Campagne d'évaluation EQueR-EVALDA
Evaluation en question-réponse

Christelle Ayache (1), Brigitte Grau (2), Anne Vilnat (2)

(1) Evaluations and Language resources Distribution Agency (ELDA)
55-57, rue Brillat Savarin 75013 Paris
ayache@elda.org
(2) Groupe LIR - Langues Information et Représentation - LIMSI
BP 133, 91403 Orsay Cedex
{brigitte.grau, anne.vilnat}@limsi.fr

Mots-clés : campagne d'évaluation, système de question-réponse, type de question, type de réponse, fichier-résultats (soumission).

Keywords : evaluation campaign, question-answering system, question type, answer type, run.

Résumé - Abstract

Cet article présente dans son ensemble la campagne d'évaluation EQueR-EVALDA. Cette campagne a bénéficié d'une aide du Ministère délégué à la Recherche dans le cadre de l'action Technolangue[1].

This paper describes the EQueR-EVALDA Evaluation Campaign. This campaign is supported by the French Ministry of Research within Technolangue.

1 Introduction

La campagne EQueR a offert un cadre d'évaluation aux systèmes de question-réponse pour la langue française, avec l'objectif d'alimenter l'activité de recherche dans le domaine en fournissant une photographie de l'état de l'art, notamment en France.

[1] L'action Technolangue est une action interministérielle destinée à mettre en place de manière pérenne une infrastructure de production et diffusion de ressources linguistiques.

EQueR a proposé deux tâches de recherche automatique de réponses : une tâche générique sur une collection hétérogène de textes – en large partie des articles de presse – et une tâche spécifique, liée au domaine médical, sur une collection de textes de cette spécialité.

L'esprit de la campagne EQueR correspondait davantage à une réflexion collective qu'à une véritable compétition ; néanmoins, aucune intervention manuelle n'a été autorisée pour la recherche et l'extraction des réponses.

2 Spécifications de la campagne

2.1 Collections de documents

Les participants ont eu accès aux collections textuelles quelques mois avant le test d'évaluation, après avoir rempli un accord d'utilisation finale des données. Les données ont été fournies sous forme de DVD ainsi que par téléchargement.

Les textes fournis étaient composés d'un balisage simple avec un identifiant de document, de titre et de paragraphe, et codés en ISO-Latin-1 (ISO-8859-1). Voici un exemple extrait du corpus au format EQueR :

```
<DOC>
<DOCID>LEMONDE95-000001</DOCID>
<LEAD1>DIMANCHE 01 JANVIER 1995 : NAISSANCE DE L'OMC,
ORGANISATION MONDIALE DU COMMERCE</LEAD1>
<TITLE>Un commerce mondial mieux réglementé</TITLE>
<P> AVEC l'année 1995, une nouvelle institution voit le
jour, qui devrait être porteuse de plus de justice
économique : l'Organisation mondiale du commerce(OMC).
Aux pays soumis à la dure concurrence internationale et à
ses coups bas, l'OMC apporte l'espoir qu'aux rapports de
force vont se substituer progressivement des
rapports</P></DOC>
```

Les textes originaux, avec leur balisage propre, ont également été mis à la disposition des participants.

Deux collections ont été élaborées : une collection pour la tâche « générale » et une collection pour la tâche « spécialisée ».

La collection générale, d'une taille d'environ 1,5 Go, était composée d'articles de presse de plusieurs années des journaux *Le Monde* et *Le Monde Diplomatique*, de dépêches de presse et de rapports d'information du Sénat français portant sur des sujets très variés.

La collection de textes de spécialité, d'une taille d'environ 140 Mo, était composée principalement 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.

