Évaluation transparente de systèmes de questions-réponses : application au focus

Sarra EL AYARI
LIMSI-CNRS, BP 133, F-91403 ORSAY
Sarra.ElAyari@limsi.fr

Résumé. Les campagnes d'évaluation ne tiennent compte que des résultats finaux obtenus par les systèmes de recherche d'informations (RI). Nous nous situons dans une perspective d'évaluation transparente d'un système de questions-réponses, où le traitement d'une question se fait grâce à plusieurs composants séquentiels. Dans cet article, nous nous intéressons à l'étude de l'élément de la question qui porte l'information qui se trouvera dans la phrase réponse à proximité de la réponse elle-même : le focus. Nous définissons ce concept, l'appliquons au système de questions-réponses QALC, et démontrons l'utilité d'évaluations des composants afin d'augmenter la performance globale du système.

Abstract. Evaluation campaigns take into account only the final results obtained by information retrieval systems. Our perspective is that of the glass box evaluation of a question-answering system. The processing of a question is accomplished by a series of components. The purpose of this article is to study the element in the sentence which holds the key information. This element is to be found again in the sentence containing the answer next to the answer itself, and is called the focus. We will begin by defining this concept. We will then applied it to the QALC question answering system. Finally we will demonstrate the pertinence of using glass box evaluations to enhance the global performance such systems.

Mots-clés : système de questions-réponses, recherche d'information, évaluation, focus.

Keywords: question answering system, information retrieval, evaluation, focus.

1 Introduction

Les systèmes de recherche d'information (RI) ont vu apparaître de nouveaux types d'outils appelés systèmes de questions-réponses (SQR). C'est la prise en compte du besoin d'information précise de l'utilisateur qui a motivé l'émergence de tels systèmes.

Un SQR peut être opposé à un moteur de recherche sur Internet comme *Google* ou *Yahoo !* sur certains points bien précis. L'utilisateur saisit sa requête en langue naturelle (LN) et non sous la forme de mots clés. En aval, le système propose la ou les réponse(s) attendue(s) par l'utilisateur, et ne lui renvoie pas quelques milliers de documents qu'il doit parcourir manuellement.

Deux utilisations différentes s'esquissent clairement entre les deux types d'outils. Tandis que les moteurs de recherche permettent de récupérer des documents sur un thème général, les systèmes de questions-réponses sont utilisés pour trouver une information précise, qui tient en quelques

mots. Par exemple, la requête, qui est donc une question, « What is the FARC ? »[1] attend la réponse suivante : *the Revolutionary Armed Forces of Colombia*.

Les conférences organisées pour évaluer les systèmes de questions-réponses prennent uniquement en compte le résultat final obtenu. Or des traitements, des hypothèses sont instaurés à différentes étapes de la résolution des questions. Nous nous intéressons à une évaluation de type boîte transparente, qui veut évaluer les composants du système isolément. Cette évaluation est appliquée à l'étude d'un élément central pour l'extraction de la réponse : le focus. Sa reconnaissance intervient au moment de l'analyse de la question ; il constitue l'élément informationnel présent dans la question situé à proximité de la réponse dans la phrase réponse.

Nous présentons ce qu'est un système de questions-réponses (2), avant de nous focaliser sur le système du LIMSI : QALC dans une perspective d'évaluation de type « glass box » (boîte transparente). Après avoir présenté la dichotomie « black box » (boîte noire) / « glass box » (3), nous appliquerons ces principes à l'étude du focus (4).

2 Systèmes de questions-réponses

2.1 Architecture d'un système de questions-réponses

Un système de questions-réponses prend une question en entrée et doit fournir une réponse courte et précise à cette question. Nous travaillons sur le système *QALC* développé au LIMSI. QALC est conçu pour « traiter des questions factuelles ou encyclopédiques portant sur n'importe quel domaine » (Ferret *et al.*, 2001). En effet, ce système travaille en domaine ouvert, ce qui implique une certaine robustesse des processus employés. Pour ce faire, le système est composé de différents modules que nous allons expliciter. La chaîne de traitement est présentée sur le schéma 1[2].

