Validation du type de la réponse dans un système de questions réponses

Arnaud Grappy, Brigitte Grau

Validation du type de la réponse dans un système de questions réponses

Arnaud Grappy [1] — Brigitte Grau[1,2]

[1] *LIMSI-CNRS*
BP 133 ORSAY CEDEX

[2] *École Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE)*
1 square de la résistance
91025 EVRY Cedex

prenom.nom@limsi.fr

RÉSUMÉ. *Les systèmes de questions réponses recherchent la réponse à une question posée en langue naturelle dans un ensemble de documents. Certaines questions attendent une réponse d'un certain type, explicité dans la question. La méthode présentée dans cet article vérifie que la réponse renvoyée correspond bien au type cherché. Pour cela elle suit une approche par apprentissage automatique en utilisant trois types de critères. Les premiers sont statistiques et fondés sur la fréquence d'apparition de la réponse avec le type dans un ensemble de documents. Les seconds relèvent de la reconnaissance des entités nommées et les derniers utilisent l'encyclopédie Wikipédia. L'évaluation globale, 80% de résultats corrects, montre l'intérêt de la méthode.*

ABSTRACT. *In open domain question-answering systems, numerous questions wait for answers of an explicit type. The method we present in this article aims at verifying that an answer given by a system corresponds to the given type. This verification is done by combining criteria provided by different methods dedicated to verify the appropriateness between an answer and a type. The first types of criteria are statistical and are based on the frequency of both the answer and the type in documents, other criteria rely on named entity recognizers and the last criteria are based on the use of Wikipedia. The method obtains good results (80%).*

MOTS-CLÉS : *système de questions réponses, type de réponses, validation de réponses*

KEYWORDS: *question answering system, answer type, answer validation*

1. Introduction

Les systèmes de questions réponses (SQR) recherchent, dans un ensemble de documents, la réponse à une question posée en langage naturel (exemple : Quel président succéda à Jacques Chirac ?). Le mécanisme d'extraction des réponses consiste, dans un premier temps, à obtenir un court passage de texte. Puis, dans un second temps, la réponse est extraite de ce passage.

Une stratégie commune à tous les systèmes consiste à déterminer le type de réponse attendu qui est relié aux types d'entités nommées reconnues dans les textes. Plus un système sait reconnaître de types d'entités nommées, meilleures sont ses performances ((Harabagiu *et al.*, 2000), (Hovy *et al.*, 2001), (Sekine *et al.*, 2002)). Toutefois, on ne peut envisager d'établir *a priori* une liste exhaustive de tous les types de réponses pouvant être demandés. D'autre part, on ne sait pas reconnaître des occurrences de tous ces types, en dehors de celles correspondant à un type d'entité nommée, même si l'ensemble classiquement défini par (Grishman *et al.*, 1995) a été étendu dans les systèmes de questions réponses avec de nouveaux types tels que des titres de film, de livre, et des sous-types plus précis tels que acteur, écrivain, chanteur, etc.

Il est donc crucial pour un SQR de pouvoir procéder à une vérification dynamique du type des réponses candidates afin de mieux filtrer ses propositions ou de valider les réponses proposées. Nous nous sommes placés dans ce dernier cadre afin d'appliquer le travail que nous proposons sur la vérification de type. La validation de réponses vérifie *a posteriori* que la réponse donnée par un système de questions réponses est valide, c'est à dire qu'elle répond bien à la question et qu'elle peut être extraite depuis le fragment de texte sélectionné. Par exemple, la question « Quel président succéda à Jacques Chirac ? » attend une entité nommée qui désigne un président. La vérification de cette contrainte de type permettra, par exemple, de voir que la réponse « Michel Rocard » extraite du passage « Michel Rocard succède à Jacques Chirac au poste de Premier Ministre » ne répond pas à la question. Ce filtrage a pour but à terme de diminuer le nombre de mauvaises réponses renvoyées par un système de questions réponses et donc de l'améliorer.

La vérification du type ne peut pas toujours être réalisée en n'exploitant que le passage contenant la réponse. En effet, une étude des passages renvoyés par des SQR (Grau *et al.*, 2008) et provenant de la campagne EQueR (Ayache *et al.*, 2006) montre que dans les passages où il manque un mot de la question (au nombre de 111 sur 500), il s'agit du mot désignant le type dans 27% des cas. De même, l'étude de corpus présentée par Grappy *et al.* (2010) qui porte sur un corpus où ont été annotés les problèmes que peuvent rencontrer les SQR pour justifier leurs réponses, tels que variations sémantiques, anaphores ou le fait qu'une information soit manquante, fait apparaître que la réponse n'est pas explicitement typée dans 11,5% des cas.

L'approche présentée dans cet article consiste à utiliser différentes méthodes pour vérifier le type de réponse, qui produiront chacune un avis, avis qui seront combinés par apprentissage pour donner une décision finale. Certaines sont d'ordre statistique et étudient la présence commune de la réponse et du type dans un ensemble de docu-

ments. D'autres, plus précises, reposent sur des systèmes de reconnaissance d'entités nommées, soit pour rejeter des réponses qui ne sont pas du bon type, soit pour constituer une base de connaissances. Un troisième type de méthode utilise l'encyclopédie Wikipédia[1] afin de détecter la présence de la réponse et du type dans des mêmes pages, soit à l'aide de patrons d'extraction exprimant la relation existant entre ces deux termes, soit en examinant la page consacrée à la réponse. Les différents critères sont ensuite combinés grâce à un arbre de décision.

La méthode est évaluée en elle-même, par l'évaluation des critères pris indépendamment les uns des autres et de leur combinaison. Elle est également évaluée dans le cadre de la validation de réponse afin de tester l'apport de cette vérification pour une méthode de validation.

2. État de l'art

Tous les systèmes de questions réponses ont recours à la reconnaissance des entités nommées. Les entités nommées sont des objets textuels (mots ou groupe de mots) pouvant être catégorisés dans des classes. Classiquement, quatre grandes classes sont utilisées : nom de personne, nom de lieu, nom d'organisation et date. Cette liste est souvent plus étendue et certains systèmes de questions réponses en utilisent jusqu'à une centaine. Le système conçu par Sekine *et al.* (2002) utilise 200 types d'entités nommées, celui créé par Hovy *et al.* (2001) en utilise 122 et celui de Harabagiu *et al.* (2000) en utilise plus encore puisqu'il s'appuie sur WordNet (Fellbaum, 1998).

Les entités nommées interviennent afin de sélectionner une réponse dont le type correspond bien à celui que la question attend. Dans un cadre de validation de réponses, cette reconnaissance des entités nommées dans les passages réponses permet, par exemple, de rejeter la réponse « Paris » pour une question attendant une personne en retour comme dans « Quel président succéda a Jacques Chirac ? ». Toutefois, comme cette vérification repose sur des types définis *a priori*, elle ne permet pas de savoir si la réponse est ou non du bon type, quand un type plus spécifique ou un nouveau type apparaît.