2.2 Questions

Cinq types de questions ont été proposés aux systèmes participants : des questions « factuelles simples » (comprenant 7 catégories : personne, organisation, date, lieu, mesure, manière et objet/autre, « Qui est le président du Chili ? », « Quand a eu lieu le festival d'Avignon ? », etc.), des questions de type « définition » (autrement dit dont la réponse attendue est une définition, comprenant deux catégories : personne et organisation, « Qu'est-ce que l'OTAN ? », « Qui est Salvador Dali ? », etc.), des questions de type « liste » (« Quelles sont les 4 religions pratiquées en Hongrie ? »), des questions de type « oui/non » (« Existe-t-il une ligne de TGV Paris-Valencienne ? ») ainsi que des « reformulations » de questions factuelles simples déjà présentes dans le jeu de test.

Des questions sans réponse possible dans les collections de documents ont été introduites au sein du corpus de questions « général ». Dans ce cas, le système devait renvoyer en réponse : la valeur « NIL ».

Le type des questions était indiqué par un codage d'identification attribué à chaque question. Les identifiants de classes étaient : F (factuelle simple), D (définition), L (liste), et B (oui/non). Un « R » (reformulation) a été ajouté à l'identifiant de classe si nécessaire. Ci-après, un exemple de codage d'une question : « GF18 Où est né Jacques Chirac ? ». Ce codage indique que la question n°18 est de type factuel simple (F) et s'applique à la tâche générale (G).

Les sources et les modes de génération des questions ont été diversifiés. Une partie a été dérivée de mots clés qui accompagnaient les articles et les dépêches de presse, une autre partie a été créée par un groupe d'utilisateurs potentiels, dont certains connaissaient le domaine du TAL. La présence d'au moins une bonne réponse a été vérifiée manuellement dans le corpus pour chaque question proposée aux participants (hormis pour les questions « NIL »).

Pour la tâche générale, ELDA a élaboré un corpus de 500 questions réparties comme suit : 407 « factuelles », 32 « définitions », 31 « listes » et 30 « oui/non ».

Pour la tâche spécialisée, l'équipe du CISMeF a élaboré un corpus de 200 questions réparties comme suit : 81 « factuelles », 70 « définitions », 25 « listes » et 24 « oui/non ».

2.3 Réponses

Pour chaque question, les systèmes pouvaient renvoyer soit une réponse courte exacte, un passage (moins de 250 caractères contigus extrait d'un document de la collection) et un identifiant de document justifiant de cette réponse et de ce passage, soit au moins un passage et un identifiant de document justifiant ce passage.

Pour chaque type de questions (sauf pour les questions de type « liste), les systèmes pouvaient renvoyer jusqu'à cinq réponses ordonnées (20 pour les questions de type « liste »). Les réponses (ordonnées) devaient être présentées dans l'ordre des questions. Concernant les réponses de type « oui/non », les systèmes devaient pouvoir justifier du passage que ce soit pour une réponse positive ou négative.

3 Evaluation

La phase d'évaluation des différents systèmes a eu lieu sur chacun des sites des participants, et a duré une semaine, du 16 au 23 juillet 2004.

3.1 Réponses courtes et passages

La majeure partie des systèmes participants ont renvoyé un passage et une réponse courte exacte (un seul groupe a fait le choix de ne pas être évalué sur les réponses courtes). Les deux types de réponses ont été évalués distinctement.

En accord avec les participants lors de l'évaluation, deux sortes de jugements ont été appliqués, l'un porte sur les réponses courtes, l'autre sur les passages.

Concernant les réponses courtes, quatre types de jugements étaient possibles, la réponse était soit « correcte » (réponse juste et la plus précise possible, c'est-à-dire sans information obsolète), soit « inexacte » (réponse juste, mais pas assez précise, soit il manquait de l'information, soit au contraire de l'information avait été ajoutée), soit « incorrecte » (la réponse n'était pas juste, elle ne contenait pas la réponse attendue), soit « non justifiée » (la réponse était juste et exacte mais le document associé à la réponse ne justifiait pas celle-ci) . Pour l'évaluation des passages, seuls deux jugements étaient possibles : le passage était « correct » s'il contenait la réponse à la question et était justifié par le document associé, sinon, il était jugé « incorrect ».

