EQueR : Evaluation de systèmes de Question-Réponse

Brigitte Grau, Anne Vilnat, Christelle Ayache

Chapitre 6

EQueR : Evaluation de systèmes de Question-Réponse

Brigitte Grau [1,2], Anne Vilnat [1] et Christelle Ayache [3].
[1] LIMSI, CNRS – [2] ENSIIE – [3] ELDA

6.1 Introduction

Un système de question-réponse (QR) permet de poser une question en langue naturelle et se donne pour but d'extraire la réponse, quand elle y figure, d'un ensemble de textes. En cela, ces systèmes traitent de recherche d'informations précises, ou factuelles, c'est-à-dire telles qu'elles puissent être spécifiées en une seule question et dont la réponse tient en peu de mots. Typiquement, ce sont des réponses fournissant des dates, ou des noms de personnalités comme par exemple « Quand est mort Henri IV ? » ou « Qui a tué Henri IV ? », mais aussi donnant des caractéristiques sur des entités ou des événements moins faciles à typer, par exemple « Comment est mort Henri IV ? » ou « De quelle couleur est le drapeau français ? ».

La recherche en question-réponse connaît un essor important depuis quelques années. On peut le constater au travers des conférences d'évaluation en recherche d'information qui proposent toutes une tâche question-réponse dorénavant, mais aussi par les conférences qui sont nombreuses à proposer ce thème dans leurs appels à propositions d'articles, et enfin via l'existence d'ateliers spécifiques à ce thème dans les grandes conférences de recherche d'information (RI) mais aussi de traitement de la langue et d'intelligence artificielle. Cela est sans doute dû à une conjonction de facteurs : 1) l'inadéquation des systèmes de recherche d'information qui proposent systématiquement une liste de documents face à différents besoins utilisateur. En effet, lorsque l'utilisateur recherche une information précise, il semble plus pertinent à la fois de pouvoir poser sa question en langue naturelle, ce qui lui permet de mieux préciser sa requête, et de ne retourner en résultat qu'un court passage contenant le réponse cherchée ; 2) l'arrivée à maturité d'un certain nombre de techniques en RI et en traitement de la langue qui permettent d'en envisager une application à large échelle, sans restriction sur le domaine traité ; 3) la possibilité de définir un cadre d'évaluation des systèmes.

Après une présentation rapide des problèmes soulevés en QR et des types de solutions qui y sont apportés, nous donnerons un aperçu de l'évaluation en QR afin de positionner la campagne EQueR.

6.1.1 Les systèmes de question-réponse

Les systèmes de QR se démarquent des systèmes de RI classique par leur entrée. En effet, le fait de poser une question permet à l'utilisateur d'expliciter son besoin alors qu'en RI les utilisateurs donnent une requête en entrée du moteur de recherche, et doivent ainsi transformer eux-mêmes leur besoin en un ensemble de termes devant figurer dans les documents. C'est pourquoi tous les systèmes de QR réalisent une analyse la plus fine possible des questions, afin d'en inférer un maximum de contraintes sur la réponse.

La recherche d'une réponse à une question peut être définie comme un problème d'appariement de la question formulée de manière déclarative, comportant un élément à instancier, la réponse, avec un passage réponse, c'est-à-dire une ou plusieurs phrases. Cet appariement repose sur le fait de disposer de passages pertinents et d'être capable de mettre en relation les éléments de la question avec ces passages, compte tenu de l'importante variabilité linguistique susceptible d'exister entre les deux formulations. La variabilité [GRA 04] peut provenir de différences lexicales, avec l'emploi de synonymes, d'hyperonymes ou d'hyponymes et de différences syntaxiques avec des paraphrases partielles ou complètes de la question. Par ailleurs, pour être considérée comme correcte, une réponse se doit d'être justifiée : l'entité réponse seule est donc insuffisante, il faut aussi être en mesure de présenter un ou plusieurs passages justifiant l'extrait choisi.

Les systèmes de QR reposent sur trois modules principaux : l'analyse des questions, la sélection et l'annotation de passages pertinents et l'extraction de la réponse (cf. **Erreur ! Source du renvoi introuvable.**). L'une des principales tâches de l'analyse de la question consiste à typer la réponse attendue. Ces types vont des types d'entités nommés tels qu'ils ont été définis dans le cadre des évaluations MUC [GRI 95] à des types beaucoup plus fins et spécifiques, qui peuvent aller jusqu'à l'ensemble des types de WordNet [FEL 98] comme dans [HAR 00], mais correspondent souvent à un ensemble délimité par les concepteurs des systèmes selon les techniques mises en oeuvre pour les retrouver dans les textes [HOV 01], [PRA 00]. Hormis la reconnaissance du type de réponse attendu, les caractéristiques retenues par l'analyse des questions diffèrent parmi les approches existantes. Celles-ci sont de nature lexicale, avec la reconnaissance de termes clés qui seront recherchés tels quels ou sous forme de variantes, et de nature syntaxique avec l'extraction de relations syntaxiques entre termes, ou même la construction de

l'arbre syntaxique de la question, qui seront appariés à tout ou partie des passages réponses. Plus rares sont les approches sémantiques.

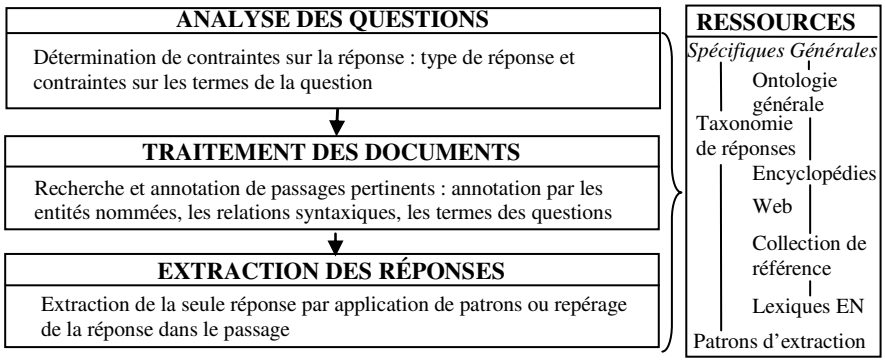

Figure 1. *Composants d'un système de question-réponse*

Le deuxième module recherche des passages pertinents. Là aussi, il existe de nombreuses approches, allant de l'application de techniques de RI classiques à la construction d'index spécifiques. La sortie de ce module consiste en différents passages annotés. Cette annotation peut avoir lieu après sélection des passages [FER 01] ou avant la recherche pour être prise en compte lors de l'indexation [LAU 05].

Le dernier module extrait le groupe de mots constituant la réponse. Les approches consistent généralement à pondérer les différents candidats, ceux-ci ayant été sélectionnés par l'application des contraintes issues de l'analyse des questions.