Schlobach *et al.* (2004) présentent une méthode traitant de la vérification du type dans les cas où la réponse est un lieu géographique. Cette spécification permet d'utiliser des bases de connaissances, sous forme d'ontologie, ainsi que WordNet. Les informations obtenues de cette manière sont combinées avec des informations statistiques traitant entre autre de la présence commune de la réponse et du type dans un ensemble de documents par un mécanisme d'apprentissage automatique. Ce système est évalué en notant l'amélioration qu'apporte cette vérification à un système de questions réponses existant.

Schlobach *et al.* (2007) le poursuivent en l'appliquant au domaine général. Les méthodologies sont assez semblables et combinent des informations fournies par Word-

1. Wikipédia : http ://fr.wikipedia.org

Net et des informations statistiques. La vérification grâce à WordNet s'effectue en cherchant un lien possible entre la réponse et le type. La méthode statistique traite entre autre de la fréquence d'apparition de la réponse, du type, du type et de la réponse ensemble, de la probabilité conditionnelle d'apparition commune du type et de la réponse en fonction de l'apparition du type ou de la réponse. La méthode utilise aussi une mesure tenant compte de l'apparition de la règle « Réponse est un Type » avec un ou deux mots pouvant séparer la réponse et le type dans un ensemble de documents. L'évaluation de la méthode permet d'obtenir une amélioration de 20% sur la mesure MRR, mesure permettant d'évaluer les systèmes de questions réponses.

La méthode présentée dans cet article est relativement proche de celles-ci. Toutefois plusieurs différences sont à observer. D'une part, notre travail est effectué sur le français, et il n'existe pas de base de donnée lexicale aussi complète et structurée que WordNet. Des tests ont été effectués en utilisant EuroWordNet, une version de WordNet pour de nombreuses langues dont le français. Ce réseau lexical n'étant pas aussi complet que WordNet, de nombreux mots y sont absents et cette ressource n'est pas réellement utilisable pour la validation du type de la réponse. De nouveaux critères reposant sur l'utilisation d'entités nommées ont donc été ajoutés. D'autre part, les évaluations menées tiennent uniquement compte de l'apport de la vérification du type sur les systèmes de questions réponses sans juger la méthode en elle-même. Dans cet article, les deux types d'évaluation sont présentés.

La vérification de type se rapproche de la recherche de relation is-a présentée initialement par (Hearst, 1992). La méthode cherche un certain nombre d'hyponymes grâce à des patrons syntaxiques du type « Réponse est type » et acquiert aussi de nouveaux patrons. Pour cela, elle part d'un ensemble de patrons, puis cherche, dans des documents, les phrases dans lesquelles ils apparaissent, ce qui permet d'obtenir un premier ensemble de couples hyperonymes-hyponymes. De nouveaux patrons sont formés en étudiant l'apparition de ces couples et de nouveaux couples sont ensuite extraits grâce à ces patrons.

Cette méthode est assez peu utilisable dans notre travail car il est impossible de répertorier tous les couples qu'un système de questions réponses pourrait rencontrer. Toutefois l'idée de chercher des relations d'hyperonymie est conservée et correspond à l'un des critères.

3. Types de réponses

Les questions qui nous intéressent sont celles qui indiquent explicitement le type de réponse attendu. Ce type est dénommé *type spécifique*. Par ailleurs, ces questions peuvent aussi permettre de déduire que la réponse sera une entité nommée d'un type connu, qui sera appelé *type EN*.

Ainsi, le type spécifique peut être vu comme la dénomination du concept dont la réponse est une instance (type « acteur » pour la question « Quel acteur joue dans Danse avec les loups ? ») ou un hyponyme (type « oiseau » pour la question « Quel

est l'oiseau le plus rapide ? »), mais il ne fait référence à aucune classification préexistante et cette dénomination est celle qui est trouvée dans les questions. On dira que la réponse est compatible avec le type (ou est du type) attendu si elle est une instance ou un hyponyme du concept de ce type comme « Kevin Costner » pour le type « acteur » ou « autruche » pour le type « oiseau ». Le type spécifique est constitué d'un à plusieurs mots, par exemple premier ministre.

Le type d'entité nommée est le nom de la classe d'entité nommée que la question attend en retour. Dans la question précédente ce type est PERSON. Alors que les types spécifiques dépendent de la formulation des questions, les types d'entités nommées sont en nombre fixe (une vingtaine), structurés suivant les cinq types classiques : personne, lieu, organisation, date et entités numériques (une description plus précise est donnée section 4.1).

L'extraction de ces deux types est faite par l'analyseur de questions du système de questions réponses FRASQUES (Grau *et al.*, 2005). Il utilise des règles portant sur des critères syntaxico-sémantiques, le type d'interrogatif ainsi que la forme syntaxique des questions. Pour un ensemble de questions, plusieurs cas sont possibles :

– la question précise un type spécifique qui n'a pas de type d'entité nommée associé comme « Quel film remporta la palme d'Or en 1994 ? » pour laquelle « film » est le type spécifique.

– la question précise un type spécifique associé à un type d'entité nommé. Deux cas sont alors possibles :

- les deux types sont égaux : la question « Dans quel lieu des massacres de musulmans ont-ils été commis ? » a « lieu » comme type spécifique et type d'entité nommée ;

- le type spécifique est plus précis que le type d'entité nommée. C'est par exemple le cas de « Quel acteur a joué dans Danse avec les Loups ? » qui attend en réponse une entité nommée de type PERSON et de type spécifique « acteur » .

Notre étude s'appuie sur un corpus d'apprentissage provenant de la campagne d'évaluation de questions réponses EQueR dans laquelle les systèmes participants devaient répondre à 500 questions. Parmi ces questions, 198 mentionnent un type spécifique et sont utilisées pour créer le corpus. De nombreuses questions reprenant le même type, la base contient 98 types spécifiques différents.

La granularité de ces types est très variable. Certains comme « lieu » et « traitement » sont très larges, d'autres comme « bisquine » (bateau de pêche à voiles) ou « parc » sont précis, que la question attende ou non une entité nommée en réponse. Ces exemples illustrent la diversité des termes et le fait que l'on ne peut prédire tous les types.

Les réponses à ces questions sont celles données par les systèmes participant à cette campagne, qui pouvaient fournir cinq réponses précises et cinq passages pour chaque question.

La base d'apprentissage contient 2720 couples réponse/type spécifique dont la moitié (1360) est valide, c.-a-d. la réponse est du type attendu. Une étude manuelle a été effectuée pour déterminer la validité des couples. Aussi, notre problème se pose ainsi : étant donné un triplet (réponse courte, type spécifique, passage), déterminer que le couple (réponse, type spécifique) est valide.

La suite de l'article présente l'ensemble des différentes méthodes utilisées pour vérifier qu'une réponse est ou non du type spécifique attendu. Celles-ci peuvent être organisées en trois catégories : l'utilisation des systèmes de reconnaissance des entités nommées, l'exploitation de Wikipédia et l'utilisation de mesures statistiques de co-occurence.

4. Utilisation de systèmes de reconnaissance d'entités nommées

La première stratégie est l'utilisation des types d'entités nommées reconnues afin de vérifier le type d'une réponse.

Deux utilisations en sont faites :

– la première pour vérifier globalement que la réponse est compatible avec le type. Par exemple, pour la question « Quel acteur a joué dans Danse avec les loups ? » et la réponse « Kevin Costner », la méthode recherche le type EN de la réponse et vérifie qu'il est compatible avec le type spécifique.

