QRISTAL, système de Questions-Réponses

Dominique Laurent, Patrick Séguéla

Synapse Développement
33 rue Maynard,
31000 Toulouse, France
{dlaurent, p.seguela}@synapse-fr.com

Mots-clés : Système de questions-réponses, recherche d'information, évaluation des systèmes de questions-réponses, extraction de réponse, recherche sur le Web, QRISTAL.

Keywords: Question Answering system, information retrieval, Question Answering evaluation, answer extraction, Web search strategy, QRISTAL.

Résumé

QRISTAL[1] (Questions-Réponses Intégrant un Système de Traitement Automatique des Langues) est un système de questions-réponses utilisant massivement le TAL, tant pour l'indexation des documents que pour l'extraction des réponses. Ce système s'est récemment classé premier lors de l'évaluation EQueR (Evalda, Technolangue[2]). Après une description fonctionnelle du système, ses performances sont détaillées. Ces résultats et des tests complémentaires permettent de mieux situer l'apport des différents modules de TAL. Les réactions des premiers utilisateurs incitent enfin à une réflexion sur l'ergonomie et les contraintes des systèmes de questions-réponses, face aux outils de recherche sur le Web.

Abstract

QRISTAL[1] is a question answering system which makes intensive use of natural language processing techniques, for indexing documents as well as for extracting answers. This system recently ranked first in the EQueR evaluation exercise (Evalda, Technolangue[2]). After a functional description of the system, its results in the EQueR exercise are detailed. These results and some additional tests allow to evaluate the contribution of each NLP component. The feedback of the first QRISTAL users encourage further thoughts about the ergonomics and the constraints of question answering systems, faced with the Web search engines.

---

[1] développé avec le soutien de l'ANVAR et de la Commission Européenne (TRUST, IST-1999-56416), cf. Amaral (2004), Laurent (2004).

[2] http://www.technolangue.net

1 Introduction

QRISTAL est un système de questions-réponses multilingue (français, anglais, italien, portugais, polonais) conçu pour extraire des réponses de documents placés sur un disque dur, ou pour extraire des réponses à partir du Web sur la base des pages ou passages retournés par des moteurs Web classiques (Google, MSN, AOL, etc.)

Le système reconnaît un grand nombre de formats (.html, .xml, .txt, .doc, .dbx, .hlp, .pdf, .ps, etc.), autorisant ainsi l'indexation de l'immense majorité des textes mais également des e-mails ou encore des fichiers d'aide.

Commercialisé depuis novembre 2004 pour la plate-forme Windows, ce système est destiné au grand public. Cependant, il est en cours d'intégration dans des applications professionnelles de recherche d'information. Chacun peut le tester sur le site www.qristal.fr, le corpus de test étant constitué du manuel de grammaire en ligne disponible sur http://www.synapse-fr.com/grammaire/GTM_0.htm

Notre système est fondé sur la technologie Cordial d'analyse syntaxique et sémantique du texte. Il se caractérise par une utilisation intensive des outils de TAL, entre autres l'analyse syntaxique, la désambiguïsation sémantique, la recherche des référents des anaphores, la détection des métaphores, la prise en compte des converses, le repérage des entités nommées ou encore l'analyse conceptuelle et thématique. L'utilisation professionnelle ou grand public a nécessité une optimisation constante des différents modules afin que le logiciel reste extrêmement rapide, l'utilisateur étant maintenant habitué à obtenir ce qui ressemble à des réponses dans un délai très court.

2 Architecture

L'architecture du système est modulaire. Le schéma général est décrit par la figure 1.