Lorsqu'un système renvoyait « NIL », il s'agissait d'évaluer cette réponse comme si on évaluait un passage. Tout d'abord, vérifier que cette question était bien supposé renvoyer « NIL » ; si c'était le cas, le passage était jugé « CORRECT » ; sinon il était jugé « INCORRECT ».

3.2 Mesures adoptées

Pour les questions de type « factuel », « définition » et « oui/non », la mesure que nous avons adoptée est la Moyenne des Réciproques du Rang (MRR). Ce critère tient compte du rang de la première bonne réponse trouvée (métrique TREC[2]). Si une bonne réponse est trouvée plusieurs fois, elle n'est comptée qu'une seule fois.

$$MRR = \frac{1}{\text{nb questions}} \sum_{i=1}^{\text{nb questions}} \frac{1}{answer_i\ rank}$$

Pour les questions de type « liste », la mesure que nous avons adoptée est la précision moyenne (*non interpolated average precision, NIAP*, métrique TREC). Ce critère tient compte à la fois du rappel (pourcentage de bonnes réponses présentes dans la liste parmi toutes les bonnes réponses à trouver) et de la précision (pourcentage de bonnes réponses trouvées parmi toutes les réponses trouvées) mais aussi de la position des bonnes réponses dans la liste.

[2] TREC, Text REtrieval Conference, http://trec.nist.gov/

$$\text{prec\_moy}(q_i) = \frac{\sum_{j=1}^{j=n} I(rep_j).prec(j)}{R} \leq 1$$

avec :

$$I(rep_j) = \begin{cases} 1 \text{ si } rep_j \text{ est une bonne réponse} \\ 0 \text{ si } rep_j \text{ est une mauvaise réponse ou une réponse déjà proposée} \end{cases}$$

et :

$$\text{prec}(j) = \frac{\sum_{k=1}^{j} I(rep_k)}{j} = \frac{\text{Nombre de bonnes réponses différentes jusqu'au rang } j}{j} \leq 1$$

4 Résultats

4.1 Tâche générale

Sept groupes ont participé à la tâche générale de la campagne EQueR. Quatre laboratoires publics : le LIMSI, l'Université de Neuchâtel, le Laboratoire d'Informatique d'Avignon en collaboration avec iSmart et le CEA-LIST/LIC2M. Ainsi que trois institutions privées : France Télécom R&D, Synapse Développement et Sinequa.

Les travaux de la plupart de ces systèmes sont présentés plus en détail dans les pages suivantes des actes de l'atelier.

Au total, douze runs (ou fichier-résultats) ont été évalués. Deux juges ont évalué les résultats pendant un mois. De nombreuses discussions et mises au point ont permis d'optimiser la cohérence inter-juges.

Parmi les 500 questions du corpus de départ, cinq comportaient des erreurs. Nous avons décidé de supprimer ces cinq questions du corpus ainsi que de l'ensemble des fichier-résultats. Les scores ont donc été calculés sur la base de 495 questions réparties comme suit : 400 « factuelles », 33 « définitions », 31 « oui/non » et 31 « listes ».

Les trois systèmes de question-réponse ayant obtenu les meilleurs résultats pour la tâche générale lors de la campagne EQueR/EVALDA 2004 sont : pour les passages, les systèmes de Synapse Développement (participant 5), de Sinequa (participant 4), et du LIMSI (participant 2) ; Pour les réponses courtes : les systèmes de Synapse Développement, du LIA (participant 6) et du LIMSI.

Les résultats ont été fournis aux participants sous forme de deux tableaux, respectivement pour les réponses courtes et les passages. Le premier (Figure 1) présente pour chaque fichier-résultats, le nombre de questions traitées, le nombre de passages (ou réponses) corrects renvoyés, ainsi que les scores obtenus pour chaque type de question et combinaison. Le second (Figure 2) présente pour chaque fichier-réponse, un détail sur les passages (ou réponses) corrects renvoyés en indiquant le nombre de passages (ou réponses) corrects par type de réponse attendue (personne, temps, lieu, organisation…).