Tous les systèmes procèdent de cette architecture, avec éventuellement des rétroactions entre modules. Ils mêlent en général des approches numériques et symboliques pour les différents composants. Signalons cependant que l'un des meilleurs systèmes ([HAR 00], [MOL 02]) utilise de manière poussée des processus de TAL et de la déduction logique. Une autre approche très efficace ([SOU 01], [SOU 02]) est fondée sur l'utilisation intensive de patrons d'extraction. Enfin, l'une des stratégies souvent mise en œuvre consiste à utiliser le Web, soit exclusivement pour y chercher des réponses [BRI 01], soit pour l'exploiter conjointement avec une base de textes ([CHA 03], [CHU 02], [CLA 01], [MAG 02]).

6.1.2 Evaluation des systèmes de QR

Le succès que remporte la tâche question-réponse dans les évaluations et sa complexité toujours croissante sont une preuve de la vitalité des recherches effectuées. Le principe de l'évaluation consiste à poser un jeu de questions aux

participants, qui doivent retourner leurs propositions de réponses extraites de documents de la collection dont ils disposent. L'évaluation de leurs soumissions est réalisée par des juges humains. Les réponses doivent être accompagnées du document qui justifie la réponse. Ainsi une réponse, exacte pour la valeur, mais qui n'est pas justifiée par le document proposé ne sera pas considérée comme une réponse correcte.

Nous allons maintenant présenter les campagnes actuelles puis les différents points qui caractérisent la campagne d'évaluation EQueR.

TREC[1], avec TREC8 (1999), fut la première conférence proposant une évaluation en question-réponse. Dès la deuxième année, le succès de la tâche s'est affirmé. La tâche a évolué chaque année pour arriver en 2002 (TREC11) à la proposition d'une seule réponse, et uniquement la réponse, à chaque question, réponse recherchée dans un corpus de 3 gigaoctets composé d'articles de journaux. Actuellement, les questions sont regroupées en séries portant sur des entités ou des événements permettant de définir un contexte.

En Europe, la campagne CLEF[2] a pour but l'évaluation de systèmes en recherche d'information qu'ils soient monolingues de langue européenne ou multilingues. La campagne CLEF a intégré en 2003 une piste question-réponse. La différence avec la campagne TREC vient des langues traitées et de l'introduction de nombreuses pistes multilingues. Les tâches monolingues comportent 200 questions auxquelles les systèmes doivent donner la réponse uniquement et les corpus sont constitués d'articles de journaux de tailles allant de 200 à 540 mégaoctets. Pour les tâches multilingues, il s'agit de rechercher les réponses dans une langue cible à des questions posées dans une langue source différente. Le français en tâche monolingue a été introduit en 2004. Les réponses sont principalement de type entité nommée, au sens large, définition, ou n'existent pas dans le corpus.

La campagne NTCIR[3] a pour but l'évaluation en recherche d'information pour les langues asiatiques, i.e. japonais et chinois, et propose des tâches monolingues et multilingues également. Les réponses, quand elles existent dans le corpus, sont uniquement de type entité nommée.

La campagne EQueR ([AYA 05], [AYA 06]) en 2004 a constitué la première évaluation de systèmes monolingues français et a proposé deux types de tâche : l'une en domaine ouvert et l'autre sur un domaine de spécialité, en l'occurrence la médecine. Le but était de voir si les méthodes de résolution sont les mêmes ou du

[1] http://trec.nist.gov
[2] http://www.clef-campaign.org
[3] http://research.nii.ac.jp/ntcir/index-en.html

moins de même nature, et si la notion de question factuelle a un sens pour un domaine de spécialité. Pour la tâche générale, des questions polaires (OUI/NON) ont été créées. Les systèmes pouvaient renvoyer jusqu'à 5 réponses, réponses longues et réponses exactes. Le but était de constituer un corpus d'étude le plus complet possible, en récoltant des passages réponses, corrects et incorrects, et des réponses seules, correctes, incorrectes ou non justifiées.

6.2 Présentation de la campagne EQueR

Deux tâches de recherche automatique de réponses ont été proposées : une tâche générique sur une collection hétérogène de textes – en majorité 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. Les participants ont reçu des jeux de questions différents pour les deux tâches. Les questions ont été élaborées en fonction des différents types de réponses attendues : questions de types « factuel », « définition », « liste fermée d'éléments » ou encore questions de type « oui/non ». Pour certaines questions, aucune réponse n'était disponible dans les collections textuelles utilisées. Les évaluateurs humains vérifiaient puis jugeaient la réponse exacte ainsi que le passage renvoyé par un système participant et ce, pour chaque question. Vérifier une réponse signifiait vérifier qu'elle était exacte et justifiée par un document.

6.2.1 Corpus de textes

Les participants ont eu accès aux collections de documents quelques mois avant le test d'évaluation. Les textes fournis étaient balisés simplement au moyen d'un identifiant de document, de titre et de paragraphe, et codés en ISO-Latin-1 (ISO-8859-1), comme le montre l'exemple Figure 2 extrait du corpus au format EQueR. Deux collections de textes ont été élaborées : une collection pour la tâche générale et une collection pour la tâche médicale.

La collection générale, d'une taille de 1,5 Go environ, est 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, ainsi traités selon différents points de vue et types de texte : articles d'actualité, articles de fond, dépêches, rapports. La collection de textes de spécialité, d'une taille de 140 Mo environ, est 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.

```
<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>
```

Figure 2. Extrait d'un article du journal « Le Monde » au format EQueR

6.2.2 Corpus de questions

Cinq types de questions ont été proposés aux systèmes participants : les questions de type « factuel simple », les questions de type « définition », les questions de type « liste », les questions de type « oui-non » et les questions sans réponse possible dans les collections de documents (questions de type « NIL »).

Questions de type « factuel simple » (F) : Ces questions attendent en réponse un fait simple correspondant à l'un des 6 sous-types définis pour l'évaluation EQueR (Tableau 1). Les questions demandant une réponse subjective (« Quel est le principal monument de Paris ? ») ou les questions dites « emboîtées » (« Où se trouve l'édifice le plus haut d'Europe ? ») n'ont pas été proposées.

Type	Sous-type	Exemple de questions
Factuel	*Personne*	Qui a écrit "La bicyclette bleue" ? (*Régine Desforges)*
	Localisation	Quelle est la capitale de la Tchétchénie ? (*Grozny)*
	Organisation	Quelle organisation veille sur les droits de l'homme ? (*l'ONU)*
	Date	Quand Staline est-il mort ? (*5 mars 1953)*
	Mesure	Combien de films Ingmar Bergman a-t-il réalisé ? (*cinquante-trois)*
	Objet / Autre	Quel est le nom actuel du Ceylan ? *Sri Lanka*

Tableau 1. *Sous-types des questions de type « factuel simple »*

Questions de type « définition » (D) : Ces questions attendent en réponse une « définition » et ont été formulées de manière à attendre une réponse courte, présente dans un document. Deux sous-types de questions « définition » ont été proposés :

– Personne : « Qui est Jacques Chirac ? » (P*résident français*) ;

– Organisation : « Qu'est-ce que l'OTAN ? » (*Organisation du Traité de l'Atlantique Nord*).

