiselles, </P>
<P>Messieurs, </P>
<P>En 1997, à la demande de M. Daniel Goulet, sénateur de l'Orne, le Sénat décidait la création d'un groupe sénatorial France-Pays du Golfe. Très rapidement des contacts furent pris avec les autorités des cinq pays couverts par le groupe sénatorial -Arabie saoudite, Bahreïn, Emirats Arabes Unis, Oman, Qatar- qui se traduisirent par l'établissement de relations étroites, amicales et de travail.
<P>Le groupe sénatorial décida de consacrer sa première mission à l'étranger aux Emirats Arabes Unis. Mal connus et pourtant placés à une position charnière dans la région, disposant d'une stabilité enviable, d'une économie prospère, menant une politique étrangère constructive, les Emirats entretiennent avec notre pays des relations spéciales marquées par une chaleur particulière. Ceci est apparu clairement lors de la visite effectuée sur place par le Président de la République, M. Jacques Chirac, les 15 et 16 décembre 1997. Ceci s'est confirmé lors de la mission de notre groupe sénatorial. </P>
...
</DOC>

Annexe I : Exemple d'article issu de la collection médicale (EQueR)

```
<FILE origin="EQueR.final.0705.d/www.hc-sc.gc.ca/hpfb-dgpsa/nhpd-
dpsn/index_e.html.txt">
<DOC>
<DOCNO>/home/u/thd/EQueR/process-0420.d/abcd/www.hc-sc.gc.ca/hpfb-
dgpsa/nhpd-dpsn/index_e.html.txt</DOCNO>
 <TITLE> Home</TITLE>
 Health Canada / Santé Canada - Government of Canada
 Français
 Contact Us
 Help
 Search
 Canada Site
 Natural Health Products
 Health Products and Food Branch Home
 Welcome to the Natural Health Products Directorate
 Green Bar
==================================================

 Welcome to the Natural Health Products Directorate

 <P>Our mission is to ensure that all Canadians have
 ready access to natural health products that are safe, effective, and
 of high quality, while respecting freedom of choice and philosophical
 and cultural diversity.
 </P>

 What's New:
 -----------
 <P>The Compliance Approach for Natural Health Products
 As of January 1, 2004, the Natural Health Products Regulations
( Regulations) come into force and apply to all natural health products
( NHPs). As we have done throughout the development of these
 regulations, Health Canada will continue to work with Canadians and
 other government bodies, to ensure a smooth implementation. While
 developing the compliance approach, we are aware of consumers' desire
 to have access to NHPs, and the need to ensure they are safe and
 effective. The approach has been developed with this in mind. Please
 visit this link for more information: Compliance Approach. </P>
...
</DOC>
</FILE>
```

Annexe J : Jeu de question-test construit manuellement

1	A quel âge est décédé le comédien et réalisateur napolitain Massimo Troisi ? Il avait quarante et un ans. LEMONDE94-000781-19940607
2	Qui est secrétaire général de l'OCDE ? Secrétaire général de l'OCDE […] Jean-Claude Paye LEMONDE94-000001-19941201
3	A quelle date la France renvoie-t-elle à Téhéran deux Iraniens ? 29 DECEMBRE 1993 LEMONDE94-000002-19940101
4	Où se trouve le siège de l'OCDE ? L'OCDE, dont le siège est à Paris LEMONDE94-000001-19941201
5	Qui est recherché par la Suisse pour l'assassinat de Kazem Radjavi ? deux Iraniens recherchés en Suisse LEMONDE94-000002-19940101
6	Combien de membres l'OCDE compte-t-elle ? les vingt-cinq membres de l'OCDE LEMONDE94-000001-19941201
7	Qui est secrétaire général de l'ONU ? Le secrétaire général de l'ONU, Boutros Boutros-Ghali LEMONDE94-000002-19941201
8	Quel pourcentage d'augmentation pour le chômage en novembre ? En novembre, le chômage a augmenté de 0,1 %, […] LEMONDE94-000004-19940101
9	Qui est ministre de l'intérieur ? Le ministre de l'intérieur […] Raymond Barre LEMONDE94-000023-19941201
10	Dans quel article du code civil trouve-t-on la mention " chacun a droit au respect de sa vie privée " ? article 9 du code civil LEMONDE94-000006-19940101
11	Qui est le premier ministre canadien ? Le premier ministre canadien, Jean Chrétien LEMONDE94-000034-19941201
12	Quel est l'archevêque de Lyon ? Mgr Ducourtray, archevêque de Lyon LEMONDE94-000007-19940101
13	Qui est directeur du mensuel Actuel ? Jean-François Bizot, directeur du mensuel Actuel LEMONDE94-000040-19941201
14	Quel est le nom scientifique du bouvreuil ? bouvreuil (Pyrrulha pyrrulha)