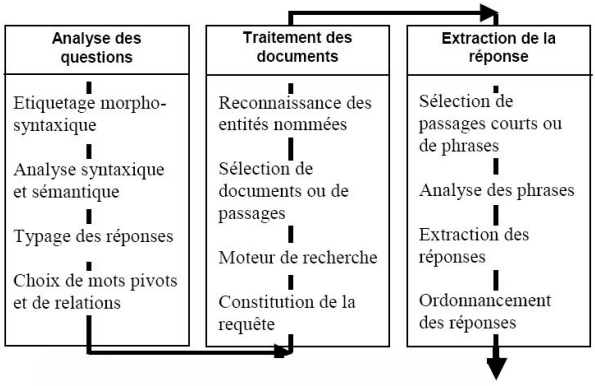

FIG. 1 – Chaîne de traitement de QALC

[1] Cette question est extraite du corpus CLEF 2005.
[2] Ce schéma est extrait de (Grau, 2004a).

Le traitement effectué par un SQR se fait en une séquence de trois étapes : l'analyse de la question, le traitement des documents sélectionnés par le moteur de recherche et l'extraction de la réponse dans les documents récupérés. Le traitement est ici plus fin que celui des moteurs de recherche : des traitements supplémentaires sont effectués en amont et en aval du moteur de recherche.

Il devient nécessaire de réaliser un travail plus fin sur la langue, qui fait appel au traitement automatique de la langue. Les informations principales dont on dispose se trouvent dans la question. Il s'agit d'en tirer le plus d'éléments pertinents possibles.

2.2 Le module d'analyse des questions

Le module d'analyse des questions permet de récupérer des informations essentielles pour identifier la réponse correspondant à la question. Nous présentons différents éléments qui importent pour le repérage de la réponse pour lesquels des traitements linguistiques peuvent être réalisés.

– Le type de réponse attendu, dans le cas où la réponse est une entité nommée (c'est-à-dire une unité d'un élément discursif qui fait référence à une personne, un lieu, une organisation, etc.). Ces entités sont repérées comme telles dans les documents extraits par le moteur de recherche. Elles sont déterminées par le pronom interrogatif utilisé ainsi que par des critères syntaxiques et sémantiques : *qui* indique que la réponse devra être une personne, *où* indique un lieu, *quand* indique une date, etc. S'il s'agit d'une question de type *Quel président français a été élu deux fois ?* le système repère une entité nommée de type personne.
– Le type sémantique de la réponse, lorsque celui-ci est explicite, est déterminé par un terme générique dans lequel est inclus le terme sur lequel la question porte. Pour *Quelle est la capitale du Togo ?* le type sémantique de la réponse sera lieu, par extension sémantique du terme *capitale*.
– L'objet de la question, à savoir le focus, qui correspond au mot de la question le plus important, défini comme étant celui que l'on doit retrouver dans la phrase réponse. Il s'agit du sujet de la question en quelque sorte.
– La catégorie de la question, déduite de son analyse syntaxique.
– Une extension sémantique de la question, qui consiste en la recherche de synonymes ou/et d'hyperonymes.

Par exemple, le traitement de la question *Qui a tué Henri IV ?* devra obtenir les informations suivantes :

- type attendu de la question : PERSONNE (entité nommée)
- objet de la question : Henri IV
- catégorie de la question : qui
- forme syntaxique de la question : PERSONNE VP[3] Focus
- forme syntaxique de la réponse :
  – REPONSE VP Focus pour ***<Réponse attendue>*** *a tué Henri IV*.
  – FOCUS VP REPONSE pour *Henri IV a été tué par* ***<Réponse attendue>***.
  – REPONSE nominalisation du verbe FOCUS pour ***<Réponse attendue>****, le tueur d'Henri IV*.
- extension sémantique : poignarder, assassiner, abattre

[3] VP est l'abréviation de verbe principal.

3 Évaluation fine des processus mis en place

3.1 Évaluation de type « black box » (boîte noire)

Les enjeux d'évaluation des systèmes de recherche d'informations ont pris une importance plus forte avec l'avènement du Web et le développement industriel des traitements textuels avec des infrastructures comme celle de la DARPA (Defense Advances Research Projects Agency), le NIST (National Institute of Standards and Technology) ou encore LDC (Linguistic Data Consortium) principalement aux Etats-Unis, et une prédominance de la langue anglaise.