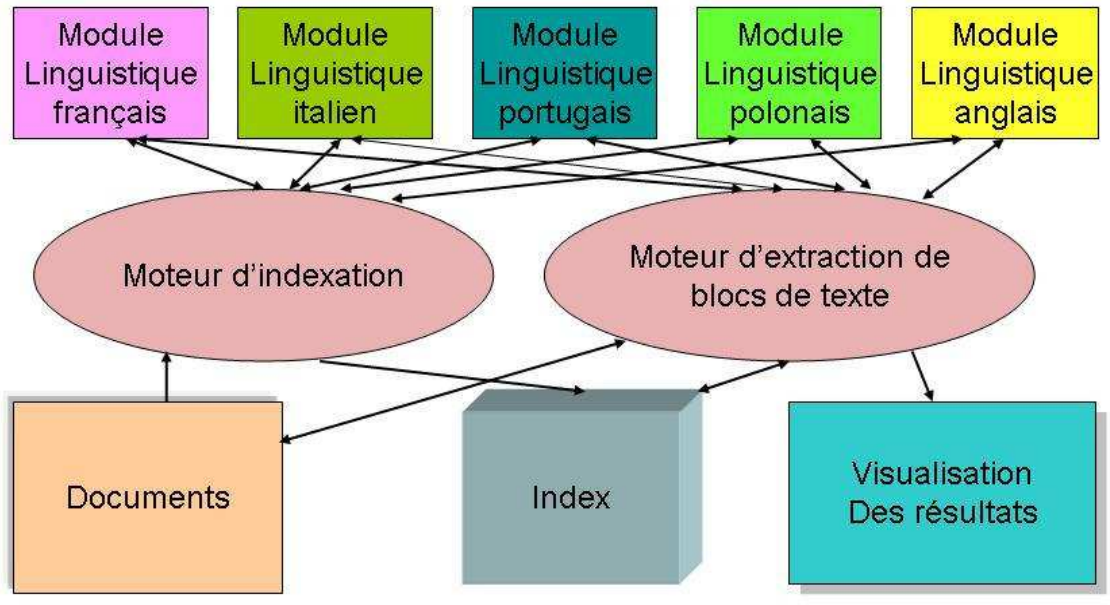

Figure 1. Architecture générale du système

Ce système est donc un moteur complet d'indexation et d'extraction de réponses. Toutefois l'indexation n'est effectuée que pour les documents "statiques", la recherche sur le Web se faisant à l'aide d'un métamoteur, par conséquent sans indexation préalable des pages.

2.1 Indexation multicritères

Au-delà du schéma général de fonctionnement du système, la figure 2 décrit le processus d'analyse linguistique effectué lors de l'indexation, lors de l'analyse de la question et lors de l'extraction de la réponse.