Identifiant du run	Nb de questions répondues [464]	Nb passages corrects	Nb passages incorrects	MRR sur tout les types sauf Listes	MRR sur Factuelles et Définitions	MRR sur Factuelles	MRR sur Définitions	MRR sur Oui/ Non	NIAP sur Listes	Nb de NIL renvoyés en rang 1	NIL	
											Préci-- sion	Rap-- pel
participant 5	464	378	86	0.7	0.71	0.7	0.74	0.67	0.29	4	1	0.8
participant 4	464	237	227	0.37	0.37	0.36	0.55	0.32	0	20	0.05	0.2
participant 2	464	210	254	0.37	0.38	0.37	0.47	0.25	0.09	69	0.01	0.2
participant 6	388	182	206	0.33	0.32	0.31	0.43	0.38	0.08	0	0	0
participant 3	464	184	280	0.31	0.31	0.3	0.43	0.35	0.08	54	0.01	0.2
participant 1	458	126	332	0.22	0.24	0.24	0.23	0.04	0	168	0.01	0.4
participant 7	464	113	351	0.18	0.17	0.17	0.17	0.38	0.13	236	0	0.4

Figure 1 : Résultats de l'évaluation tâche générale pour les passages

Identifiant du run	Passages corrects											
	Nb Définitions [33]		Nb Factuelles [400]							oui non [31]	Total	
	org [19]	pers [14]	lieu [65]	man [26]	mes [77]	org [28]	autre/ objet [67]	pers [95]	date [42]		Total [464]	%
participant 5	16	14	52	20	65	25	57	71	36	22	378	81.46
participant 4	14	13	38	9	32	17	40	41	23	10	237	51.07
participant 2	8	11	34	4	32	20	19	46	28	8	210	45.25
participant 6	10	7	32	0	31	9	7	49	25	12	182	39.22
participant 3	9	10	30	6	26	11	22	38	23	11	186	40.08
participant 1	6	6	20	2	20	7	12	35	14	4	126	27.15
participant 7	10	0	14	8	12	8	33	13	3	12	113	24.35

Figure 2 : Résultats de l'évaluation tâche générale pour les passages selon le type de réponse attendu

lieu = lieu, localisation, mes = mesure, org = organisation, pers = personne. man = manière, autre/objet = objet ou autre, date = date, temps

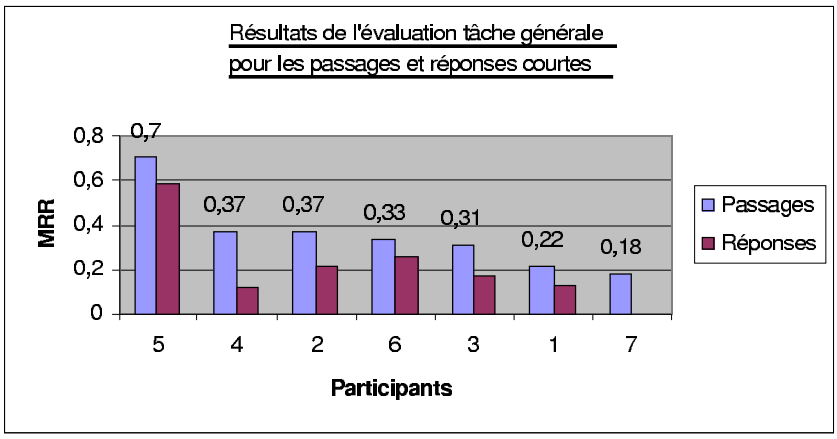
Identifiant du run	Nb de questions répondues [464]	Nb passages corrects	MRR sur tout les types sauf Listes	MRR sur Factuelles et Définitions	MRR sur Factuelles	MRR sur Définitions	MRR sur Oui/ Non	NIAP sur Listes
participant 5	464	312	0.58	0.58	0.57	0.69	0.67	0.71
participant 6	388	139	0.25	0.24	0.24	0.27	0.38	0.02
participant 2	463	131	0.22	0.22	0.24	0	0.25	0.02
participant 3	464	106	0.17	0.16	0.16	0.13	0.35	0
participant 1	333	80	0.13	0.15	0.16	0.01	0.04	0
participant 4	195	76	0.12	0.13	0.09	0.58	0	0

