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

Arnaud Grappy, Brigitte Grau[1]

LIMSI-CNRS
BP 133 ORSAY CEDEX
prenom.nom@limsi.fr

RÉSUMÉ. *Dans le cadre de la recherche de réponse à une question posée en langue naturelle dans des textes, de nombreuses questions attendent une réponse d'un certain type. Par exemple la question « Quel président succéda à Jacques Chirac ? » attend en réponse une entité du type président. La méthode présentée dans cet article vérifie que la réponse renvoyée est du bien type cherché. Pour cela elle suit une approche par apprentissage automatique en utilisant trois types de critères. Les premiers sont statistiques et calculent entre autre la fréquence d'apparition de la réponse avec le type dans un ensemble de documents. Les seconds utilisent des entités nommées et les derniers l'encyclopédie Wikipédia.*

ABSTRACT. *In open domain question-answering systems, numerous questions wait for answers of an explicit type. For example, the question "Which president succeeded to Jacques Chirac?" requires an instance of president as answer. 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 compute the presence rate 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.*

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

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

1. Et Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE)

1. Introduction

Les systèmes de questions réponses (QR) recherchent, dans un ensemble de documents, la réponse à une question posée en langage naturel (exemple : Quel président succéda a 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, et d'autre part, on ne sait pas reconnaître des instances de tous ces types : seules les entités nommées répondent à ce critère, même si l'ensemble classiquement défini ((Grishman *et al.,* 1995)) a été étendu pour les systèmes de QR avec de nouveaux types tels que des titres de film, de livre, et des sous-types plus précis tels que acteur, écrivain, chanteur, etc.

C'est pourquoi, il est crucial pour un système de QR de pouvoir procéder à une vérification dynamique du type de 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 QR 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, une question comme « 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 permet 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. A ce propos, on peut citer (Grau *et al.,* 2008) qui présente une étude des passages issus de la campagne EQueR où il manquait un mot de la question (111 passages), et dans 27% desquels le mot désignant le type était absent.

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 cooccurrence de la réponse et du type dans un ensemble de documents. 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 comme base de

connaissances. Un troisième type de méthode utilise l'encyclopédie Wikipédia[1] afin de détecter la cooccurrence de la réponse et du type dans des pages, soit à l'aide de patrons d'extraction exprimant une relation 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.

L'article commence donc par présenter un état de l'art des travaux portant sur la vérification du type des réponses. La seconde partie détaille la méthode. La troisième concerne son évaluation sur des corpus en français et l'article se termine par les conclusions et perspectives.

2. État de l'art

Tous les systèmes de questions réponses utilisent des entités nommées. Les entités nommées sont des objets textuels (mots ou groupe de mots) catégorisables 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 (Sekine *et al.,* 2002) utilise 200 type d'entités nommées, le système (Hovy *et al.,* 2001) en utilise 122 et le système (Harabagiu *et al.,* 2000) en utilise plus encore puisqu'il s'appuie sur WordNet.

Les entités nommées sont utilisées afin de sélectionner une réponse dont le type correspond bien à ce que la question attendait. 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 réponse 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) présente 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 cooccurrence 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.

Ce travail a été ensuite poursuivi et présenté dans (Schlobach *et al.,* 2007) appliqué au domaine général. Les méthodologies sont assez semblables puisque ce dernier combine des informations fournies par WordNet et des informations statistiques. La vérification grâce à WordNet s'effectue en cherchant un lien possible entre la réponse

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

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. 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, comme nous travaillons sur le français, il ne nous est pas possible d'utiliser WordNet et de nouveaux critères sont donc créés comme l'utilisation d'entités nommées. 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 seront présentés.

La vérification de type se rapproche de la recherche de relation is-a présentée initialement dans (Hearst, 1992). La méthode cherche un certain nombre d'hyponymes grâce à des patrons sémantiques du type « Réponse est type » ainsi que les patrons eux même. Pour ceci, 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 recherchés en étudiant l'apparition de ces couples et de nouveaux couples sont obtenus 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.

