

Campagne EVALDA/EQueR Evaluation en Question-Réponse

Contact : Christelle Ayache (ayache@elda.org)

Coordinateurs : Khalid Choukri (choukri@elda.org), Brigitte Grau (grau@limsi.fr)



Sommaire

1. Historique du document	-----p.3
2. Résumé	-----p.3
3. Spécifications	
3.1. Introduction	-----p.6
3.2. Collections de documents	-----p.6
3.3. Questions	-----p.8
3.4. Réponses	-----p.11
3.5. Evaluation	-----p.12
3.6. Exemples de jugements pour les réponses courtes	-----p.13
3.7. Mesures adoptées / scoring	-----p.15
4. Résultats EQueR pour la tâche générale	
4.1. Participants	-----p.17
4.2. Jugement, évaluation des résultats	-----p.17
4.3. Calcul des scores	-----p.18
4.4. Présentation des résultats	-----p.19
5. Résultats EQueR pour la tâche médicale	
5.1. Participants	-----p.22
5.2. Jugement, évaluation des résultats	-----p.22
5.3. Calcul des scores	-----p.23
5.4. Présentation des résultats	-----p.24
6. Package d'évaluation EQueR	
6.1. Pourquoi un package d'évaluation ?	-----p.26
6.2. Contenu du package d'évaluation EQueR	-----p.26
7. Conclusion	-----p.27

1. Historique du document

Version	Date	Auteur(s)	Commentaires
1.0	07/03/05	Christelle Ayache	Premier draft
1.1	16/03/05	Christelle Ayache	Ajout commentaires participants
1.2	14/04/05	Christelle Ayache	Version finale

2. Résumé

Ce rapport 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 Technolangue¹.

La problématique des systèmes de question-réponse se situe à l'intersection de plusieurs domaines, dont notamment la recherche d'information et le traitement de la langue naturelle et l'apprentissage automatique, avec l'apprentissage de critères de sélection et de classification d'extraits. Alors qu'à l'heure actuelle certains pensent que les moteurs de recherche documentaire ont tendance à stagner, des avancées en question-réponse profitent largement à cette activité.

La campagne d'évaluation EQueR, en offrant un cadre d'évaluation à des systèmes complets de question-réponse, offre aussi une opportunité d'évaluer l'apport de modules de différentes natures par leur impact sur la tâche complète (moteur de recherche, modules de traitement automatique des langues, etc.).

La campagne d'évaluation EQueR s'est déroulée en six phases principales :

1. La première phase visait à spécifier et produire les ressources linguistiques nécessaires à la campagne d'évaluation.
2. La seconde phase visait à mettre en place l'environnement scientifique et technique indispensables aux tests d'évaluation.
3. La troisième phase consistait à faire exécuter les tests par les participants.
4. La quatrième phase consistait à recueillir et analyser les résultats.
5. La cinquième phase consistait à organiser l'atelier de clôture de la campagne
6. la sixième phase consistait à fusionner les résultats en vue de la production d'un corpus validé.

Les retombées escomptées de cette campagne d'évaluation EQueR sont multiples.

En premier lieu des retombées scientifiques sont envisagées pour la promotion de la recherche dans le domaine de la recherche d'information mais aussi une avancée sur la problématique des

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 (<http://www.technolangue.net/>)

systèmes de question-réponse et sur leur évaluation ; des retombées technologiques pour le perfectionnement et l'amélioration des systèmes évalués dans le cadre de cette campagne ; mais aussi, des retombées économiques pour la mise à disposition des corpus et outils créés.

Différents acteurs ont participé à cette campagne :

- **Coordinateurs** : Khalid Choukri / Christelle Ayache (ELDA), Brigitte Grau (LIMSI)
- **Participants** : AP/HP-Paris XIII, France Télécom R&D, LIMSI, Sinequa, Synapse Développement, LIA-iSmart, l'Université de Neuchâtel, CEA-LIST/LIC2M.
- **Support** : Systal-Pertimm (moteur de recherche) ; l'équipe du CISMef de Rouen, Catalogue et Index des Sites Médicaux Francophones (constitution du corpus de questions spécialisées et jugement des résultats des systèmes participants à la tâche médicale).

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.

Pour la tâche générale, ELDA a élaboré un corpus de 500 questions.

Pour la tâche spécialisée, l'équipe du CISMef a élaboré un corpus de 200 questions.

Ces deux corpus ont été constitués en tenant compte des différents types de questions prévus pour cette campagne, à savoir : « factuel », « définition », « oui/non » et « liste ».

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.

Pour chaque question les systèmes pouvaient renvoyer soit, une réponse courte exacte, un passage (moins de 250 caractères contigus) 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.

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.

De plus, chaque participant pouvait soumettre jusqu'à deux runs par tâche, c'est-à-dire que chaque participant avait la possibilité de faire tourner une deuxième fois son système après avoir modifier quelques spécificités sur celui-ci.

Les résultats ont été jugés par deux évaluateurs humains pendant 1 mois et compilés en interne au mois de septembre. Les résultats ont été fournis aux participants le 1er octobre 2004.

Le classement des 3 premiers systèmes a été établi en fonction du taux de MRR global (Moyenne des Réciproques du Rang calculée sur l'ensemble des types de question, hormis sur les questions de type « liste »). Ce classement est indiqué pour les deux tâches (générale et médicale) et respectivement pour les passages et pour les réponses courtes.