Questions de type « liste » (L) : Ces questions attendent un nombre bien précis de réponses (nombre indiqué dans la question). Cependant pour ce type de questions, les systèmes pouvaient renvoyer jusqu'à 20 réponses par questions.

Questions de type « oui/non » (B) : Ces questions attendent en réponse « oui » ou « non » accompagnée d'un passage justifiant cette réponse. Pour ces questions, seule la première ligne-réponse a été prise en compte pour l'évaluation.

Questions « NIL » : Quelques questions sans réponse possible dans les corpus ont été introduites au sein des questions de type « général ». Dans ce cas, le système devait renvoyer « NIL » pour que sa réponse soit jugée « correcte » (« NIL » signifiant « il n'y a pas de réponse dans le corpus »).

Corpus / **Type Question**	**Général [500]**	**Médical [200]**
Factuel	407	81
Définition	32	70
Liste	31	25
Oui / Non	30	24

Tableau *2.* *Répartition du nombre de questions par type*

Un jeu de questions spécifiques a été fourni pour chacune des deux tâches, les questions ayant été catégorisées selon les mêmes sous-classes (hormis les questions NIL que nous ne retrouvions que dans le corpus de questions « général »). Dans l'exemple ci-dessous le codage « GF18 » indique que la question n°18 attend une réponse de type factuel simple (F) et s'applique à la tâche générale (G).

EXEMPLE : GF18 Où est né Jacques Chirac ?

Plusieurs sources et plusieurs modes de génération de questions ont été utilisés. Une partie des questions a été engendrée à partir de mots clés extraits de certains articles et de certaines dépêches de presse. Une autre partie a été créée par un groupe d'utilisateurs potentiels en fonction des sous-types manquants. Pour valider chaque

question, il a fallu vérifier la présence d'au moins une bonne réponse par question dans le corpus. 500 questions ont été créées pour la tâche générale, 200 questions pour la tâche médicale. La répartition du nombre de questions par type a été contrôlée pour les deux tâches (cf. Tableau 2). De plus, les participants qui le désiraient ont pu partir d'un ensemble de textes associés à chaque question par l'organisateur et sélectionnés en utilisant le moteur de recherche PERTIMM[4].

6.2.3 Évaluation EQueR

6.2.3.1 Fichiers de soumission

Les participants avaient le choix de se faire évaluer ou non sur les réponses courtes. Les passages étaient, quant à eux, systématiquement évalués. Ils pouvaient soumettre au maximum deux fichiers de soumission par tâche. Pour les questions de type « factuel » et « définition », les participants pouvaient renvoyer jusqu'à 5 réponses par question. Ces réponses ordonnées devaient être présentées les unes en dessous des autres dans l'ordre des questions. Pour les questions de type « Liste », 20 lignes réponses étaient autorisées, une seule réponse pour les questions de type « oui-non ». Chaque ligne-réponse dans un fichier de soumission comprenait 5 champs séparés par une tabulation :

Identifiant de question	Identifiant du participant	Identifiant du document	Réponse exacte	Passage
GF1	elda04g1	LEMONDE94-000001	Paris	aura lieu à Paris ; la capitale de la France va accueillir

– **Identifiant de question :** tel qu'il était fourni en entrée dans le jeu de test.

– **Identifiant du participant :** il indiquait le nom du participant (4 caractères), l'année, la tâche (G pour « Générale », M pour « Médicale ») et le numéro du fichier de soumission (1 ou 2).

– **Identifiant du document :** tel qu'il était fourni dans les corpus, indiqué par la balise <DOCID>. S'il s'agissait d'une question sans réponse, les systèmes devaient renvoyer « NIL » à cet emplacement.

– **Réponse exacte :** ce champ pouvait contenir « NUL » si le participant ne souhaitait pas se faire évaluer sur les réponses courtes ou pouvait rester vide s'il s'agissait d'une question sans réponse.

– **Passage :** une contrainte a été mise en place au départ du projet, les passages ne devaient pas dépasser 250 caractères pour pouvoir être évalués.

[4] http://www.pertimm.fr

6.2.3.2 Jugement humain des résultats

S'agissant d'une évaluation sur le français, il était important que les fichiers fussent jugés par des Français natifs. 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 était du ressort de l'équipe d'évaluation. La règle fondamentale appliquée lors de l'évaluation était : « une réponse est considérée correcte si et seulement si elle est justifiée par le document qui lui est associé ».

Pour l'évaluation des réponses courtes, quatre jugements étaient possibles :

– CORRECT : la réponse est juste et précise (sans aucune information obsolète) et est justifiée par le document associé ;

– INCORRECT : la réponse n'est pas juste, elle ne correspond pas du tout à la réponse attendue ;

– INEXACT : la réponse exacte ou une partie de la réponse est présente mais la réponse n'est pas assez précise (soit il manque une partie de l'information, soit la réponse exacte est noyée dans trop d'informations) ;

– NON SUPPORTÉ (par le document) : la réponse est juste et précise mais le document associé ne justifie pas du tout la réponse renvoyée (la réponse n'est pas présente dans le document, le document parle d'un tout autre sujet…).

Pour l'évaluation des passages, seuls deux jugements étaient possibles :

– CORRECT : le passage contient la réponse juste et précise et est justifié par le document associé ;

– INCORRECT : la réponse n'est pas présente dans le passage, elle ne correspond pas du tout à la réponse attendue.

6.2.3.3 Mesures adoptées

Deux métriques d'évaluation standards ont été adoptées : la Moyenne des Réciproques du Rang (MRR) et la Précision moyenne (NIAP). La Moyenne des Réciproques des Rangs a été calculée pour les questions de type « factuel », « définition » et « oui-non » ; la Précision moyenne pour les questions de type « liste ».

La Moyenne des Réciproques du Rang (MRR, cf. Figure 3) tient compte de la première bonne réponse trouvée et de son rang (métrique TREC). Si une réponse est trouvée plusieurs fois, elle n'est comptée qu'une seule fois. La Précision moyenne (NIAP, cf. Figure 4) 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.

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

Figure 3. *Formule MRR*

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

avec :

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

et :

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

Figure 4. *Formule NIAP*

6.2.4 Résultats de l'évaluation

6.2.4.1 Tâche générale

Sept groupes ont participé à l'évaluation EQueR pour la tâche générale :

– quatre laboratoires publics : le LIMSI-CNRS, l'Université de Neuchâtel, le Laboratoire d'Informatique d'Avignon et le CEA-LIST ;

– trois institutions privées : France Télécom R&D, Synapse Développement et Sinequa.