3. Types de réponses

Afin de déterminer le type de réponse attendu, chaque question est analysée par un analyseur robuste dont les règles reposent sur des critères syntaxico-sémantiques permettant de déterminer des groupes syntaxiques et de les typer sémantiquement éventuellement, le type d'interrogatif ainsi que la forme syntaxique des questions. Cette analyse est celle du système de questions réponses Frasques (Grau *et al.,* 2005). Elle permet d'obtenir deux types différents :

– **le type spécifique** correspond au nom de type présent dans la question. Par exemple, pour la question « Quel acteur a joué dans Danse avec les Loups ? », le type spécifique sera acteur ;

– **le type d'entité nommée** qui correspond à l'entité nommée que la question attend en retour. Dans la question précédente ce type est PERSON.

Différents cas de figure se présentent :

– le type spécifique est identique au type d'entité nommée ;

– le type spécifique est plus précis que le type d'entité nommée ;

– le type spécifique ne correspond pas à un type d'entité nommée.

Notre étude s'appuie sur un corpus d'apprentissage issu de la campagne d'évaluation de questions réponses EQueR (Grau, 2005). Celle-ci demandait à ses participants de répondre à 500 questions. Parmi ces questions, 198 explicitent un type général et sont utilisées pour créer le corpus. De nombreuses questions reprennent le même type, aussi la base contient-elle 98 noms de types différents.

La granularité de ces types est très variable. Certains comme « lieu » sont très larges, d'autres comme « bisquine » (sorte de petits bateaux) sont très précis. Cette particularité se manifeste pour des types d'entités nommées (exemple « parc ») aussi bien que pour des types ne correspondant pas à des entités nommées (exemple « traitement »). Cela illustre bien le fait que l'on ne peut prédire tous les types.
Les réponses à ces questions correspondent à celles données par les participants à cette campagne, qui pouvaient donner 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, la réponse est effectivement du type attendu.

La suite de l'article présente l'ensemble des différentes méthodes utilisées. Celles-ci peuvent être organisées en trois catégories : l'utilisation des entités nommées, l'exploitation de Wikipédia et l'utilisation de mesures statistiques.

4. Utilisation de systèmes de reconnaissance d'entités nommées

La première stratégie est l'utilisation des entités nommées reconnues afin de vérifier le type d'une entité. Cette vérification permet d'une part de rejeter certaines réponses clairement reconnues comme mauvaises, i.e. d'une classe d'entité nommée différente, d'autre part elle permet de détecter certaines réponses comme correctes. Rappelons que nous nous situons dans le cadre général où étant donnés une question, une réponse et un passage, nous voulons décider si la réponse est du type spécifique attendu.

4.1. *Filtrer les réponses*

La première utilisation des entités nommées est le rejet des réponses dont le type d'entité nommée ne correspond pas à celui que la réponse 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 une personne.

Nous retenons les types d'entités nommées attendues déterminées par le module d'analyse des questions du système Frasques (Grau *et al.,* 2005). 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 quatre types classiques (personne, lieu, organisation, date). Par exemple, le type lieu regroupe un type d'entité « ville » et un type « pays ». A ces classes d'entités sont ajoutées des entités permettant de reconnaître les expressions numériques (longueur, vitesse, etc.) et une autre, NomPropre, qui étiquette tous les termes non étiquetés contenant un nom propre. Ce dernier type permet de pallier les absences de reconnaissance des entités nommées en utilisant un type plus global. 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. Dans ce cas la réponse est vue comme mauvaise et la valeur NON est renvoyée.

– La question attend une entité nommée en réponse et celle-ci est d'un type comparable : soit le même, soit de la même catégorie par exemple Lieu ou NomPropre pour pays. Dans ce cas la réponse est vue comme étant presque du type et la valeur OUI est renvoyée.
Cette vérification ne permet d'avoir qu'une idée globale de la validité de la réponse. En effet pour une question dont le type spécifique est « président », une réponse reconnue comme une personne satisfait ce module et par exemple la réponse « Michel Rocard » est vue comme vraie.