Figure 3 : Résultats de l'évaluation tâche générale pour les réponses courtes

Identifiant du run	Réponses courtes correctes											
	Nb Définitions [33]		Nb Factuelles [400]							oui non [31]	Total	
	org [19]	pers [14]	lieu [65]	man [26]	mes [77]	org [28]	autre/ objet [67]	pers [95]	date [42]		Total [464]	%
participant 5	14	14	46	5	63	19	28	67	34	22	312	67.24
participant 6	8	1	24	0	21	5	5	46	17	12	139	29.95
participant 2	0	0	24	1	24	9	4	39	22	8	131	28.23
participant 3	3	4	15	0	19	9	7	26	12	11	106	22.84
participant 1	1	0	15	1	14	4	3	27	11	4	80	17.24
participant 4	13	11	14	0	10	0	2	18	8	0	76	16.37

Figure 4 : Résultats de l'évaluation tâche générale pour les réponses courtes selon le type de réponse attendu

Pour une plus grande visibilité des résultats, nous avons fourni aux participants les résultats de la tâche générale sous forme de graphe.



4.2 Tâche spécialisée

Cinq groupes ont participé à la tâche spécialisée dans le domaine médical. Trois laboratoires publics : l'Université de Neuchâtel, le CEA-LIST/LIC2M et AP/HP en collaboration avec Paris XIII. Ainsi que deux institutions privées : France Télécom R&D et Synapse Développement.

Les travaux de la plupart de ces systèmes sont présentés plus en détail dans les pages suivantes des actes de l'atelier.

Au total, sept fichier-résultats ont été évalués. Un juge spécialiste de l'équipe du CISMeF (Catalogue et Index des Sites Médicaux Francophones) du CHU de Rouen a évalué les résultats. Les scores ont été calculés sur la base de 200 questions réparties comme suit : 81 « factuelles », 70 « définitions », 24 « oui/non » et 25 « listes ».

Les trois systèmes de question-réponse ayant obtenu les meilleurs résultats pour la tâche spécialisée lors de la campagne EQueR/EVALDA 2004 sont : pour les passages, les systèmes de Synapse Développement (participant 4), de l'Université de Neuchâtel (participant 2), et *ex-aequo* les systèmes de AP/HP-Paris XIII (participant 3) et de France Télécom R&D

(participant 1) ; pour les réponses courtes : le système de Synapse Développement, et *ex-aequo* les systèmes de AP/HP-Paris XIII et de l'Université de Neuchâtel.

Les résultats ont été fournis aux participants sous la forme d'un seul tableau, respectivement pour les réponses courtes et les passages. Il présente pour chaque fichier-résultats le nombre de questions traitées, le nombre de passages (ou réponses) corrects renvoyés, ainsi que les scores obtenus pour chaque type de question et combinaison.

