20024557

REVISE, un outil d'évaluation précise des systèmes de questions-réponses

Sarra El Ayari [*], Brigitte Grau [*], Anne-Laure Ligozat [**]

LIMSI-CNRS [*]
Bât 508
F-91403 Orsay cedex (France)

Laboratoire IBISC - Université d'Evry [**]
Tour Evry 2, 523 place des terrasses de l'agora
91000 Evry

sarra.elayari, brigitte.grau@limsi.fr
aligozat@ibisc.univ-evry.fr

RÉSUMÉ. *Des campagnes d'évaluations sont organisées chaque année pour évaluer des systèmes de questions-réponses sur la validité des résultats fournis. Pour les équipes, il s'agit ensuite de réussir à mesurer la pertinence des stratégies développées ainsi que le fonctionnement des composants. À ces fins, nous décrivons un outil générique d'évaluation de type boîte transparente qui permet à un système produisant des résultats intermédiaires d'évaluer ses résultats. Nous illustrerons cette démarche en testant l'impact d'une nouvelle définition de la notion de focus.*

ABSTRACT. *Evaluation campaigns for question answering systems aim at evaluating their final results, i.e. the number of right answers. Then, in order to improve these systems, researchers try to evaluate each component, to improve them as well as to improve the global strategy. In order to help for these precise evaluations, we have conceived and developed a glass-box evaluation framework that works from the intermediary results provided by the different components. We will exemplify its capacities by showing how to measure a change in the determination of a question feature, the focus.*

MOTS-CLÉS : *Évaluation boîte transparente, évaluation boîte noire, systèmes de questions-réponses, plate-forme d'évaluation.*

KEYWORDS: *Glass-box evaluation, black-box evaluation, question answering systems, evaluation framework.*

L'objet, pages 8 à LMO

1. Introduction

Évaluer un système modulaire est une tâche complexe qui suppose de prendre en compte à la fois les résultats finaux obtenus par le système ainsi que les résultats intermédiaires produits par chacun des modules. Les campagnes d'évaluation de systèmes de questions-réponses organisent chaque année de nouvelles tâches afin de permettre aux équipes de tester leurs systèmes sur des problèmes de plus en plus complexes : les corpus vont d'articles journalistiques bien écrits au web avec des formats et des styles très divers, et l'évaluation peut porter sur des passages ou sur des réponses de plus en plus précises, voire multiples pour des questions dont la réponse attend une liste d'éléments. L'évaluation s'effectue sur le nombre de bonnes réponses renvoyées par le système, donc sur les résultats finaux, ce que l'on appelle une évaluation boîte noire. Pour améliorer ces systèmes complexes, chacun des participants évalue en interne la pertinence de ses composants, c'est-à-dire mène une évaluation de type boîte transparente, de son système. Si cette pratique est courante, il n'y a pas d'outils génériques permettant d'évaluer ce qui se passe à l'intérieur des systèmes (Moldovan *et al.*, 2003). Le problème principal est que leur architecture dépend des stratégies utilisées : chaque système a ses stratégies propres et de ce fait sa propre architecture.

Nous proposons un outil d'évaluation de type boîte transparente qui permet à la fois d'évaluer les sorties produites par les composants, mais aussi de tester des stratégies sans toucher au système lui-même. Nous présenterons le système FRASQUES développé au LIMSI sur lequel nous travaillons (2), puis notre outil d'évaluation (3), que nous illustrerons par l'évaluation de la pertinence du terme pivot (le focus) utilisé pour extraire la réponse attendue (4).