– La question attend une entité nommée en réponse et la réponse est une entité nommée dont un rapprochement avec celle attendue ne peut être fait. Par exemple la réponse « 300 » pour une question attendant un lieu. Dans ce cas, la réponse sera reconnue comme n'étant pas du bon type et la valeur NON est renvoyée.

L'évaluation de cette méthode sur le corpus d'apprentissage permet d'obtenir les résultats présentés dans le tableau 1. Ce tableau présente plus spécifiquement le lien entre la valeur obtenue et le fait que la réponse soit ou non du type. Les résultats marqués en gras correspondent aux bonnes décisions.

valeur donnée	# 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. *Résultats de la méthode travaillant par entités nommées*

Ce tableau montre que lorsque la réponse est vue comme n'étant pas du bon type d'entité nommée alors la réponse n'est généralement pas du type spécifique attendu par la question (à 71 %). Quand la réponse est bien du type d'entité nommée, la différence de pourcentage montre qu'il est délicat de faire une supposition quant à la correspondance de la réponse avec le type spécifique attendu (seul 63 % de réponses sont du type spécifique quand il y a correspondance au niveau des entités nommées).

Le tableau 2 présente une autre évaluation de cette méthode. Il montre entre autre que le rappel est assez bas, ce qui est dû au grand nombre de réponses dont la valeur est INCONNU.

précision	rappel	f-mesure
0,65	0,45	0,53

Tableau 2. *Évaluation de la stratégie*

Comme la plupart des systèmes de questions réponses utilisent une vérification par entité nommée assez semblable, les résultats obtenus par cette méthode seront des résultats de base qu'il faudra améliorer.

4.2. *Valider des réponses*

Les entités nommées peuvent aussi être utilisées comme des bases de connaissances. En effet, les entités nommées reconnues dans un grand corpus permettent de construire des listes de mots correspondant aux types reconnus. Il semble donc intéressant de tester la présence ou l'absence de la réponse dans les listes correspondant au type cherché. Sa présence indiquerait qu'elle est bien un hyponyme du type. Afin d'avoir des informations pertinentes, il est nécessaire de collecter un grand nombre d'entités nommées.

Le module d'entités nommées du système de questions réponses RITEL (Rosset *et al.,* 2005) a été utilisé dans ce but. Il contient un ensemble de 274 types pouvant être assez précis comme « religion » ou « fleuve ». Comme le nombre de type d'entités nommées est limité il y aura toujours des types spécifiques ne correspondant pas à une entité nommée. La méthode teste la présence de la réponse dans la liste correspondant au type. Trois cas sont alors possibles :

– Il n'y a pas de correspondance entre le type spécifique de la réponse et l'un des types d'entité nommée. Dans ce cas le module ne peut pas savoir si elle est du type et renvoie donc INCONNU.

– La réponse ne se trouve pas dans la liste d'instances associées au type. Dans ce cas elle est vue comme n'étant pas de ce type et la réponse NON est renvoyée.

– La réponse se trouve dans la liste correspondant au type. Elle est donc considérée comme étant de ce type et la réponse OUI est renvoyée.

Les tableaux 3 et 4 permettent d'évaluer la méthode. Ils montrent que quand une correspondance de type est vue, alors la valeur est très souvent bonne (précision 0,75). Le faible rappel s'explique par le fait que seuls peu de couples réponse/type sont évalués (43 %).

valeur donnée	#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 3. *Résultats de la méthode RITEL*

précision	rappel	f-mesure
0,75	0,32	0,45

Tableau 4. *Évaluation de la validation par entités nommées*

5. Utilisation de Wikipédia

5.1. *Recherche dans des pages particulières*

Cette méthode part du constat que comme Wikipédia[2] est 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 de ce type.

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

– Aucune page ne contient la réponse comme titre, dans ce cas rien ne peut être déduit grâce à cette méthode et la valeur INCONNU est renvoyée.

– La page correspondant à la réponse contient bien le type. Ceci implique que la réponse a de fortes chances qu'elle soit de ce type et la valeur OUI est renvoyée.

