FRASQUES, le système du groupe LIR, LIMSI

B. Grau (1), G. Illouz (1), L. Monceaux (2), P. Paroubek (1), O. Pons (3), I. Robba (1), A. Vilnat (1)

(1) Groupe LIR – LIMSI
BP 133, 91403 Orsay Cedex
{grau, illouz, pap, robba, vilnat}@limsi.fr
(2) LINA – Université de Nantes
2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03
Laura.Monceaux@lina.univ-nantes.fr
(3) CEDRIC-IIE – CNAM
IIE, 18 allee Jean Rostand, 91025 Evry Cedex
pons@cnam.fr

Mots-clés : système de question-réponse, évaluation
Keywords: question-answering system, evaluation

Résumé Le système FRASQUES qui a participé à l'évaluation EQueR est présenté ici en comparaison avec notre système QALC dédié à l'anglais. Ses résultats sont commentés et une évaluation des différents modules est exposée.

Abstract We present our system FRASQUES in comparison to QALC our system for English. The results that FRASQUES obtained at EqueR are presented, and an evaluation of its modules is given.

1 Présentation de FRASQUES

Comme QALC notre système pour l'anglais, (Ferret *et al.* 2002), FRASQUES s'organise en quatre principaux modules présentés dans la figure 1 : l'analyse des questions, la sélection des documents par un moteur de recherche, et le traitement des documents pour en extraire les phrases et les réponses finales. Nous allons tout d'abord présenter globalement notre système en précisant comment s'est faite l'adaptation de QALC au français, puis nous en donnerons ses résultats, et ce pour les différents modules.

L'analyse des questions est réalisée en deux étapes. L'analyseur XIP de Xerox (Aït-Mokthar *et al.* 2002) construit les segments syntaxiques et établit les relations entre eux. A partir de ces

données, des informations telles que le focus ou le type attendu de la réponse sont calculées. Ce module a été réécrit pour traiter les questions en français, mais les règles de reconnaissance ont été transposées depuis l'anglais de manière très directe.

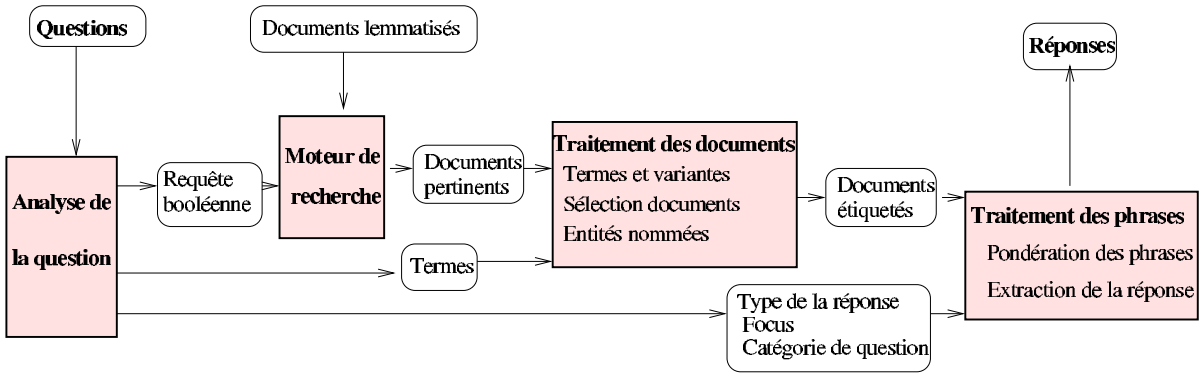

Figure 1 : Le système FRASQUES

Le moteur de recherche utilisé, Lucene[1], est un moteur booléen ; il nous a permis d'indexer le corpus, qui a été au préalable lemmatisé par le Tree Tagger[2] et l'analyseur morphologique de XIP. QALC, pour sa part, faisait appel à un moteur vectoriel utilisant le stemming. Lors de l'interrogation, Lucene reçoit un ensemble de requêtes constituées des mots non vides de la question. Les requêtes les plus larges, n'étant utilisées que si les plus précises ne retournent pas assez de documents. Les documents trouvés par Lucene, sont traités par Fastr[3] qui permet de reconnaître des variantes morphologiques, syntaxiques et sémantiques des termes simples et composés de la question et de pondérer les documents selon les termes trouvés. Les documents sont ainsi réordonnés et un sous-ensemble est extrait sur lequel on applique alors le module d'extraction des entités nommées. Cette partie de la chaîne de traitement est la même pour FRASQUES et QALC, seules diffèrent les ressources qui lui sont données.

Enfin le module d'extraction de la réponse est appliqué ; il procède différemment selon que la question attend ou non pour réponse une entité nommée. Les phrases sont pondérées en fonction du taux de présence des mots de la question dans la phrase, et de la présence ou non de l'entité nommée attendue. Dans FRASQUES, comme dans la plupart des systèmes de QR, les questions à entités nommées obtiennent de meilleurs résultats que les autres, car, de par leur nature, les entités nommées sont plus facilement repérées dans les documents. L'extraction de la réponse exacte a été réalisée différemment dans FRASQUES, puisque nous avons utilisé l'analyseur SCOL[4], afin d'appliquer des patrons d'extraction. Ces patrons, écrits sous forme de règles, s'appuient sur un étiquetage morpho-syntaxique des phrases spécialisant les caractéristiques de la question qui sont présentes dans la phrase sous leur forme initiale ou sous forme de synonyme.