– la seconde s'appuie sur des listes d'entités nommées de différents types et vérifie que la réponse en fait partie. Pour la question précédente, elle cherche Kevin Costner dans la liste correspondant aux acteurs, si elle existe.

Notons que ces deux méthodes sont complémentaires. Dans la seconde tous les types ne peuvent être couverts mais quand ils le sont, elle est plus précise que la première. La première, en revanche, est applicable sur plus d'exemples.

4.1. *Filtrer les réponses*

La première utilisation des entités nommées est une utilisation globale qui indique qu'une réponse n'est pas compatible avec le type si le type EN de la réponse n'a pas de relation avec celui que la question attend. Par exemple, la question « En quelle année eut lieu la révolution russe ? » attend une date en réponse. L'utilisation de ce module permettra de rejeter la réponse « Alexandre Issaievitch Soljenitsyne » qui est de type PERSON.

Les types d'entités nommées sont ceux déterminés par le module d'analyse des questions du système FRASQUES. Les passages réponses, et donc la réponse proposée qui en est extraite, sont analysés aussi par un module de ce système afin de les annoter selon les mêmes types d'entités nommées. Le système permet de reconnaître une vingtaine d'entités nommées structurées suivant les cinq types classiques

(personne, lieu, organisation, date, entités numériques). Par exemple, le type LIEU regroupe les types « VILLE » et « PAYS ». Les entités numériques permettent de typer les expressions numériques (longueur, vitesse, etc.) et le type, NomPropre, a été ajouté pour étiqueter tous les termes contenant un nom propre non étiquetés plus finement. Ce dernier type permet de pallier l'absence de reconnaissance d'entités nommées en utilisant un type plus global. Parmi les triplets à évaluer, quatre cas sont possibles :

– la question n'attend pas en réponse une entité nommée. C'est par exemple le cas de la question « Quel oiseau est le plus rapide d'Afrique ? ». Dans ce cas aucune information ne peut être donnée par ce critère et la valeur INCONNU sera donnée au couple réponse/type.

– la question attend une entité nommée et la réponse n'en est pas une. Par exemple la réponse « bateau » pour une question attendant une personne. Dans ce cas la réponse est vue comme mauvaise et la valeur NON est renvoyée.

– la question attend une entité nommée en réponse et la réponse est une entité nommée incompatible avec le type. Par exemple la réponse « 300 » pour une question attendant un LIEU. Dans ce cas, la réponse sera considérée incompatible avec le type attendu et la valeur NON est renvoyée.

– la question attend une entité nommée en réponse et celle-ci est d'un type compatible avec le type EN attendu : soit le même, soit un type plus précis, soit une catégorie plus générale par exemple LIEU ou NomPropre pour PAYS. Dans ce cas, le système considère que la réponse est compatible avec le type attendu et la valeur OUI est renvoyée.

Cette vérification ne permet d'avoir qu'une idée globale de la validité de la réponse et la réponse de l'exemple donné en introduction (« Michel Rocard ») est vue ici comme un « président ».

Afin d'évaluer ce critère, deux évaluations sont faites :

– la première, précise, indique la proportion de réponses OUI et NON données correctement ;

– la seconde, globale, mesure la précision (proportion de bonnes valeurs parmi celles fournies par le système), le rappel (la proportion de bonnes valeurs parmi celles attendues) et la f-mesure qui permet de combiner ces deux mesures. La précision est différente du rappel quand des couples réponse/type ne sont pas évalués par le critère.

$$précision = \frac{\#\ valeurs\ correctes\ données}{\#\ valeurs\ données} \qquad rappel = \frac{\#valeurs\ correctes\ renvoyées}{\#\ valeurs\ attendues}$$

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

Ces deux évaluations seront aussi appliquées aux critères présentés par la suite.

Le tableau 1 évalue cette première méthode de manière précise. La première ligne indique, parmi les décisions validant le couple réponse/type (valeur « OUI »), 63% de réponses correctes, les réponses sont bien du type attendu, et un taux d'erreurs de

37% de réponses déclarées compatibles alors qu'elles ne sont pas du type attendu. La deuxième ligne montre que lorsque la réponse est vue comme incompatible avec le type EN (valeur « NON »), alors elle ne correspond pas non plus au type spécifique attendu par la question à 71%, alors que cette décision est erronée pour 29% d'entre elles. La troisième ligne traite du cas où la détection ne peut être faite (31%).

Décision	# Réponses du type spécifique	#Réponses pas du type spécifique
OUI (1411)	**885 (63%)**	526 (37%)
NON (457)	132 (29%)	**325 (71%)**
INCONNU (852)	344 (40%)	508 (60%)

Tableau 1. *Matrice de confusion de la méthode utilisant les entités nommées comme filtre*

Lorsque la réponse est évaluée, la précision de la méthode est de 0,65, et puisque 31% des couples (réponse, type) ne sont pas évalués, le rappel est plus bas et vaut 0,45.

Comme la plupart des systèmes de questions réponses mettent en œuvre une vérification par entité nommée assez semblable, les résultats obtenus par cette méthode constitueront les résultats de base qu'il s'agira d'améliorer.

4.2. *Valider des réponses*

Les entités nommées reconnues dans un grand corpus permettent de construire des listes de termes pour chaque type reconnu. Il semble donc intéressant de tester la présence ou l'absence de la réponse dans les listes correspondant au type cherché pour valider ou invalider son type. Afin de disposer d'informations pertinentes, il est nécessaire de collecter un grand nombre de types possibles d'entités nommées. Le module d'entités nommées du système de questions réponses RITEL (Rosset *et al.*, 2006) a été utilisé dans ce but. Il traite un ensemble de types pouvant être assez précis comme « religion » ou « fleuve » et reconnaît 70 types spécifiques. Comme le nombre de types d'entités nommées est fini, il y aura malgré tout toujours des types spécifiques ne correspondant pas à une entité nommée. La méthode testant la présence de la réponse dans la liste associée au type, trois cas sont possibles :

– il n'y a pas de liste associée au type. Dans ce cas le module ne peut pas savoir si elle correspond au type et renvoie INCONNU.

– la réponse ne se trouve pas dans la liste des termes du type cherché. Dans ce cas la valeur NON est renvoyée.

– la réponse se trouve dans la liste correspondant au type cherché. La réponse est considérée compatible avec le type attendu et la valeur OUI est renvoyée. Cette vérification permet notamment de valider que « Pulp Fiction » est un film.

Le tableau 2 permet d'évaluer cette méthode. Il montre que lorsqu'une correspondance de type est reconnue, la valeur est très souvent correcte (77%), et la précision

globale est de 0,75. En revanche, comme peu de données (43%) sont évaluées, le rappel est bas, 0,32.

Décision	#Réponses du type spécifique	#Réponses pas du type spécifique
OUI (656)	**506 (77%)**	150 (23%)
NON (515)	138 (27%)	**377 (73%)**
INCONNU (1549)	716 (46%)	833 (54%)

Tableau 2. *Matrice de confusion de la méthode utilisant les entités nommées comme base de connaissance*

5. Utilisation de Wikipédia

5.1. *Recherche dans des pages particulières*

Wikipédia étant une encyclopédie, chacune de ses pages définit l'élément qui constitue son titre. Cela permet de formuler l'hypothèse suivante : si le type spécifique est trouvé dans la page Wikipédia associée à la réponse, cette dernière a de fortes chances d'être une instance ou un hyponyme de ce type.