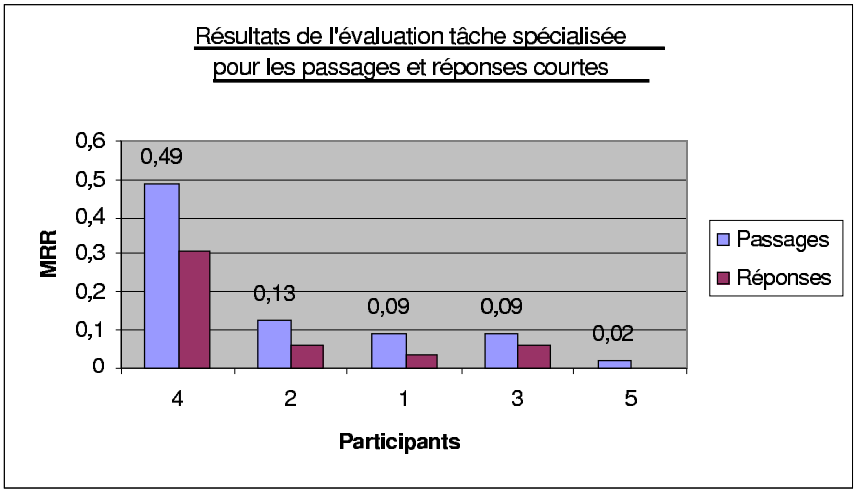
Identifiant du run	Nb de questions répondues [175]	Nb passages corrects	Nb passages incorrects	% passages corrects	MRR sur tout les types sauf Listes	MRR sur Factuelles et Définitions	MRR sur Factuelles	MRR sur Définitions	MRR sur Oui/ Non	NIAP sur Listes
participant 4	175	110	65	62.85	0.49	0.51	0.42	0.62	0.37	0.02
participant 2	175	27	148	15.42	0.13	0.13	0.23	0.02	0.08	0.02
participant 1	166	23	143	13.14	0.09	0.09	0.11	0.07	0.04	0
participant 3	112	16	96	9.14	0.09	0.05	0.02	0.08	0.33	0.01
participant 5	175	7	168	4	0.02	0.02	0.04	0	0	0

Figure 5 : Résultats de l'évaluation tâche spécialisée pour les passages

Identifiant du run	Nb de questions répondues [175]	Nb passages corrects	Nb passages incorrects	% passages corrects	MRR sur tout les types sauf Listes	MRR sur Factuelles et Définitions	MRR sur Factuelles	MRR sur Définitions	MRR sur Oui/ Non	NIAP sur Listes
participant 4	175	110	65	62.85	0.49	0.51	0.42	0.62	0.37	0.02
participant 2	175	27	148	15.42	0.13	0.13	0.23	0.02	0.08	0.02
participant 3	166	23	143	13.14	0.09	0.09	0.11	0.07	0.04	0
participant 1	112	16	96	9.14	0.09	0.05	0.02	0.08	0.33	0.01

Figure 6 : Résultats de l'évaluation tâche spécialisée pour les réponses courtes

Pour une plus grande visibilité des résultats, nous avons fourni aux participants les résultats de la tâche médicale sous forme de graphe :



4.3 Analyse des résultats

En premier lieu, nous avons pu constater de meilleurs résultats pour la tâche générale que pour la tâche spécialisée : les meilleurs scores des systèmes pour la tâche générale s'échelonnent entre 0,7 et 0,18 (selon la métrique adoptée : MRR, Moyenne des Réciproques du Rang, cf. paragraphe 3.2) alors que pour la tâche médicale les résultats s'échelonnent entre 0,49 et 0,02 (toujours d'après la métrique adoptée, MRR, cf. paragraphe 3.2). Ceci s'explique peut-être par la spécificité des textes liés au domaine médical contenu dans cette tâche. Mais ceci s'explique peut-être aussi en raison du délai de livraison du corpus médical (le corpus pour la tâche médicale n'a été distribué que quelques semaines avant l'évaluation).

De plus, l'ensemble des systèmes ont obtenu un meilleur score lors de l'évaluation des passages que lors de l'évaluation des réponses courtes. En effet, il paraît plus difficile pour un système d'extraire une réponse courte exacte et précise qu'un passage un peu plus long dans lequel finalement il est plus probable que se trouve la réponse attendue.

Si l'on compare l'ensemble des systèmes participants on s'aperçoit qu'ils allient tous plus ou moins massivement des technologies de Traitement Automatique des Langues. Pourtant, au vu des résultats, un système obtient des résultats nettement supérieurs aux autres participants, et ce, pour les deux tâches, générale et spécialisée. L'équipe de Synapse Développement présentant ses travaux dans les actes de la conférence principale TALN, on pourra y trouver des éléments d'explication.