2. Architecture d'un système de questions-réponses

Comment répondre de façon automatique à une question ? C'est le défi qu'essayent de relever les systèmes de questions-réponses. Contrairement aux moteurs de recherche, de tels systèmes permettent à un utilisateur de poser une question en langage naturel, et lui fournissent une réponse précise : *Quelle est la nationalité d'Ayrton Sena ?* attendra la réponse *espagnol*. Le système FRASQUES (Grau *et al.*, 2006) est composé de modules que sont l'analyse des questions, le moteur de recherche et l'extraction de réponses. Ces quatre modules fonctionnent de façon linéaire, les informations extraites lors de l'analyse de la question (la catégorie, le type général, les entités nommées, le focus, les termes et leurs variations sémantiques) sont ensuite utilisées par le moteur de recherche pour créer une requête qui fournira des documents contenant les mots de la question, ou bien leurs variations. Les informations extraites lors de l'analyse de la question *De quelle organisation Javier Solana était-il secrétaire général ?* seraient la **catégorie** (*quel*), le **type général** (*organisation*), l'**entité nommée** (*Javier Solana*), le **type d'entité attendu** (*organisation*), le **focus** (*secrétaire général*) et le **verbe principal** (*être*). Ces informations seront également nécessaires pour la sélection des phrases réponses candidates, qui se voient attribuer un poids en fonction de leur similarité avec les éléments extraits de la question. Enfin, lors de l'extraction de la

réponse précise, le focus, le verbe principal et l'entité nommée attendue jouent un rôle déterminant grâce à des patrons d'extraction de la réponse définis sur ces éléments. L'évaluation de la pertinence de chacun de ces composants n'est possible que si l'on a accès aux résultats intermédiaires qu'ils produisent, afin de juger de leur apport réel et des phénomènes qui posent problème.

3. REVISE, un outil pour visualiser et évaluer

3.1. *Etat de l'art*

Évaluer finement l'apport des composants d'un système suppose de mesurer la contribution de chacun des modules par rapport aux résultats globaux obtenus par le système. De ce point de vue, l'évaluation de type boîte transparente n'est pas en opposition avec une évaluation boîte noire, mais complémentaire pour obtenir un diagnostique complet (Sparck Jones, 2001). Ces deux méthodes dépendent avant tout de ce que l'on veut évaluer. Selon (Gillard *et al.*, 2006), l'état de l'art des évaluations réalisées sur des systèmes de questions-réponses montre le manque de lisibilité des apports des composants sur les résultats finaux et la nécessité d'une étude plus approfondie de chacun des composants.

Dans la littérature, on trouve deux courants liés à l'évaluation des différents modules d'un système. Le premier consiste à enlever un composant et à mesurer les résultats obtenus. Il devient alors également possible de le remplacer par un autre, en mesurant à nouveau l'apport ou la perte obtenus (Costa *et al.*, 2006), (Tomas *et al.*, 2005). Le deuxième courant, assez novateur dans le domaine, est illustré par le système de questions-réponses JAVELIN (Nyberg *et al.*, 2003). JAVELIN est un système de questions-réponses qui intègre un module permettant de contrôler l'exécution du processus, ainsi que les informations qui sont utilisées. Il a été conçu de façon à permettre une évaluation de type boîte transparente. Il permet également de tester différentes stratégies qui peuvent ensuite être intégrées au système. Néanmoins, il n'existe aucun outil générique pour effectuer ce type d'évaluation.

Notre stratégie consiste à étudier et éventuellement à modifier les résultats intermédiaires créés par les composants et insérer ces nouveaux résultats dans le processus de traitement sans modifier le système lui-même. L'outil que nous avons conçu permet d'observer les données, de les modifier et d'évaluer l'impact des modifications sur le système.