– La page ne contient pas le type. Dans ce cas elle n'est très probablement pas de ce type. La valeur NON est donc renvoyée.

Les tableaux 5 et 6 présentent les résultats obtenus par cette méthode.

valeur donnée	#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 5. *Résultats de la méthode*

Le premier tableau montre que la méthode peut être fiable quand elle renvoie OUI (la réponse est du type à 74%) mais beaucoup moins quand elle renvoie NON (la

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

précision	rappel	f-mesure
0,68	0,32	0,43

Tableau 6. *Évaluation de la méthode*

réponse n'est pas du type à 61%) . Ceci peut s'expliquer par le fait que le type peut être remplacé dans la page de la réponse par l'un de ses synonymes.
Les résultats montrent également que peu de réponses sont évaluées : beaucoup de réponses n'ont pas une page Wikipédia qui leur est consacrée.
Le principe de cette vérification semble donc correct mais ne permet pas de couvrir tous les cas possibles. Les critères suivants poursuivent cette idée en essayant d'élargir la 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 dans lesquelles apparaissent le type et la réponse indiquent que la réponse est du type. Par exemple, **Réponse est un Type**, indique que la réponse est bien du type. Cinq grands types de règles, issus d'une analyse des corpus, sont utilisés :

– **RÉPONSE être déterminant TYPE** (exemple Nicolas Sarkozy est le président).

– **TYPE RÉPONSE** (le 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)

Afin de savoir si la réponse est bien du type, pour chaque couple réponse/type spécifique, chacune des règles est instanciées avec les bonnes valeurs, TYPE prend la valeur du type spécifique et RÉPONSE celui de la réponse. Puis une requête est calculée grâce à l'ensemble de ces règles et fournie au moteur de recherche Lucene (Hatcher *et al.,* 2004). Celui-ci cherche la présence de l'une d'entre elle dans les documents Wikipédia. Si un document est trouvé alors la réponse est considérée comme étant du type et la valeur OUI est renvoyée sinon la réponse est considérée comme n'étant pas du type et la valeur NON est donnée.

Les tableaux 7 et 8 montrent les résultats obtenus par cette méthode. Ils montrent tout d'abord que la méthode peut être assez fiable quand la réponse OUI est renvoyée (73 % des réponses sont bien du type). Ces tableaux montrent 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.

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

Tableau 7. *Résultats de la méthode*

précision	rappel	f-mesure
0,66	0,66	0,66

Tableau 8. *Évaluation de la méthode*

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, quelque soit le document ou la relation les liant. Cette utilisation peut ainsi indiquer que si le type et la réponse cooccurrent souvent dans des 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, un premier système travaillant par apprentissage a été créé. Ce système reprend un ensemble de critères décrits dans (Schlobach *et al.,* 2004) et (Schlobach *et al.,* 2007). Les critères utilisés 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 ne contenant que la réponse ou que le type. Ce critère permet de détecter les cas où la réponse apparaît fréquemment accompagnée du type.

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

– **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.

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 [3].

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

Comme cette méthode travaille par apprentissage, son évaluation ne peut se faire sans base de test. Celle-ci est décrite dans la section suivante et les résultats, présentés dans le tableau 9, montrent que la méthode obtient une précision, un rappel et une f-mesure de 0,68. Ces bons résultats montrent l'intérêt de cette méthode.

Une seconde méthode consiste à changer l'ensemble des documents. Au lieu d'utiliser Wikipédia, cela peut être intéressant de tester les données sur les documents dans lesquelles la réponse a été extraite. En effet, parfois la réponse est une entité très précise ayant eu une brève apparition historique et n'est pas assez significative pour être présente dans Wikipédia mais se trouvera dans les articles de journaux. Ces documents correspondent aux articles du journal Le Monde allant de l'année 1992 à l'an 2000. Cette recherche obtient des résultats légèrement supérieurs à ceux obtenus par la recherche dans les pages Wikipédia (f-mesure 0,70).