Concernant la tâche générale, nous avons trouvé intéressant de faire connaître aux participants les résultats en fonction du type de réponse attendu. Ainsi, ils ont pu se rendre compte, sur quel type de question et de réponse, leur système avait été le plus performant lors de l'évaluation. Tous systèmes confondus, lors de l'évaluation des passages, les meilleurs résultats obtenus concernent les questions de type « définition », puis les questions de type « factuel » simple, les questions de type « oui/non » et enfin les questions de type « liste » pour lesquelles les systèmes ont rencontré le plus de difficultés. Concernant spécifiquement les questions de type « définition », les systèmes ont obtenu de meilleurs résultats lorsque la réponse attendue était une organisation plutôt qu'une personne. Concernant les questions de type « factuel » simple, les systèmes ont obtenu de meilleurs résultats lorsque la réponse attendue était de type « lieu », « organisation », « personne » ou « date » plutôt que « manière », « mesure » ou « objet ».

Pour la tâche générale, lors de l'évaluation des passages, le meilleur système a obtenu 81,46 % de bonnes réponses contre 51,07 % pour le deuxième système. Lors de l'évaluation des réponses courtes, la moyenne baisse avec 67,24 % de bonnes réponses pour le meilleur système et seulement 29,95 % pour le deuxième.

Pour la tâche spécialisée, les résultats baissent encore. Le meilleur système, lors de l'évalution des passages, a obtenu 62,85 % de bonnes réponses contre 15,42 % pour le deuxième système. Et lors de l'évaluation des réponses courtes, le meilleur système obtient seulement 40,57 % de bonnes réponses contre 7,42 % pour le deuxième.

Nous constatons bien une frontière entre les résultats du premier système et ceux des autres systèmes.

5 Conclusion et perspectives

En conclusion, cet article a décrit les principaux aspects de la première campagne d'évaluation de systèmes de question-réponse en France : EQueR.

Cette campagne a été un véritable succès avec la participation et l'intérêt croissant d'une très large majorité des acteurs académiques et industriels du domaine (au total, 7 participants français et 1 participant suisse). Certains participants n'avaient jamais fait d'évaluation question-réponse auparavant et jamais autant de groupes français n'avaient participé à une évaluation question-réponse de la sorte.

Concernant le domaine de l'évaluation, EQueR a innové avec un nouveau type de question, les questions de type « oui/non », qui ont suscitées beaucoup d'intérêt de la part des participants. EQueR a gagné aussi en proposant une tâche question-réponse dans un domaine spécialisé, ce qui a permis d'attirer d'autres participants intéressés plus particulièrement par le domaine médical.

Enfin, EQueR s'européanise avec la campagne d'évaluation CLEF[3] qui, depuis l'année dernière, offre une tâche spécialisée pour l'évaluation des systèmes de question-réponse en Europe. ELDA joue le rôle de coordinateur pour le français dans la campagne européenne CLEF ainsi que celui de distributeur pour l'ensemble des ressources européennes. Au vu des résultats de la campagne EQueR, nous pouvons constater que pour les meilleurs systèmes, les résultats sont comparables avec les résultats des meilleurs systèmes de la campagne CLEF 2004. Concernant la campagne CLEF 2005 qui débutera très prochainement, notre expérience de par la campagne EQueR a été très enrichissante aussi bien pour constituer les corpus de questions que pour discuter de la façon dont seront compilées les données, etc.

Notre souhait est de pouvoir voir en la campagne européenne CLEF l'avenir d'une campagne très enrichissante comme EQueR en France.

Références

AYACHE C. (2005), Rapport final de la campagne EVALDA/EQueR, Evaluation en Question-Réponse, http://www.technolangue.net/article61.html.

VALIN A., MAGNINI B., et AL.(2004), Overview of the CLEF 2004 Multilingual Question Answering Track, Actes de *Cross Language Evaluation Forum.*

VOORHEES E., HARMAN D., (1999), Overview of the Eight Text REtrieval Conference (TREC8), *National Institute of Standards and Technology*, page 1.

[3] CLEF, Cross Language Evaluation Forum, www.clef-campaign.org