En ce qui concerne les systèmes de questions-réponses, on distingue essentiellement deux campagnes internationales que sont TREC[4] (Text REtrieval Conference) et CLEF[5] (Cross Language Evaluation Forum), ainsi qu'une campagne française : EQUER[6] (Évaluation en Question Réponse). Nous pouvons également citer NTCIR, où l'évaluation porte uniquement sur des systèmes qui traitent les langues japonaise et chinoise.

La première campagne d'évaluation en questions-réponses a eu lieu en 1999 : il s'agit de la huitième édition de TREC, qui ne portait jusqu'alors que sur les systèmes de RI. Le déroulement de la tâche s'est complexifiée au fil des années. Si les premières campagnes proposaient 200 questions, pour lesquelles les participants pouvaient proposer cinq réponses par question sur des questions uniquement factuelles, il s'agit désormais de plus de 500 questions, qui peuvent être factuelles, mais qui peuvent également porter sur des listes, des questions définitionnelles et des scénarios (questions liées aux précédentes). Enfin, certaines questions n'ont pas de réponse dans les documents du corpus, et le système doit pouvoir l'indiquer. Une seule réponse courte est exigée, et celle-ci doit être accompagnée d'une justification (extrait de phrase permettant de valider la réponse proposée). De la même façon, la taille du corpus a augmenté de 528 000 documents en 1999 à 1 033 000 en 2005.

Ces campagnes permettent une évaluation globale des systèmes entre eux. Il s'agit de structures d'évaluation qui permettent de stimuler la recherche avec de nouveaux enjeux chaque année. Il devient alors possible de faire un constat mesuré de l'avancé technologique de ces systèmes chaque année. Un autre aspect important réside dans la mise au point d'étalons de référence pour évaluer ces systèmes avec des processus bien définis et des méthodes de comparaisons qui se veulent objectives.

Néanmoins, si la nécessité de ces campagnes d'évaluation n'est plus à démontrer, on peut tout de même avancer quelques réserves quant à leur évaluation car seul le résultat final obtenu par chacun des systèmes est pris en compte (classement selon les résultats obtenus). Elles ne permettent pas d'évaluer ce qui se passe à l'intérieur des systèmes de façon précise.

3.2 Évaluation de type « glass box » (boîte transparente)

Contrairement aux campagnes présentées, une évaluation de type boîte transparente donne un accès plus fin aux résultats produits par le système. Nous travaillons sur des systèmes composés de plusieurs modules qui s'enchaînent les uns après les autres (traitement de la question, des

[4]http ://trec.nist.gov/
[5]http ://clef.isti.cnr.it/
[6]http ://www.technolangue.net/article195.html/

documents puis extraction de la réponse). On voit alors la pertinence de telles évaluations : le résultat d'un module est pris en entrée par celui qui suit, et le résultat obtenu en aval dépend alors de la qualité des traitements effectués. Ces nouvelles formes d'évaluation deviennent nécessaires (Sparck Jones, 2001).

Il s'agit donc d'observer les résultats produits par les différents modules isolément, sans avoir à lancer le processus dans sa globalité, lequel ne sera pas forcément révélateur des problèmes existants. Une évaluation sélective et partitionnée permet de mesurer l'efficacité réelle de tel ou tel processus, afin de maximiser l'équilibre entre temps de traitement et efficacité. En effet, certains traitements peuvent être longs à s'exécuter sans pour autant être significatifs au niveau des résultats. C'est sur ce point qu'une évaluation interne permet d'améliorer un système. Évaluer de façon pointue permet de visualiser de façon précise ce qui n'est pas correct, mais aussi d'évaluer la « rentabilité » de certains traitements.

Plus précisément, dans le cadre qui nous intéresse, analyser les résultats obtenus par le système va permettre la redéfinition de certaines notions comme le focus, définitions qui seraient trop larges ou bien trop strictes par apport à l'utilisation qui en est fait. De plus, une mise en place d'un procédé d'évaluation transparente permet de modifier certains traitements, d'en ajouter ou bien d'en enlever et de tester la pertinence de ces modifications.

Nous allons appliquer cette méthode évaluative à l'étude de la notion de focus dans le module d'analyse de la phrase dans le système de question-réponse QALC.