Toutes tâches confondues et tous systèmes confondus, les résultats obtenus par les systèmes oscillent entre 0.02 pour le système ayant obtenu les moins bons résultats et 0.7 pour le meilleur système.

Pour la tâche générale, les résultats varient entre 0.18 pour le système ayant obtenu les moins bons résultats et 0.7 pour le meilleur système.

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, de Sinequa, et du LIMSI.

pour les réponses courtes : les systèmes de Synapse Développement, du LIA (Laboratoire Informatique d'Avignon) et du LIMSI.

Concernant la tâche médicale, les résultats se situent entre 0.02 pour le moins bon système et 0.49 pour le meilleur.

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, de l'Université de Neuchâtel, et ex-aequo les systèmes de AP/HP-Paris XIII et de France Télécom R&D.

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.

Ce premier rapport de projet EQueR comporte l'ensemble des spécifications ainsi que les résultats fournis aux participants

3. Spécifications

3.1. 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.

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.

Les participants ont reçu des jeux de questions différents pour les deux tâches. Les questions ont été élaborées en tenant compte de différents types de réponses attendues : questions de types « factuel », « définition », « liste fermée d'éléments » ou encore de type « oui/non ». Certaines questions n'avaient pas de réponse dans les collections textuelles utilisées.

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.

Les évaluateurs humains vérifiaient puis jugeaient la réponse exacte ET le passage retourné par un système participant et ce, pour chaque question. Vérifier une réponse signifiait vérifier qu'elle soit correcte et justifiée par un document.

3.2. 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. Par exemple, voici un extrait de document source du journal Le Monde :

```

<DOC>
<DOCNO>LEMONDE95-000001</DOCNO>
<DOCID>LEMONDE95-000001</DOCID>
<ACCOUNT>365857</ACCOUNT>
<GENRE>BULLETIN</GENRE>
<DATE>19950102</DATE>
<LMDOC>LLY</LMDOC>
<SUBJECTS>2</SUBJECTS>
<ETA1>BASE</ETA1>
<FAB>Q311294:0100</FAB>
<SUBJECTS>FRET</SUBJECTS>
<NUM1>950102-2-001-00</NUM1>
<NAMES>GATT,OMC</NAMES>
<PUM1>QUO</PUM1>
<REFERENCE1>2-001-00</REFERENCE1>
<SEC1>ETR</SEC1>
<SU21>BULLETIN</SU21>
<PAGE>1</PAGE>
<LEAD1>DIMANCHE 01 JANVIER 1995 : NAISSANCE DE L'OMC,
ORGANISATION MONDIALE DU COMMERCE</LEAD1>
<TITLE>Un commerce mondial mieux réglementé</TITLE>
<TEXT> 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
... </TEXT></DOC>

```

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. Les fenêtres temporelles couvertes par les différentes collections ont été contrôlées, dans le but d'assurer au mieux la couverture des sujets des questions, qui ont été ainsi traités selon plusieurs points de vue et types de rédaction : articles d'actualité, articles de fond, dépêches, rapports.

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.

3.3. Questions

Typologie des questions

La typologie des questions fournies aux systèmes participants a été arrêtée comme suit :

- 1) Questions factuelles simples (temps, lieu, personne, objet, mesure...)
Ex : « Où est né Jacques Chirac ? »
- 2) Questions dont la réponse attendue est une « liste »
Ex : « Quel sont les pays signataires du traité de Schengen ? »
- 3) Questions dont la réponse attendue est une « définition »
Ex : « Qu'est-ce que le SMIC ? »
- 4) Questions de type booléen, c'est-à-dire, dont la réponse attendue est « oui » ou « non »
Ex : « Est-ce que Jean-Paul II a visité la Chine ? »
- 5) Reformulations de questions factuelles simples
Ex : « Dans quelle ville Jacques Chirac est-il né ? »

Pour chacune de ces questions une justification (passage au minimum) est attendu.

3.3.1 Les « Factuelles »

Les réponses simples aux questions factuelles portaient sur des dates, des mesures de durée, distance ou dimension, des lieux, des personnes, des organisations, la manière ou le mode de déroulement d'événements (« Comment... ? »), des entités concrètes ou abstraites. Les unités de mesure qui qualifient des quantités (par exemple, « m² ») devaient être incluses dans les réponses aux questions de type « Combien... ? ».

Les questions qui demandent une réponse subjective (« Quel est le principal monument de Paris ? ») ou une réponse à liste ouverte (« Quels sont les peintres surréalistes ? ») n'ont pas été proposées, ainsi que les questions « emboîtées » (« Où se trouve l'édifice le plus haut d'Europe ? »).

3.3.2 Les « Définitions »

Les questions dont la réponse est une définition avaient pour objet une personne ou une organisation et ont été formulées de manière à attendre une réponse courte, présente dans un

document du corpus, par exemple « Qui est Jacques Chirac ? » > « président français » ou « Qu'est-ce que le SMIC ? » > « Salaire Minimum Interprofessionnel de Croissance ».

Compte tenu des difficultés rencontrées dans TREC 2003 et à l'instar de la campagne CLEF-QA 2004, les définitions de concepts (« Qu'est-ce que l'art brut ? ») qui demandent l'assemblage d'informations à partir de plusieurs documents n'ont pas été proposées.

3.3.3 Les « Reformulations »

Une partie des questions étaient des reformulations de questions déjà présentes dans le jeu de test. Les réponses étaient à extraire, *a priori*, dans le même passage.