Pour ce travail nous avons utilisé la version de Wikipédia de novembre 2006 retenue pour la campagne de questions réponses CLEF (Forner *et al.*, 2009).

La méthode teste la présence du type, pris sous sa forme textuelle, dans les pages Wikipédia ayant pour titre la réponse ou dont le titre contient la réponse. Trois cas sont alors possibles :

– aucun titre de page ne peut être associé au type. Dans ce cas, rien ne peut être déduit par cette méthode et la valeur INCONNU est renvoyée. Ces cas correspondent par exemple à des personnes ayant eu un rôle très ponctuel dans le temps comme « Alfred Henninger » qui est un meurtrier.

– la page correspondant à la réponse contient bien le type. Cela implique que la réponse a de fortes chances d'être du type cherché et la valeur OUI est renvoyée. Ainsi « Jacques Chirac » est un « maire ».

– la page ne contient pas le type. Dans ce cas la réponse ne correspond très probablement pas à ce type. La valeur NON est donc renvoyée. C'est par exemple le cas de « Bethléem » pour le type « planète ».

Le tableau 3 présente les résultats obtenus par cette méthode. Il montre que lorsque la méthode voit la réponse comme correspondant au type, c'est souvent le cas (74%). En revanche, d'avantages d'erreurs sont commises quand la réponse est vue comme ne correspondant pas au type (seules 61% des décisions sont correctes). Cela peut s'expliquer par le fait que le type peut être remplacé dans la page de la réponse par un synonyme ou un terme faisant référence à un type plus général. Les résultats montrent également que peu de réponses sont évaluées : beaucoup de réponses n'ont pas de page

Décision	#Réponses du type spécifique	# Réponses pas du type spécifique
OUI (661)	**491 (74%)**	170 (26%)
NON (589)	228 (39%)	**361 (61%)**
INCONNU (1470)	641 (43%)	829 (57%)

Tableau 3. *Matrice de confusion de la méthode utilisant les pages Wikipédia correspondant à la réponse*

Wikipédia qui leur est consacrée. D'un point de vue global, la décision est correcte dans 68% des cas. Le nombre élevé de réponses non évaluées entraîne un rappel de 0,32.

Notons que ces résultats sont moins bons que ceux présentés dans (Grappy *et al.*, 2008). Cela s'explique par un changement d'ensemble d'exemples. Les données beaucoup moins nombreuses, une centaine d'exemples, correspondaient à des connaissances encyclopédiques. De plus les données avaient préalablement été filtrées afin de ne conserver que les réponses correspondant au type d'entité nommée attendu par la question.

Le principe de cette vérification produit des résultats intéressants mais ne permet pas de couvrir tous les cas possibles. Les critères suivants poursuivent dans cette voix afin d'étendre sa couverture.

5.2. *Utilisation de patrons d'extraction*

Le critère suivant utilise lui aussi Wikipédia mais ne limite pas la recherche à certaines pages. L'idée est que certaines structures de phrases permettent d'exprimer qu'une entité est d'un type donné, comme « Réponse est un Type ». Cinq types de règles, issus d'une analyse du corpus, ont été définis :

– ***RÉPONSE* être déterminant *TYPE***, avec de nombreuses variantes du verbe être et du déterminant, (exemple Nicolas Sarkozy est le président).

– ***TYPE RÉPONSE*** (président Nicolas Sarkozy)

– ***RÉPONSE*, déterminant *TYPE*** (Nicolas Sarkozy, le président)

– ***RÉPONSE* (déterminant *TYPE*** (Nicolas Sarkozy (le président de la république)

– ***RÉPONSE* : déterminant *TYPE*** (Nicolas Sarkozy : le président)

La méthode permettant de savoir si la réponse correspond au type attendu commence par instancier les variables *TYPE* et *RÉPONSE* par leurs valeurs. Ainsi, pour vérifier que Johnny Depp est un acteur, *TYPE* prend la valeur acteur et *RÉPONSE* la valeur Johnny Depp. Cela est fait pour chacune des règles, ce qui amène à des phrases comme « Johnny Depp est un acteur », « Johnny Depp sera l'acteur », etc.

Ces phrases sont ensuite cherchées dans les pages Wikipédia grâce au moteur de recherche Lucene (Hatcher *et al.*, 2004). Si l'une de ces phrases est trouvée, alors le couple (réponse, type) est validé et la valeur OUI renvoyée, sinon la réponse est considérée incompatible avec le type attendu et la valeur NON est donnée. Le choix de Wikipédia comme corpus tient compte de sa spécificité. En effet, Wikipédia étant une encyclopédie, elle est plus à même de contenir des phrases de définitions qu'un corpus de journaux par exemple.

Le tableau 4 montre les résultats obtenus par cette méthode. Il montre tout d'abord que les résultats sont plutôt bons quand la réponse OUI est renvoyée (73% des réponses correspondent bien au type). Il montre également que toutes les valeurs peuvent être évaluées par cette méthode, contrairement aux précédentes. Ce qui se traduit par un rappel et une précision égaux et valant 0,66.

Décision	#Réponses du type spécifique	# Réponses pas du type spécifique
OUI (974)	**713 (73%)**	261 (26%)
NON (1746)	647 (37%)	**1099 (63%)**

Tableau 4. *Matrice de confusion de la méthode utilisant des patrons d'extraction*

6. Recherche en corpus

Le troisième type de critère est d'ordre statistique. Il se place dans un cadre plus général et s'intéresse à l'apparition de la réponse et du type dans un ensemble de documents, quels que soient le document ou la relation les liant. On suppose que si le type et la réponse apparaissent souvent dans les mêmes documents, alors ils sont liés.

Afin de mettre en évidence les relations entre la fréquence d'apparition du type et de la réponse et le fait que la réponse soit du type attendu, un système par apprentissage a été créé. Ce système reprend un ensemble de critères décrits dans Schlobach *et al.* (2004) et Schlobach *et al.* (2007) et sont les suivants :

– **les proportions d'apparition** : le rapport entre le nombre de documents contenant le type et la réponse et le nombre de documents contenant la réponse ou le nombre de documents contenant le type. Ce critère permet de détecter les cas où la réponse apparaît fréquemment accompagnée du type.

$$C1 = \frac{nb\ documents(réponse + type)}{nb\ documents(réponse)} \quad [1]$$

$$C2 = \frac{nb\ documents(réponse + type)}{nb\ documents(type)} \quad [2]$$

– **la mesure PMI** (*Pointwise Mutual Information*), mesure classique de statistique de la force d'association de deux termes, correspond au rapport entre la fréquence d'apparition commune et le produit des fréquences d'apparition du type et de la réponse.

$$C3 = \frac{Fréquence(réponse + type)}{Fréquence(réponse) * Fréquence(type)} \quad [3]$$

– **les fréquences d'apparition** du type, de la réponse et de l'ensemble type+réponse. Ces critères complètent le précédent dans le sens où ils permettent de dissocier différents cas comme celui où la réponse ou le type apparaissent très rarement de celui où ils sont au contraire très fréquents. Ces différentes possibilités entraînent des valeurs différentes de la mesure PMI.

$$C4 = \frac{nb\ documents(réponse)}{nb\ documents} \quad [4]$$

$$C5 = \frac{nb\ documents(type)}{nb\ documents} \quad [5]$$

$$C6 = \frac{nb\ documents(type + réponse)}{nb\ documents} \quad [6]$$

Les résultats fournis par ces différents critères sont ensuite combinés grâce à une combinaison d'arbres de décision par la méthode bagging (cf. section 7) fournie par le système WEKA [2].