3.2. *Description de notre interface*

REVISE est l'acronyme de *Recherche, Extraction, VISualisation et Evaluation*, termes qui résument ce que permet de faire notre outil. Recherche et extraction sont effectuées par la base de données qui contient les résultats intermédiaires produits par les composants du système. La visualisation des données se fait grâce à un export en

XML lié à des fichiers XSLT qui permettent la mise en relief de certains phénomènes. Enfin, l'évaluation est au cœur de cet outil grâce à la possibilité de modifier les résultats intermédiaires du système et de les ré-injecter dans le système. La figure 1 illustre les points d'évaluation transparente effectués sur le système FRASQUES, que nous avons décrit précédemment (Grau *et al.*, 2006). Des formulaires PHP sont également

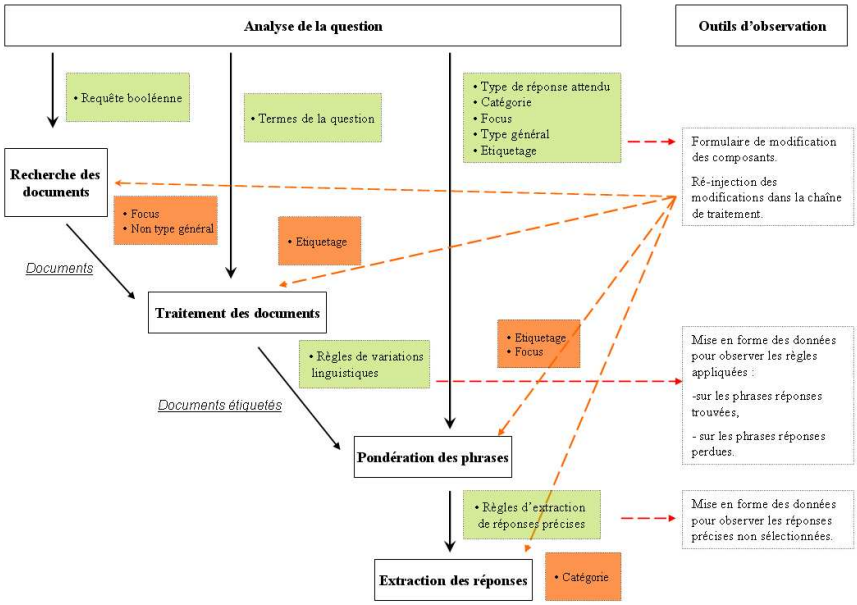

Figure 1. *REVISE, un outil d'évaluation transparente*

intégrés, qui permettent de modifier certains résultats dans la base de données. Les nouveaux résultats sont ensuite exportés et ré-injectés dans la chaîne de traitement avec le format adéquat.

3.2.1. *Base de données*

Les résultats intermédiaires produits par les différents composants sont décrits en XML dans un fichier contenant l'ensemble des résultats des processus effectués. La structure de ce fichier a conduit à la définition d'un schéma de base de données relationnelle. Les résultats sont stockés dans la base de données qui permet ainsi de représenter et rapprocher les différentes entités manipulées dans la chaîne, pour chaque test, c'est-à-dire l'ensemble des informations produites par une version donnée de la chaîne, pour un jeu de question et un corpus donnés : question vs passages réponses, caractéristiques de la question issues de l'analyse vs caractéristiques des phrases réponses (termes, poids, entités nommées), réponse trouvée ou non. Cette base contenant différents tests sous la forme de tables, il est possible de les comparer entre eux, de façon à prendre en compte différentes versions du système. L'intérêt de cette base de données est de donner la possibilité de faire des requêtes (prédéfinies ou libres)

sur ces résultats, et de ne sélectionner que ce qui intéresse l'utilisateur. Par exemple, on peut visualiser les informations liées aux questions de certaines catégories ou bien celles pour lesquelles on ne trouve pas la réponse, ou encore sur les questions pour lesquelles le focus est erroné. Le fait de pouvoir trier les données en fonction de critères précis, comme ceux énoncés précédemment, permet une évaluation plus fine des phénomènes.