Cependant, la génération des questions de type « reformulation » s'est avérée être une tâche compliquée. En effet, certaines de ces questions n'ont pas été correctement formulées, ce qui a posé un problème pour pouvoir les inclure dans l'analyse des résultats en tant que type de questions à part entière.

Finalement, il s'avère que, lors du calcul des scores, nous n'avons pas pris en compte ce type de questions comme étant des reformulations, mais tout simplement comme étant des questions de type « factuel », « définition », « liste » ou « oui/non » comme elles l'étaient à l'origine.

3.3.4 Les « Booléennes »

Des questions directes et indirectes ont été également proposées. Il s'agissait pour les systèmes de renvoyer en réponse courte « OUI » ou « NON » accompagnée du passage justifiant cette réponse. Pour ces questions, seule la première ligne-réponse de chacun des systèmes a été prise en compte.

3.3.5 Les « Listes »

Les questions de type « liste » attendaient un nombre bien précis de réponses (nombre indiqué dans la question). Cependant, les systèmes pouvaient renvoyer jusqu'à 20 lignes-réponses par question.

3.3.6 « NIL »

Cinq 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 : 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),
- B (oui/non).
- R (reformulation) a été ajouté à l'identifiant de classe si nécessaire.

Un jeu de questions spécifiques a été fourni pour chacune des deux tâches. Les questions ont été catégorisées selon les mêmes sous-classes (factuel, définition, etc.) dans les deux jeux de questions.

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 était 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 dans le corpus pour chaque question proposée aux participants.

Nombre de questions

Les quotas des différents types de questions ont été contrôlés.

Pour la tâche question-réponse dans le corpus « général », les participants ont reçu un jeu de 500 questions dont 100 reformulations, notamment :

- 407 factuelles (dont 98 reformulations) → dont 5 questions NIL sans réponse dans le corpus
- 32 définitions (dont 1 reformulation)
- 31 listes (dont 1 reformulation)
- 30 oui/non (pas de reformulation)

Pour la tâche question-réponse dans le corpus « médical », les participants ont reçu un jeu de 200 questions dont 51 reformulations, notamment :

- 81 factuelles, réponse simple (dont 38 reformulations)
- 25 listes (dont 0 reformulation)
- 70 définitions (dont 13 reformulations)
- 24 oui/non (dont 0 reformulation)

3.4. Réponses (format de sortie...)

Concernant les réponses de type « Factual » ou « Définition », les participants pouvaient nous soumettre :

- soit des passages textuels (250 caractères contigus max.) contenant la réponse,
- soit des réponses courtes et précises et des passages textuels (250 caractères contigus max.)

Pour chaque question, les systèmes pouvaient nous soumettre jusqu'à cinq réponses ordonnées (20 pour les questions de type « liste »). Les réponses (ordonnées) devaient être présentées les unes en dessous des autres dans l'ordre des questions.

Le format des réponses renvoyé en sortie par les systèmes participants représentait 5 champs séparés par une tabulation (tab) :

Identifiant de la question	Identifiant du run	Identifiant du document	Réponse exacte	Passage
GF1	elda04g1	LEMONDE95-000001	Paris	aura lieu a Paris ; la capitale de la France va accueillir

1. **l'identifiant de la question** : tel qu'il était fourni en entrée dans le jeu de test ;
2. **l'identifiant du run** : il indiquait le nom du participant (séquence de quatre caractères), l'année, la tâche (G ou M) et le numéro de « run » soumis par le participant (1 ou 2). Cet identifiant ne changeait pas à l'intérieur d'un jeu de résultats soumis (un « run ») et correspondait aussi au nom du fichier qui contenait le jeu de résultats (par ex. elda04g1.txt).
3. **l'identifiant du document** : tel qu'il était fourni dans les corpus, indiqué par la balise <DOCID>. Certaines questions n'avaient pas de réponse dans les corpus. Dans ce cas, la référence au document était NIL par défaut.
4. **la réponse exacte** : elle pouvait contenir NUL si le run n'était censé fournir que les passages (au choix du participant). Elle était vide si l'identifiant du document est NIL (aucune réponse trouvée dans le corpus).
5. **le passage** : une contrainte a été mise en place, pour pouvoir être évalués les passages ne devaient pas dépasser 250 caractères.

Les réponses de type « liste » ont été plafonnées à 20 éléments, 1 réponse par ligne. Elles pouvaient être réparties, *a priori*, dans plusieurs paragraphes mais pas dans plusieurs documents.

Concernant les réponses de type « oui/non », les systèmes devaient pouvoir justifier du passage que ce soit pour une réponse positive ou négative.

3.5. Evaluation

S'agissant d'une évaluation sur la langue française, il était important que les fichiers soient jugés par des juges dont la langue native était le français.

Règle fondamentale : Une réponse était considérée correcte lorsqu'elle était justifiée par un document.

Les résultats ont fait l'objet d'un contrôle manuel pour déterminer si une réponse pouvait être correcte et, éventuellement, précise. Le jugement de la pertinence des réponses a été du ressort de l'équipe d'évaluation.

Evaluation réponse courte / passage :

Dans un même run (ou « fichier-réponse »), à la fois la réponse courte et le passage pouvaient être évalués. Le passage a été systématiquement évalué. Les participants devaient simplement indiquer pour chaque run s'ils souhaitaient se faire évaluer sur les réponses courtes ou non.

2 types d'évaluation pour les réponses :

Réponse courte :

CORRECTE : si la réponse était juste et précise (sans aucune information obsolète). La réponse exacte devait être la plus petite partie correcte de ce qui était fourni dans le document.