4 Application à l'étude de la notion de focus

Le focus est un élément informationnel fort dans un énoncé (qu'il soit écrit ou oral), qui est utilisé pour aider à l'extraction de réponses dans QALC.

4.1 La notion de focus dans la littérature

La notion de focus prend ses origines dans la linguistique, et notamment en syntaxe et en phonologie. On parle de focus et d'opération de focalisation avec des niveaux d'analyse aussi variés que la syntaxe, la sémantique, la phonologie ou la phonétique. Ces disciplines considèrent que le focus est un élément informationnel important de la phrase. Il peut se manifester par une mise en valeur intonative à l'oral, et est l'objet mis en exergue dans les phrases clivées (*C'est cette poupée que je veux*).

Partant de ce constat, Wendy Lehnert a été la première à appliquer ce concept de focus à l'étude des questions pour les systèmes de questions-réponses (Lehnert, 1978). Elle définit alors le focus comme le concept de la question qui représente le besoin d'information exprimé par la question.

Plusieurs systèmes de questions-réponses ont intégré la reconnaissance du focus à leur traitement de la question, et en ont donné une définition.

– Pour (Ferret *et al.*, 2002) :
  « l'élément important de la question, celui qui devra se trouver à proximité de la réponse ».
– Pour (Plamondon *et al.*, 2002) :
  « une portion de la question qui doit obligatoirement figurer près du candidat-réponse [...].

Par exemple, le focus de la question *What was the monetary value of the Nobel Peace Prize in 1989 ?* serait Nobel Peace Prize car l'hypothèse est faite que la réponse correcte devrait se trouver la proximité de l'expression Nobel Peace Prize ou d'une expression sémantiquement apparentée ». Leur système est XR3 [7], premier système de question-réponse développé à l'Université de Montréal.

– Pour (Mendes & Moriceau, 2004) :
« l'élément le plus important de la question i.e. le focus ».

Ces trois définitions, qui viennent de différentes équipes de recherche, mettent en relief l'intérêt de la reconnaissance d'un terme qui doit se trouver dans la réponse : le focus. Elles montrent le lien syntaxique qui peut exister au sein de la phrase réponse entre le focus et la réponse à la question.

4.2 Une (re)définition du focus

4.2.1 Intérêt du focus

Le système QALC se situe dans une perspective d'approche robuste, où peu de connaissances sémantiques sont utilisées. De ce fait, nous essayons de déduire de façon automatique le plus de caractéristiques de la question. Le focus apparaît alors comme un élément important à repérer, nécessaire à la sélection des documents ainsi que pour l'extraction de la réponse, dont il se situe à proximité.

4.2.2 Définition du terme

Le focus est un terme pivot pour extraire la réponse attendue, qui doit apparaître à proximité de la réponse (Ferret *et al.*, 2002). Le système recherche le focus dans la phrase réponse, puis applique des patrons d'extraction par rapport à sa position dans la phrase réponse. Ce terme focus est un élément important pour l'extraction de la réponse courte attendue.

Dans notre approche, il constitue le plus souvent le sujet de la question. En effet, pour *When was the treaty on Conventional Forces in Europe signed ?* le focus est le sujet de la question : *treaty*. Pour des questions de type *Which EU conference adopted Agenda 2000 in Berlin ?* il s'agit alors du complément d'objet direct *Agenda 2000*. Considérer le focus comme le sujet ou l'objet d'un verbe sont des choix effectués afin d'implémenter cette notion de focus.

Dans notre expérience, le focus correspond à un groupe nominal présent dans la question, qui doit apparaître à proximité de la réponse. Nous différençons le focus du type général, qui renseigne sur le type de la réponse attendue. Cette différenciation est très claire si nous l'illustrons d'exemples :

– Which genes cause cancer ?
L'information recherchée est un type de gène. *Genes* constitue le type général de la question. Par contre, le focus, c'est-à-dire le terme autour duquel la réponse à la question s'articule, est *cancer*.

[7] Il s'agit de l'acronyme de eXtraction de Réponses Rapide et Robuste.

– Which US Army Division provided the paratroopers who took part in the invasion of Haiti ?
  Le focus repéré est *paratroopers* et *US Army Division* constitue ici le type général.
  Ce premier est un indice pour trouver la réponse, l'autre permet de renseigner sur la nature de l'information recherchée.