3.2.2. *Visualisation*

Un autre intérêt de notre outil est la possibilité d'interroger la base de données et de visualiser les résultats de façon lisible et organisée. En effet, grâce aux technologies XML et XSLT, les résultats sont extraits en XML et la visualisation est présentée en XHTML. En plus de l'interopérabilité de ces langages, il devient alors possible de jouer sur les traits à faire ressortir pour faire sens. Si la visualisation propre de données est importante, il devient également crucial de permettre l'émergence de formes grâce à des jeux de couleurs, pour mettre en évidence des phénomènes. Le fait de colorer le focus dans les phrases réponses candidates va permettre immédiatement de voir dans quels contextes il apparaît, de même pour les autres mots de la question. La figure 2 montre la mise en relief du focus et de la réponse attendue dans les différentes phrases réponses candidates extraites par le système FRASQUES.

Figure 2. *Exemple de visualisation de données*

3.2.3. *Evaluation*

Enfin, en lien avec les deux points développés ci-dessus, REVISE permet d'évaluer les stratégies utilisées par un système. Par exemple, il est possible de tester différentes définitions d'un critère, et de ce fait mesurer la pertinence des définitions les unes par

rapport aux autres, sans modifier le système de questions-réponses. Ayant accès aux résultats produits à différents niveaux de la chaîne de traitement, l'entrée et la sortie de chacun des modules peuvent être contrôlées, validées et le cas échéant modifiées afin de tester une hypothèse. Nous illustrerons ce point dans la suite de l'article.

3.3. *Généricité de l'outil*

REVISE est un outil qui travaille uniquement sur les résultats intermédiaires produits par le système. C'est-à-dire que les résultats obtenus par chacun des composants doivent être sauvegardés dans un même fichier. Si l'exécution de la chaîne de traitement est décrite de manière explicite, il est possible de modifier ces résultats et relancer des processus sans toucher à la chaîne de traitement. Ainsi, notre système de questions-réponses est décrit dans un fichier XML et on peut aisément déterminer des points de reprises. Il devient alors possible de relancer le processus avec les résultats modifiés au point de reprise désiré. De ce fait, tout système qui produit des résultats intermédiaires peut utiliser REVISE. Afin d'illustrer ce que notre outil permet de faire, nous avons pris appui sur une évaluation de type boîte transparente effectuée sur le système de questions-réponses QRISTAL[1] dans l'article (Laurent *et al.*, 2006) : « QRISTAL est un système de questions-réponses multilingue (français, anglais, italien, portugais, polonais, tchèque) conçu pour extraire des réponses à partir de documents placés sur un disque dur, ou pour extraire des réponses à partir du web sur la base de pages ou passages retournés par des moteurs web classiques (Google, MSN, AOL, etc.) ». Cet article, écrit après participation à différentes campagnes d'évaluation (Equer[2], Clef05 et Clef06[3]) montre la nécessité pour les équipes de mesurer les performances des systèmes de façon précise, indépendamment des résultats fournis par les campagnes.

Les résultats produits par le système sont analysés, au niveau de l'analyse (syntaxique et sémantique) des questions (extraction des informations nécessaires au bon fonctionnement du système), de la sélection des blocs de réponses et de celle des phrases réponses. Il s'agit essentiellement d'observation des données, ainsi que de mesurer les taux de bonne mise en oeuvre des composants par le nombre final de bonnes réponses. Afin de tester l'impact de ces composants, chacun des modules du système a été déconnecté. Ceci a révélé l'importance de la catégorisation des questions, notamment lors de l'extraction des réponses. Les auteurs énoncent tout de même le problème de déconnection de composants tels que l'analyseur syntaxique, lequel est indispensable au bon fonctionnement de QRISTAL. REVISE permet de réaliser le même type d'évaluation de façon automatisée : les résultats produits sont stockés dans la base de données, ce qui permet de lancer toutes les requêtes désirées :

1. QRISTAL est l'acronyme de *Questions-Réponses Intégrant un Système de Traitement Automatique des Langues*.
2. Voir http ://www.technolangue.net/article.php3 ?id_article=195.
3. Voir http ://www.clef-campaign.org/.