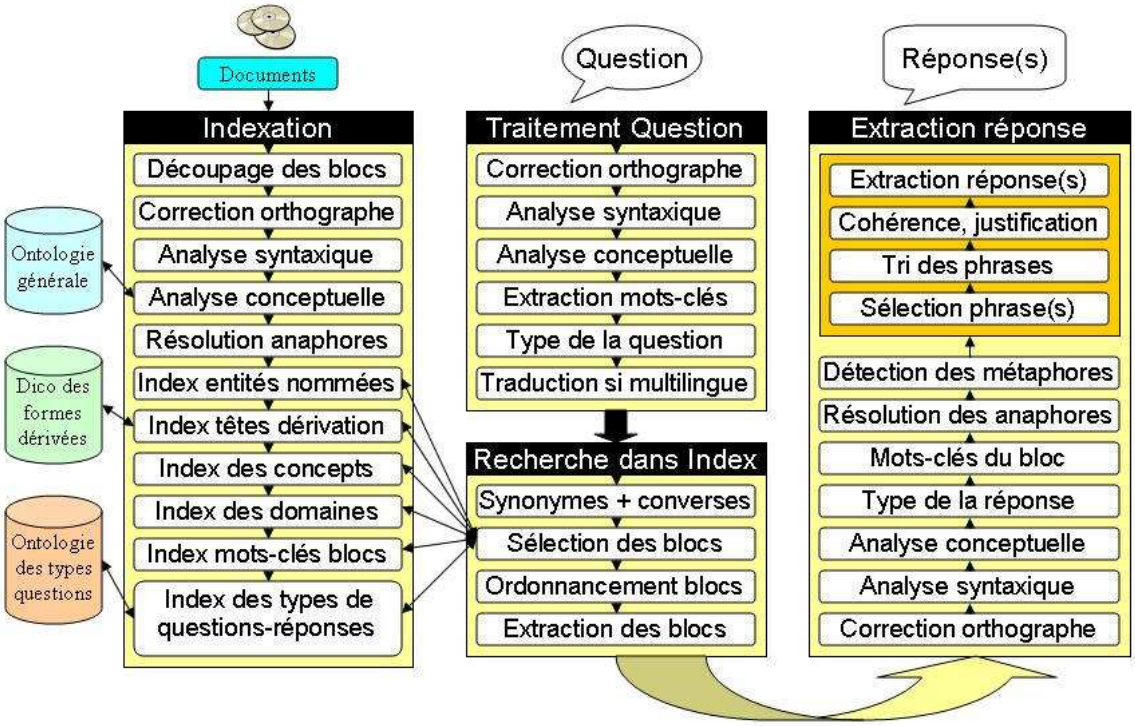

Figure 2. Processus d'analyse linguistique lors de l'indexation

Les textes sont convertis en Unicode puis découpés en blocs de longueur fixe, actuellement un kilo-octet. Cette découpe permet de réduire la taille de l'index car seul le nombre d'occurrences d'un lemme donné par bloc est sauvé dans l'index, ce nombre constituant par ailleurs un indice précieux de la pertinence de chaque bloc lors de la recherche d'un lemme donné dans l'index. En fait le mot "lemme" est ici fautif car sont indexées des têtes de dérivation. Par exemple "symétriques", "asymétrie", "dissymétriques", "symétriseraient" ou "symétrisable" seront indexés dans la même entrée "symétrie", réduisant de fait encore la taille de l'index sur disque.

Chacun des blocs de texte est analysé syntaxiquement et sémantiquement. À partir des informations issues de cette analyse, plusieurs index sont constitués :

- index des têtes de dérivation (les têtes pouvant être des sens de mots comme "symétrie"),
- index des noms propres (si ces noms propres figurent dans nos dictionnaires),

- index des expressions (connues de nos dictionnaires d'expressions d'environ 50 000 entrées, comme "frein moteur", "prendre l'air", "par inadvertance"...),
- index des entités nommées (repérées dans le texte, comme "George W. Bush" ou "Société Nationale des Chemins de fer Français"),
- index des concepts (sur deux niveaux de l'ontologie générale : 256 catégories, comme la visibilité et 3 387 sous-catégories, comme l'éclairage ou la transparence),
- index des domaines (186 domaines, comme l'aéronautique, l'agriculture, etc.),
- index des types de questions-réponses (distance, vitesse, définition, causalité...),
- index des mots-clés du texte.

Le processus d'indexation est similaire pour chacune des langues, ce qui permet de disposer de données indépendantes de la langue (numéros de concepts dans l'ontologie, numéro de domaine, type de question-réponse) qui fournissent des bases intéressantes pour retrouver dans une langue des réponses à des questions posées dans une autre langue.

En français, le taux de désambiguïsation grammaticale correcte (distinction nom-verbe-adjectif-adverbe) est supérieur à 99%, le taux de désambiguïsation sémantique est d'environ 90% pour 9 000 mots polysémiques et environ 30 000 sens pour ces mots (nombre de sens nettement inférieur à celui du Larousse par exemple, les dictionnaires d'expressions couvrant déjà un grand nombre de sens idiomatiques). La vitesse d'indexation varie entre 200 et 400 Mo par heure avec un Pentium 3 GHz, selon la taille et le nombre des fichiers à indexer.

L'indexation des types de questions est sans doute l'un des aspects les plus originaux de notre système. Lors de l'analyse des blocs à indexer, les réponses éventuelles sont repérées, par exemple un nom de fonction pour une personne ("boulanger", "ministre", "directeur de cabinet"...) une date de naissance ("né le 28 avril 1958"), un lien de causalité ("dû à l'accumulation de neige", "en raison du gel"...) ou de conséquence ("entraînant de graves perturbations", "facilitant la gestion du trafic"...), et le bloc est alors indexé comme pouvant fournir une réponse du type repéré.

