# Systèmes de questions-réponses: vers la validation automatique des réponses

Anne-Laure Ligozat, Brigitte Grau, Isabelle Robba, Anne Vilnat

# Systèmes de questions-réponses : vers la validation automatique des réponses

Anne-Laure LIGOZAT, Brigitte GRAU, Isabelle ROBBA, Anne VILNAT
LIMSI-CNRS - BP 133, 91403 Orsay Cedex
prenom.nom@limsi.fr

**Résumé.** Les systèmes de questions-réponses (*SQR*) ont pour but de trouver une information précise extraite d'une grande collection de documents comme le Web. Afin de pouvoir comparer les différentes stratégies possibles pour trouver une telle information, il est important d'évaluer ces systèmes. L'objectif d'une tâche de validation de réponses est d'estimer si une réponse donnée par un *SQR* est correcte ou non, en fonction du passage de texte donné comme justification. En 2006, nous avons participé à une tâche de validation de réponses, et dans cet article nous présentons la stratégie que nous avons utilisée. Celle-ci est fondée sur notre propre système de questions-réponses. Le principe est de comparer nos réponses avec les réponses à valider. Nous présentons les résultats obtenus et montrons les extensions possibles. À partir de quelques exemples, nous soulignons les difficultés que pose cette tâche.

**Abstract.** Question answering aims at retrieving precise information from a large collection of documents, typically the Web. Different techniques can be used to find relevant information, and to compare these techniques, it is important to evaluate question answering systems. The objective of an Answer Validation task is to estimate the correctness of an answer returned by a QA system for a question, according to the text snippet given to support it. We participated in such a task in 2006. In this article, we present our strategy for deciding if the snippets justify the answers. We used a strategy based on our own question answering system, and compared the answers it returned with the answer to judge. We discuss our results, and show the possible extensions of our strategy. Then we point out the difficulties of this task, by examining different examples.

**Mots-clés :** systèmes de questions-réponses, validation de réponses.

**Keywords:** question answering, answer validation.

## 1 Introduction

Les systèmes de questions-réponses (*SQR* par la suite) ont pour but de trouver une information précise dans une grande collection de documents. L'hypothèse sous-jacente au développement de tels systèmes est que les utilisateurs préfèrent en général recevoir une réponse précise à la question qu'ils se posent plutôt qu'un ensemble de documents à explorer, comme le proposent habituellement les moteurs de recherche (Voorhees, 1999). Cependant, pour être considéré comme fiable par un utilisateur, un *SQR* doit être capable de donner des éléments permettant d'évaluer ses réponses. L'objectif d'un système ne doit donc pas seulement être de trouver

les réponses, mais aussi de les exprimer d'une façon qui permette à l'utilisateur de savoir s'il peut avoir confiance en ces réponses. Ces éléments de justification donnent à l'utilisateur un moyen de vérifier que la réponse fournie correspond bien à l'information qu'il cherche, et ainsi de donner une valeur de vérité à cette réponse, en supposant que l'utilisateur a des connaissances « standard ».

Une bonne justification doit être concise et complète. Le but est de ne fournir que les extraits de documents qui permettent à l'utilisateur de retrouver toutes les informations qu'il a données, sans avoir à lire un document entier. Voici un exemple d'une telle justification.

> Question : Quand a eu lieu la chute du mur de Berlin ?
>
> Réponse : **en 1989**
>
> Justification (passage d'un document) : Cette ère de la dissuasion, fondée sur l'équilibre de la terreur entre deux grands blocs antagonistes, est remise en question **en 1989**, avec la chute symbolique du mur de Berlin.

## 2 Validation de réponses

(Lin & Pantel, 2001) soulignent la possible distance linguistique entre les questions et leurs réponses accompagnées de leur justification, en prenant l'exemple de la phrase *« Stendhal a écrit 'La chartreuse de Parme' en 1838 »* justifiant la réponse *« Stendhal »* à la question *« Qui est l'auteur de 'La chartreuse de Parme' ? »*. Ils définissent les liens entre une question et sa réponse justifiée par le terme d'*inférence*. Ils proposent alors de définir des *règles d'inférence* pour reconnaître par exemple la relation entre « X a écrit Y » et « X est l'auteur de Y ». Ces règles correspondent plus ou moins à ce qui est appelé *paraphrases* ou *variantes* dans d'autres travaux (Jones & Tait, 1984; Fabre & Jacquemin, 2000).