Deux corpus de documents ont été utilisés. Le premier est la Wikipédia et le second les articles du journal « Le Monde » de 1992 à 2000. Ces articles sont utilisés car les réponses ont été extraites de ces journaux et ils contiennent donc toutes les réponses et notamment celles n'apparaissant pas dans la Wikipédia telles que les réponses ayant eu une brève importance historique quand les questions portent sur un fait divers.

Comme cette méthode travaille par apprentissage, son évaluation ne peut se faire sans base de test. Celle-ci est décrite dans la section suivante. Les résultats, présentés dans le tableau 7, montrent que les méthodes obtiennent de bons résultats puisqu'en utilisant la Wikipédia la f-mesure est de 0,68 et en utilisant Le Monde elle est de 0,72. Dans les deux cas, la précision est égale au rappel puisque toutes les réponses sont évaluées. Afin de voir l'effet de la base de documents, les tableaux 5 et 6 présentent les matrices de confusion obtenues par la méthode en utilisant tout d'abord la Wikipédia puis Le Monde.

Ces deux tableaux permettent de voir que la méthode utilisant le corpus Le Monde obtient de meilleurs résultats que ceux utilisant la Wikipédia (72% contre 68%). Cela est dû à un meilleur traitement des réponses qui sont du type attendu.

2. WEKA : http ://sourceforge.net/projects/weka/

Décision	#Réponses du type spécifique	#Réponses pas du type spécifique
OUI (822)	**542 (66%)**	280 (34%)
NON (725)	220 (31%)	**505 (69%)**

Tableau 5. *Matrice de confusion de la méthode utilisant des mesures statistiques sur Wikipédia*

Décision	#Réponses du type spécifique	# Réponses pas du type spécifique
OUI (634)	**479 (76%)**	155 (24%)
NON (913)	283 (31%)	**630 (69%)**

Tableau 6. *Matrice de confusion de la méthode utilisant des mesures statistiques sur Le Monde*

7. Combinaison des critères

Après avoir créé l'ensemble des critères, l'étape finale consiste à les combiner, ce qui est fait par une méthode d'apprentissage de WEKA qui permet d'utiliser un grand nombre de classifieurs. Celui qui est choisi pour cette étude est une combinaison d'arbres de décision grâce à la méthode bagging.

Les arbres de décision regroupent un ensemble de cas ayant des similarités communes. Pour ce faire, ils recherchent parmi les critères la valeur permettant de répartir au mieux les données, celle qui induit le moins d'erreurs. Une fois ce critère trouvé, les données sont séparées. Cette étape est répétée sur les groupes trouvés jusqu'à ce que les données soient réparties pour le mieux.

La méthode bagging permet d'utiliser un ensemble d'arbres de décision. Chaque arbre peut, en effet, donner des résultats différents suivant les séparations effectuées. Les différents résultats sont réunis par vote afin de fournir une réponse globale. Le système utilise cinq arbres de décision et le poids associé à chacun d'eux est le même. Le résultat renvoyé est donc celui obtenu par la majorité des arbres.

Les critères utilisés sont ceux présentées précédemment :

1) le filtre sur les entités nommées,

2) la validation du type grâce aux entités nommées,

3) la présence du type dans la page Wikipédia de la réponse,

4) l'application de règles syntaxiques (« RÉPONSE est un TYPE ») dans les pages Wikipédia,

5) les critères statistiques calculés sur Wikipédia :

- le rapport entre le nombre d'apparitions de la réponse et du type ensemble et le nombre d'apparitions du type ou de la réponse,

- la fréquence d'apparition du type, de la réponse et de l'ensemble type+réponse,

- la mesure PMI,

6) les critères statistiques calculés sur les articles du journal Le Monde.

8. Évaluation

Les sections précédentes ont présenté un certain nombre de méthodes ainsi que leurs évaluations. Dans cette section, nous combinons ces méthodes par apprentissage et effectuons trois évaluations :

– l'évaluation des critères pris séparément sur la base de test ;

– l'évaluation de la combinaison des méthodes ;

– l'apport de la vérification du type à un système de validation de réponses, décrit section 9.

La base de test est construite à partir des données fournies par la campagne de validation de réponses AVE 2006 (Peñas *et al.*, 2006). La section 9 explique plus en détail cette tâche. Les données de cette campagne permettent d'utiliser 1547 paires réponse/type spécifique dont la moitié (762) est valide, i.e. la réponse est du type spécifique. Ces paires sont issues de 90 questions et correspondent à 47 types spécifiques différents.

8.1. *Évaluation des critères*

La première évaluation porte sur chacun des critères pris séparément. Le tableau 7 présente ces résultats en terme de précision, rappel et f-mesure.

Critère	Précision	Rappel	F-mesure
1) Filtre utilisant les entités nommées	0,69	0,54	0,60
2) Validation grâce aux entités nommées	**0,80**	0,32	0,45
3) Recherche dans la page Wikipédia de la réponse	0,72	0,46	0,57
4) Utilisation de patrons syntaxiques	0,70	0,70	**0,70**
5) Critères statistiques sur Wikipédia	0,68	0,68	**0,68**
6) Critères statistiques sur Le Monde	0,72	**0,72**	**0,72**

Tableau 7. *Résultats des critères*

Ce tableau montre que les résultats obtenus par chacune des méthodes sur la base de test sont assez semblables à ceux obtenus sur la base d'apprentissage. La méthode consistant à vérifier la présence du type dans la page Wikipédia associée à la réponse obtient des résultats meilleurs sur la base de test, surtout au niveau du rappel (0,46 contre 0,32). Cela s'explique par la répartition des données. En effet, les données de

la base de test contiennent plus souvent des noms de personne et il existe donc plus de pages Wikipédia ayant comme titre ces noms.

Nous pouvons aussi remarquer que les méthodes obtenant les meilleurs résultats sont l'utilisation des patrons syntaxiques(f-mesure 0,70) et les critères statistiques en utilisant le corpus Le Monde (f-mesure 0,72). La méthode de validation utilisant les entités nommées obtient la meilleure précision (0,80) qui indique peu d'erreurs quand elle voit une réponse comme correspondant au type attendu par la question.

8.2. *Évaluation totale*

Après avoir vu les résultats obtenus par chaque critère, l'étape suivante consiste à les combiner afin d'obtenir une seule valeur pour chaque réponse : elle correspond ou non au type.

Pour mieux évaluer la méthode, il faut comparer ses résultats à ceux obtenus par d'autres méthodes. La plupart des systèmes de questions réponses utilisant la détection du type d'entité nommée comme filtre, les résultats obtenus par cette méthode sont utilisés comme baseline. Le tableau 8 présente les résultats obtenus.