La typologie comprend actuellement 86 types, dont des types factuels (dimension, surface, poids, vitesse, pourcentage, température, prix, nombre d'habitants, nom d'œuvre, etc.) mais aussi de nombreux types non factuels (forme, possession, jugement, but, causalité, opinion, comparaison, classification, etc.). Lors de l'évaluation EQueR (voir §3), 492 questions sur 500 ont été classées selon cette typologie avec seulement 6 erreurs (exemple d'erreur, sur la question 391, "Quels sont les cinq nouveaux membres non-permanents du Conseil de Sécurité de l'ONU ?", le type repéré est "noms de personnes" au lieu de "noms de pays").

L'index des mots-clés du texte est également une originalité de notre système. Il est rendu nécessaire par la découpe des textes en blocs. Du fait de cette découpe, des blocs spécifiques peuvent ne pas contenir les sujets du texte bien que les phrases de ces blocs portent sur ces sujets. L'index des mots-clés permet d'ajouter une plus-value aux textes portant a priori sur le sujet recherché, lequel sujet peut être une notion, une personne, un événement, etc.

2.2 Extraction de réponse

Lorsque l'utilisateur pose sa question, celle-ci est analysée, syntaxiquement et sémantiquement. Le type de question-réponse est déterminé. Relevons cependant ici que l'analyse

sémantique de la question est plus complète que l'analyse effectuée sur les textes car l'énoncé est généralement court. De ce fait, la désambiguïsation sémantique est plus incertaine, par manque de contexte. Si l'utilisateur a la possibilité de "forcer" tel ou tel sens de mot, cette option reste peu utilisée. Aussi calculons-nous un poids pour chaque sens possible et ce poids entre en compte dans la recherche dans l'index (exemple : sens 1 à 20%, sens 2 à 65%, sens 3 à 5% de probabilité). Ainsi les erreurs éventuelles de désambiguïsation sémantique sont tempérées par ces coefficients qui permettent de "remonter" des blocs correspondant à d'autres sens mais ayant d'autres caractéristiques recherchées, par exemple des réponses potentielles au type de la question, un nom propre identique, un même thème, etc.

Après analyse de la question, si le corpus visé est sur disque, les différents index sont consultés et les blocs les mieux placés pour ces index sont réanalysés. Sur le Web, des requêtes sont générées vers les moteurs Web classiques. Dans les pages de résultats retournées par les moteurs, les bribes ("snippets") ou les pages indiquées par les liens (pour les questions non factuelles) sont analysées.

Comme indiqué figure 2, l'analyse des blocs sélectionnés est similaire à l'analyse effectuée lors de l'indexation ou lors de l'analyse de la question avec, par exemple, une désambiguïsation sémantique des mots polysémiques. Toutefois cette analyse se double d'un calcul de poids pour chacune des phrases, le poids étant fonction du nombre de mots et entités nommées trouvés dans cette phrase, de la présence ou non du type de réponse correspondant à la question, de l'accord entre les thèmes et domaines. C'est ce poids qui permettra ensuite le classement des réponses.

Après analyse, les phrases ou paragraphes semblant répondre à la question sont triés et une analyse complémentaire extrait les entités nommées ou les groupes de mots (parfois des propositions ou des listes) qui correspondent le mieux aux réponses. Cette extraction se fait en fonction des caractéristiques des entités nommées ou sur la base de nature syntaxique des groupes ou propositions.

Pour une question sur un corpus fermé, le temps de réponse est d'environ 3 secondes avec un Pentium 3 GHz. Sur le Web, les premières réponses sont généralement fournies au bout de 2 secondes, un affinage progressif ayant lieu et pouvant durer jusqu'à une quinzaine de secondes, selon le paramétrage utilisateur (nombre de moteurs, nombre de pages analysées, etc.)

3 Évaluation EQueR

QRISTAL a été évalué dans le cadre d'EQueR, campagne d'évaluation des systèmes de Questions-Réponses du projet EVALDA (voir GRAU 2004). Le projet EVALDA et, plus généralement, les projets Technolangue, ont été initiés par les Ministères français de l'Industrie, de la Recherche et de la Culture.

La campagne EQueR a été organisée par l'ELDA (Evaluations and Language resources Distribution Agency, www.elda.org) entre janvier 2003 et décembre 2004. Très similaire dans ses principes aux campagnes TREC-QA (USA) ou NTCIR (Japon), elle a pris la forme de deux tests distincts :

- 500 questions générales, principalement factuelles, sur un corpus journalistique et administratif de 1,5 Go.
- 200 questions, souvent non factuelles, sur un corpus médical d'articles scientifiques et de pages Web d'environ 50 Mo.