Le lien entre question et réponses correspond à la notion de *textual entailment* telle qu'elle est définie par Pascal Recognizing Textual Entailement Challenge [1] (RTE). *L'implication textuelle* est définie comme une tâche de décision qui à partir de deux fragments de texte, estime si d'un point de vue sémantique on peut déduire l'un de l'autre. Ainsi le passage de texte suivant (appelé *justification*) : *« Yoko Ono a inauguré une statue de bronze représentant son mari décédé, John Lennon, pour compléter le changement de nom officiel de l'aéroport de Liverpool qui devient l'aéroport John Lennon de Liverpool »* implique la phrase *« Yoko Ono est la veuve de John Lennon »* (appelée *hypothèse* dans le contexte de l'implication textuelle). Dans RTE, les participants reçoivent des paires justification-hypothèse de ce type et doivent ensuite décider si les hypothèses peuven ! t ou non être déduites des justification. Cette tâche est similaire à la tâche de réponses aux questions en ce qui concerne les questions booléennes (attendant *oui* ou *non* en réponse), car répondre à ces questions revient en fait à décider si la justification de la réponse implique la réponse.

En 2006, un nouvel exercice de validation des réponses, AVE [2], a été introduit dans la campagne de questions-réponses de CLEF. Le but de cet exercice est d'une part d'améliorer les performances des *SQR*, en développant des méthodes automatiques d'évaluation des réponses, et d'autre part de rendre le jugement humain semi-automatique à la condition que l'exercice produise des méthodes fiables d'évaluation. Pour cet exercice, les organisateurs ont produit un

[1] `http://www.pascal-network.org/Challenges/RTE`

[2] Answer Validation Exercise, `http://nlp.uned.es/QA/AVE/`

corpus à partir des réponses des participants à la tâche de questions-réponses et des passages de texte donnés comme justification. Les participants avaient alors pour tâche de décider pour chaque réponse si elle était correcte ou non en fonction du passage justificatif.

Les premiers travaux de validation automatique de réponses ont eu lieu au cours de la campagne AVE en 2006 ; cependant, les campagnes d'implication textuelle RTE avaient déjà proposé ce type de tâche.

Voici un exemple de couple (hypothèse, justification) d'AVE :

> Hypothèse : Yasser Arafat était **leader de l'Organisation de Libération de la Palestine** [3]
>
> Justification : Le président Clinton a fait appel personnellement au **leader de l'Organisation de Libération de la Palestine** Yasser Arafat et aux Palestiniens mercredi pour qu'ils reprennent les pourparlers en faveur de la paix avec Israël

Ici l'hypothèse est une reformulation de la question *« Qui était Yasser Arafat ? »* dans laquelle a été insérée une réponse proposée par un système *« leader de l'Organisation de Libération de la Palestine »*.

Dans AVE, le corpus de paires justification-hypothèse a été construit semi-automatiquement à partir des réponses obtenues par les participants lors de QA@CLEF 2006, campagne d'évaluation des *SQR*. Le corpus contient environ 3000 paires. Les participants à AVE ont été évalués sur leur capacité à prédire si une réponse (attestée par des juges humains) était correcte ou non. Ils avaient donc pour chaque paire deux possibilités de réponse : OUI ou NON.

Les résultats ont été évalués par la précision, le rappel et la f-mesure qui ont été calculés de la façon suivante :

$$précision = \frac{\#paires\ jugées\ OUI\ correctement}{\#jugées\ comme\ OUI}, rappel = \frac{\#paires\ jugées\ comme\ OUI\ correctement}{\#paires\ OUI}$$

$$\text{et } f\text{-}mesure = \frac{2*précision*rappel}{précision+rappel}$$

# 3 Travaux en validation de réponses

(Peñas *et al.*, 2006) présentent le déroulement de la première campagne AVE. 11 groupes ont participé à ce premier essai en soumettant 38 runs dans 7 langues différentes. L'anglais et l'espagnol étaient les langues les plus représentées avec respectivement 11 et 9 runs soumis. 2 groupes ont proposé des runs dans les 7 langues : ce sont les universités de Twente et d'Alicante.