1) Analyse syntaxique de la question / étiquetage
Notre outil stocke les informations liées à la question dans une base de données. Une requête sur la base permet ensuite de visualiser ce qui a été extrait. Si l'extraction n'est pas satisfaisante, il est tout à fait possible de modifier les champs erronés.

2) Analyse sémantique de la question
De la même façon, les synonymes étant des résultats produits par le système, nous pouvons observer leur pertinence et leur bonne application en les visualisant dans les phrases sélectionnées. Nous pouvons aussi visualiser les mots de la question qui ne figurent pas dans les passages réponses. Ainsi, on peut étudier à la fois la pertinence des synonymes, mais aussi le défaut de couverture des lexiques.

3) Analyse des passages extraits
L'ordonnancement des passages est également accessible, avec les scores attribués.

4) Analyse des phrases réponses
Même chose pour les phrases réponses candidates, sur lesquelles on peut également observer finement l'application ou non des patrons d'extraction de la réponse précise.

Notre outil offre, en terme de visualisation, un accès à tous les éléments évalués jugés importants au sein du système QRISTAL, avec une normalisation de la visualisation et la possibilité de modifier ces résultats et de relancer le traitement. L'étude d'une question en particulier est tout à fait possible en indiquant son numéro ainsi que le numéro de processus qui nous intéresse.

4. Etude du focus

Dans cette partie, nous allons montrer comment REVISE peut servir à étudier un paramètre particulier d'un système, le focus. Après avoir redéfini le terme focus, nous utiliserons notre outil pour valider la nouvelle définition.

4.1. *Redéfinition du terme focus*

Le terme focus est un élément de la question, qui est le terme pivot pour extraire la réponse attendue, terme censé apparaître à proximité de la réponse (Ferret *et al.*, 2002). Il s'agit de l'élément central de la question, auquel se rattache directement la réponse. Si l'on représente la question sous la forme d'un graphe, c'est le focus qui en sera le nœud central. Comme le montre la figure 3, la représentation de la phrase *Quand le pont de Normandie a-t-il été inauguré ?* serait : Sa définition est essentielle à l'extraction de la réponse dans notre système, car il permet d'expliciter des patrons d'extraction modélisant l'expression en langue de la relation existant entre le focus et la réponse. Le critère principal de la reconnaissance de ce focus était qu'il s'agit du sujet du verbe principal de la question.

Nous avons mené une expérience sur le focus ainsi reconnu (El Ayari, 2007), afin de tester si cette hypothèse était vérifiée par le système QALC, version anglaise

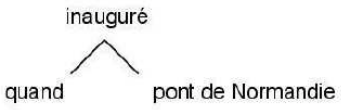

Figure 3. *Graphe de la question*

du système FRASQUES. Il en ressort que des questions n'ont pas de focus, car on ne s'appuyait que sur les entités nominales, et que le choix systématique d'un groupe nominal induit le fait que l'écriture de patrons d'extraction précis selon le type de relation qui doit être vérifié entre focus et réponse est plus difficile, car le groupe nominal choisi peut varier pour un même type de relation. Nous avons donc précisé la notion de focus et nous allons montrer comment cette nouvelle définition peut être évaluée sans avoir à modifier l'analyse des questions pour mettre en oeuvre sa reconnaissance automatique. Si le focus a toujours été défini comme une entité, nous proposons l'idée qu'il peut également s'agir d'un événement. Les données que nous présentons sont tirées de la campagne d'évaluation CLEF07[4].

4.1.1. *Le focus est un événement*

Sa reconnaissance repose sur le sens du verbe, qui doit être assez fort pour traduire un événement. Si le verbe possède un complément, le focus sera composé.

– Quand débuta le procès de Paul Touvier ?

– Le procès de Paul Touvier s'ouvrira le **17 mars** devant la cour d'assises des Yvelines (région parisienne), a-t-on appris officiellement jeudi.

