SQuAr : Prototype de Moteur de Questions Réponses

Eric Blaudez (1,2), Eric Crestan (1,2) et Claude de Loupy (1,3)

(1) Sinequa Labs,
51-54, rue Ledru-Rollin,
92400 Ivry-sur-Seine, France
{*blaudez*, *loupy*, *crestan*}@sinequa.com

(2) Laboratoire Informatique d'Avignon,
B.P. 1228 Agroparc, 339 Chemin des Meinajaries,
84911 Avignon Cedex 9, France Adresse2

(3) Laboratoire MoDyCo - UMR 7114, Université de Paris 10,
Bâtiment L, 200, avenue de la République
92001 Nanterre Cedex, France

Mots-clés : Moteur de question réponse, Évaluation
Keywords: Question Answering, Evaluation

Résumé : Le système SQuAr est conçu autour de trois modules. Un module est chargé de l'analyse fine de la question. Le second retrouve les documents contenant potentiellement des réponses grâce au moteur de recherche *Intuition*. La dernière étape consiste à extraire les passages contenant une réponse correcte par un calcul de distance d'édition entre question/reformulations et les passages. De plus, les reformulations des questions en forme affirmative servent de patron d'extraction sur le corpus EQueR.

1 Introduction

Depuis le lancement de la première campagne d'évaluation des moteurs de question-réponse (*MQR*) en 1999 dans le cadre des campagnes TREC (Voorhees, Harman, 1999), l'engouement de la communauté pour cette tâche n'a cessé de croître. Au-delà du mode conventionnel de recherche documentaire, les MQR offrent un cadre complet faisant potentiellement intervenir tous les *corps de métier* du TAL. Alors que le nombre de participants aux campagnes TREC-QA est en nette régression depuis quelques années, la première campagne d'évaluation des MQR en français, EQueR (Évaluation en Questions-Réponses), a été menée dans le cadre du programme Evalda[1].

[1] Le projet EVALDA est financé par le Ministère français en charge de la Recherche, dans le cadre du programme Technolangue (http://www.technolangue.net/article20.html)

Le système SQuAr[2] est basé sur une approche en trois phases. Lors de la première phase, les questions sont analysées via des règles linguistiques pour déterminer le type de question et pour repérer les parties essentielles de celles-ci. Les règles sont formalisées sous la forme d'une cascade de 340 transducteurs pour 166 types de question identifiée. De ces types de question découlent des types de réponse attendue (généralement les entités nommées). L'analyse de la question permet également de générer la requête qui sera utilisée pour retrouver des documents susceptibles de contenir une réponse valide. De plus, pour les questions ayant eu une analyse complète, des reformulations sont générées. Le principe est de générer à partir d'une structure interrogative, des patrons d'extraction sous forme affirmative. Ces patrons ont été créés manuellement pour chaque type de question. La seconde phase permet d'extraire des documents de la base EQueR à partir des requêtes générées lors de la phase d'analyse de la question. Les requêtes comportent différents niveaux de contrainte suivant les éléments reconnus dans la question. Seul les 5 premiers documents retournés par le moteur ont été pris en compte à cause de la longueur du traitement d'extraction. La troisième et dernière phase consiste à extraire les passages répondant aux questions. Pour cela, une extraction d'entités est réalisée à l'aide d'une chaîne de transducteurs. Puis, les passages pertinents sont sélectionnés sur un critère d'une distance entre le passage et les reformulations proposées. La distance de *Levenshtein* (Wu, Manber, 1992) a été adaptée au besoin d'une recherche de réponse, en ajoutant une notion de distance entre entités. Les passages obtenant les meilleurs scores sont retournés en premier lieu. Les réponses courtes sont quant à elles extraites en sélectionnant la première entité qui correspond au type attendu. Cependant, quelques problèmes techniques n'ont pas permis d'extraire correctement ces dernières.

2 Système SQuAr

2.1 Analyse de la question

La phase la plus importante pour un MQR est l'analyse de la question. Cela est primordial dans le sens où une mauvaise analyse a de grandes chances d'engendrer des réponses incorrectes. Son rôle principal consiste à déterminer le type de la question et donc le type de la réponse attendue. Cette étape, bien qu'indispensable pour une extraction de réponse exacte, l'est moins pour une extraction par passage. Toutefois, la contrainte de la présence d'une entité candidate dans le passage est un indice fort.

Afin de déterminer le type de la question et d'identifier les éléments clés de celle-ci pour former la requête, des règles linguistiques ont été développées sous la forme d'une cascade de 340 transducteurs. 166 types de questions sont aussi distingués. Différents niveaux d'analyse résultent de cette première phase :

- Analyse complète : La question a été totalement analysée. Le type de question (donc le type de réponse attendue) a été déterminé et les éléments ont été extraits de la question pour former une requête ;
- Analyse incomplète : La question a été partiellement analysée, seul le type de la question a été déterminé avec une certaine latitude dans le type de réponse

[2] Sinequa's QUestions-AnsweRing system.

attendue. La requête sera formée directement à partir de la question après nettoyage des mots outils ;

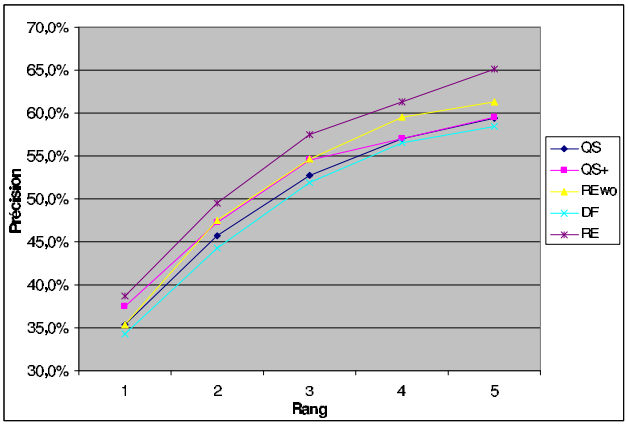
- Échec d'analyse : La question n'a pu être analysée. La requête sera là aussi formée en utilisant les termes présents dans la question.