INEXACTE : si la réponse était juste mais pas assez précise (soit il manquait quelque chose, soit, au contraire, de l'information avait été ajoutée, il ne s'agissait donc pas de la réponse la plus exacte possible).

INCORRECTE : si la réponse n'était pas juste, elle ne correspondait pas à la réponse attendue.

NON JUSTIFIEE (par le document) : si la réponse était correcte mais le document retourné ne justifiait pas de la réponse (si on ne retrouvait pas la réponse dans le document par exemple).

Passage :

CORRECTE : si le passage contenait une réponse juste et précise.

INCORRECTE : si le passage contenait une réponse incorrecte.

Remarque : Lorsqu'un système renvoyait NIL (pas de réponses dans l'ensemble du corpus), il s'agissait d'évaluer cette réponse comme si on évaluait un passage.

Tout d'abord, vérifier que cette question était bien supposée renvoyer NIL ; si c'était le cas, alors le passage était jugé « CORRECT » ; si ce n'était pas le cas, alors le passage devait être jugé « INCORRECT ».

3.6. Exemples de jugements pour les réponses courtes

Sachant toujours que :

Une réponse était considérée correcte que lorsqu'elle était justifiée par un document. C'est-à-dire que pour chaque question, il ne fallait pas s'attendre à une réponse unique et absolue. Du moment qu'un document paraissait justifier correctement la réponse à une question, cette réponse était à considérer juste.

- **Question-réponse de type « FACTUEL » :**

Qui a épousé Bill Gates à Hawaï ?

Melinda CORRECTE (si justifiée par document)
 Melinda French CORRECTE (si justifiée par document)
 sa collègue Melinda French CORRECTE (si justifiée par document)
 a épousé sa collègue Melinda French INEXACTE

Quel âge a l'abbé Pierre ?

81 ans CORRECTE (si justifiée par document)
 quatre-vingt-cinq ans CORRECTE (si justifiée par document)
 âgé de 81 ans INEXACTE

Combien y a-t-il d'habitants à Saint-Chéron ?

4000 CORRECTE (si justifiée par document)
 4000 hab. CORRECTE (si justifiée par document)
 4000 habitants CORRECTE (si justifiée par document)

Quel est le record du monde du 100 mètres ?

Ici, la question n'était pas précise, on ne pouvait savoir s'il s'agissait du 100 mètres en course à pied ou en natation par exemple. Il suffisait donc qu'un document justifie d'une bonne réponse pour qu'elle soit jugée CORRECTE, qu'il s'agisse d'un record de course à pied ou de natation.

48s42 CORRECTE (si justifiée par document)

- **Question-réponse de type « DEFINITION » :**

Que signifie CGT ?

Confédération générale du travail CORRECTE (si justifiée par document)
 Confédération générale des travailleurs CORRECTE (si justifiée par document)

Quelle est la définition de « chimiothérapie » ?

Thérapeutique par les substances chimiques CORRECTE (si justifiée par document)
 un traitement complémentaire au traitement chirurgical CORRECTE (si justifiée par document)

- **Question-réponse de type « LISTE » :**

Chacune des questions « LISTE » créée dans les corpus de questions EQueR avait, *a priori*, l'ensemble de ses réponses dans un seul document.

Citez quatre infractions militaires.

Insoumission	CORRECTE (si justifiée par document)
capitulation	CORRECTE (si justifiée par document)
insubordination	CORRECTE (si justifiée par document)
désertion	CORRECTE (si justifiée par document)

- **Question-réponse de type « OUI/NON » :**

Les réponses aux questions « oui/non » devaient être justifiées par un passage pouvant approuver la réponse courte OUI ou NON.

Existe-t-il une ligne de TGV Valenciennes-Paris ?

Rép : « OUI » CORRECTE (si justifiée par passage)
 Passage : « Les 170 passagers d'un TGV Valenciennes-Paris en avaient été quitte pour la peur de leur vie le 21 décembre vers 07H30 du matin. » CORRECTE

3.7. Mesures adoptées

- Questions de type « factuel », « définition », et « oui/non » :

La mesure que nous avons adoptée était la Moyenne des Réciproques du Rang (MRR).

Ce critère tient compte du rang de la première bonne réponse trouvée (métrique TREC). Si une bonne réponse est trouvée plusieurs fois, elle n'est comptée qu'une seule fois.

Une réponse NIL était acceptée seulement si elle se présentait en première position. Les systèmes ne trouvant pas la bonne réponse en rang 1 étaient désavantagés en fonction de cette mesure.

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

Exemple de calcul pour la MRR :

Soit trois questions, et pour chacune d'entre elles, les 5 réponses ordonnées suivantes (1 étant une réponse correcte, 0 une réponse incorrecte) :

	Réponse 1	Réponse 2	Réponse 3	Réponse 4	Réponse 5
Question 1	0	0	1	1	1
Question 2	0	1	0	1	1
Question 3	1	0	0	1	1

la moyenne des réciproques du rang vaut :

$$MRR = \frac{1}{3} * \left(\frac{1}{3} + \frac{1}{2} + \frac{1}{1} \right) = \frac{1}{3} * \frac{11}{6} = \frac{11}{18}$$

- **Questions de type « liste » :**

La mesure que nous avons adoptée pour les questions de type « liste » était 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.

$$\text{prec_moy}(q_i) = \frac{\sum_{j=1}^{j=n} I(\text{rep}_j) \cdot \text{prec}(j)}{R} \leq 1$$

avec :

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

et :

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

Exemples de calcul pour la précision moyenne :

Soit la question q pour laquelle 3 (bonnes) réponses sont à trouver.