Cette question demande une précision sur l'événement *débuter le procès*. C'est cet événement qui sera le focus de la question, et qui permettra l'extraction de la réponse précise attendue par la question. La réponse est complément du verbe s'ouvrir. La notion de début est essentielle, pour sélectionner entre des phrases qui parleraient du procès, mais donneraient d'autres dates : fin, interruption, etc..

4.1.2. *Le focus est une entité*

Si le verbe n'a qu'un sens relatif, c'est alors l'entité sur laquelle la question est posée qui sera notre focus. Cette entité pourra être exprimée *en intension* ou *en extension*.

– De quelle organisation Javier Solana était-il secrétaire général ?

– Javier Solana a officiellement été nommé mardi secrétaire général de l'**OTAN**, mettant fin à une vacance de plusieurs semaines.

Cet exemple illustre une entité exprimée en extension, c'est-à-dire nommée : *Javier*

4. Voir http ://www.clef-campaign.org/.

Solana, et l'autre en intension, c'est-à-dire qu'on fait référence à quelqu'un par une description : *secrétaire général*. Dans ce cas, le focus sera l'entité exprimée en intension (secrétaire général), car c'est sur la fonction que la question est posée. On voit effectivement que la réponse est liée à ce terme focus dans la phrase réponse.

– A quel parti appartient Thérèse Aillaud ?

– Placée dans la même situation que M. Siffre, le député et maire sortant de Tarascon, Thérèse Aillaud (**RPR**), n'enlève que le tiers des suffrages exprimés alors qu'elle avait été réélue en 1989 dès le premier tour de scrutin avec près de 63% des voix.

Ici, il n'y a ni verbe porteur d'un événement, ni entité exprimée en intension. Le focus sera *Thérèse Aillaud*, entité sur laquelle la question est posée.

4.1.3. *Quelques particularités en fonction des catégories de questions*

Dans le domaine des systèmes de questions-réponses, les questions sont souvent classées en fonction de leur pronom interrogatif et du type d'entité attendu (date, lieu, personne ou autre). Nous avons défini des catégories en fonction du type de focus exprimé et de la relation recherchée. Les principales catégories sont :

– Définition : *Qu'est-ce que l'accélération centrifuge ?*

– Combien : *Combien y a-t-il eu de mariages en Grande-Bretagne en 1993 ?*

– Quand : *Quand a eu lieu la chute du régime communiste en Afghanistan ?*

– Où : *Où se situent les îles Marquises ?*

– Quel : *Quelle était la nationalité d'Ayrton Senna ?*

– Qui : *Qui est Michael Jackson ?*

– Instance : *Citer le nom d'un corps céleste.*

Pour les questions de type Combien, nous avons deux valeurs à extraire de la question que sont l'unité et le terme focus.

– Combien de puits ont dû être fermés suite à la rupture d'un oléoduc en Sibérie ?

– La rupture d'un oléoduc dans le gisement de Samotlor, dans l'ouest de la Sibérie, a contraint les autorités à fermer **52** puits pour empêcher toute extension de la pollution, a rapporté lundi la télévision russe.

On voit que l'unité peut constituer un indice fort pour extraire la réponse attendue, de même que l'événement dont il est question. La réponse attendue est encadrée par ces deux éléments dans la phrase réponse, ce qui tend à justifier les deux stratégies. Dans ce cas-là, des patrons d'extraction sont constitués autour de l'unité et autour du focus.

En ce qui concerne les questions de type Instance, il s'agit de questions qui donnent la définition (ou description) de la réponse : elles donnent le type de la réponse. Il n'y a donc pas de focus, et une stratégie différente pour extraire la réponse, axée sur la vérification de ce type, est mise en place. Par exemple : *Quelle est la plus grande banque du Japon ?* ou *Que s'est-il produit en Algérie dans la nuit du 17 au 18 août 1994 ?* Nous n'indiquons donc pas de terme focus pour ces questions, sa définition

ne pouvant s'appliquer ici. Nous allons illustrer le fonctionnement de REVISE sur la notion de focus qui vient d'être définie.

