Le LIA à EQueR

L. Gillard, P. Bellot, M. El-Bèze

Laboratoire d'Informatique d'Avignon (LIA)
339 ch. des Meinajaries, BP 1228 ; F-84911 Avignon Cedex 9 (France)
{ laurent.gillard, patrice.bellot, marc.elbeze } @ univ-avignon.fr

Résumé Cet article présente un système de Question Réponse pour le français développé dans le cadre de notre participation à la campagne d'évaluation EQueR 2004.

Abstract This paper describe a Question Answering system for French which was developed for our participation to the EQueR 2004 evaluation campaign.

1 Introduction

Un système de Question Réponse (sQR) permet, à partir d'une question exprimée en langue naturelle, d'obtenir automatiquement une réponse concise à cette question. Ainsi, la campagne EQueR est la première campagne d'évaluation sur le français de ces systèmes. Elle se propose d'établir un référentiel permettant leur émulation et leur comparaison, et rejoint en cela les objectifs des campagnes comme TREC-QA (précurseur du domaine, sur l'anglais) ou CLEF-QA (langues européennes).

Dans cet article, nous présentons le sQR que nous avons développé pour EQueR. L'architecture du système correspond à un découpage séquentiel en modules, dont les principaux concernent l'étiquetage des questions, la recherche de segments de documents et l'extraction d'une réponse avec ou sans exploitation de bases de connaissances. Chacun de ces constituants donnera lieu à une description succincte avant la présentation des résultats.

2 Architecture générale du sQR : par composants

2.1 Analyse des questions

Le sQR du LIA, comprend un composant d'étiquetage hiérarchique des questions, capable d'effectuer un appariement entre une question et un ou plusieurs types d'entités réponses attendues (*ERa*). La hiérarchie utilisée a été inspirée par celle proposée par (Sekine *et al.*, 2002), dont elle est un sous ensemble. Ce sous-ensemble a été choisi selon une observation de la fréquence des questions associées à une entrée de cette hiérarchie dans les questions françaises proposées lors des campagnes d'évaluation QR de CLEF. Concrètement, cet étiquetage se déroule après une étape d'uniformisation, à base de règles et de lexiques, permettant de réduire différentes variantes à une même écriture. Cela permet de diminuer le nombre de règles d'étiquetage, mais également d'en faciliter l'écriture (qui est manuelle).

2.2 Pré-traitements : Filtrage du corpus et Reconnaissance des Entités

Le filtrage du corpus consiste à restreindre, à partir de la question ou d'une variante, l'espace de recherche de la collection des documents à un sous ensemble de celle-ci. Pour EQueR, il était proposé comme facilité, et pour chaque question, les 100 premiers documents retournés par le moteur de recherche de la société Pertimm. Le sQR décrit ici n'utilise qu'une partie de ces listes de documents : en effet, seules les collections « Le Monde » et « Le Monde Diplomatique » ont été considérées suite à différents problèmes d'ingénierie.

Avec un objectif similaire de réduction des possibilités, la tâche d'étiquetage des entités permet dans un texte, de marquer certaines de ses parties comme des réponses candidates. Etant donné que notre sQR ne fonctionne que par extraction des entités (pas de processus de synthèse ou de raisonnement), la qualité et la granularité de cet étiquetage est crucial. Il est effectué par deux outils dont la couverture est complémentaire : le premier construit à partir de la plateforme GATE (Cunningham et al., 2002) dans le cadre d'une collaboration avec la société iSmart ; le second étant également à base de transducteurs et de lexiques. Au final, le nombre d'entités nommées ou génériques reconnues est d'environ 70, et entre en correspondance avec un sous ensemble des *ERa* issues de l'analyse des questions. Cette reconnaissance d'entités est effectuée à la fois sur les questions et sur les documents, de même qu'un étiquetage syntaxique obtenu grâce au TreeTagger (Schmid, 1994).

2.3 Recherche de passages

Afin de mieux localiser les entités candidates, notre sQR effectue une recherche de passages dans les documents préalablement filtrés et étiquetés.