Les **500** questions générales se décomposaient en :

- **407** questions factuelles simples (ex: *Comment s'appelle le fils de Juliette Binoche ?)*
- **31** questions dont la réponse est une liste (ex.: *Quels sont les trois pays qui bordent la Bosnie-Herzégovine ?)*
- **32** questions dont la réponse est une définition (ex.: *Qu'est-ce que la NSA ?*)
- **30** questions binaires, à réponse oui ou non (ex.: *La carte d'identité existe-t-elle au Royaume-Uni ?*)

La métrique utilisée pour noter les résultats était le MRR (*Mean Reciprocal Rank*, voir http://trec.nist.gov/data/qa.html), c'est-à-dire 1 pour une réponse exacte en première position, 1/2 pour une réponse exacte en seconde position, 1/3 pour une réponse exacte en troisième position, etc. Seules 5 réponses étaient prises en compte, sauf pour les questions binaires ou une seule réponse justifiée était acceptée. Pour les questions dont la réponse était une liste, la métrique utilisée était le NIAP (*Non Interpolated Average Precision,* voir MONZ, 2003).

Chaque participant pouvait fournir deux fichiers de résultats avec deux grilles d'évaluation : le mode "passages" dans lequel, sur un passage d'au maximum 200 caractères, la réponse était jugée juste si elle était contenue dans ce passage; le mode "réponses exactes" où il fallait donner la réponse exacte et uniquement celle-ci.

Notre système de Questions-Réponses évalué pour EQUER était une version bêta de QRISTAL, et Synapse Développement participait là à sa première campagne d'évaluation de moteurs de questions-réponses.

Sur la tâche générale (500 questions), les résultats des sept participants ont été les suivants :

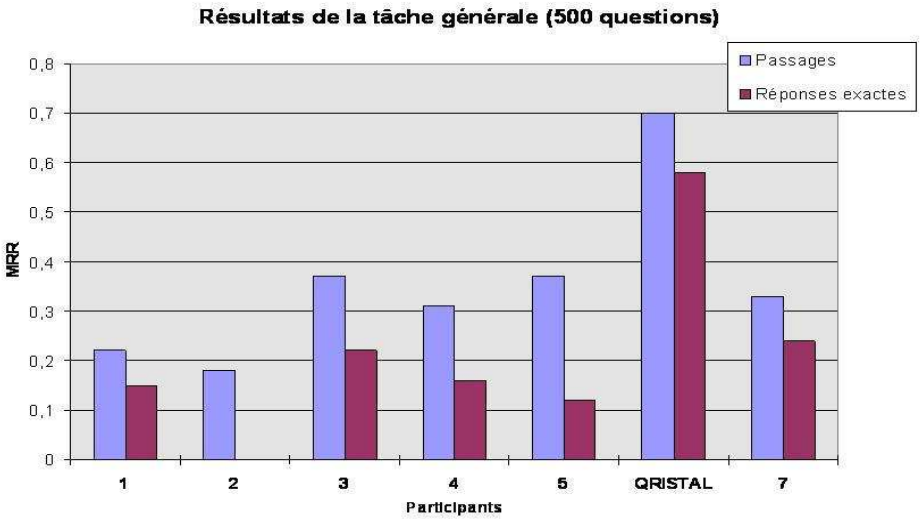

Figure 3. Résultats de la tâche générale

Sur la tâche spécialisée (200 questions sur un corpus médical), les résultats des cinq participants ont été les suivants :