Au total, 12 fichiers-résultats ont été évalués. 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 également réalisé une évaluation croisée sur deux fichiers-résultats (chacun a évalué 2 fichiers déjà évalués par l'autre juge), puis la cohérence de leurs jugements respectifs a été calculée. Un taux de désaccord inférieur à 5% ayant été constaté, leurs jugements ont pu, de ce fait, être validés. Lors de l'évaluation des fichiers-résultats par les

juges, deux champs s'ajoutaient automatiquement aux fichier-résultats bruts envoyés par les participants (cf. Tableau 3). Ces deux champs apparaissaient ensuite dans les fichiers en première et deuxième position :

– champ 1 : le jugement de la réponse courte représenté par un chiffre (-1 à 3) ;

– champ 2 : le jugement du passage représenté par un chiffre (-1 à 1).

Valeur du jugement	Signification
-1	réponse ou passage non jugé
0	réponse ou passage « correct »
1	réponse ou passage « incorrect »
2	réponse « inexacte »
3	réponse « non supportée » par le document

Tableau *3. Correspondance Chiffre-Jugement*

Sur les 500 questions du corpus général envoyées aux participants, 5 d'entre elles comportaient des erreurs (date, incompréhension, orthographe…). Les scores ont donc été calculés sur la base de 495 questions réparties comme suit :

– 400 questions « Factuelles » ;

– 33 questions « Définition » ;

– 31 questions « Oui-Non » ;

– 31 questions « Liste ».

Les trois systèmes de Question-Réponse ayant obtenu les meilleurs résultats pour la tâche générale sont :

– pour les passages : les systèmes de Synapse Développement, de Sinequa et du LIMSI ;

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

Les résultats ont été fournis sous forme de deux tableaux distincts. Le premier tableau présentait le nombre de questions traitées, le nombre de passages (ou réponses) corrects renvoyés, ainsi que les scores obtenus pour chaque type de questions et de combinaison pour chaque fichier de soumission. Le second tableau représentait un détail sur les passages (ou réponses) corrects renvoyés et indiquait le nombre de passages (ou réponses) corrects par type de réponse (personne, temps, lieu, organisation…) pour chaque fichier de soumission.

Par souci de clarté, nous présenterons dans ce chapitre les deux graphiques explicitant les résultats pour la tâche générale.

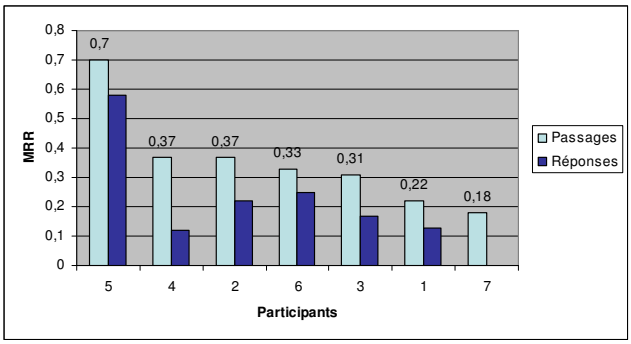

Figure 5. *Résultats pour la tâche générale : passages et réponses courtes*

La **Figure 5** donne les résultats globaux, hors questions listes ; la Figure 6 détaille les résultats sur les passages pour les trois types de question. Il est intéressant de constater que les différents systèmes ne se comportent pas globalement de la même manière selon le type de question.

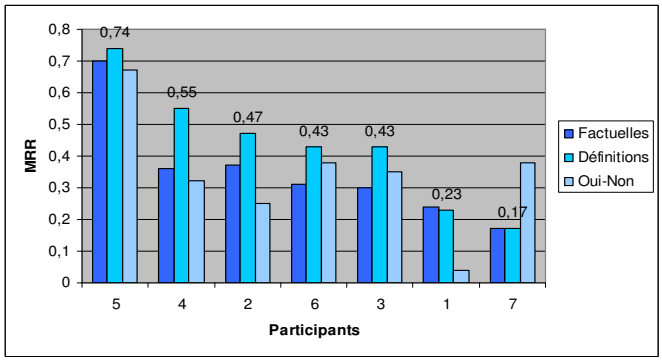

Figure 6. *Résultats pour la tâche générale : Factuel, Définition, Oui-Non*

6.2.4.2 Tâche « Médicale »

Cinq groupes ont participé à l'évaluation EQueR pour la tâche médicale :

– trois laboratoires publics : l'Université de Neuchâtel, le CEA-LIST et AP/HP-Paris XIII ;

– deux institutions privées : France Télécom R&D et Synapse Développement.

Au total, 7 fichiers-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. Aucun jugement de cohérence n'a été établi.

Les trois systèmes de Question-Réponse ayant obtenu les meilleurs résultats pour la tâche médicale 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 : les systèmes de Synapse Développement, et *ex-aequo* les systèmes de AP/HP-Paris XIII et de l'Université de Neuchâtel.

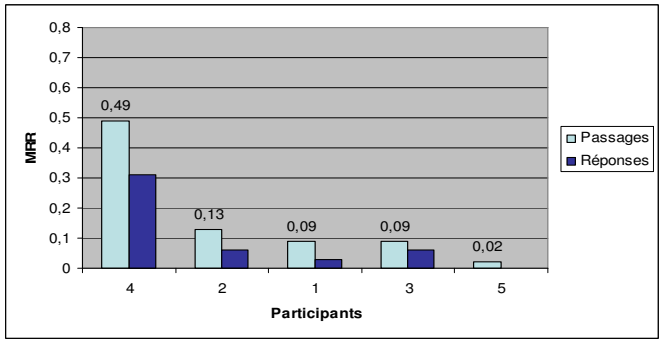

Figure 7. Résultats pour la tâche Médicale : passages et réponses courtes

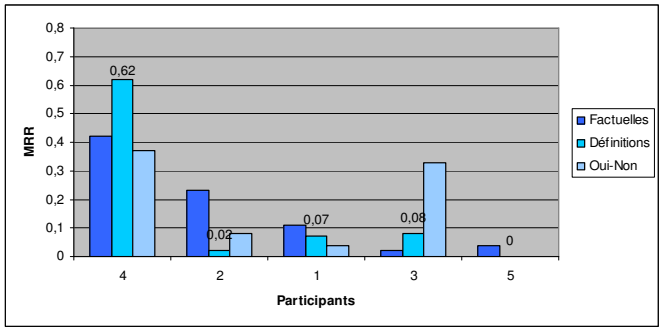

Figure 8. *Résultats pour la tâche Médicale : Factuel, Définition, Oui-Non*

Les résultats ont été présentés sous forme d'un tableau unique indiquant 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 pour

chaque fichier de soumission. Comme pour la tâche générale, nous présentons dans ce chapitre les deux mêmes graphiques explicitant les résultats pour la tâche médicale (cf Figure 7 et Figure 8).