Dans chaque langue, les paires *justification-hypothèse* ont été construites à partir des soumissions à la tâche questions-réponses de la campagne CLEF 2006. De ce fait, le pourcentage de paires positives, négatives et non évaluées [4] peut-être variable d'une langue à l'autre, ce qui ne permet pas réellement la comparaison des systèmes ayant participé dans des langues distinctes. Voici par exemple les pourcentages pour les 3 langues où les différences sont les plus importantes :

[3] Dans nos exemples, la réponse est écrite en **gras**.

[4] Les paires non évaluées de AVE proviennent de runs qui n'ont pu être évalués lors de la campagne QA@CLEF. En anglais et en portugais ce nombre est très élevé : 35% et 40%.

– en hollandais, OUI : 10%, NON : 86%, NON ÉVALUÉES : 4% ;
– en anglais, OUI : 10%, NON : 55%, NON ÉVALUÉES : 35% ;
– en espagnol, OUI : 28%, NON : 68%, NON ÉVALUÉES : 4% ;

Différentes approches ont été adoptées dans cette campagne. L'approche logique obtient les meilleurs résultats (Tatu *et al.*, 2006) [5], soulignons qu'elle est très souvent accompagnée de connaissances linguistiques : elles servent à transformer les éléments textuels en représentation logique. Au moins 3 équipes ont utilisé logique et connaissances linguistiques. Les approches qui utilisent de l'apprentissage sont également au nombre de 3 et l'une d'entre elle s'est attaquée aux 7 langues proposées. Elles utilisent des corpus déjà annotés comme ceux des campagnes RTE. Une approche, qui a participé elle aussi dans les 7 langues, adopte une méthode fondée sur les paraphrases : celles-ci sont engendrées automatiquement à partir de corpus bilingues alignés. Deux approches au moins utilisent des connaissances linguistiques sans faire référence à l'utilisation de la logique. Partant du constat qu'en espagnol 75% des questions de la campagne QA@CLEF étaient factuelles, une approche s'est fondée uniquement sur la reconnaissance d'entités nommés.

(Tatu *et al.*, 2006) utilisent un mécanisme de reconnaissance des entités nommées, un analyseur syntaxique et un analyseur sémantique pour transformer le passage justificatif et l'hypothèse en une représentation logique qu'ils qualifient de riche. Les représentations sont ensuite soumises à COGEX, qui détermine si oui ou non la justification implique l'hypothèse. La plupart des erreurs commises par ce système sont dues à une mauvaise syntaxe des hypothèses (celles-ci sont construites automatiquement), qui entraîne la construction de représentations logiques erronées. Néanmoins ce système obtient les meilleurs résultats dans les 2 langues dans lesquelles il a participé. En anglais, il obtient une f-mesure de 0.4393 et en espagnol une f-mesure de 0.6063.

(Ferrandez *et al.*, 2006) dérivent également une forme logique à partir du passage justificatif et de l'hypothèse. Pour cela, ils utilisent l'analyseur de Lin, MINIPAR (Lin, 2005), et obtiennent une représentation des phrases sous la forme d'un ensemble de relations de dépendances. Les relations sont ensuite transcrites dans des formes logiques, puis une mesure de similarité est calculée, celle-ci produit un poids sémantique utilisé pour juger si le passage justificatif implique ou non l'hypothèse. Ils ont soumis des runs dans toutes les langues et obtenu les meilleurs résultats en français (f-mesure : 0.4693) et en italien (f-mesure : 0.4066).

Pour leur participation à AVE, (Kouylekov *et al.*, 2006) ont adopté une approche fondée sur la notion de distance : ils essayent d'effectuer un *mapping* entre le contenu de l'hypothèse et la justification. Ils soulignent que plus ce *mapping* est direct plus il est probable que la justification implique l'hypothèse. Le *mapping* consiste ici en une séquence d'opérations d'édition, chacune ayant un coût. Les opérations (insertion, suppression, substitution) sont appliquées sur les arbres de dépendances du passage justificatif et de l'hypothèse. Quand le coût total de ces opérations est en dessous d'un seuil fixé, le passage justificatif est considéré comme impliquant l'hypothèse. Malgré différents problèmes dans la mise en place de ces modules, ils ont obtenu la 3ème place en anglais avec une f-mesure de 0.3776.