---

[1] http://jakarta.apache.org/lucene/docs/index.html

[2] http://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html

[3] http://www.limsi.fr/Individu/jacquemi/FASTR/

[4] http://www.sfs.nphil.uni-tuebingen.de/Staff-Old/abney

La campagne EQueR proposait des questions booléennes ainsi que des questions dont la réponse devait être une liste de réponses. La stratégie mise en œuvre pour répondre à ces questions était très simple : si la phrase réponse comportait tous les noms de la question et le verbe principal, la réponse était positive. En ce qui concerne les listes, le choix de la réponse finale consistait à reclasser les réponses extraites, en favorisant le plus grand ensemble de réponses dans un même document et en fixant un nombre de réponses proposées en fonction du nombre de réponses demandées dans la question.

Nous allons maintenant donner les résultats de FRASQUES et les évaluations de chacun des modules réalisés. Pour cela, nous avons recensé pour chaque question les différentes réponses données par les participants et l'organisateur, et nous avons testé la présence de ces réponses dans les documents et passages retournés par notre système[5].

2 Sélection des documents

En ce qui concerne la sélection des documents, le moteur de recherche ne retourne de documents contenant la réponse que pour 73 à 76% des questions pour les deux tests soumis. Cela s'explique par plusieurs facteurs : imprécision de la sélection des mots-clefs des questions, qui sont retenus uniquement en fonction de leur étiquette morpho-syntaxique ; erreurs de lemmatisation problèmes de référence... Ces difficultés étaient déjà présentes dans QALC. La différence entre les deux tests tient au fait que pour le deuxième, le nombre de documents retournés était limité à 200.

La sélection par Fastr de 50 documents en fonction des reformulations des multi-termes (et de synonymes mono-termes pour le test 2) trouvés n'entraîne pas de perte de bons documents. Le second test ayant retenu les synonymes mono-termes, on pourrait s'attendre à ce que son rappel soit meilleur, mais il n'en est rien. Ce phénomène peut s'expliquer par le fort degré de ressemblance des questions avec les phrases réponses, et par le bruit introduit par la recherche de « mauvais » synonymes.

	Réponses longues	MRR	MRR2	Réponses courtes	MRR	MRR2	Phrases Rang 1-5	MRR
Test 1	210 (42%)	0,37	0,38	131 (26%)	0,22	0,22	253 (60%)	0,48
Test 2	187 (37%)	0,32	0,33	118 (24%)	00,2	0,2		
Trec9	393 (56%)		0,407					
Trec11				139 Rang 1 (28%)				

Tableau 1 : Résultats officiels de FRASQUES à EqueR et comparaison avec QALC à TREC

[5] Le nombre des questions est ici ramené à 450 : les questions booléennes et les questions dont la réponse est une liste n'ayant pas été prises en compte.

3 Pondération des phrases et extraction de la réponse

Dans le tableau 1, qui donne les résultats officiels, nous avons comparé nos résultats à ceux de QALC. Il est peu surprenant de voir que les systèmes obtiennent des performances similaires. En effet, les résultats officiels à Trec11 (165 bonnes réponses sur 500 questions) provenaient de la fusion de 2 sources de réponses et augmentaient le nombre de réponses correctes (Berthelin *et al.* 2003). En outre, nous avons appliqué les patrons de réponses sur les 5 premières phrases du test 1, et obtenons alors un résultat de 60% de bonnes réponses contre 42% de réponses longues dues à des phrases tronquées par erreur.

Des résultats issus d'EqueR, nous avons mené deux études. La première mesure le taux de présence des termes à l'identique et celui des synonymes, dans nos phrases réponses et dans celles des participants (Grau et al . 2005). On voit que très peu de synonymes sont présents dans les phrases retenues et cela amène à poser la question du type de connaissance à utiliser. La seconde étude (réalisée par A.-L. Ligozat, groupe LIR) vise à déterminer les phénomènes responsables de l'extraction d'une réponse erronée quand on dispose d'une phrase correcte. Aussi, nous avons retenu notre ensemble de réponses longues correctes, pour lesquelles nous avons recherché la phrase d'origine. Il en ressort que parmi 74 questions, 25 réponses incorrectes sont dues à l'application d'un mauvais patron, 33 à un mauvais étiquetage de la réponse attendue : 15 ont une mauvaise étiquette, 11 sont dues à une absence d'étiquette et 7 à un étiquetage présent à tort, 11 EN sont absentes des documents et 5 erreurs diverses.

4 Conclusion

Même s'il est encore perfectible, FRASQUES notre système mis au point dans le cadre d'EqueR, apporte une brique fondamentale à notre système multilingue MUSCAT qui jusqu'à présent ne possédait que peu de modules véritablement multilingues. En dehors des résultats et des comparaisons possibles avec les participants, la finalisation de systèmes est un des apports majeurs d'une campagne d'évaluation telle EQueR.

Références

AÏT-MOKTHAR S., CHANOD J.P, ROUX C., (2002), Robustness beyond shallowness : incremental deep parsing, *Journal of Natural Language Engineering*, Vol. 8, n°3-2.

BERTHELIN J.B., DE CHALENDAR G., ELKATEB-GARA F., FERRET O. GRAU B., HURAUL-PLANTET M., MONCEAUX L., ROBBA I., VILNAT A. (2003), Getting reliable answers by exploiting results from several sources of information, Actes de CoLogNET-ElsNET Symposium (Question and Answers: Theoretical and Applied Perspectives), Amsterdam.

FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G., JACQUEMIN C., MONCEAUX L., ROBBA I., VILNAT A. (2002) How NLP Can Improve Question Answering, *Knowledge Organization*, Vol. 29 (2002), N°3-4, pages 135-155

GRAU B., LIGOZAT A. L., ROBBA I., VILNAT A., ELKATEB-GARA F., ILLOUZ G., MONCEAUX L. PAROUBEK P., PONS O., (2005) De l'importance des synonymes pour la sélection de passages en question-réponse, Actes de CORIA, Grenoble.