Une requête est constituée d'un ensemble « d'objets » provenant de la question : les lemmes des mots, à l'exception des mots outils, les étiquettes d'entités présentes dans la question ainsi que celles des *ERa*. Ensuite, un score normalisé (F 1) est calculé à partir d'une distance moyenne μ (évaluée en nombre de mots) entre une occurrence d'un objet, et les autres objets de la requête (ou de leur plus proche occurrence en cas de présence multiple), cela dans chacun des documents issus du filtrage et pour chacun des objets.

$$score(o_i) = \frac{\log[\mu(o_i) + (\text{tailleRequête} - \text{nbObjetsTrouvés}) * \textit{pénalité}]}{\text{tailleRequête}} \qquad \text{(F 1)}$$

La pénalité est fixée empiriquement afin de plus ou moins favoriser le nombre d'objets communs entre la question et le texte par rapport à la distance moyenne qui les sépare.

Le score d'une phrase correspond au meilleur score des objets qu'elle contient. Pour chaque phrase, un « passage » est constitué à partir de cette dernière mais aussi de la phrase qui la précède et qui la suit, lorsqu'elles existent (cela afin d'essayer de compenser une éventuelle perte d'information sur les phrases courtes ou utilisant des référents trans-phrase). Le score d'un passage est le score de sa phrase centrale. Les meilleurs passages (au plus 1 000) pour l'ensemble des documents trouvés sont proposés en entrée de l'étape suivante.

2.4 Sélection de la Réponse

Enfin, la sélection d'une réponse est l'étape ultime de notre sQR. Pour cela, il dispose des passages ordonnées, suivant le score présenté en (2.3), contenant des entités (2.2), et d'entités réponses attendues (sortie de l'étiquetage 2.1).

Pour certaines questions, il dispose également de l'appui d'un module de base de connaissances (BC), sorte de couples pré-enregistrés de QR utilisés pour crédibiliser une entité candidate. En effet, les réponses contenues dans ces BC sont sous la forme d'expressions régulières et permettent l'extraction d'une réponse dans un passage déjà sélectionné comme un « support susceptible d'être correct ». Ce module avait été construit suite au constat que certaines questions assimilables à des thématiques de culture générale (telles que les capitales géographiques), et aux réponses peu variables, étaient relativement fréquentes dans les campagnes QR. Cette observation nous avait conforté dans le choix de mettre en place un tel composant lors de notre participation à TREC-11 (Bellot et al., 2002). Il a été « francisé » et sa couverture légèrement augmentée : ainsi, d'environ 20% sur le jeu des questions traduites de TREC-11, il permet d'atteindre au mieux 12% de couverture sur EQueR.

Cependant, l'essentiel de la sélection d'une réponse repose sur une autre approche : l'approximation que le candidat optimal peut être extrait à partir d'une « compacité moyenne » des mots. Par compacité moyenne, nous entendons une mesure normalisée du nombre de mots communs entre une question et un passage à l'intérieur d'une fenêtre glissante centrée sur une entité candidate d'un type compatible avec l'une des entités réponses attendues. Bien que sa formulation soit différente, cette mesure est de même nature que celle utilisée pour la sélection des passages.

3 Résultats

Le détail des métriques d'évaluation ainsi que celui des résultats pour tous les participants sont présentés dans le rapport final de l'évaluation (Ayache *et al.*, 2005). Aussi, dans cette section n'est envisagée qu'une partie des résultats obtenus par notre système.

Le tableau 1 présente le nombre de réponses correctes retournées par type (questions définitoires, *« Qu'est-ce que l'Unscom ? » ;* q. factuelles, *« Où se trouve le siège d'Adidas en France ? » ;* et q. booléennes, *« Est-ce que Bernard Dort a rencontré Bertolt Brecht ? »*) à la fois pour les évaluations « passages » et « courtes » pour les 2 soumissions (*« run »*) officielles effectuées lors de notre participation : les différences se situent au niveau du prétraitement c.à.d. dans le nombre de documents pris en compte et provenant de l'étape de filtrage du corpus, au plus 30 pour le *« run1 »* et au plus 100 pour le *« run2 »* ; ainsi que dans la finesse de l'étiquetage des entités, le premier, est basé sur une hiérarchie moins fine que le second. Cependant, ce *« run2 »* était entaché d'un problème de format qui n'a été corrigé qu'après l'échéance d'EQueR, aussi, bien que plus abouti, il a donné des résultats moins satisfaisant qu'attendu. Le *« run2 corrigé »* est également présenté, il a été évalué à l'aide de motifs construits d'après l'ensemble des réponses jugées et soumises par tous les participants.