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 présente dans le logiciel 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, celui 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 traitant de la recherche en corpus plus les résultats de chacune des autres méthodes. Ces critères sont donc :

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 (Pointwise Mutual Information),

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

En faisant un premier apprentissage avec la base d'apprentissage égale à la base de test, 90 % de réponses sont bien classées (étant du type spécifique ou ne l'étant pas).

8. Évaluation

Les sections précédentes ont présenté la méthode ainsi qu'une première évaluation. Afin de mieux comprendre l'intérêt de cette méthode, il faut maintenant l'évaluer, ce qui est fait de trois manières différentes :

– La première évalue les critères pris séparément sur la base de test ;

– La seconde évalue la combinaison des méthodes ;

– La dernière étudie l'intérêt de la vérification du type pour les systèmes de questions réponses.

La base de test est construite à partir des données fournies par la campagne AVE 2006 (Penas *et al.,* 2006). Dans celle-ci, un ensemble de questions étaient proposées aux participants ainsi que leurs réponses potentielles, associées à des passages justificatifs, et la tâche consistait à détecter si la réponse est correcte ou non. Les réponses et les passages sont ceux produits par des systèmes de questions réponses.

Les données de cette campagne permettent d'utiliser 1547 paires réponses/types spécifiques dont la moitié (763) 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 9 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	0,70	0,70	**0,70**

Tableau 9. *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). Ceci s'explique par la répartition des données. En effet, les données de la base de test contiennent plus souvent des personnes et donc il existe plus de pages Wikipédia ayant comme titre ces noms.

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 est ou non du 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 10 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 10. *Résultats globaux*

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

Une étude distinguant les cas où la réponse a un type d'entité nommée associé 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 tableau 11 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 ont un type d'entité nommée (78 %).

base de test	proportion de réponses correctes
avec EN (1205)	82 %
sans EN (342)	74 %

Tableau 11. *Vérification du type en fonction des entités nommées*

Valeur donnée	#réponses du type spécifique	#réponses pas du type spécifique
OUI	**603 (80%)**	149 (20%)
NON	159 (20%)	**636 (80%)**

Tableau 12. *Matrice de confusion*

Le tableau 12 présente la matrice de confusion de la méthode. Il montre qu'elle obtient des résultats similaires quelque soit la valeur retournée (OUI ou NON).

Avant de clore cette section, voyons quelques résultats obtenus :

– Hosni Moubarak est bien vu comme un président

– Yasser Arafat est vu comme n'étant pas un président

– Krypton est bien vu comme une planète

– Bethléem est bien vu comme n'étant pas une planète.

– Barings n'est malheureusement pas vu comme étant une « grande banque ». Cela est surement dû à la présence de l'adjectif.

– et le dow jones est vu, à tort, comme étant une entreprise. Ces deux mots sont liés et se trouvent donc régulièrement ensemble dans des documents ce qui a entrainé la mauvaise classification.

Les résultats peuvent aussi être rapprochés de ceux cherchant toutes les entités nommées présentes dans un corpus de texte. La campagne MUC (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 ce qui s'explique par le grand nombre de types que nous avons utilisés.

8.3. *Apport pour un système de questions 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. Pour ce faire, nous avons utilisés les résultats de la campagne AVE 2006. Les réponses à évaluer sont fournies par des systèmes de questions réponses recherchant une réponse aux différentes questions. Parmi ces réponses très peu sont correctes : 289 sur 1457 soit 20 %. Le tableau 13 présente une correspondance entre la vérification du type et la validité de la réponse. Le fait d'améliorer ces résultats montre donc l'apport de notre méthode pour des systèmes de questions réponses.

Ce tableau montre aussi que si la réponse est vue comme n'étant pas du type alors elle est très souvent mauvaise (92 %) ; en revanche rien ne peut être dit quand la réponse est vue comme étant du type. Le très grand nombre de mauvaises réponses implique que les réponses vues comme étant du type sont encore le plus souvent in-

réponse du type	# réponses correctes	# réponses mauvaises
OUI (698)	**233 (34 %)**	465 (66 %)
NON (759)	56 (8 %)	**703 (92 %)**

Tableau 13. *Rapport entre la vérification du type et la validité des réponses*

correctes (66 %). En effet, ce seul critère est insuffisant pour décider de la validité d'une réponse.