La sortie de cette analyse se présente sous un format xml, intégrant à la fois les éléments de syntaxe et les éléments d'analyse. Celle-ci est utilisée afin de générer les requêtes pour le moteur de recherche d'un côté et les reformulations pour interroger directement la base et extraire les réponses de l'autre.

Les reformulations correspondent en fait à une transformation des questions à l'affirmatif. Cette transformation permet d'obtenir des séquences plus proches des passages répondant à la question, en terme de structure syntaxique. Par exemple, la question « *Qui est le président de la Zambie ?* » sera reformulée sous la forme des patrons suivants : « *président de la Zambie, {NPP}* » ou encore « *{NPP}, qui est le président de la Zambie* ».
L'élément entre '{…}' correspond au type de l'entité recherchée, dans le cas présent, une personne. La position de cette entité est également importante car elle définit une dépendance syntaxique de la réponse avec son contexte.

2.2 Sélection des documents candidats

Il serait inconcevable de traiter l'ensemble du corpus pour trouver la réponse à chaque question. Pour palier ce problème, nous avons pris le pari de n'analyser que les 5 premiers documents retournés par le moteur. Toutefois, les requêtes précises issus de l'analyse de la question, combinées avec le moteur de recherche sémantique de Sinequa, *Intuition*, permettent de maximiser les chances d'avoir un « *document pertinent*[3] » en tête de liste. Les requêtes générées à partir de l'analyse sont plus ou moins contraintes suivant la complétude de l'analyse. Plus l'analyse est « parfaite », plus la requête résultante sera contrainte (recherche avec présence stricte d'un mot, d'une séquence, ajout de mots supports pour la question…).



L'évaluation de la recherche de documents effectuée après la campagne EQueR montre, en moyenne, que l'utilisation des requêtes évoluées (*RE*, issues de l'analyse

[3] Nous nous référons, par *document pertinent*, aux documents contenant une réponse valide à la question considérée.

des questions) permet d'obtenir plus de document pertinent parmi les 5 premiers retournés. A 5 documents, notre approche utilisant le moteur Intuition permet de gagner près de 12% de précision en plus.

2.3 Extraction des réponses

Une fois les documents « pertinents » trouvés et les entités de chacun de ces documents détectées, la sélection des passages répondant à la question est effectuée. Un score est calculé pour chaque passage de 250 caractères de chaque document. Ils sont ensuite triés suivant ce score et les *N* meilleurs passages sont présentés comme réponses. Afin d'apparier les questions ou reformulation avec les passages, la distance de Levenshtein (Wu, Manber, 1992) a été employée. Cependant, celle-ci a été modifiée pour pourvoir calculer la distance par rapport aux mots et non au caractère comme communément utilisée. De plus, un thésaurus et un dictionnaire de synonyme permettent de prendre en compte les proximités sémantiques ou de domaine entre un mot de la question et un mot lié des passages.

Un second mode d'extraction de réponse possible a été mis au point à travers l'extraction par les reformulations. Elles sont utilisées comme des patrons d'extraction et appliqué directement sur les 100 premiers documents retournés par le moteur de recherche. Le poids de ces réponses est prépondérant sur celles extraites par le calcul de distance.

3 Évaluation

Lors de la campagne EQueR, les questions de type liste n'ont pas été traitées avec SQuAr. De plus, les efforts ont été menés seulement sur l'extraction des réponses sous forme de passages de 250 caractères, au détriment des réponses courtes. Cela s'explique par un maque de temps concédé pour ce développement.

	Définition	Factuelle	Booléenne	*Total*
Nb. Question	33	400	31	**464**
Nb. Correct	19	225	14	**237**

Ces résultats nous ont permis de terminer deuxième sur les 7 participants pour l'évaluation des passages avec 237 réponses correctes sur 464 questions, avec moyenne du rang inverse pour le système de 0,37.

4 Bibliographie

LEVENSHTEIN V. I. (1965), *Binary codes capable of correcting deletions, insertions and reversals.* Doklady Akademii Nauk SSSR 163(4) p845-848.

MANIGOT L., PELLETIER B. (1997), *Intuition, une approche mathématique et sémantique du traitement d'informations textuelles*. Actes de Fractal'1997. pp. 287-291.

VOORHEES E., HARMAN D. (1999), *Overview of the Eighth Text REtrieval Conference (TREC-8)*, National Institute of Standards and Technology, pp. 1.

WU S. AND MANBER U. (1992), *Agrep: a fast approximate pattern-matching tool.* Actes de USENIX Technical Conference, pp. 153-162.