Mais cette définition du focus est-elle suffisante ? C'est ce que nous allons tenter de mesurer en analysant les résultats produits par notre système.

4.3 Observation des données

Nous effectuons dans cette étude une évaluation transparente du module d'analyse des questions en évaluant la reconnaissance ou non du focus et en observant en aval le nombre de réponses correctes obtenues pour les 200 questions dont nous disposons. Nous travaillons sur le corpus de CLEF 2005, campagne à laquelle le LIMSI a participé avec le système QALC. Nous disposons de 200 questions, réparties en 16 catégories.

Après avoir défini ce que nous considérons comme le focus, nous avons constitué une base de données afin d'observer en détail le traitement du focus effectué par le système de questions - réponses QALC.

Pour l'instant, le système reconnaît comme focus le premier groupe nominal rencontré qui est différent du type général. Il s'agit de faire une approximation de la notion de sujet.

Voici un exemple de question qui contient un type général et un focus : *In which year did the Islamic Revolution take place in Iran ?* où *year* constitue le type général et *Islamic Revolution* le focus. L'analyse de la question aboutit bien à la reconnaissance de ces deux éléments. La réponse extraite est correcte : *since the 1979 Islamic Revolution*. Cet exemple nous permet de légitimer la prise en compte de ces deux informations dans l'analyse des questions. Elles sont distinctes et nécessaires à la résolution des questions.

Nous avons effectué une étude de corpus manuelle afin d'observer les cas où la reconnaissance du focus pose problème au système.

4.3.1 Focus non repéré

Dans certains cas, le système ne repère pas le focus. Par exemple, *Which institution initiated the European youth campaign against racism ?* est une question pour laquelle aucun focus n'est spécifié. En effet, comme le premier groupe nominal repéré est le type général, le système ne recherche pas de focus. Il n'y a pas de réponse correcte trouvée pour cet exemple, alors que l'identification de *campaign* (pour se limiter à la tête du focus) aurait pu permettre au système une analyse des réponses plus complète.

4.3.2 Focus erroné

Certains focus identifiés ne correspondent pas à notre définition et génèrent des réponses erronées. C'est le cas pour *Which EU conference adopted Agenda 2000 in Berlin ?* où le système identifie *EU conference* comme focus, alors qu'il s'agit du type général. Or, le type général ne peut pas toujours être utilisé pour extraire des réponses. Il s'agit d'un terme générique - le plus

souvent un hyperonyme - sur lequel porte la question et qui n'est pas forcément présent tel quel dans la réponse.

Si le système attribue comme focus le type général d'une question, il ne pourra logiquement pas trouver de réponse correcte. Dans *Which Russian city is twinned with Glasgow ?* nous voyons bien que *Russian city* ne peut être un terme pivot autour duquel rechercher la réponse. La réponse attendue est une ville russe, mais la phrase réponse ne comportera pas forcément cette information.

Quand le système se trompe de focus, en le confondant avec le type général, il ne trouve pas de réponse.

4.3.3 Focus difficile à déterminer

D'autres questions posent problème quant à la définition même d'un focus. En effet, si l'on prend l'exemple *What newspaper was found in Kiev in 1994 ?* le type général est *newspaper* mais nous n'avons pas de terme focus. Certaines questions sont plus difficiles à traiter comme *According to which government did radioactivity from Chernobyl stop at the Franco-German border ?* Doit-on se focaliser sur *government*, *radioactivity* ou encore *Franco-German border* ? En fonction de la définition donnée du focus, *government* désigne le type attendu de la réponse : nous recherchons l'instance d'un gouvernement. Par contre, en ce qui concerne le focus, le terme doit être lié syntaxiquement à la formulation de la réponse. *Radioactivity* apparaît alors comme un bon candidat. Il s'agit là encore du sujet de la question, qui devra se trouver à proximité de la réponse attendue.