Méthode	Précision	Rappel	F-mesure
Filtre utilisant les entités nommées	0,69	0,54	0,60
Combinaison de méthodes	0,80	0,80	0,80

Tableau 8. *Résultats globaux*

On peut voir que 80% des données sont bien classées. Ce pourcentage élevé montre que la méthode choisie est efficace. Nous pouvons aussi voir que les résultats obtenus par la combinaison de méthodes sont très nettement supérieurs à ceux reposant juste sur un module de reconnaissance des entités nommées, et qu'ils sont également supérieurs à toutes les méthodes appliquées isolément.

Afin de mieux comprendre les résultats, une étude distinguant les cas où la réponse est associée à un type d'entité nommée des cas où elle n'en n'a pas, a été menée. L'idée était de savoir si les mêmes phénomènes se rencontraient dans les deux cas. Le

Base de test	Proportion de couples correctement classés
Avec EN (1205)	82%
Sans EN (342)	74%

Tableau 9. *Vérification du type en fonction des entités nommées*

tableau 9 montre que les données sont mieux classées quand un type d'entité nommée est associé à la réponse. Cela peut s'expliquer par le fait que la plupart des réponses de la base d'apprentissage sont associés à un type d'entité nommée (78%) et que les critères portant sur les entités nommées sont efficaces.

Le tableau 10 présente la matrice de confusion de la méthode. Il montre qu'elle obtient des résultats similaires quelle que soit la valeur retournée (OUI (80%), NON (80%)).

Décision	#Réponses du type spécifique	#Réponses pas du type spécifique
OUI	**603 (80%)**	149 (20%)
NON	159 (20%)	**636 (80%)**

Tableau 10. *Matrice de confusion de la méthode globale*

Avant de clore cette section, voyons quelques résultats obtenus :

– Hosni Moubarak est un président ;

– Yasser Arafat n'est pas un président ;

– Krypton est une planète ;

– Bethléem n'est pas une planète ;

– Barings n'est malheureusement pas une « grande banque ». Cela doit être dû à la présence de l'adjectif ;

– le Parti Socialiste est, à tort, un président. Cela est sûrement dû au rapport fréquent entre ces deux termes.

Les résultats peuvent aussi être rapprochés de ceux des systèmes de recherche d'entités nommées cherchant toutes les entités nommées présentes dans un corpus de texte. La campagne MUC présentée par (Grishman *et al.*, 1995) a permis d'évaluer ces systèmes en se focalisant sur la recherche des types d'entités nommées personne, lieu, organisation, date, expressions de temps, pourcentage et unité monétaire. Le système ayant eu les meilleurs résultats à cette campagne obtient une f-mesure de 0,93. Ces résultats sont supérieurs à ceux obtenus par notre système. Toutefois cette campagne s'intéresse seulement à sept types d'entités, d'un niveau de granularité supérieur à celui de notre système, qui peut de plus détecter tout nouveau type apparaissant dans une question. Cette différence introduit une différence de résultats car il semble plus facile de montrer qu'un nom correspond à une personne que de montrer que c'est un acteur.

9. Apport pour un système de validation de réponses

Après avoir évalué le système en lui même, l'étape suivante consiste à voir en quoi il permettrait d'améliorer les systèmes de questions réponses en améliorant la validation de réponse. La validation de réponse est une tâche consistant à détecter à partir de triplets (question, réponse, passage justificatif) si la réponse est bien valide c'est-à-dire si elle est correcte et justifiée par le passage justificatif. Le passage justificatif est le passage de texte duquel la réponse a été extraite. Voyons un exemple :

– Q : Quel club de football remporta le championnat de France en 2010 ?

– R : L'Olympique de Marseille

– P : L'Olympique de Marseille a décroché son neuvième titre de champion de France en 2010.

Dans ce cas, si l'on sait que l'Olympique de Marseille est un club de football, on peut considérer que la réponse est valide puisqu'elle est correcte et le passage la justifie. Mais en l'absence de cette information, rien ne permet dans le passage de la justifier, ce qui montre l'intérêt de disposer d'une méthode pour mieux valider les réponses proposées.

Ainsi que nous l'avons déjà dit, les campagnes d'évaluation AVE (Answer Validation Exercise) (Peñas *et al.*, 2006) permettent d'évaluer les systèmes de validations de réponses et les données proviennent de différents systèmes de questions réponses qui ont proposé plusieurs réponses et passages justificatifs pour les différentes questions. L'évaluation y est faite uniquement sur les réponses déclarées valides en utilisant la précision, le rappel et la f-mesure. Aussi, allons nous évaluer nos résultats selon cette méthode.

9.1. *Lien entre la validité du type et la validité de la réponse*

Notre première étude traite du rapport entre la validité du type de la réponse et la validité de la réponse. Cette problématique a été étudiée dans (Rodrigo *et al.*, 2006) afin de montrer l'utilité des entités nommées pour la validation de réponses. Les auteurs présentent un système détectant une réponse comme valide si toutes les entités nommées de la question sont impliquées par une entité nommée du passage. Une entité nommée en implique une autre si elle la contient. Ce travail a obtenu d'assez bons résultats, notamment un bon rappel (0,68). Toutefois la précision est assez basse (0,43), ce qui indique qu'il faut d'autres critères permettant de détecter des réponses non valides.

Le tableau 11 présente une correspondance entre la vérification du type et la validité de la réponse pour les réponses de la base de test. Notons que celle-ci contient beaucoup plus de réponses non valides (80%) que de réponses valides. Comme certaines réponses n'ont pas été évaluées en terme de validité, seules 1457 des 1547 réponses précédentes sont utilisées pour ce travail.

Réponse du type	# Réponses valides	# Réponses non valides
OUI (702)	**236 (34%)**	466 (66%)
NON (755)	53 (7%)	**702 (93%)**
TOTAL (1457)	289 (20%)	1168 (80%)

Tableau 11. *Rapport entre la vérification du type et la validité des réponses*

Ce tableau montre que si la réponse est vue comme ne correspondant pas au type cherché alors elle est très souvent non valide (93%). En revanche rien ne peut être dit quand la réponse est considérée comme étant du type attendu ; le nombre élevé d'erreurs (66%) indique que les réponses vues comme étant du type attendu sont le plus

souvent incorrectes, cette seule information étant insuffisante pour décider de la validité d'une réponse qui dépend aussi du fait que le passage contient bien l'information donnée dans la question. Ce tableau montre également que la proportion de réponses vues comme n'étant pas du type parmi les réponses valides n'est pas négligeable avec 19% d'erreurs (53/289). Les informations contenues dans ce tableau permettent toutefois de voir que cette méthode peut être très utile pour détecter les réponses invalides. En évaluant le système à l'aide des mesures utilisées pour la validation de réponses, la méthode obtient une précision de 0,34 et un rappel de 0,81.