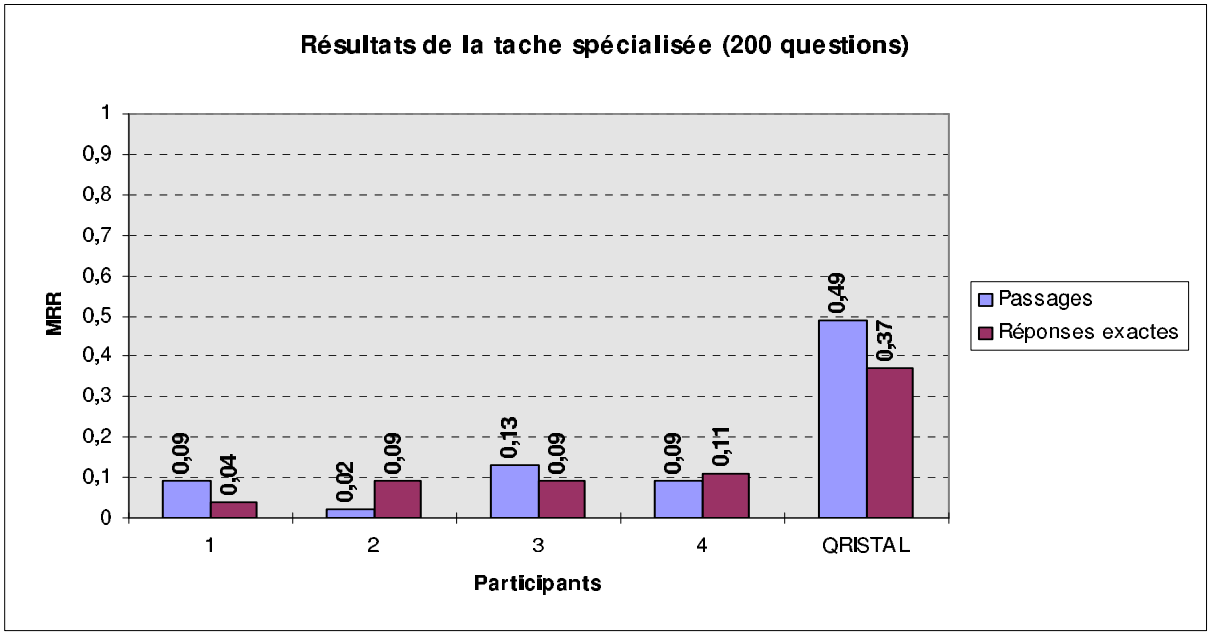

Figure 4. Résultats de la tâche spécialisée

Ces deux graphiques mettent en évidence le bon niveau de performance de QRISTAL, qui obtient les meilleurs résultats parmi les sept systèmes en compétition, sur les deux tâches. Son niveau de performance assez voisin de celui du meilleur moteur américain de TREC (MRR de 0,58 contre 0,68, voir Harabagiu, 2002 et Voorhees, 2003) ou du meilleur moteur japonais de NTCIR (MRR de 0,58 contre 0,61) sur la tâche générale.

Si l'on considère les différentes catégories de questions, QRISTAL présente des résultats homogènes selon les catégories, contrairement aux autres participants (figure 5).