4.2. *Validation de l'hypothèse*

Nous rappelons que le terme focus sert essentiellement à extraire les réponses précises. La réponse devant être liée à ce terme, ces deux termes sont souvent proches dans la réponse et cela peut être mesuré par la distance en mots entre les deux entités. Pour tester la pertinence de cette nouvelle définition, nous avons comparé l'ancienne définition du focus à la nouvelle, en terme de proximité avec la réponse. Nous allons détailler les différentes étapes de notre méthodologie :

1) le choix manuel[5] du focus des questions en fonction de la définition
Pour ce faire, nous avons effectué une requête sur la base de données, en spécifiant le corpus (Clef07). Un formulaire nous a permis de modifier le focus extrait par FRASQUES.

2) la sélection des phrases réponses qui contiennent ce focus et une réponse possible
Une requête sur la table contenant les réponses a été créée, en filtrant sur la présence du terme focus et d'une réponse possible (les réponses sont également stockées dans une table). La visualisation est permise par des scripts XSLT qui permettent de choisir le format d'affichage des résultats ainsi que les codes couleur à utiliser.

3) le calcul de la distance (en mots) entre focus et réponse
Un nouveau script a été ajouté à l'interface pour calculer automatiquement ces distances. Les résultats sont stockés dans la base de données de façon à rester accessibles.

Cette méthodologie a permis de comparer les définitions, afin de mesurer laquelle est plus pertinente que l'autre pour extraire la réponse précise. Le tableau 1 présente les résultats du calcul automatique de la distance moyenne en mots entre un focus et une réponse, pour chacune des définitions. Nous avons classé les résultats par catégories de questions. La nouvelle définition du focus apparaît plus pertinente que l'ancienne,

Ancien focus	Nouveau focus
8 mots	4 mots

Tableau 1. *Résultats de la distance focus/réponse*

avec une distance moyenne de quatre mots, soit moitié moins que l'ancienne. **345 phrases réponses** sur les données de Clef07 contiennent une réponse et le focus de la question à laquelle cette phrase répond. Le tableau 2 présente le nombre de phrases réponses par distance. Nous nous sommes arrêtés à une distance maximale de quatre mots, une distance syntaxique supérieure rend inutilisable les patrons d'extraction.

5. Cette étape est réalisée manuellement pour tester la pertinence de la nouvelle définition du focus, sans avoir à modifier l'analyse des questions.

Les distances 0 et 1 comptabilisent un peu plus de la moitié des phrases réponses

Distance (mots)	Nb phrases	Catégories les plus fréquentes
0	129	Combien, Définition
1	48	Quel, Quand
2	39	Quel
3	22	Quel
4	24	Quel

Tableau 2. *Distances les plus fréquentes*

(177). Si l'on rajoute les distances 2, 3 et 4 on obtient un score de 75% des phrases réponses, ce qui est très encourageant pour l'extraction des réponses. REVISE a permis ici de vérifier notre nouvelle définition du terme focus, grâce à l'observation et la modification des résultats de l'analyse des questions, la sélection des phrases réponses extraites de la base de données qui nous intéressaient ainsi que le calcul des distances de mots entre focus et réponses. Nous avons modifié manuellement les focus extraits par le système, données préalablement insérées dans la base, à l'aide d'un formulaire PHP. Nous avons ensuite sélectionné les phrases réponses grâce à une requête pré-enregistrée, et calculé automatiquement la distance en mots entre le focus et la réponse. Notre base de données a également permis d'effectuer des calculs, comme la moyenne des distances, de façon à pouvoir faire notre évaluation. La sélection du focus, qu'il s'agisse de la première ou de la deuxième, a été faite en fonction des mots de la question uniquement. C'est-à-dire qu'aucune variation syntaxique ni sémantique n'a été prise en compte.