5 Conclusion

La perspective de cette première analyse est de regarder plus finement la définition du focus, de façon à maximiser les résultats obtenus. Il sera intéressant de voir comment formaliser le focus pour répondre aux difficultés soulevées dans cet article, afin que le système le retrouve automatiquement. Cela suppose d'affiner la définition du focus, et de modifier les patrons d'extraction de la réponse qui lui sont liés. A plus long terme, nous nous intéressons au développement d'une méthodologie d'évaluation transparente des systèmes de questions-réponses, de façon à affiner les traitements, et essentiellement à proposer des traitements différents selon les questions. Il s'agira de revoir la typologie effective des questions en fonction d'éléments comme la structure syntaxique de la réponse ou encore le type de focus. Surtout, nous pensons également qu'une étude fine des réponses obtenues ainsi que des réponses qu'il aurait fallu obtenir pourra nous permettre de valider ou encore une fois d'affiner notre définition du focus, ainsi que de mesurer la pertinence de cette information dans notre système.

Nous nous proposons de plus d'affiner notre définition du focus en fonction des réponses obtenues. Il serait intéressant de pouvoir par la suite réaliser une étude comparative avec un autre système de questions-réponses : RITEL[8] (Rosset *et al.*, 2006) afin de tester la validité de nos hypothèses.

En nous inscrivant dans une démarche d'évaluation transparente de systèmes de questions-

[8]Plus d'informations sur http ://ritel.limsi.fr/.

réponses, nous nous sommes intéressée à l'étude du focus, élément essentiel pour l'extraction de la réponse attendue de la question. Cette première observation nous a permis de voir que la redéfinition du concept utilisé permet d'affiner les résultats et de les améliorer. Cette démarche n'est pas possible lors d'évaluations de type boîte noire, or elle est plus que nécessaire pour l'amélioration des systèmes, en particulier des questions-réponses qui sont formés de plusieurs composants qui interagissent entre eux de manière séquentielle. Ce type d'évaluation apparaît bel et bien comme complémentaire aux évaluations de type boîte noire, afin d'améliorer finement les traitements effectués dans l'optique, par la suite, de l'obtention d'une meilleure performance globale.

Remerciements

Un énorme merci à Anne-Laure Ligozat pour son aide, ses encouragements et ses bons conseils. Grand merci aussi à Brigitte Grau et Benoît Habert pour leur aide et leurs relectures.

Références

BERTHELIN J.-B., GRAU B. & HURAULT-PLANTET M. (2001). Two levels of evaluation in a complex NL system. *Workshop on Evaluation for Language and Dialogue Systems*.

FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G. & JACQUEMIN C. (2001). Utilisation des entités nommées et des variantes terminologiques dans un système de question-réponse. *Actes de TALN*.

FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G., MONCEAUX L., ROBBA I. & VILNAT A. (2002). Recherche de la réponse fondée sur la reconnaissance du focus de la question. *Actes de TALN*.

GRAU B. (2004a). Evaluation des systèmes de question-réponse. In *Évaluation des systèmes de traitement de l'information*, chapter 3, p. 77–98. Hermès.

GRAU B. (2004b). Les systèmes de question-réponse. In *Méthodes avancées pour les systèmes de recherche d'informations*, chapter 10, p. 189–218. Hermès.

HARABIGIU S. & MOLDOVAN D. (2003). Question Answering. *Revue Computational Linguistics*.

HARABIGIU S., MOLDOVAN D., PASCA M. & SURDEANU M. (2003). Performance Issues and Error Analysis in an Open-Domain Question Answering System. *ACM Transactions on Informations Systems*.

LEHNERT W. (1978). *The Process of Question Answering : A Computer Simulation of Cognition*. John Wiley & Sons Inc.

MENDES S. & MORICEAU V. (2004). L'analyse des questions : intérêt pour la génération des réponses. *Workshop Question-Réponse*.

PLAMONDON L., KOSSEIM L. & LAPALME G. (2002). The quantum question answering system at trec-11. In E. M. VORHEES & D. K. HARMAN, Eds., *Proceedings of the Eleventh Text Retrieval Conference (TREC-2002)*, p. 750–757, Gaithersburg, Maryland : NIST.

ROSSET S., GALIBERT O., GABRIEL I. & MAX A. (2006). Interaction et recherche d'information : le projet RITEL. *Revue TAL*.

SPARCK JONES K. (2001). Automatic language and information processing : rethinking evaluation. In *Natural Language Engineering*, chapter 7, p. 1–18.

VOORHEES E. M. & HARMAN D. K. (2005). *TREC : Experiment and Evaluation in Information Retrieval*. MIT Press.