La vérification du type peut s'appliquer comme un filtre consistant à éliminer les réponses n'étant pas du type. Dans ce cas, la proportion de réponses correctes de la base de test passe de 20% à 34 %. Toutefois le nombre de réponses correctes diminue lui aussi passant de 289 à 233, ce qui constitue une diminution de 19 %.

9. Conclusion et perspectives

Nous avons vu dans cet article une méthode permettant de vérifier qu'une réponse est ou non du type spécifique attendu par une question. Cette méthode est fondée sur un apprentissage à partir de critères assez différents comme la vérification d'entités nommées, des mesures statistiques d'apparition dans des corpus ou l'utilisation de l'encyclopédie Wikipédia. La méthode obtient de bons résultats qui témoignent de son efficacité. Les résultats concernant les systèmes de questions réponses sont mitigés sur le plan qualitatif puisque si la précision augmente significativement, cela s'accompagne d'une baisse du nombre de bonnes réponses total proposées. Cela signifie qu'il faut étudier plus précisément la manière d'utiliser cette méthode pour sélectionner ou classer des réponses : en tant que filtre (ce qui a été fait dans cet article) ou en tant que critère se combinant à ceux qui existent dans les systèmes de QR.

Le système (Grappy *et al.,* 2008) a étudié la validité des réponses par une méthode par apprentissage. Il utilise une vérification simple du type de la réponse. La vérification du type des réponses pourrait être être une indication contribuant à la mesure de confiance qu'a le système en ses résultats.

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. Une telle validation permettrait de comprendre automatiquement le sens de la question. Dans ce cas, la vérification du type constitue un premier travail, une des nombreuses informations est vérifiée, qu'il faut encore poursuivre pour valider entièrement les réponses. [4]

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

10. Bibliographie

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

Grau B., « EQueR, une campagne d'évaluation des systèmes de question/réponse », *Journée Technolangue/Technovision (ASTI'2005)*, 2005.

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., *L'évaluation des technologies de traitement de la langue : les campagnes Technolangue*, Traité IC2, série Cognition et traitement de l'information, Lavoisier, chapter 6, évaluation de systèmes de question-réponse, 2008.

Grishman R., Sundheim B., « Design of the MUC-6 evaluation », *MUC*, p. 1-11, 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*, Association for Computational Linguistics, Morristown, NJ, USA, p. 292-298, 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 », *In Proceedings of the 14th International Conference on Computational Linguistics*, p. 539-545, 1992.

Hovy E., Gerber L., Hermjakob U., Lin C.-Y., Ravichandran D., « Toward semantics-based answer pinpointing », *HLT '01 : Proceedings of the first international conference on Human language technology research*, Association for Computational Linguistics, Morristown, NJ, USA, p. 1-7, 2001.

Penas A., Rodrigo A., Sama V., Verdejo F., « Overview of the Answer Validation Exercise 2006. », *CLEF*, 2006.

Rosset S., Galibert O., Max A., « Interaction et recherche d'information : le projet RITEL », *Traitement Automatique des Langues*, 2005.

Schlobach S., Ahn D., de Rijke M., Jijkoun V., « Data-driven Type Checking in Open Domain Question Answering », *J. of Applied Logic*, vol. 5, n° 1, p. 121-143, 2007.

Schlobach S., Olsthoorn M., Rijke M. D., « Type Checking in Open-Domain Question Answering », *In Proceedings of European Conference on Artificial Intelligence*, IOS Press, p. 398-402, 2004.

Sekine S., Sudo K., Nobata C., « Extended Named Entity Hierarchy », *Proceedings of* 3^{rd} *International Conference on Language Resources and Evaluation (LREC'02)*, Canary Islands, Spain, p. 1818-1824, May, 2002.