[5] Tous les articles évoqués dans ce paragraphe ne seront pas tous référencés, mais ils sont rassemblés dans les notes de travail du workshop CLEF 2006 et sont consultables à l'adresse `http://www.clef-campaign.org/2006/working_notes/CLEF2006WN-Contents.html`

Comme cela a été dit dans l'introduction, il existe une forte connexion entre AVE et RTE. La proposition à l'origine d'AVE était que l'on pouvait reformuler la tâche de validation de réponse comme un problème d'implication textuelle. Et, plusieurs groupes ont d'ailleurs participé aux deux évaluations en utilisant la même approche.

En 2006, a été organisé le second RTE. (Bar-Haim *et al.*, 2006) soulignent les particularités des deux systèmes qui ont obtenu les meilleurs résultats. L'un a utilisé de façon extensive des connaissances sémantiques, l'autre a favorisé l'utilisation de grands corpus d'entraînement.

Dans notre travail, nous ne faisons pas l'hypothèse d'une source de connaissances sémantiques existante qui permettrait des déductions logiques. Aussi, nous reposons nous sur des critères linguistiques, qui peuvent être vérifiés en domaine ouvert, et qui permettent d'exprimer des relations sémantiques entre le sens des mots.

# 4 Valider des réponses avec un *SQR*

Notre objectif était d'utiliser notre propre *SQR* pour le français : FRASQUES, et d'utiliser ses résultats, c'est-à-dire à la fois les réponses extraites et les types d'informations de la questions présentes dans les justifications, pour évaluer la pertinence des justifications par rapport aux hypothèses.

## 4.1 FRASQUES : notre système de questions-réponses pour le français

Nous présentons tout d'abord brièvement FRASQUES avant de présenter comment il a été adapté pour la tâche de validation.

Le système se divise en 4 composants :
- Analyse de la question : ce premier module effectue l'analyse syntaxique de la question pour en détecter certaines de ses caractéristiques telles que :
  - ses mot-clés, utilisés ultérieurement lors de la recherche des documents,
  - le type attendu de la réponse, qui peut-être une entité nommée (une personne, un pays, une date...) ou un type général comme *conférence* ou *adresse*,
  - le focus de la question, que nous définissons comme le terme de la question qui sera vraisemblablement présent dans la phrase contenant la réponse,
  - le verbe principal de la question.
- Sélection des documents : le moteur de recherche Lucene [6] cherche dans la collection les documents pertinents.
- Traitement des documents : ce module utilise Fastr [7] pour reconnaître les variantes linguistiques des termes de la question : par exemple, « monnaie de l'Europe » sera reconnue comme une variante de « monnaie européenne ». Ensuite, les entités nommées du document sont étiquetées, nous utilisons environ une vingtaine de type d'entités nommées. Les phrases contenant au moins une variante des termes de la question sont gardées.
- Extraction de la réponse : ce dernier module extrait les réponses précises des phrases candidates. La stratégie d'extraction dépend du type attendu de la réponse. Si la réponse est une entité nommée, l'entité nommée qui est du type attendu et qui est la plus proche des mots de

[6] Moteur de recherche entièrement écrit en Java `http://lucene.apache.org/`
[7] `http://www.limsi.fr/Individu/jacquemi/FASTR/`

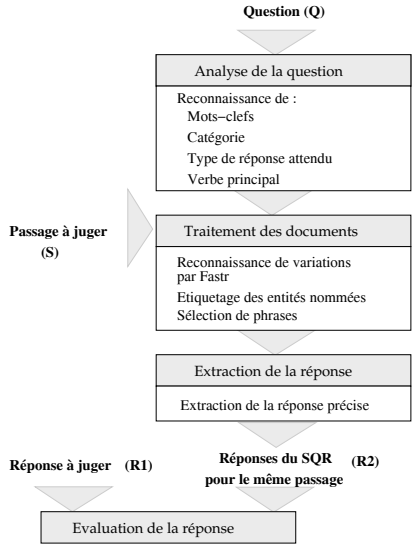


FIG. 1 – Architecture du système de validation de la réponse

la question est sélectionnée. Sinon, des patrons d'extraction sont utilisés, ils sont écrits dans le format Cass [8], un analyseur syntaxique qui est utilisé ici pour extraire la réponse plutôt que comme analyseur. Ces patrons expriment la position possible de la réponse par rapport au focus ou au type attendu de la réponse.