	LEMONDE94-000008-19940101
15	Qui est président de la commission des affaires étrangères de l'Assemblée nationale ? président de la commission des affaires étrangères de l'Assemblée nationale, Valéry Giscard d'Estaing LEMONDE94-000053-19941201
16	Quel est le nom de l'archevêque de Lyon ? Mgr Ducourtray, archevêque de Lyon LEMONDE94-000007-19940101
17	Quel paquebot s'est retrouvé en flamme le 30 NOVEMBRE 1994 dans l'océan Indien ? Le paquebot " Achille-Lauro " en flammes dans l'océan Indien LEMONDE94-000056-19941201
18	Combien de passagers se trouvaient à bord du paquebot « Achille-Lauro » le 30 novembre 1994 ? Les 579 passagers du paquebot italien Achille-Lauro LEMONDE94-000056-19941201
19	Qui est primat de l'Eglise catholique d'Irlande ? Le cardinal Cahal Daly, primat de l'Eglise catholique d'Irlande LEMONDE94-000057-19941201
20	Avec qui Pierre Marthelot se marrie-t-il ? avec Miette, la femme de sa vie LEMONDE94-000014-19940101
21	Combien de détenus se sont évadés de la prison de Tocuyito au Vénézuela ? l'évasion spectaculaire de 107 détenus LEMONDE94-000063-19941201
22	Qui est le président sud-africain ? Le président sud-africain, Nelson Mandela LEMONDE94-000069-19941201
23	Quel est le ministre de l'économie espagnole ? Le ministre de l'économie espagnol, Pedro Solbes LEMONDE94-000017-19940101
24	Quel nom porte le ministre de l'économie espagnole ? Le ministre de l'économie espagnol, Pedro Solbes LEMONDE94-000017-19940101
25	Quel jour la Banque d'Espagne a-t-elle destitué son conseil d'administration ? Banque d'Espagne mardi 28 décembre LEMONDE94-000017-19940101

26	Quel est le gouverneur de la Banque d'Espagne ? le gouverneur de la Banque d'Espagne, Luis Angel Rojo LEMONDE94-000017-19940101
27	Combien de morts lors de l'attentat à la grenade à Bujumbura ? cinq morts dans un attentat à la grenade à Bujumbura LEMONDE94-000070-19941201
28	A quelle date est survenue la mort de Joe French ? JEUDI 30 DECEMBRE 1993 LEMONDE94-000019-19940101
29	A quelle maladie a succombé Joe French ? Joe French, […], est mort, jeudi 30 décembre, d'un cancer LEMONDE94-000019-19940101
30	Quel âge a Laurent Garnier ? Aujourd'hui, à vingt-cinq ans, il est notre disc-jockey le plus célèbre LEMONDE94-002036-19941216
31	Sous quel surnom est connue Marie-Claude Peyronnet-Masson ? plus connue sous le nom d'" Ulla " LEMONDE94-000003-19940301
32	Qui est connu sous le pseudo de "Ulla" ? Marie-Claude Peyronnet-Masson, plus connue sous le nom d'" Ulla " lorsqu'elle défendait les droits des prostituées LEMONDE94-000003-19940301
33	Qui est le président roumain ? le président roumain; Ion Iliescu LEMONDE94-002047-19941216
34	Qui est le chef de la diplomatie israélienne ? Le chef de la diplomatie israélienne, Shimon Peres LEMONDE94-002062-19941216
35	Qui est président de l'Assemblée nationale au Burundi ? Le président de l'Assemblée nationale, Jean Minani LEMONDE94-002065-19941216
36	Quelle est l'augmentation du prix à la consommation aux Etats-Unis en une année ? Prix à la consommation : + 2,7 % en un an. LEMONDE94-002068-19941216
37	Sur quel fleuve se situe le barrage des Trois Gorges ? barrage des Trois Gorges, sur le fleuve Yangzy. LEMONDE94-002071-19941216