9.2. *Intégration de la validation du type dans un système de validation de réponses*

Un premier système de validation de réponses avait été réalisé, fondé sur une méthode par apprentissage (Grappy *et al.*, 2008). Nous avons voulu l'étudier en intégrant la validation du type proposée dans cet article. Le corpus choisi pour cette étude est un sous-ensemble du corpus d'AVE 2006 dans lequel les réponses triviales non valides ont été supprimées. Ces cas de non validité sont :

– le passage ne contient pas la réponse ;

– la réponse est contenue dans la question ;

– un certain nombre de questions contiennent une information de date comme « Quel groupe suédois sortit son premier album en 1994 ? ». La vérification porte sur cette date et vérifie que la date est contenue dans le passage justificatif ou qu'elle correspond à la date de création de ce passage ;

– l'entité nommée attendue en réponse ne correspond pas à celle obtenue ; ce cas se rencontre par exemple quand la question attend une personne en réponse et qu'une date est retournée.

Le corpus a été subdivisé en base d'apprentissage et base de test. L'apprentissage est effectué par une combinaison d'arbres de décision grâce à la méthode bagging. Les critères, principalement d'ordre lexical, sont présentés ci-dessous.

9.2.1. *Critères lexicaux*

Le premier critère considère le taux de mots de la question présents dans le passage, à l'identique ou sous forme de variante. Si un passage possède suffisamment de termes analogues à ceux de la question, il a de fortes chances de parler du thème de la question, et donc de contenir la réponse cherchée.

Le rapprochement de termes n'est effectué que sur certaines catégories morpho-syntaxiques ainsi les déterminants, prépositions et adverbes ne sont pas considérés.

L'importance à accorder aux termes communs varie selon leur catégorie morpho-syntaxique. Ainsi un nom propre commun au passage et à la question semble plus important qu'un adjectif. Les seconds critères calculent donc la proportion de noms propres, noms communs, verbes et adjectifs communs au passage et à la question.

Chacune de ces catégories correspond à un critère. La proportion d'expressions numériques communes au passage et à la question est aussi considérée.

L'analyse de la question permet également de détecter certains éléments importants de la question. Elle reconnaît :

– **le focus :** par définition, cet élément devrait être repris dans le passage pour exprimer la réponse. Dans la question « Quel est le sport pratiqué par Zinédine Zidane ?», le focus est « Zinédine Zidane » ;

– **le type spécifique de la réponse** ;

– **les multitermes :** un multiterme est un ensemble de mots consécutifs reconnus comme étant liés, comme « prix Nobel ». Trouver un multiterme de la question dans le passage signifie souvent que les mots présents sont utilisés dans le même sens. Ces termes sont reconnus dans les passages par FASTR[3] (Jacquemin, 1999) qui permet également de reconnaître les variations de ces termes.

Un ensemble de critères indiquant la présence ou l'absence de chacune de ces caractéristiques dans le passage est créé.

9.2.2. *Utilisation d'un système de questions réponses*

Un autre critère tient compte de l'utilisation du système de question réponses FRASQUES. Celui-ci cherche une réponse à la question dans le passage justificatif. Puis, une comparaison est effectuée entre la réponse obtenue et la réponse à évaluer. Si la réponse renvoyée est similaire à celle à juger, elle a de bonnes chances d'être correcte. Le critère est donc cette correspondance.

9.2.3. *Proximité des termes*

Le dernier critère étudie la proximité des mots de l'hypothèse, qui correspond à la forme déclarative de la question à laquelle la réponse est ajoutée, dans le passage. L'idée étant que si les mots y sont proches alors ils entretiennent le même type de relation dans le passage et dans la question.

La méthode recherche la plus longue chaîne de mots consécutifs mais non ordonnés présents dans le passage et l'hypothèse. Deux mots sont dits consécutifs s'ils sont adjacents, séparés par des items autorisés (virgule, déterminant ...) ou séparés par un unique mot considéré comme bonus. La valeur du critère est la proportion de mots de l'hypothèse présents dans cette chaîne.

9.2.4. *Vérification du type de la réponse*

Dans ce premier système de validation, la vérification du type spécifique de la réponse était faite dans les pages Wikipedia de titre analogue à la réponse. La vérification par rapport au type EN attendu intervenait dans l'étape de filtrage pour détecter les incompatibilités, et dans la validation de la réponse par application de FRASQUES,

3. http ://www.limsi.fr/Individu/jacquemi/FASTR/

puisque dans le cas d'entités nommées attendues, FRASQUES choisit une réponse dans le passage compatible avec ce type.

Nous allons décrire maintenant l'étude menée pour tenir compte de la vérification du type proposée dans cet article dans ce système de validation de réponse.

9.3. *Intégration de la vérification du type dans le système de validation de réponse*

L'intégration de la vérification du type peut être envisagée de deux façons :

– la première consiste à utiliser la vérification du type comme un filtre. Les réponses ne correspondant pas au type sont décrétées fausses. La validité des autres est détectée par application de la méthode de validation expliquée ci-dessus ;

– la seconde ajoute la vérification du type comme un critère fourni au classifieur, l'idée étant que la méthode de classification détectera la meilleure utilisation de cette vérification.

Les bases d'apprentissage et de test sont extraites de la base préalablement présentées. Seules les questions mentionnant explicitement un type de réponse attendu sont retenues afin d'étudier l'impact de la validation du type. La base d'apprentissage contient 488 exemples dont 186 sont valides. La base de test contient 302 exemples dont 80 sont valides.

Le tableau 12 présente les résultats des deux ajouts possibles de la vérification du type à la validation de réponse, l'utilisation comme un critère et l'utilisation comme un filtre. Les résultats à améliorer sont ceux obtenus en ne considérant pas le type mais en faisant un apprentissage sur les autres critères (méthode baseline).

L'étape de filtre, avant l'apprentissage, reconnaît 111 réponses comme n'étant pas du type attendu. Parmi elles 11 (10%) sont malheureusement valides.

Méthode	Précision	Rappel	F-mesure
Baseline	0,50	0,62	0,55
Comme critère	0,51	0,62	0,56
Comme filtre	0,59	0,56	0,57

Tableau 12. *Intégration de la vérification du type*

Ce tableau montre que l'ajout du type n'améliore que peu les résultats globalement. En effet l'utilisation comme critère conserve les mêmes résultats que ceux n'utilisant pas ce critère, mais cela n'est pas surprenant car cette vérification ne constitue pas le critère principal pour justifier une réponse (la proximité des contenus informationnels entre question et passage prédominent).

L'utilisation de la vérification du type de la réponse comme filtre montre une meilleure précision (0,5 vs 0,59). La proportion de réponses correctement vues comme

valides est donc plus élevée. Toutefois le rappel diminue lui aussi ce qui est dû aux 10% d'erreurs de la méthode de vérification du type.

Une étude des résultats a montré que la méthode n'utilisant pas le type détecte 49 faux positifs (la réponse est vue comme valide à tort). Ce sont ces faux positifs que la vérification du type pourrait faire diminuer. Or, sur ces 49 fausses réponses seules 17 (35%) peuvent effectivement être résolues en utilisant la vérification du type. Sur ces 17 cas, 14 (82%) sont effectivement bien résolus. Ce faible nombre explique que ce critère est peu utilisé quand la vérification est utilisée comme critère et que les résultats globaux restent analogues. Des tests similaires ont été effectués avec des bases d'apprentissage et de tests différents, notamment en considérant également des questions sans type spécifique, mais ces modifications ne changent pas les résultats.