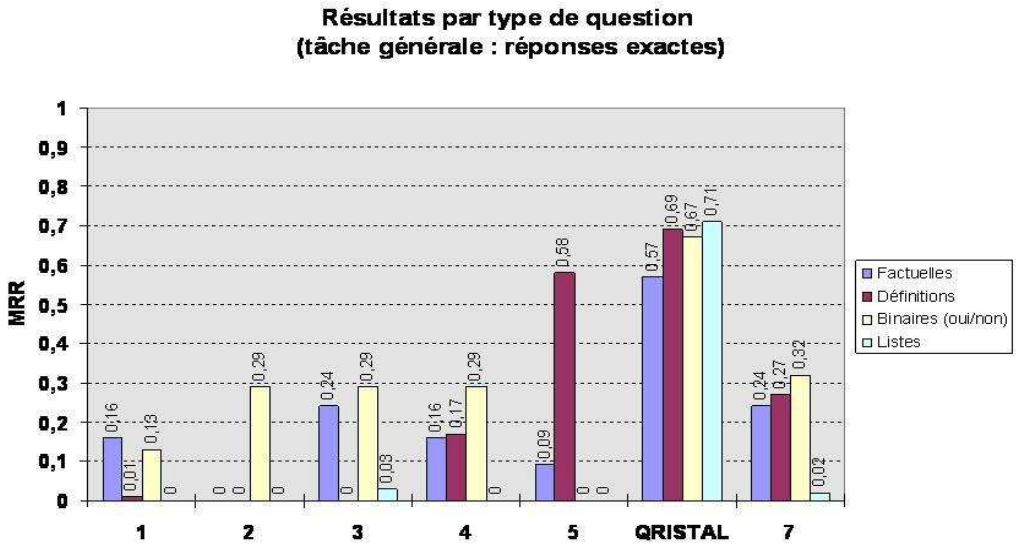

Figure 5. Résultats par type de question

Ces résultats correspondent au meilleur run fourni, celui correspondant à une passe complète d'indexation et d'extraction de réponse dans QRISTAL. Nous avions fourni un autre fichier de résultats dans lequel les textes analysés pour extraction de réponse étaient les textes retournés par un indexeur classique de recherche d'informations. Dans ce second fichier de résultats, les MRR du moteur de Synapse étaient de 0,64 (contre 0,70) pour les passages et de 0,48 (contre 0,58) pour les réponses exactes. Cette différence, significative statistiquement, indique que notre processus d'indexation multicritères est payant, car il permet de remonter de meilleurs textes à de meilleures positions qu'un moteur de recherche d'information classique.

Nous avons par ailleurs effectué un autre test afin d'évaluer l'impact de notre typologie de questions-réponses. En forçant à "inconnu" le type de toutes les questions analysées et en ne prenant pas en compte l'index des types de questions-réponses, nous avons alors obtenu un MRR de 0,46 contre 0,70 pour les passages, démontrant ainsi la pertinence de ce type de variable d'indexation et d'analyse, d'ailleurs déjà largement utilisé en questions-réponses, en particulier par les participants de TREC-QA.

4 Utilisation de QRISTAL

Les réactions des premiers utilisateurs de QRISTAL (quelques centaines après deux mois de commercialisation) permettent de tirer nombre d'enseignements sur la perception des systèmes de questions-réponses, les attentes et les illusions tant sur ces outils que sur les moteurs classiques de recherche sur le Web.

En matière de recherche sur disque dur, les réactions sont assez rares et très souvent positives. La vitesse de recherche est bonne et le fait que le moteur trouve des dérivés ou des synonymes des mots de la question améliore nettement le taux de documents retrouvés. Ainsi, sur un corpus de dépêches d'un mois de l'AFP, les moteurs classiques se révèlent incapables de retrouver le nom du président du Pérou sur l'interrogation "président Pérou" alors que QRISTAL retrouve "Alejandro Toledo" dans "le chef de l'état péruvien, Alejandro Toledo" par un double processus de synonymie et de dérivation.

Compte tenu du très grand nombre de pages indexées sur le Web, ce type d'absence de réponse ne se pose qu'avec de petits corpus, pas sur le Web, sauf pour des questions moins générales, où peu de pages contiennent la réponse. Dans ces cas-là, l'utilisateur conclura en général qu'"il n'y a pas de réponses sur le Web", ne percevant pas que ce silence résulte souvent de l'absence de traitement linguistique des moteurs classiques !

En matière de recherche sur le Web, les moteurs existants donnent le sentiment à l'utilisateur qu'ils disposent de la réponse quasi instantanément. En fait, ces moteurs fournissent des bribes de texte et des liens vers des pages, en aucun cas la réponse exacte à la question posée. Il faut au mieux lire quelques fragments, au pire ouvrir quelques pages, pour espérer obtenir une réponse. Ce processus demande toujours plusieurs secondes, d'autant que la découpe de bribes par les moteurs associe parfois des données issues de phrases différentes (ainsi une demande associant "surface" et un nom de pays donnera rarement la superficie du pays mais plutôt des tailles d'appartements situés dans ce pays). Et ceci n'est valable que pour des questions factuelles ou portant sur des personnes, les questions du type "pourquoi" ou "comment" étant habituellement hors de portée des moteurs de recherche du Web.

Utiliser les moteurs de recherche sur le Web suppose par ailleurs la maîtrise de la syntaxe de ces moteurs (rarement identique). Or cette maîtrise est peu commune. Certes, la plupart des utilisateurs de Google savent qu'il vaut mieux ne saisir que quelques mots, si possible des noms, mais peu connaissent l'usage des guillemets. Pour ces très nombreux utilisateurs, la saisie de questions en langage naturel est un atout de poids.