38	Quel est le titre du dernier roman de Françoise Cérésa ? François Cérésa dans son dernier roman, la Femme aux cheveux rouges LEMONDE94-002201-19941216
39	Combien de vétérans américains vont réitérer leur saut en parachute ? La trentaine de vétérans américains qui se préparent à réitérer, le 5 juin prochain, leur saut en parachute sur la Normandie LEMONDE94-000006-19940502
40	Qui est gouverneur de la Banque de France ? Jean-Claude Trichet, gouverneur de la Banque de France LEMONDE94-002118-19941217
41	Comment s'appel le premier ministre hongrois ? le premier ministre hongrois, Gyula Horn LEMONDE94-002125-19941217
42	En combien de morceaux est conservé le cerveau de Albert Einstein ? le cerveau d'Einstein ? Conservé en deux cents morceaux LEMONDE94-001638-19940514
43	Qui est Gyula Horn ? le premier ministre hongrois LEMONDE94-002125-19941217
44	A quelle date ont lieu les élections européennes ? élections européennes du 12 juin LEMONDE94-001661-19940514
45	Qui est le leader du Parti démocratique de la gauche en Italie ? Massimo D'Alema, le leader du Parti démocratique de la gauche LEMONDE94-002651-19941222
46	Quelle amende Monsieur Max Théret a-t-il dû payer dans l'affaire Pechiney ? 2,5 millions de francs d'amende LEMONDE94-003306-19940527
47	Combien de jeans la marque Levi's Strauss confectionne-t-elle par année ? Levi's Strauss, à lui seul, en fabrique plus de 80 millions par an LEMONDE94-000112-19940502
48	Combien d'années dure la concession du Festival de jazz de Paris ? Festival de jazz de Paris, reprend le flambeau en vertu d'une concession de quatre ans LEMONDE94-000010-19941001
49	Quel pourcentage d'audience les radios publiques ont-elles entre 1993-1994 ? Les radios publiques ne totalisent plus, en 1993-1994, que 43 % de l'audience LEMONDE94-000015-19941001

50	Quelle voiture est lancée au Salon de Paris en 1948 sur des projets d'avant-guerre ? la " Dodoche ", qui sera lancée au Salon de Paris en 1948 sur des projets d'avant-guerre, LEMONDE94-000851-19941006
51	Quel est le prix de revient du kilowattheure (kWh) ? 50 centimes ATS.940721.0062
52	Quelle est la progression des ventes de véhicules en septembre ? Les constructeurs automobiles révélaient, quant à eux, pour le mois de septembre, des ventes en progression de 7,4 % LEMONDE94-000737-19941007
53	Combien a coûté la fugue d'Otto et de César au couple Pellanda ? La fugue d'Otto et de César a coûté quelque 17 000 francs au couple Pellanda. ATS.940721.0064
54	Combien de collaborateurs emploie ABB ? ABB emploie près de 206 000 collaborateurs de par le monde ATS.941214.0105
55	Quel déficit prévoit le budget 1995 des CFF ? CFF: il a accepté le budget 1995 des CFF. Celui-ci prévoit un déficit de 312 millions de francs pour 6,7 milliards de charges. ATS.941214.0107
56	Combien de compagnies sont représentées à Orly ? à Orly, où 51 compagnies sont représentées. LEMONDE94-001961-19940517
57	Combien d'actions Viniprix a acquis Carrefour en décembre 1994 ? Du 9 au 22 décembre inclus, Carrefour a acquis sur le marché 58 actions #Euromarché et 1 775 actions Viniprix aux prix respectifs de 4 000 francs et 2 380 francs par titre. LEMONDE94-002956-19941226