Les erreurs restantes pour la validation de réponses correspondent aux cas où la réponse est du bon type, où il y a bien de nombreux termes de la question dans le passage, mais où ceux-ci ne sont pas liés par les relations syntaxiques présentes dans la question, comme par exemple :

– Q : Qui est le président du Canada ?

– R : Clinton

– P : La visite du président Clinton au Canada.

De ces tests nous pouvons retenir que la vérification du type permet bien de rejeter les réponses n'étant pas du type attendu. Toutefois, la vérification du type n'améliore que peu la validation de réponses sur cette base de test. Il faut en effet noter que cette base est construite à partir de résultats renvoyés par des SQR, qui ont déjà sélectionné la réponse sur le critère du type attendu, et son impact se trouve de ce fait affaibli.

La prochaine étape sera d'appliquer la validation de réponses avec cette vérification de type en l'intégrant dans un système de questions réponses afin de déterminer les réponses à renvoyer. Le principe que nous voulons tester est de sélectionner des candidats réponses dans les passages sur des critères les moins restrictifs possibles, et appliquer le système de validation pour sélectionner les réponses que le système propose. Dans ce cadre, la vérification de type gardera tout son intérêt car elle n'interviendra pas après un filtrage par les entités nommées.

10. Conclusion et perspectives

La validation de réponse détecte si une réponse obtenue par un système répond bien à la question posée. Dans ce cadre, nous avons créé une méthode permettant de vérifier qu'une réponse correspond à un type spécifique attendu par une question. Cette méthode est fondée sur un apprentissage à partir de différents critères. Le premier utilise les entités nommées pour détecter une non-correspondance en rejetant les réponses dont le type d'entité nommée ne correspond pas à celui attendu par la question. Le deuxième utilise également des entités nommées et vérifie que la réponse apparaît dans une liste d'entités correspondant au type. Un autre critère provient de la

recherche de la réponse dans la page Wikipédia correspondant au type. L'encyclopédie Wikipédia est aussi utilisée pour le critère suivant, fondé sur la recherche des structures de phrases exprimant la relation entre la réponse et le type attendu. Les derniers critères sont d'ordre statistique et tiennent compte de la co- occurence de la réponse et du type dans des documents.

Les critères ont tout d'abord été évalués séparément. L'évaluation de leur combinaison, par apprentissage, montre que la méthode est efficace et obtient de bons résultats (80%). La dernière étude traite de l'apport de cette méthode dans un système de validation de réponses par apprentissage. Le fait de rejeter, dans un premier temps, les réponses ne correspondant pas au type semble améliorer la proportion de réponses valides reconnues comme valides mais cela reste à confirmer.

Dans un futur travail, ce système sera utilisé dans un système de questions réponses. Dans ce cas, plusieurs réponses seront extraites par passage et le système de validation de réponse intégrant la validation du type permettra alors de ne proposer que des réponses considérées comme valides. Ce travail permettra d'évaluer directement l'apport de cette vérification pour un système de questions réponses.

Ce travail pourrait également prendre sa place dans un système de validation décomposant les questions afin de détecter un certain nombre d'informations à vérifier. Par exemple, pour valider la réponse « Pierre Béregovoy » à la question « Quel ministre se suicida en 1993 ? » il faut montrer qu'il est ministre, qu'il s'est suicidé et que l'action a lieu en 1993. Dans ce cas, la vérification du type constitue un premier travail qu'il faut encore poursuivre pour valider entièrement les réponses. [4]

11. Bibliographie

Ayache C., Grau B., Vilnat A., « EQueR : the French Evaluation campaign of Question-Answering Systems », *Proceedings of the fifth conference on International Language Resources and Evaluation (LREC'06)*, 2006.

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

Forner P., Peñas A., Agirre E., Eneko, Alegria I., Forăscu C., Moreau N., Osenova P., Prokopidis P., Rocha P., Sacaleanu B., Sutcliffe R., Sang E. T. K. S., « Overview of the Clef 2008 multilingual question answering track », *CLEF'08 : Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, 2009.

Grappy A., Grau B., Ferret O., Grouin C., Moriceau V., Robba I., Tannier X., Vilnat A., Barbier V., « A Corpus for Studying Full Answer Justification », *Proceedings of the Seventh conference on International Language Resources and Evaluation*, 2010.

Grappy A., Ligozat A.-L., Grau B., « Evaluation de la réponse d'un système de question-réponse et de sa justification », *COnférence en Recherche d'Infomations et Applications - CORIA*, 2008.

4. Ce travail fut partiellement financé par OSEO dans le cadre du programme Quæro.

Grau B., Illouz G., Monceaux L., Paroubek P., Pons O., Robba I., Vilnat A., « FRASQUES, le système du groupe LIR, LIMSI », *Atelier EQueR, Conférence (TALN'05)*, 2005.

Grau B., Vilnat A., Ayache C., « EQueR : évaluation de systèmes de question-réponse », *in* S. Chaudiron, K. Choukri (eds), *L'évaluation des technologies de traitement de la langue : les campagnes Technolangue*, Traité IC2, Paris, Hermes, chapter 6, 2008.

Grishman R., Sundheim B., « Design of the MUC-6 evaluation », *Proceedings of the 6th Conference on Message Understanding*, 1995.

Harabagiu S. M., Paşca M. A., Maiorano S. J., « Experiments with open-domain textual Question Answering », *Proceedings of the 18th conference on Computational linguistics*, 2000.

Hatcher E., Gospodnetic O., *Lucene in Action (In Action series)*, Manning Publications Co., Greenwich, CT, USA, 2004.

Hearst M. A., « Automatic Acquisition of Hyponyms from Large Text Corpora », *Proceedings of the 14th International Conference on Computational Linguistics*, 1992.

Hovy E., Gerber L., Hermjakob U., Lin C.-Y., Ravichandran D., « Toward semantics-based answer pinpointing », *Proceedings of the first international conference on Human language technology research*, 2001.

Jacquemin C., « Syntagmatic and paradigmatic representations of term variation », *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999.

Peñas A., Rodrigo A., Sama V., Verdejo F., « Overview of the answer validation Exercise 2006 », *7th Workshop of the Cross-Language Evaluation Forum*, 2006.

Rodrigo Á., Peñas A., Herrera J., Verdejo F., « The Effect of Entity Recognition on Answer Validation », *7th Workshop of the Cross-Language Evaluation Forum*, 2006.

Rosset S., Galibert O., Illouz G., Max A., « Interaction et recherche d'information : le projet RITEL », *Traitement Automatique des Langues (TAL), numéro spécial Répondre à des questions, volume 46 :3*, 2006.

Schlobach S., Ahn D., de Rijke M., Jijkoun V., « Data-driven type checking in open domain question answering », *J. Applied Logic, volume 5 :1*, 2007.

Schlobach S., Olsthoorn M., de Rijke M., « Type Checking in Open-Domain Question Answering », *Proceedings of European Conference on Artificial Intelligence*, 2004.

Sekine S., Sudo K., Nobata C., « Extended Named Entity Hierarchy », *Proceedings of* 3^{rd} *International Conference on Language Resources and Evaluation (LREC'02)*, 2002.