Étant donné que notre nouvelle définition du focus repose essentiellement sur la notion d'événement, c'est le verbe de la question qui est sélectionné. Il est fréquent que l'événement exprimé par le verbe dans la question soit sous une forme nominalisée dans la réponse. Les synonymes sont autant de termes que nous n'avons pas comptabilisés ici. La prise en compte de ces variations lors de l'extraction des réponses devrait augmenter les résultats obtenus en terme de présence du focus. La prochaine étape consiste à établir des critères de reconnaissance automatique de la nouvelle définition du focus, entre entité et événement. Nous utiliserons notre outil pour mesurer l'impact de son intégration au système au niveau de l'extraction des réponses précises.

5. Conclusion

L'intérêt d'une évaluation de type boîte transparente n'est plus à prouver. Mais il est difficile de créer une méthodologie pour évaluer la contribution de chacun des différents composants d'un système. Notre outil, REVISE, facilite l'observation des résultats produits par le système. Il permet également de réaliser des évaluations de type boîte transparente. De plus, sans modifier le système, il devient possible de tester ses stratégies de recherche, qu'il s'agisse de l'analyse des questions ou encore de l'extraction des réponses, en ne modifiant que les résultats produits, et en relançant

le système à un endroit particulier de la chaîne de traitement. De la sorte, n'importe quel système produisant des résultats intermédiaires devrait être en capacité d'utiliser REVISE.

6. Bibliographie

Costa L., Sarmento L., « Component Evaluation in a Question Answering System », *Actes de la 5e conférence Language Resources and Evaluation Conference (LREC)*, Gênes, Italie, 24-26 mai, 2006.

El Ayari S., « Evaluation transparente de systèmes de questions-réponses : application au focus », *Actes de ReciTAL*, 2007.

Ferret O., Grau B., Hurault-Plantet M., Illouz G., Monceaux L., Robba I., Vilnat A., « Recherche de la réponse fondée sur la reconnaissance du focus de la question », *Actes de Traitement automatique du langage naturel (TALN)*, 2002.

Gillard L., Bellot P., El-Beze M., « Question Answering Evaluation Survey », *Actes de la 5e conférence Language Resources and Evaluation Conference (LREC)*, Gênes, Italie, 24-26 mai, 2006.

Grau B., « Evaluation des systèmes de question-réponse », *Évaluation des systèmes de traitement de l'information*, Hermès, chapter 3, p. 77-98, 2004.

Grau B., Ligozat A.-L., Robba I., Vilnat A., Monceaux L., « FRASQUES : A Question-Answering System in the EQueR Evaluation Campaign », *Actes de la 5e conférence Language Resources and Evaluation Conference (LREC)*, Gênes, Italie, 24-26 mai, 2006.

Laurent D., Nègre S., Séguéla P., « QRISTAL, le QR à l'épreuve du public », *Actes de Traitement automatique du langage naturel (TALN)*, 2006.

Moldovan D., Paca M., Harabagiu S., Surdeanu M., « Performance issues and error analysis in an open-domain question answering system », *Actes de ACM Transactions on Informations Systems*, 2003.

Nyberg E., Mitamura T., Callan J., Carbonell J., Frederking R., Collins-Thompson K., Hiyakumoto L., Huang Y., Huttenhower C., Judy S., Ko J., Kupse A., Lita L. V., Pedro V., Svoboda D., Durme B. V., « The JAVELIN Question-Answering System at TREC 2003 : A Multi-Strategy Approach with Dynamic Planning », *Actes de Text Retrieval Conference (TREC)*, 2003.

Sparck Jones K., « Automatic language and information processing : rethinking evaluation », *Natural Language Engineering*, chapter 7, p. 1-18, 2001.

Tomas D., Vicedo J. L., Saiz M., Izquierdo R., « Building an XML framework for Question Answering », *Actes de Cross Language Evaluation Forum (CLEF)*, 2005.