Il est a noté que l'aspect séquentiel du sQR a également entraîné différents silences au niveau de chacun de ses composants, ce qui s'est traduit par un nombre final de questions répondues d'au mieux 407 réponses sur l'ensemble des 464 questions de l'évaluation (*cf.* avant dernière colonne du tableau 1). Nous excluons volontairement les questions « listes » dans ces décomptes, notre approche embryonnaire de celles-ci étant à améliorer.

En terme de score, pour les réponses « courtes » – par opposition à celles « passages », c.à.d. moins localisées et contenues dans un bloc de 250 caractères – le MRR (Moyenne des réciproques du rang, elle correspond à la moyenne des inverses du rang de la première bonne

réponse parmi les 5 autorisées, ou zéro en cas d'absence d'une réponse correcte) est de 0,25 sur le « *run1* » est de 0,23 sur le « *run2* » (*cf.* dernière colonne du tableau 1).

A titre de comparaison, sur les réponses « courtes », le meilleur système obtient un MRR de 0,58 et répond correctement à 312 sur 464 questions. Notre sQR se positionne en seconde position sur les réponses « courtes », mais en 5ième sur les 7 systèmes participants dans le cas des réponses « passages ».

	# Q Définitoires (33)	**# Q Factuelles (400)**	**# Q Booléennes (31)**	**# réponses CORRECTES**	**soit %**	**# réponses répondues (464)**	**MRR**
PASSAGE – run1	17	153	12	182	39,2	388	0,33
COURTE – run1	9	118	12	139	29,5	388	0,25
PASSAGE– run1	14	137	11	162	34,9	354	0,29
COURTE – run2	8	111	11	130	27,6	354	0,23
PASSAGE – run2Corr.	14	194	11	219	47,2	407	0,39
COURTE - run2Corr.	4	155	11	170	36,6	407	0,29

Tableau 1 : Nombre de réponses correctes par soumission et par type ; MRR global

4 Conclusion et perspectives

Nous avons présenté les grandes lignes d'un système de QR pour le français principalement basé sur des métriques de compacité moyennes pour la sélection des passages et des réponses ; ainsi que les résultats obtenus lors de la campagne EQueR.

En outre, le fait de disposer d'un sQR complet et d'un corpus de référence en français permettra d'autres expériences, l'évaluation précise (en cours) de chacun des composants de ce système, et ainsi de répondre à la question concernant la probable amélioration de ses performances, actuellement plus qu'acceptables si on considère le classement, largement perfectibles si l'on tient compte de la précision.

Références

AYACHE C., CHOUKRI K., GRAU B. (2005). Campagne EVALDA/EQueR Evaluation en Question-Réponse. *http://www.technolangue.net/IMG/pdf/rapport_EQueR_1.2.pdf*

BELLOT P., CRESTAN E., EL-BÈZE M., GILLARD L., DE LOUPY C. (2002). Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC-11 Question Answering Track. *in Proceedings of the 11th Text REtrieval Conference.* Gaithersburg, Maryland, USA pp. 398-406.

CUNNINGHAM H., MAYNARD D., BONTCHEVA K., TABLAN V. GATE: (2002). A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).* Philadelphia, July 2002. *http://gate.ac.uk/sale/acl02/acl-main.pdf*

ISMART, *http://ismart.fr/*.

PERTIMM, *http://www.pertimm.fr*.

SEKINE S., SUDO K., NOBATA C. (2002). Extended Named Entity Hierarchy. *In Proceedings of the LREC-2002 Conference*, pp. 1818–1824.

SCHMID H. (1994). Probablistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of the First International Conference on New Methods in Natural Language Processing (NemLap-94)*, Manchester, U.K., pp. 44-49.