6.2.4.3 Analyse des résultats et mises en perspectives avec les autres campagnes

Dans un premier temps, nous avons pu constater que les résultats pour la tâche générale étaient meilleurs que pour la tâche spécialisée. Ils s'échelonnent entre 0,7 et 0,18 (MRR) pour la tâche générale et entre 0,49 et 0,02 (MRR) pour la tâche médicale. Dans ce dernier cas, la faiblesse des résultats peut s'expliquer par la terminologie spécifique utilisée dans les textes médicaux. De plus, l'ensemble des systèmes a obtenu de meilleurs résultats lors de l'évaluation des passages que lors de l'évaluation des réponses courtes. En effet, il est plus difficile pour un système d'extraire une réponse courte exacte et précise qu'un passage un peu plus long dans lequel il est finalement plus probable de trouver 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 à ceux des autres participants, et ce, pour les deux tâches, générale et spécialisée. Les différentes présentations du système de Synapse Développement par l'équipe elle-même ont montré que les nombreux lexiques associés à leur système, ainsi que les nombreuses sous-catégories spécifiques utilisées faisaient toute la différence avec les autres systèmes.

Concernant la tâche générale (500 questions), 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 des types de questions et de réponses sur lesquelles leur système avait été le plus (ou le moins) 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 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'évaluation 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.

6.2.4.4 Apports de la campagne EQueR

Cette campagne d'évaluation a été un véritable succès de par la participation et l'intérêt croissant d'une très large majorité des acteurs académiques et industriels du domaine. 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 questions, les questions de type « oui/non », qui ont suscité beaucoup d'intérêt de la part des participants. En proposant une tâche Question-Réponse dans un domaine spécialisé, EQueR a attiré d'autres participants intéressés plus particulièrement par le domaine médical.

EQueR a également permis de développer un outil d'aide à l'évaluation, QASTLE[5] (Question Answering TooL for Evaluation), permettant aux juges de visualiser sur un même écran, la question ainsi que la réponse, le passage et le document à évaluer. Il met également en valeur dans le document l'ensemble des mots de la question. Cet outil a depuis été adapté et utilisé pour l'évaluation des résultats de CLEF pour le français.

EQueR s'européanise avec la campagne d'évaluation CLEF qui, depuis trois années maintenant, offre une tâche spécialisée pour l'évaluation des systèmes de Question-Réponse en Europe. ELDA[6], l'agence pour l'évaluation et la distribution de ressources linguistiques, joue le rôle de coordinateur pour le français dans la campagne européenne CLEF ainsi que celui de fournisseur de données 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. La campagne européenne CLEF est devenue, en quelque sorte, l'avenir d'une campagne très enrichissante comme EQueR en France.

---

[5] QASTLE est disponible en libre téléchargement : http://www.elda.org/qastle
[6] http://www.elda.org

6.3 Présentation des approches des participants

Nous allons ici dégager les principales caractéristiques des systèmes participant à cette campagne, que ce soit dans la tâche générale ou dans la tâche médicale. Par rapport aux grandes étapes mentionnées au schéma Figure 1, nous allons situer les approches des sept groupes qui ont participé à la tâche générale et des cinq groupes qui ont participé à la tâche médicale, sachant que parmi eux seul l'AP-HP, participant conjointement avec Paris XIII, n'a pas participé à la tâche générale.

6.3.1 Analyse des questions

La finalité générale de cette étape est identique pour la plupart des systèmes : déterminer les éléments importants de la question permettant la recherche des passages pertinents, et fournir des informations sur le type de réponse attendue, en rattachant la question à une classe plus ou moins précise. Toutefois, suivant les systèmes, cette tâche inclut des techniques de TAL plus ou moins développées.

L'Université de Neufchâtel [PER 05] se contente ainsi de passer un étiqueteur, qui donne les caractéristiques morpho-syntaxiques des mots de la question, ainsi que le lien vers un concept connu et une éventuelle entité nommée. Le but est de catégoriser les questions en six classes d'une part et d'autre part de construire une requête. Le CEA-LIST [BAL 05] analyse la question en lui appliquant une série de patrons morpho-syntaxiques. L'originalité de leur approche repose sur la constitution de cet ensemble de patrons à partir d'un corpus de questions, en alignant les paires les plus proches. Reconnaître la question consiste à identifier son type. La technique mise en œuvre ne reposant pas sur des connaissances linguistiques particulières, elle a été appliquée aussi bien au corpus général qu'au corpus médical.

Sinequa [BLA 05] procède à une analyse des questions via des règles linguistiques pour déterminer le type des questions et en extraire les parties essentielles. 166 types de questions sont ainsi identifiés. De ces types vont découler les types de réponses attendues. Lorsque l'analyse aboutit, elle conduit à une reformulation sous la forme affirmative qui sera utilisée pour chercher les réponses. France-Telecom utilise des règles de reformulation des questions afin de regrouper des formes syntaxiques différentes correspondant à des questions similaires. Celles-ci sont ensuite transformées en patrons de réponses attendues, comme le fait également Sinequa. Le système développé par le LIA [GIL 05] effectue un étiquetage hiérarchique des questions après une étape d'uniformisation à base de règles et de lexiques permettant de regrouper les variantes de questions, et ainsi de réduire le nombre de règles d'étiquetage écrites manuellement. Le but de cet étiquetage est d'apparier la question avec des entités-réponses attendues.

Le système développé par le LIMSI [GRA 05] procède à une analyse syntaxique de la question afin d'en déduire les éléments importants, utiles pour formuler la requête, ainsi que pour sélectionner les réponses. Cette analyse a également pour but de déterminer à quelle classe de question appartient celle qui est analysée. Le système développé par Synapse [LAU 05] est celui qui fait appel aux plus grand nombre d'informations linguistiques. L'analyse se fait en utilisant l'analyseur syntaxique développé par Synapse, incluant un outil de correction orthographique. Une base de concepts est utilisée pour ensuite faire une analyse conceptuelle et extraire les mots-clés de la question ainsi que son type.

L'AP-HP et Paris XIII [DEL 05] se sont intéressés uniquement à la tâche médicale. Leur analyse de la question permet elle aussi de déterminer la catégorie de la question, le type d'entité nommée attendue s'il y a lieu, ainsi que l'ensemble des mots-clés. La spécialisation du domaine leur permet alors d'utiliser le MeSH pour reconnaître des mots-clés spécifiques du domaine.

6.3.2 Traitement des documents

Plusieurs participants ont eu recours aux documents proposés, trouvés à l'aide du moteur de recherche PERTIMM (LIA, CEA-LIST), d'autres ont utilisés leur propre moteur, mettant ainsi en jeu une stratégie propre.