## 4.2 Le système de validation des réponses

Le système de validation des réponses utilise trois de ces quatre composants, ce que montre la figure 1. L'entrée du système est une paire justification-hypothèse, ainsi que la question d'origine *Q* et la réponse à juger *R1*. La question est d'abord analysée puis le composant qui traite les documents est appliqué à la justification. Le module d'extraction de la réponse extrait les réponses *R2* des passages justificatifs. Enfin, la paire hypothèse-justification est évaluée en tenant compte des différentes informations de l'hypothèse trouvées dans l'extrait et de la réponse trouvée par FRASQUES. Le système retourne OUI si elle est considérée comme justifiée, NON dans le cas contraire. Un score de confiance est également attribué à chaque jugement.

L'algorithme de décision se déroule en 2 étapes. La première a pour but de détecter les erreurs les plus triviales, par exemple une réponse qui serait complètement incluse dans la question ou qui ne serait pas présente dans la justification. Dans le cas où la question contient une date, le contexte temporel de la question et l'extrait sont comparés. Pour l'instant, le contexte est formé par les dates reconnues comme telles présentes dans la description du document ou dans le passage. S'ils sont contradictoires, la paire est rejetée.

[8] `http://www.sfs.nphil.uni-tuebingen.de/~abney/`

La seconde étape consiste en des vérifications plus complexes. Dans un cas idéal, un passage justificatif correct correspond à la reformulation de la question sous forme déclarative avec la réponse qui y est donnée. Chaque terme de la question, ou de l'hypothèse, figure dans le passage, liés par les mêmes relations.

En ce qui concerne les termes, dans la grande majorité des cas, le passage justificatif ne comporte pas tous les termes de la questions sous leur forme d'origine : ils subissent des variations de différentes natures : flexionnelles, morphologiques, syntaxiques, sémantiques ou des combinaisons de ces variations si on recherche des groupes nominaux complexes. Dans FRASQUES, ces variations sont reconnues par Fastr. Parmi les termes de la question, certains jouent un rôle plus important. Il en est ainsi de l'objet de la question, que nous appelons focus dans FRASQUES. Le focus correspond à l'entité sur laquelle porte la question, que l'on en cherche une caractéristique ou une définition. Aussi, selon les types de question, le focus n'est pas toujours présent, mais s'il l'est, il doit figurer dans le passage justificatif. Un autre terme qui, s'il est présent, a une grande importance, est le type de réponse attendu, quand ce type n'est pas un nom d'entité nommée. Ce type est nommé type général. Ainsi, dans « De quel parti politique Lionel Jospin est-il membre ? » Le focus est « Lionel Jospin » et le type général est « parti politique ». Lorsqu'il est présent dans le passage réponse, le type général sera souvent placé à proximité de la réponse ou même fera partie de celle-ci, comme dans « Lionel Jospin, membre du parti socialiste ».

Lorsqu'il s'agit du verbe, celui-ci a tendance à subir plus de variations que les termes nominaux ; il est souvent exprimé par une préposition ou un verbe proche mais non synonyme. C'est le cas par exemple si on demande « qui a réalisé un film » et que la réponse est exprimée par « le film de X ... » ou « quelle entreprise a changé son nom » et la réponse est donnée par « le groupe X a adopté le nom de la filiale ... ». On retrouve ici les variations traitées par (Lin & Pantel, 2001). Ne disposant pas de telles ressources, nous avons considéré que l'absence du verbe n'influerait pas sur la décision finale.

Enfin les derniers types de termes jouant un rôle primordial sont les noms propres : ils sont toujours présents dans le passage et subissent peu de variations, sauf en ce qui concerne les noms de pays souvent repris par l'adjectif correspondant, comme dans « qui est le président de l'Egypte » avec « le président égyptien » repris dans le passage.

En ce qui concerne les relations entre termes, celles-ci seront souvent vérifiées par leur manifestation en langue, c'est-à-dire par un ensemble de relations syntaxiques. Nous avons vu que certains travaux s'appuient sur une notion de distance syntaxique. Mais pour cela, il est nécessaire de disposer d'une analyse complète des phrases. Afin de ne pas reposer sur cette hypothèse souvent non vérifiée, nous avons choisi de ne vérifier que certaines relations en les exprimant sous forme de patrons d'extraction. Ces relations sont celles qui lient la réponse avec certains éléments de la phrase : le focus ou le type général.