- Si la liste à évaluer présente les 3 réponses à trouver dans les 3 premières positions, la précision moyenne vaut :

$$(1/1 + 2/2 + 3/3) / 3 = 1.$$

- Si les 3 réponses à trouver sont en positions 3, 4 et 5, la précision moyenne vaut :

$$(1/3 + 2/4 + 3/5) / 3.$$

- Si seulement 2 bonnes réponses sont trouvées et qu'elles se trouvent en positions 1 et 2, la précision moyenne vaut :

$$(1/1 + 2/2) / 3 = 2/3.$$

- Si seulement 1 bonne réponse est trouvée et qu'elle se trouve en position 5, la précision moyenne vaut :

$$(1/5) / 3.$$

4. Résultats EQueR pour la tâche « générale »

4.1. Participants

7 groupes participants :

- 4 laboratoires publics :
- LIMSI
 - Université de Neuchâtel
 - Laboratoire Informatique d'Avignon – iSmart
 - CEA-LIST/LIC2M (uniquement pour les passages)
- 3 institutions privées :
- France Télécom R&D
 - Synapse Développement
 - Sinequa

4.2. Jugement, évaluation des résultats

- Au total :
- 12 runs (ou fichiers-résultat) à évaluer
 - 500 questions par run
 - 5 réponses possibles par question

Deux juges ont évalué les résultats pendant un mois. De nombreuses discussions et mises au point ont été engagées, pour un maximum de cohérence entre eux deux.

Les deux juges ont réalisé une évaluation croisée sur deux runs (chacun a évalué « à vide » 2 runs que l'autre juge avait déjà évalués), puis nous avons calculé la cohérence entre ces runs.

→ Nous avons obtenu moins de 5% de désaccord entre les juges, de ce fait, nous avons pu valider leurs jugements.

Lors de l'évaluation, les fichiers-résultat de chaque participant ont été complétés. Deux champs ont été remplis en première et deuxième positions :

- **champ 1** : jugement de la réponse courte.

-1 : réponse courte non jugée. (NUL : le participant ne souhaitait pas se faire évaluer pour une réponse courte).
 0 : réponse jugée « correcte »
 1 : réponse jugée « incorrecte »
 2 : réponse jugée « inexacte » (contenant seulement une partie de la réponse, ou trop bruitée par exemple)
 3 : réponse jugée « non supportée par le document »

- **champ 2** : jugement du passage

-1 : passage non jugée (si la taille du passage dépasse 250 caractères par exemple)
 0 : passage jugé « correct »
 1 : passage jugé « incorrect »

4.3. Calcul des scores

Sur les 500 questions, 5 d'entre elles comportaient des erreurs à l'origine :

- GF44 Combien de personnes ont visité le Musée océanographique de Monaco en 1997 ? (date)
- GF79 Quelle est la préfecture d'Aomori ? (incompréhension)
- GF88 Qui remplace Noel Coppin à "La Croix" ? (mal orthographié)
- GF101 À quelle société appartient le satellite Asisat-2 ? (mal orthographié)
- GF145 Quelle fonction occupe Philippe Dépernet au sein de Prisunic ? (erreur prénom)

Nous avons pris la décision de supprimer ces 5 questions du corpus ainsi que de l'ensemble des runs. Les scores ont donc été calculés sur la base de **495 questions** réparties comme suit :

FACTUEL	400	(dont 5 questions NIL, sans réponse dans le corpus)
lieu	65	
manière	26	
mesure	77	
organisation	28	
personne	95	
date	42	
autre/objet	67	
DEFINITION	33	
definition-organisation	19	
definition-personne	14	
OUI-NON	31	
LISTE	31	

Les différents types de questions :

GF/GRF → questions de type Factuel (GRF : il s'agissait de reformulations de questions de type Factuel. Pour le calcul des scores, nous les avons incluses avec les GF).

GD/GRD → questions de type Définition (GRD : il s'agissait de reformulations de questions de type Définition. Pour le calcul des scores, nous les avons incluses avec les GD).

GB → questions de type Booléen (Oui-Non).

GL → questions de type Liste.

Pour évaluer les questions de types Factuel, Définition et Oui-Non :

MRR : la Moyenne des Réciproques du Rang (Mean reciprocal answer Rank) *cf. paragraphe 3.7.*

Pour évaluer les questions de type Liste :

NIAP : la précision moyenne (Non Interpolated Average Precision) *cf. paragraphe 3.7.*

4.4. Présentation des résultats

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 (Laboratoire Informatique d'Avignon, participant 6) et du LIMSI.

Nous avons présenté les résultats comme suit :

Pour les passages et pour les réponses courtes respectivement 2 tableaux :

- le premier présente pour chaque run, 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 présente pour chaque run, un détail sur les passages (ou réponse) corrects renvoyés en indiquant le nombre de passages (ou réponses) corrects par type de réponse (personne, temps, lieu, organisation...).