C'est le cas de l'Université de Neufchâtel. Un prétraitement des documents supprime les mots outils puis un enracineur élimine les suffixes flexionnels des mots. Un moteur de recherche probabiliste (modèle Okapi) a ensuite été utilisé afin de ramener les documents les plus pertinents. En se limitant aux 10 premiers documents, ils ont évalué avoir la réponse dans 80% des cas.

Sinequa a utilisé leur moteur de recherche Intuition pour extraire des documents. Plus l'analyse précédemment décrite est précise, plus la requête permet de retrouver des documents pertinents bien classés. En se limitant aux 5 premiers documents retournés par Intuition, Sinequa a évalué disposer de 65% de documents pertinents lorsque l'analyse de la question a permis de constituer une requête évoluée. France Telecom a recherché les documents en effectuant une requête booléenne (moteur open source Swish-e) sur une double indexation des documents : mots lemmatisés et mots laissés fléchis. Après avoir lemmatisé les documents, le LIMSI-CNRS a utilisé le moteur *Lucene* pour rechercher les documents les plus pertinents. La requête est exprimée sous forme booléenne en relâchant éventuellement les contraintes de façon à recueillir 200 documents.

Synapse procède certainement au traitement le plus approfondi sur les documents avant de procéder à la recherche de passages pertinents. Ainsi, les

documents sont découpés, corrigés et analysés à la fois syntaxiquement et conceptuellement par Cordial. Les anaphores sont résolues, et les entités nommées sont étiquetées. Ils sont ensuite indexés à partir de ces entités nommées, des concepts reconnus, des têtes de dérivation identifiées et des types de question-réponse. Les synonymes et les converses des mots jugés significatifs lors de l'analyse de la question sont alors recherchés dans ces index, de façon à extraire les blocs les plus pertinents. La recherche s'est faite sur les 40 premiers documents obtenus par le moteur de recherche propre de QRISTAL (un autre test a été effectué par comparaison sur les 40 premiers documents produits par PERTIMM).

L'AP-HP a fait subir aux documents un pré-traitement très important en le découpant en phrases avant de l'étiqueter morpho-syntaxiquement et de créer une base de données autour de chacun des prédicats ou mots pleins reconnus dans ces phrases. Les entités nommées sont également identifiées à partir de patrons et sont stockées dans la base de données. C'est l'interrogation de cette base de données qui sert alors de moteur de recherche. Le choix de la meilleure phrase se fait ensuite en tenant compte de l'importance de la thématique de la question dans le document duquel la phrase est extraite, ainsi que de la présence d'entités nommées ou de vocabulaire précis. Ce système tire ainsi parti de connaissances plus spécifiques venant du domaine médical.

6.3.3 Classement des extraits et choix des réponses

Après avoir obtenu des documents pertinents, les différents systèmes ont procédé à un classement des extraits constituant ces documents. Les principes sur lesquels reposent ces classements sont assez proches les uns des autres. Les différents passages sont assortis d'un score qui est ensuite utilisé pour retenir les meilleurs passages.

Pour le LIA, ce score est obtenu à partir des scores des entités faisant partie de la requête, faisant appel à la distance de chaque entité avec les autres. Une phrase est évaluée en fonction du score des entités qu'elle contient. Un passage est ensuite constitué de trois phrases, se voyant attribué le score de la phrase centrale. Ce score permet de classer les meilleurs passages. La réponse est cherchée dans les passages les mieux notés, en tenant compte de l'entité attendue, telle qu'elle a été définie lors de l'analyse de la question, et de bases de connaissances complémentaires qui viennent éventuellement confirmer le choix de la bonne entité quand il s'agit de couples question-réponse « fréquents (comme les noms des capitales, par exemple).

Poursuivant sa stratégie minimaliste, le CEA-LIST délimite des extraits dans les documents en se fondant sur la densité des mots de la question présents dans chaque extrait. Chaque extrait est alors noté selon un score qui dépend de cette densité et de

la présence d'entités nommées du type attendu. La réponse est ensuite trouvée au sein de cet extrait en réitérant la notation précédente liée à la densité de mots de la question, ou d'entités nommées du type attendu, dans une fenêtre dont la taille est identique à la longueur de la réponse recherchée. Quand le score est trop faible, la réponse est déclarée absente de la base de documents. Globalement, cette technique se révèle bonne quand la réponse est trouvée, mais trop « pessimiste » quand à la présence d'une réponse. Ce travail est effectué à l'identique pour la tâche médicale.

Sinequa calcule également un score pour chacun des passages de la taille requise au sein des documents retenus. Ce score est lié à la distance avec les éléments de la question (que ce soit les mots eux-mêmes, des synonymes ou des mots du même domaine). La reconnaissance de reformulation de la question à l'affirmatif sur les 100 premiers documents ramenés par le moteur est une stratégie qui est utilisée et qui prime sur tout autre quand elle s'applique (elle offre une meilleure garantie de réponse correcte), mais elle n'est pas très fréquente.

France Télécom considère que les passages pertinents sont ceux qui contiennent tous les mots-clés identifiés lors de l'analyse de la question (ou au moins une grande majorité d'entre eux quand ils sont très nombreux). Sur ces phrases sont appliquées des règles de reformulation ou de transformation pour trouver la réponse. Aucune ressource particulière n'a été ajoutée pour la tâche médicale. Lors de l'analyse des résultats, il s'avère que la seconde partie donne de bons résultats alors que la sélection des passages perd trop de passages pertinents.

L'Université de Neufchâtel fait appel à l'analyseur FIPS développé au LATL à Genève pour analyser les passages retenus et reconnaître les entités nommées qui sont présentes. Chaque phrase possède un score tenant compte de la proportion de termes de la question jugés pertinents par rapport au nombre de termes jugés pertinents dans la phrase. Les dix meilleures phrases sont alors retenues afin d'y rechercher plus précisément la réponse du type attendu.

Le LIMSI-CNRS reprend les documents obtenus par le moteur de recherche pour les reclasser en tenant compte de variantes possibles des termes de la question et ne retenir que les 50 meilleurs documents. Ceux-ci sont ensuite étiquetés en entités nommées et chacune des phrases les composant sont notées en fonction de poids liés au nombre de mots de la question et d'entité nommée du type attendu qui s'y trouvent. Si la question n'attend pas une réponse du type entité nommée, des patrons de syntaxe locale liée au type de la question sont appliqués afin de localiser la réponse dans ces phrases.

Synapse reprend les documents analysés pour y chercher les meilleures phrases en ayant identifié des métaphores éventuelles. Les phrases sont triées en fonction des éléments identifiés lors de l'analyse de la question. La réponse est extraite en

s'appuyant sur des critères de cohérence entre ce qui est trouvé dans la phrase globalement bien notée et le type de réponse qui est attendu. C'est la qualité des traitements d'analyse combinée à la richesse des ressources utilisées qui permet à ce système d'avoir de très bons résultats.