L'élément prépondérant, malgré tout, reste la réponse : est-elle du type attendu ou non ? Lorsque ce type est une entité nommée, la vérification consistera à retrouver une entité nommée d'un type adéquat. Lorsque celui-ci est désigné par le type général, ou bien il figure à proximité de la réponse, ou bien il est implicite et la réponse en est une instance. Cette relation d'instanciation pourrait être inférée par l'utilisation de ressources externes, par exemple Wikipedia, qui possède un grand nombre de catégories et de définitions leur correspondant.

La mise en oeuvre de ces critères de justification donne lieu dans notre système à un calcul de 2 scores qui permet ensuite de conclure positivement ou négativement. Le premier score

porte sur l'évaluation de la correspondance entre la réponse trouvée par notre système *R2* et la réponse proposée dans l'hypothèse *R1*. Si FRASQUES trouve une réponse différente, alors la paire hypothèse-justification est réfutée. Si les 2 réponses sont proches ou s'il n'y a pas de réponse trouvée par FRASQUES, la décision va être conditionnée par la présence des différents termes que nous avons privilégiés. Le score attribué à l'évaluation de la qualité de la réponse sera positif pour une réponse exacte ou approchée, et négatif quand la distance est assez grande, par exemple, l'approximation d'une date ou d'une quantité par un nombre.

Le deuxième score évalue les termes présents. Il est calculé en combinant le nombre de critères présents et leurs valeurs. Il est négatif si aucun des critères n'est trouvé, et positif sinon.

Un passage constitue une justification acceptable :
– si *R2* est absente et le score des termes est positif. Ce dernier fournit le score final,
– si *R2* = *R1* et il y a des critères présents. Dans ce cas le score final est la valeur maximale des deux critères,
– si les 2 scores vont dans des sens opposés, on prend le meilleur des deux, s'il est positif.
Regardons l'exemple suivant :

> Justification : Trois candidats, Tony Blair, Margaret Beckett et John Prescott, se disputeront la succession de John Smith à la tête du parti travailliste, a annoncé le **Labour** jeudi, à l'issue du processus de nominations des candidats par les députés du parti. M. Blair, ministre de l'Intérieur du cabinet fantôme représentant l'a
>
> Hypothèse : le parti politique de Tony Blair, le **LABOUR** .

Dans ce passage, tous les termes de la question sont présents (*Tony Blair, parti politique*), mais le type de la réponse *Labour* n'est pas étiqueté par notre *SQR* comme une entité nommée de type organisation. De ce fait, l'algorithme de décision reçoit 2 scores opposés, dans cet exemple prenant en compte le non-marquage de *Labour* comme une organisation, il répond négativement.

## 4.3 Résultats

Le corpus d'évaluation contenait 3266 paires, parmi lesquelles 202 paires n'ont pas été jugées. Les hypothèses étant formées automatiquement, elles comportaient beaucoup d'erreurs de syntaxe, aussi nous sommes-nous fondés uniquement sur les questions, l'hypothèse ne nous permettant que d'extraire la réponse.

Lors de notre participation à AVE, beaucoup d'erreurs restaient dans nos programmes, qui ont été corrigés depuis. Les organisateurs ayant fourni les valeurs de validation attendues pour chaque paire du corpus, nous avons pu réévaluer notre chaîne. Un examen approfondi de ces résultats nous a permis de constater qu'il y avait certaines erreurs sur ces valeurs, notamment en ce qui concerne les réponses positives : des réponses exactes aux questions n'étaient pas du tout validées par le passage justificatif. Nous avons corrigé 82 d'entre elles dans le corpus.