Résultats de l'évaluation tâche générale pour les PASSAGES

Identifiant du run	Nombre de questions répondues [464]	Passages corrects (#)	Passages incorrects (#)	MRR sur GF, GD, GRF, GRD, GB	MRR sur GF, GD, GRF, GRD	MRR sur GF, GRF	MRR sur GD, GRD	MRR sur GB	NIAP (précision moyenne) sur GL	Nombre de NIL renvoyés en rang 1	NIL	
											Précision	Rappel
participant 5	464	378	86	0.7	0.71	0.7	0.74	0.67	0.29	4	1	0.8
participant 5	464	344	120	0.64	0.65	0.64	0.71	0.54	0.17	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 4	464	217	247	0.35	0.36	0.33	0.66	0.32	0	19	0	0
participant 2	464	210	254	0.37	0.38	0.37	0.47	0.25	0.09	69	0.01	0.2
participant 2	464	187	277	0.32	0.33	0.34	0.31	0.12	0.06	62	0	0
participant 6	354	162	192	0.29	0.28	0.28	0.37	0.35	0.07	0	0	0
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 3	464	183	281	0.29	0.3	0.28	0.49	0.22	0.02	44	0.02	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

Table 1 : Résultats de l'évaluation des runs pour les passages par participant.

MRR = Moyenne des Réciproques du Rang (Mean Reciprocal answer Rank)

NIAP = Précision moyenne (Non Interpolated Average Precision)

GF/GRF : question générale de type "Factuel"

GD/GRD : question générale de type "Définition"

GB : question générale de type "Booléen" (OUI/NON)

GL : question générale de type "Liste"

Identifiant du run	Passages corrects											
	Définition (#) [33]		factuel (#) [400]							oui non [31]	total	
	org [19]	pers [14]	lieu [65]	man [26]	mes [77]	org [28]	autre/objet [67]	pers [95]	date [42]		# [464]	%
participant 5	16	14	52	20	65	25	57	71	36	22	378	81.46
participant 5	14	12	51	15	58	24	58	62	32	18	344	74.13
participant 4	14	13	38	9	32	17	40	41	23	10	237	51.07
participant 4	14	13	33	9	30	15	36	36	19	12	217	46.76
participant 2	8	11	34	4	32	20	19	46	28	8	210	45.25
participant 2	6	9	35	2	31	18	18	40	24	4	187	40.30
participant 6	9	5	35	0	11	9	7	51	24	11	162	34.91
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 3	10	8	26	5	27	13	20	43	24	8	184	39.65
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

Table 2 : Résultats de l'évaluation des runs, uniquement pour les passages corrects, selon le type de réponse attendu.

Le type de réponse attendu pour toutes les questions est représenté par une abréviation, telle que :

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

Résultats de l'évaluation tâche générale pour les REPONSES COURTES

Identifiant du run	Nombre de questions répondues [464]	Réponses courtes correctes (#)	MRR sur GF, GD, GRF, GRD, GB	MRR sur GF, GD, GRF, GRD	MRR sur GF, GRF	MRR sur GD, GRD	MRR sur GB	NIAP (précision moyenne) sur GL
participant 5	464	312	0.58	0.58	0.57	0.69	0.67	0.71
participant 5	464	259	0.48	0.48	0.46	0.64	0.54	0.36
participant 6	354	130	0.23	0.22	0.22	0.24	0.35	0.02
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 2	464	118	0.2	0.2	0.22	0	0.12	0.03
participant 3	464	106	0.17	0.16	0.16	0.13	0.35	0
participant 3	464	105	0.17	0.16	0.16	0.17	0.22	0
participant 1	333	80	0.13	0.15	0.16	0.01	0.04	0
participant 4	168	51	0.07	0.08	0.07	0.16	0	0
participant 4	195	76	0.12	0.13	0.09	0.58	0	0

Table 3 : Résultats de l'évaluation des runs pour les réponses courtes.

MRR = Moyenne des Réciproques du Rang (Mean Reciprocal answer Rank)

NIAP = Précision moyenne (Non Interpolated Average Precision)

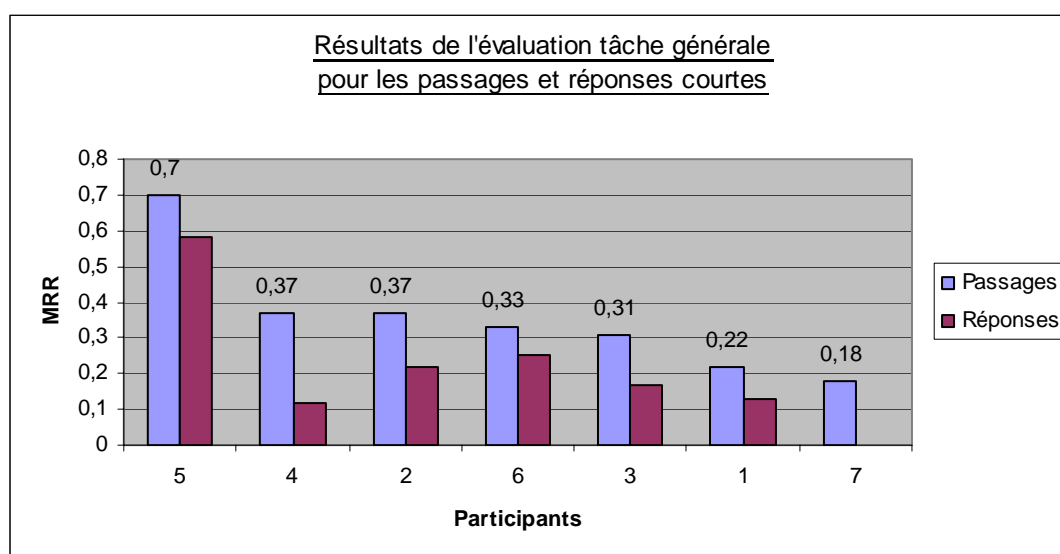
Identifiant du run	Réponses courtes correctes											
	définition (#) [33]		factuel (#) [400]							oui non [31]	total	
	org [19]	pers [14]	lieu [65]	man [26]	mes [77]	org [28]	autre/objet [67]	pers [95]	date [42]		# [464]	%
participant 5	14	14	46	5	63	19	28	67	34	22	312	67.24
participant 5	13	12	42	1	53	16	21	55	28	18	259	55.81
participant 6	7	1	23	0	11	7	4	47	19	11	130	28.01
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 2	0	0	24	1	23	10	5	32	19	4	118	25.43
participant 3	3	4	15	0	19	9	7	26	12	11	106	22.84
participant 3	5	4	12	0	20	9	8	27	12	8	105	22.62
participant 1	1	0	15	1	14	4	3	27	11	4	80	17.24
participant 4	0	8	8	0	10	0	1	14	10	0	51	10.99
participant 4	13	11	14	0	10	0	2	18	8	0	76	16.37