6.4 Analyse des résultats

Les résultats obtenus par les participants sont du même ordre de grandeur que les résultats obtenus par les participants aux campagnes CLEF monolingues. Afin d'étudier plus en détail les questions posées pour l'évaluation EQueR et de comparer plus finement les campagnes, nous avons effectué un certain nombre de mesures sur les questions et les réponses proposées par les participants. Ces mesures consistent à comparer le vocabulaire utilisé pour poser la question au vocabulaire présent dans les passages réponses proposés par les candidats, qu'ils contiennent la réponse ou non. Afin de comparer deux campagnes proches, ces mesures ont aussi été réalisées sur le corpus français de la tâche question-réponse de CLEF06.

6.4.1 Analyse de la difficulté des questions

La comparaison des mots contenus dans les passages réponses avec les mots des questions a été réalisée de la façon suivante :

– Sélection des questions factuelles (hors définitions, listes, booléeenes) ;

– Lemmatisation et catégorisation des mots des questions et des passages proposés par les systèmes avec le TreeTagger[7] ;

– Conservation des catégories nom, nom propre, verbe, adjectif, nombre et adverbe parmi les lemmes des questions ;

– Elimination des mots non significatifs parmi ces derniers : les auxiliaires, les verbes modaux, les verbes introducteurs de la demande (citer, nommer, etc.), les mots permettant de désigner le type de la réponse attendue (an, année, jour, type, mois, personne, lieu, pays, etc.) ;

– Recherche des lemmes restant de la question dans les passages. Les passages proposés par tous les participants à EQueR ont été considérés, mais en supprimant les doublons. Pour cela, seuls les passages appartenant à des documents différents ont été étudiés. Des 21 878 passages pour les 500 questions, après suppression des doublons, 4213 sont corrects et répondent à 325 questions et 13 131 sont incorrects. 95 questions n'ont pas de réponse. Comme quelques questions, au nombre de 5, n'avaient pas de réponse dans le corpus, les questions factuelles restées sans réponse sont au nombre de 90 (22%).

[7] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Ce processus a été appliqué entre questions et passages évalués comme corrects, c'est-à-dire contenant la réponse correcte, ainsi qu'entre questions et passages incorrects. Nous avons retenu, par question, le passage contenant le maximum de mots, puis rectifié manuellement quelques erreurs de comparaison, notamment lorsque l'écriture des noms propres composés diffère dans la question et la réponse (des mots sont reliés par un trait d'union dans les réponses par exemple).

Le nombre de mots conservés pour les questions va de 1 à 11, avec une moyenne de 4 mots par question. Quand on regarde les types de mots absents dans les 111 passages où un seul mot est manquant (cf. Figure 9), les phénomènes les plus présents sont :

– Absence du mot désignant le type de concept attendu (pour une entité nommée ou un concept autre) : 30 cas (27%) ;

– Absence du verbe principal : 24 cas (22%) ;

– Date absente : 4 cas (4%) ;

– Présence d'un synonyme ou d'une variation morphologique sur le verbe qui est nominalisé : 30 cas (27%) ;

– Présence d'anaphore : 5 cas (5%).

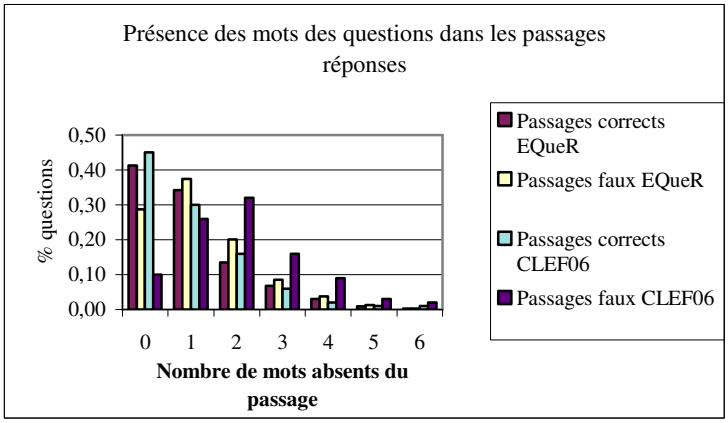

Figure 9. *Difficulté des questions*

Il y a quelques cas de relations sémantiques moins évidentes que la synonymie (causalité par exemple). Les autres cas, au nombre de 16, sont les passages où l'information absente du passage est différente du type, de la date ou du verbe principal. Si l'on regarde les passages incorrects, on voit qu'il y a tout de même 29% de questions qui ramènent un passage contenant tous les mots de la question et 37% avec un seul mot absent. Si l'on regarde les types des mots manquants dans ces

derniers, c'est un verbe dans 22% des cas et dans 16% environ, c'est le mot désignant le type. Même si la réponse à de nombreuses questions (75% des questions factuelles) figure dans des passages ayant presque tous les mots de la question, il n'en demeure pas moins la nécessité de bien les sélectionner en ajoutant d'autres critères, car 66% des erreurs répondent à ces mêmes critères.

Afin de donner un point de comparaison, l'étude des mots absents a été menée sur les résultats des deux meilleurs systèmes ayant participé à CLEF06 pour la tâche monolingue en français. L'étude porte sur toutes les questions, CLEF ne distinguant pas les questions factuelles des questions de définition, dont les passages réponses contiennent généralement le mot de la question. 58 questions sur 200 restent sans réponse (29%). On ne retrouve pas tout à fait la même répartition, même si la tendance générale reste la même. Le plus faible nombre de passages faux contenant la plupart des mots de la question (36%) peut peut-être s'expliquer par la taille du corpus : 1,5 gigaoctets pour EQueR et le tiers pour CLEF, ce qui ne favorise pas ce type de phénomène. En regardant les mots manquants dans les passages réponses de CLEF, les passages corrects omettent le verbe à 43% et les passages incorrects à 26%. Les systèmes ayant participé à CLEF06, en français, avaient aussi participé à EQueR, avec une version moins performante de leur système.

6.5 Conclusion

L'évaluation EQueR est allée dans le même sens que les évaluations existantes en question-réponse. Toutefois, nous avons eu le souci de produire le plus de données réutilisables pour mettre au point des systèmes. Ainsi, l'évaluation permettait aux systèmes de proposer cinq réponses par question sous forme courte, i .e. la réponse exacte, et longue, c'est-à-dire de 250 caractères au maximum. De cette manière, toutes les évaluations sont possibles : en comptant seulement les réponses au rang 1, on reproduit les évaluations CLEF. Les systèmes ont produit des données qui peuvent être utilisées comme corpus d'étude.

EQueR a introduit de nouvelles questions, les questions booléennes. Avec celles-ci, on rejoint la problématique de validation et justification des réponses présentes dans les évaluations AVE[8] (*Answer Validation Exercice*) à CLEF et RTE[9] (*Recognizing Textual Entailment*) du réseau PASCAL.