La table 1 présente nos résultats sur la version officielle, les corrections apportées au corpus ne modifiant pas les ordres de grandeur des résultats. La première ligne donne le nombre de paires évaluées positivement et négativement par les juges humains. La seconde ligne contient tous nos résultats et la suivante le nombre de nos résultats corrects. La dernière ligne contient le rappel et la f-mesure correspondant à ces résultats en utilisant la formule exposée dans la section 2.

| | # OUI | # NON | Total |
|---|---|---|---|
| Évalués par les organisateurs | 705 | 2359 | 3064 |
| Tous nos résultats | 142 | 2922 | 3064 |
| Nos résultats corrects | 82 | 2266 | 2348 |
| Précision | 0.58 | 0.77 | |
| Rappel | 0.12 | 0.96 | |
| F-mesure | 0.2 | 0.85 | |

TAB. 1 – AVE results at CLEF 2006

Parmi nos réponses NON, nous distinguons celles qui sont sûres des autres : ce sont les réfutations décidées lors de la première étape présentées dans la section 4.2. La réponse est considérée comme étant non justifiée et ce de façon sûre, donc avec un score de confiance élevé. Nous avons trouvé 1637 paires de « NON » sûrs. Parmi elles, 1415 étaient bien jugées, la précision pour ces réponses est donc de 0,87.

La seconde observation est que notre système a plus de facilités pour réfuter les justifications plutôt que pour les accepter. La précision et le rappel de nos réponses négatives sont bons. Et nous nous trompons rarement quand nous donnons des réponses OUI, mais nous en trouvons très peu, notre rappel est donc très faible sur ces réponses.

Certaines erreurs pourraient être corrigées en approfondissant les vérifications des relations portant sur la réponse. Par exemple, pour la question *« Quel est le nom de la femme de George W. Bush ? »*, une des hypothèses construites était *« **Norman Schwarzkopf**, la femme de George W. Bush. »*. On pourrait alors interroger le Web avec la requête *femme de George W. Bush* et constater que la ou les réponses obtenues sont fortement incompatibles avec ***Norman Schwarzkopf***.

# 5 Conclusion

Nous avons présenté une stratégie de validation des réponses issues d'un *SQR*. Cette stratégie est fondée sur FRASQUES, notre propre *SQR* monolingue : l'hypothèse et l'extrait sont analysés par FRASQUES et nous utilisons des critères qui permettent de détecter si l'extrait justifie ou non la réponse. Dans notre évaluation des paires hypothèses-extrait, nous distinguons avec une bonne précision les cas dans lesquels l'extrait ne justifie pas la réponse. Des possibilités d'extension de notre stratégie, utilisant des ressources externes et nous permettant d'acquérir de nouvelles connaissances ont également été présentées.

Cette première expérience en validation de réponses constitue une étape vers la validation semi-automatique en questions-réponses. Elle nous permettra à terme d'améliorer les performances de notre *SQR* puisque certains des critères que nous utilisons pour la validation n'y avaient pas été mis en œuvre.

# Références

BAR-HAIM R., DAGAN I., DOLAN B., FERRO L., GIAMPICCOLO D., MAGNINI B. & SZPEKTOR I. (2006). The second pascal recognising textual entailment challenge. In *The Second*

*PASCAL Challenges Workshop on Recognising Textual Entailment*.

FABRE C. & JACQUEMIN C. (2000). Boosting Variant Recognition with Light Semantics. In *Proceedings of 18th International Conference on Computational Linguistics (COLING-2000)*, Sarrebrück, Allemagne.

FERRANDEZ O., TEROL R. M., MUNOZ R., MARTINEZ-BARCO P. & PALOMAR M. (2006). A knowledge-based textual entailment approach applied to the qa answer validation at clef 2006. In *Workshop CLEF 2006*, Alicante, Spain.

JONES K. S. & TAIT J. I. (1984). Automatic Search Term Variant Generation. *Journal of Documentation*, p. 50–66.

KOUYLEKOV M., NEGRI M., MAGNINI B. & COPPOLA B. (2006). Towards entailment-based question answering : Itc-irst at clef 2006. In *Workshop CLEF 2006*, Alicante, Spain.

LIN D. (2005). Dependancy-based evaluation of minipar. In *Workshop on the Evaluation of Parsing Systems*, Southampton, UK.

LIN D. & PANTEL P. (2001). Discovery of inference rules for question-answering. *Natural Language Engineering*, **7**(04), 343–360.

PEÑAS A., RODRIGO A., SAMA V. & VERDEJO F. (2006). Overview of the answer validation exercise 2006. In *Workshop CLEF 2006*, Alicante, Spain.

TATU M., ILES B. & MOLDOVAN D. (2006). Automatic answer validation using cogex. In *Workshop CLEF 2006*, Alicante, Spain.

VOORHEES E. M. (1999). TREC-8 Question Answering Track Evaluation. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)* : Department of Commerce, National Institute of Standards and Technology.