Reste que de nombreux traitements pouvant permettre d'améliorer la qualité des résultats finaux sont inenvisageables, tout simplement parce que l'utilisateur ne supporte pas d'attendre une réponse plus de trois à quatre secondes. Ainsi nous avions implémenté un dispositif de validation de la réponse en allant interroger à nouveau les moteurs avec les mots de la question et les mots de chacune des différentes réponses possibles (Magnini, 2002). Ce dispositif permettait d'obtenir une amélioration nette : le nombre de bonnes réponses fournies en première position passait de 47% à 58%. Mais le processus de validation demandait six à dix secondes supplémentaires et a donc dû être désactivé.

Améliorer un système de questions-réponses suppose donc une gestion extrêmement serrée du temps machine requis pour chacun des traitements. De sorte qu'il faut choisir avant tout les traitements offrant le meilleur rapport amélioration des résultats/temps machine utilisé.

5 Conclusion

QRISTAL est le premier moteur de questions-réponses commercialisé, auprès du grand public comme des professionnels. Ses résultats lors de l'évaluation EQUER montrent que l'usage intensif de technologies TAL pour l'analyse de la question et des textes indexés, ainsi que pour l'extraction de la réponse, donne de bons résultats, puisque le système se classe premier parmi les sept systèmes évalués.

Ces résultats, même s'ils sont du niveau des meilleurs prototypes internationaux, peuvent toutefois être encore considérés comme insuffisants, particulièrement lors de recherches sur le Web où l'absence d'indexation, l'utilisation d'un métamoteur (donc des résultats renvoyés par les moteurs) et des impératifs de rapidité, rendent plus incertaine l'extraction de réponses correctes dans de nombreux cas.

Même s'il paraît très vraisemblable que, dans quelques années, les moteurs booléens actuels seront remplacés par des moteurs en langage naturel, démontrer les avantages de ce type d'outil et renverser quelques illusions sur les moteurs de recherche actuels demandera du temps, ne serait-ce que parce que les prescripteurs sont souvent des experts en recherche booléenne !

Remerciements

Les auteurs remercient vivement Bruno Wieckowski et l'ensemble des ingénieurs et linguistes ayant participé au développement de QRISTAL

Références

AMARAL C., LAURENT D., MARTINS A., MENDES A., PINTO C. (2004), Design & Implementation of a Semantic Search Engine for Portuguese, *Proceedings of the Fourth Conference on Language Resources and Evaluation.*

CLARKE C. L. A., CORMACK G. V., LYNAM T. R. (2001), Exploiting Redundancy in Question Answering, *Proceedings of 24th Annual International ACM SIGIR Conference (SIGIR 2001)*, p. 358-365.

GRAU B.. (2004), L'évaluation des systèmes de question-réponse, *Évaluation des systèmes de traitement de l'information,* TSTI, p. 77-98, éd. Lavoisier.

HARABAGIU S., MOLDOVAN D., CLARK C., BOWDEN M., WILLIAMS J., BENSLEY J. (2002), Answer Mining by Combining Extraction Techniques with Abductive Reasoning, *Proceedings of The Twelfth Text Retrieval Conference (TREC 2003).*

LAURENT D., VARONE M., AMARAL C., FUGLEWICZ P. (2004), Multilingual Semantic and Cognitive Search Engine for Text Retrieval Using Semantic Technologies, *First International Workshop on Proofing Tools and Language Technologies*, Patras, Grèce.

MAGNINI B., NEGRI M., PREVETE R., TANEV H. (2002), Is It the Right Answer? Exploiting Web Redundancy for Answer Validation, *Proceedings of the 40th Annual Meeting of the ACL*, p. 425-432

MONZ C. (2003), From Document Retrieval to Question Answering, *ILLC Dissertation Series 2003-4*, ILLC, Amsterdam.

VOORHEES E. M.. (2003), Overview of the TREC 2003 Question Answering Track, NIST, 54-68 (http://trec.nist.gov/pubs/trec12/t12_proceedings.html).