Table 4 : Résultats de l'évaluation des runs, uniquement pour les réponses courtes correctes, selon le type de réponse attendu.

Le type de réponse attendu pour toutes les questions est représenté par une abréviation, telle que:

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

Résultats de l'évaluation tâche générale sous forme de graphe pour les PASSAGES et REPONSES COURTES :



5. Résultats EQueR pour la tâche « médicale »

5.1. Participants

5 groupes participants :

- 3 laboratoires publics :
- Université de Neuchâtel
 - CEA-LIST/LIC2M (uniquement pour les passages)
 - AP/HP-Paris XIII
- 2 institutions privées :
- France Télécom R&D
 - Synapse Développement

5.2. Jugement, évaluation des résultats

Au total :

- 7 runs (ou fichiers-résultat) à évaluer
- 200 questions par run
- 5 réponses possibles par question

Un juge spécialiste de l'équipe du CISMef (Catalogue et Index des Sites Médicaux Francophones) du CHU de Rouen a évalué les résultats au mois de septembre.

Aucun jugement de cohérence n'a été établi.

Les fichiers-résultat de chaque participant ont été complétés. Deux champs ont été remplis en première et deuxième positions :

- **champ 1** : jugement de la réponse courte.

-1 : réponse courte non jugée. (NUL : le participant ne souhaitait pas se faire évaluer pour une réponse courte).
 0 : réponse jugée « correcte »
 1 : réponse jugée « incorrecte »
 2 : réponse jugée « inexacte » (contenant seulement une partie de la réponse, ou trop bruitée par exemple)
 3 : réponse jugée « non supportée par le document »

- **champ 2** : jugement du passage

-1 : passage non jugé (si la taille du passage dépasse 250 caractères par exemple)
 0 : passage jugé « correct »
 1 : passage jugé « incorrect »

5.3. Calcul des scores

Les scores ont été calculés sur la base de **200 questions** réparties comme suit :

FACTUEL	81
DEFINITION	70
OUI-NON	24
LISTE	25

Les différents types de questions :

MF → questions de type Factuel.

MRF → il s'agit de reformulations de questions de type Factuel. Pour le calcul des scores, nous les avons incluses avec les GF.

MD → questions de type Définition.

MRD → il s'agit de reformulations de questions de type Définition. Pour le calcul des scores, nous les avons incluses avec les GD.

MB → questions de type Booléen (Oui-Non).

ML → questions de type Liste.

Pour évaluer les questions de type Factuel, Définition et Oui-Non :

MRR : la Moyenne des Réciproques du Rang (Mean reciprocal answer Rank) *cf. paragraphe 3.7.*

pour évaluer les questions de type Liste :

NIAP : la précision moyenne (Non Interpolated Average Precision) *cf. paragraphe 3.7.*

5.4. Présentation des résultats

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.

Nous avons présenté les résultats comme suit :

Pour les passages ainsi que pour les réponses courtes, un seul tableau :

- il présente pour chaque run, 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.

Résultats de l'évaluation tâche spécialisée pour les PASSAGES

Identifiant du run	Nombre de questions répondues [175]	Passages corrects (#)	Passages incorrects (#)	% passages corrects	MRR sur MF, MD, MRF, MRD, MB	MRR sur MF, MD, MRF, MRD	MRR sur MF, MRF	MRR sur MD, MRD	MRR sur MB	NIAP (précision moyenne) sur ML
participant 4	175	110	65	62.85	0.49	0.51	0.42	0.62	0.37	0.02
participant 4	175	102	73	58.28	0.47	0.48	0.41	0.57	0.41	0.02
participant 2	175	26	149	14.85	0.11	0.13	0.19	0.05	0.04	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