La tâche portant sur un domaine de spécialité a été moins bien suivie, car elle demandait de reconsidérer certains des processus de la tâche générale, notamment la recherche des entités nommées. Par ailleurs, il a été plus difficile de trouver des

[8] http://nlp.uned.es/QA/ave/
[9] http://www.pascal-network.org/Challenges/RTE/

questions factuelles à poser, les médecins posant plus naturellement des questions de nature différente (comment, pourquoi, etc.).

Il aurait été intéressant d'aller encore plus loin et de produire des résultats permettant l'évaluation des différents modules d'un système. Mais, si on peut penser que l'évaluation de la sélection de passage peut être faites de manière analogue à l'évaluation des réponses, il faudrait néanmoins essayer de retrouver dans le corpus tous les documents pertinents (il reste 22% de questions sans réponse), car ceux-ci ne peuvent être tous retournés par l'ensemble des participants. Etant donné le faible nombre de documents sélectionnés par les systèmes, et la précision que leur évaluation demande, on ne peut appliquer la technique utilisée en RI classique qui consiste à se fonder sur les meilleurs documents des systèmes participants et ne demande pas d'intervention humaine. En ce qui concerne l'évaluation de l'analyse des questions, elles pourrait être faite sur le type de réponse attendu, mais là on se heurte à la grande diversité des types définis par les systèmes, et on ne peut se ramener aux types standards de MUC (*Message Understanding Evaluation Conference*).

6.6 Bibliographie

[AYA 05] Ayache C., "Rapport final de la campagne EQueR/EVALDA, Évaluation en Question-Réponse, disponible sur le site Web Technolangue : http://www.technolangue.net/article 61.html, 2005

[AYA 06] Ayache C., Grau B., Vilnat A., "EQueR/EVALDA, the French Evaluation campaign of Question-Answering system", *Proceedings of LREC*, Genoa, Italy, 2006.

[BAL 05] Balvet A., Embarek M. et Ferret O., « Minimalisme et question-réponse : le système Œdipe », *Actes de TALN*, Dourdan, 2005.

[BLA 05] Blaudez E., Crestan E. et de Loupy C., « SQuAr : Prototype de moteur de Questions-Réponses », *Actes de TALN*, Dourdan, 2005.

[BRI 01] Brill E., Lin J., Banko M., Dumais S., Ng A., "Data-Intensive Question Answering", *Proceedings of TREC10*, Gaithersburg, MD, 2001

[CHA 03] de Chalendar G., Ferret, O., Grau, B., ElKateb F., Hurault-Plantet, M., Monceaux L., G., Robba I., Vilnat A., « Confronter des sources de connaissances différentes pour obtenir une réponse plus fiable », *actes de la conférence TALN*, Nancy, 2003

[CHU 02] Chu-Caroll J., Prager J., Welty C., Czuba K., Ferucci D., "A Multi-Strategy and Multi-Source Approach to Question Answering", *Proceedings of TREC11*, Gaithersburg, MD, 2002.

[CLA 01] Clarke C.L.A., Cormack G.V., Lynam T.R., Li C.M.,McLearn G.L., "Web Reinforced Question Answering (MultiText Experiments for TREC 2001)", *Proceedings of TREC10*, Gaithersburg, MD, 2001

[DEL 05] Delbecque T., Zweigenbaum P., Berroyer J.-F. et Poibeau T., « Le système STIM/LIPN à EQueR 2004, tâche médicale », *Actes de TALN*, Dourdan, 2005.

[FEL 98] Fellbaum C., *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, 1998.

[FER 01] Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Jacquemin, C., "Document selection refinement based on linguistic features for QALC, a question answering system", Proceedings of the Euroconference on *Recent Advances in Natural language Processing (RANLP),* Tsigov Chark, Bulgaria, 2001.

[GIL 05] Gillard L., Bellot P. et El-Bèze M., Le LIA à EQueR », *Actes de TALN*, Dourdan, 2005.

[GRA 04] Grau**,** B., Systèmes de question-réponse, dans *Méthodes avancées pour les systèmes de recherche d'information*, dir. Ihadjadène, Hermes, chap. 10, pp. 189-218, 2004.

[GRA 05] Grau**,** B., Illouz, G., Monceaux, L., Paroubek, P., Pons O., Robba I., Vilnat, A., FRASQUES, le système du groupe LIR, LIMSI, *Atelier EQueR de TALN*, Dourdan, 2005.

[GRI 95] Grishman R., Sundheim B., "Design of the MUC6 evaluation", *Proceedings of MUC-6*, Morgan Kauffmann Publisher, Columbia, MD, 1995.

[HAR 00] Harabagiu, S., Pasca, M., Maiorano, J., "Experiments with Open-Domain Textual Question Answering". *Proceedings of Coling*, Saarbrucken, Germany, 2000.

[HOV 01] Hovy, E. , Hermjacob, U. & Lin C-Y., Ravichandran, D. , "Towards Semantics-Based Answer Pinpointing", *Human Technology Conference (HLT)*, San Diego, 2001.

[LAU 05] Laurent, D., Nègre, S., Ségula, P., QRISTAL, le QR à l'épreuve du public, *n° spécial de la revue TAL, Répondre à des questions*, dir. B. Grau et B. Magnini, Volume 46, Numéro 3, 2005

[MAG 02] Magnini B., Negri M., Prevete R., Tanev H., "Mining Knowledge from Repeated Co-occurrences: DIOGENE at TREC 2002", *Proceedings of the Text retrieval conference, TREC11,* Gaithersburg, MD. NIST Eds, 2002

[MOL 02] Moldovan, D., Harabagiu S., Girju R., Morrarescu P., Lacatusu F., Novishi A., Badulescu A., Bolohan O., "LCC Tools for Question Answering", *Proceedings of the Text retrieval conference, TREC11,* Gaithersburg, MD. NIST Eds, 2002

[PER 05] 2005 Perret L., « *Extraction automatique d'information : génération de résumé et question-réponse* », Thèse de doctorat, Université de Neuchâtel, 2005.

[PRA 00] Prager J., Brown E., Radev D. R., Czuba K., "One Search Engine or two for Question-Answering", *proceedings of TREC9*, Gaithersburg, MD, p 235-240, 2000.

[SOU 01] Soubbotin, M. M., Soubbotin, S. M., "Patterns of Potential Answer Expressions as Clues to the Right Answers". *Proceedings of the Text retrieval conference, TREC 10,* Gaithersburg, MD. NIST Eds. , 2001.

[SOU 02] Soubbotin, M. M., Soubbotin, S. M., "Use of patterns for Detection of Likely Answer Strings: a Systematic Approach", *Proceedings of the Text retrieval conference, TREC 11,* Gaithersburg, MD. NIST Eds., 2002