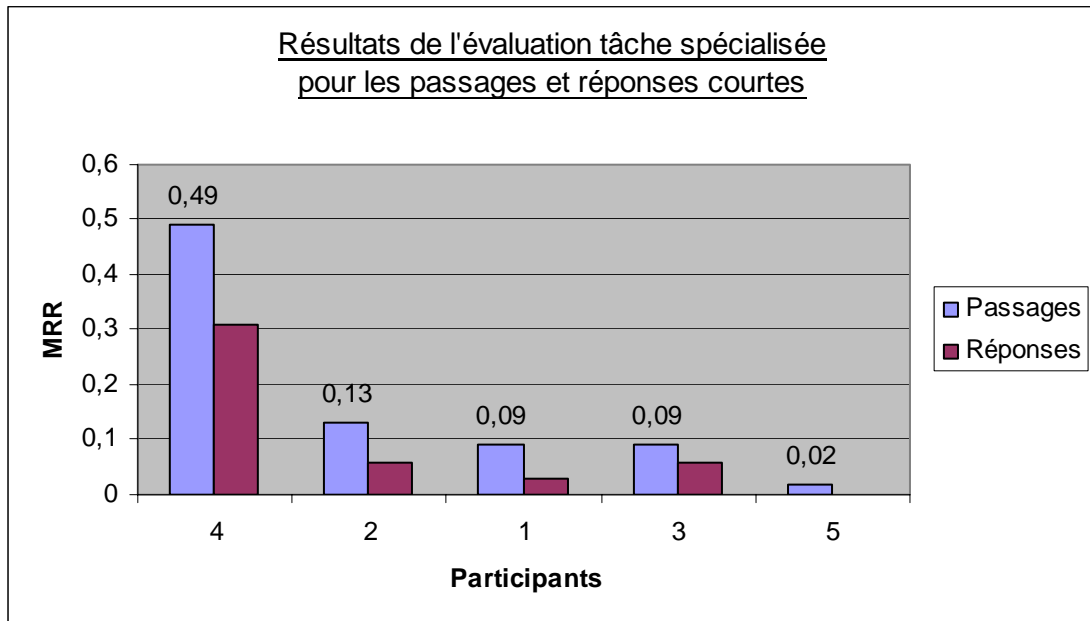
Table 1 : Résultats de l'évaluation des runs pour les passages.

Résultats de l'évaluation tâche spécialisée pour les REPONSES COURTES

Identifiant du run	Nombre de questions répondues [175]	Réponses courtes correctes (#)	% réponses correctes	MRR sur MF, MD, MRF, MRD, MB	MRR sur MF, MD, MRF, MRD	MRR sur MF, MRF	MRR sur MD, MRD	MRR sur MB	NIAP (précision moyenne) sur ML
participant 4	174	71	40.57	0.31	0.3	0.3	0.31	0.37	0
participant 4	175	59	33.71	0.27	0.25	0.25	0.24	0.41	0.01
participant 2	175	8	4.57	0.03	0.03	0.05	0	0.04	0.01
participant 2	175	13	7.42	0.06	0.06	0.11	0	0.08	0
participant 3	35	12	6.85	0.06	0.02	0	0.05	0.33	0
participant 1	117	6	3.42	0.03	0.02	0.03	0.01	0.04	0

Table 2 : Résultats de l'évaluation des runs pour les réponses courtes.

Résultats de l'évaluation tâche spécialisée sous forme de graphe pour les PASSAGES et REPONSES COURTES :



6. Package d'évaluation EQueR

6.1. Pourquoi un package d'évaluation ?

Depuis le départ du projet, il a été convenu avec l'ensemble des participants de constituer un package final d'évaluation.

Ce package d'évaluation comprendra l'ensemble des données relatives au projet et fournies aux participants lors de la campagne :

- l'ensemble des spécifications de la campagne
- les corpus :
 - les deux corpus de textes (tâche générale et tâche médicale)
 - les deux corpus de questions (500 questions pour la tâche générale et 200 questions pour la tâche médicale)
 - pour chaque question des deux corpus les 100 premiers identifiants fournis par le moteur de recherche Pertimm
 - les deux sous-corpus Pertimm créés à partir des identifiants de documents renvoyés par le moteur de recherche
 - l'ensemble des résultats fournis aux participants
- les outils :
 - logiciel d'aide à l'évaluation des résultats dans le cadre d'une évaluation de systèmes de question-réponse (avec documentation détaillée)
 - logiciel d'évaluation automatique (à réaliser)

Il permettra ainsi à n'importe quel industriel ou académique n'ayant pas participé à la campagne par exemple, de faire tourner son système exactement dans les mêmes conditions. Il pourra ainsi, s'il le souhaite, comparer ses résultats avec les propres résultats de la campagne (tout en relativisant ses résultats étant donné les avancées qui seront réalisées dans ce domaine).

6.2. Contenu du package d'évaluation EQueR

Lors de la réunion de fin de projet EQueR, avec le soutien de l'ensemble des participants, ELDA se propose de réaliser 4 versions du « package final d'évaluation » :

- une 1^{ère} version comportant exactement les mêmes données afin de pouvoir reproduire à l'identique la campagne EQueR04.
- une seconde, comportant les mêmes données que dans EQueR04, mais seulement avec un logiciel d'évaluation automatique (plus besoin de juger les résultats de façon manuelle), créé à partir des réponses renvoyées lors de la campagne EQueR04.
- une troisième version comportant un corpus de textes « propre », et des résultats comparatifs des 3 meilleurs systèmes.
- une quatrième composée d'un corpus propre et d'un logiciel d'évaluation automatique

La première version du package final sera très bientôt mise à disposition par ELDA.

7. Conclusion

En conclusion, ce rapport 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.

Dans ce rapport, les résultats sont fournis de façon brute. Une analyse plus détaillée sera très bientôt disponible dans un article spécifique.

Enfin, EQueR s'europanise avec la campagne d'évaluation CLEF (Cross Language Evaluation Forum : www.clef-campaign.org) 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.