Fusion des réponses de systèmes de questions-réponses

Arnaud Grappy, Brigitte Grau, Sophie Rosset

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

RÉSUMÉ. *Les réponses données par plusieurs systèmes de questions-réponses proviennent de l'application de stratégies différentes, et de ce fait permettent de répondre à des questions différentes. La combinaison de ces systèmes vise alors à accroître le nombre total de questions résolues. Cet article présente la combinaison de trois systèmes : QAVAL, qui s'appuie sur un module de validation de réponses et deux versions du systèmes RITEL qui s'appuie sur une analyse multi-niveaux appliquée aux questions et aux documents. La fusion des résultats est effectuée de différentes manières : en fusionnant les passages, à la sortie des systèmes par vote ou fusion en tenant compte du poids ou du rang des réponses proposées et par un mécanisme d'apprentissage sur les caractéristiques des réponses.*

ABSTRACT. *Question answering systems answer correctly to different questions because they are based on different strategies. In order to increase the number of questions which can be answered by a single process, we propose solutions to combine three question answering systems, QAVAL and two versions of RITEL. QAVAL is based on an answer validation method and RITEL develops a multi-level analysis of questions and documents. In order to merge the systems results, we developed different methods either by merging passages before answer ordering, or by merging end-results. The fusion of end-results is realized by voting, merging, and by a machine learning process on answer characteristics.*

MOTS-CLÉS : *fusion de réponses systèmes de questions-réponses réordonnancent de réponses*

KEYWORDS: *answer fusion, question answering system, system combination, answer reranking*

CORIA 2012, pp. 99–110, Bordeaux, 21-23 mars 2012

1. Introduction

Les systèmes de questions-réponses (QR) fournissent des réponses courtes extraites de textes à des questions posées en langue naturelle et portant sur tout domaine. Si la plupart des systèmes suivent une architecture pipeline comportant trois grands modules : analyse des questions, extraction de passages et extraction de réponses, les composants de chacun sont spécifiques. De ce fait, leurs performances, même si elles sont globalement comparables, reflètent de grandes disparités lorsque l'on compare les questions qui sont résolues par chacun d'eux. Aussi, afin de tirer partie des différences des systèmes, une nouvelle problématique a émergé en QR, à savoir comment combiner les réponses correctes données par les systèmes.

Cet article porte sur ce problème, et propose plusieurs méthodes de combinaison. Comme nous disposons de deux systèmes, QAVAL et RITEL, nous proposons de fusionner des résultats en interne, en exploitant des résultats intermédiaires et les stratégies propres à chaque système pour classer des candidats réponse, et de manière externe, à la sortie des systèmes. Ces systèmes fournissant différents éléments associés aux réponses comme leur score et leur rang, nous pourrons nous appuyer sur ces valeurs afin d'en proposer une combinaison. De plus, de manière à exploiter au mieux les caractéristiques des systèmes, nous avons mis en œuvre un apprentissage fondé sur les caractéristiques des réponses : nature du type de réponse attendu, redondance de la réponse, rang, score. Nous verrons que plusieurs de ces stratégies permettent une augmentation des résultats et qu'un apprentissage permet d'augmenter le nombre de bonnes réponses au rang 1 de près de 20 %.

Cet article présente dans un premier temps un état de l'art, puis après la description des systèmes utilisés, nous détaillerons les différentes méthodes de fusion proposées. Celles-ci seront évaluées sur le même corpus que celui sur lequel les systèmes ont été évalués individuellement.

2. État de l'art

De nombreuses approches traitent de la fusion des sorties de différents systèmes de questions réponses. (Jijkoun *et al.,* 2004) combinent ainsi le résultat de différentes méthodes d'extraction de réponses. La première exploite une base de connaissances, la deuxième des patrons d'extraction et une troisième les n-grammes les plus proches des mots de la question. La méthode de fusion proposée tient compte de la similarité des réponses. Deux réponses sont dites similaires si elles sont égales, si l'une est contenue dans l'autre ou si la distance d'édition est inférieure à une valeur seuil donnée. Une pondération des réponses est alors faite en sommant les scores de confiance des différents systèmes, ces scores pouvant être normalisés par une méthode dépendant du système. Les réponses sont ordonnées en fonction du poids obtenu et une amélioration du nombre de bonnes réponses en première position de 31 % est observée.

(Tellez-Valero *et al.,* 2010) combinent la sortie de différents systèmes de questions réponses en utilisant un module de validation de réponses. Ce module suit une ap-

proche par apprentissage qui combine différents critères tenant compte de la catégorie de la question, du type de réponse attendu, de restrictions sur la réponse, de la compatibilité entre la réponse et la question, du fait que la réponse soit extraite par des systèmes différents ainsi que de la proportion de termes de la question présent dans le passage. Cette fusion est appliquée aux systèmes de CLEF sur l'espagnol et les résultats de la combinaison dépasse ceux du meilleur système, le MRR[1] passant de 0,62 à 0,73.

(Aceves-Pérez *et al.,* 2008) utilisent des mesures traditionnelles de fusion de réponses appliquées aux systèmes de questions réponses multilingues. En effet, un simple interclassement des réponses et un tri suivant le score de confiance avec ou sans normalisation sont utilisées. Ils ont montré que la meilleure combinaison tenait compte du score de confiance normalisé et dépasse les résultats obtenus par un système monolingue, le MRR sur 5 réponses allant de 0,64 à 0,75. Leur article présente également la fusion des passages avant l'extraction des réponses en utilisant les mêmes méthodes et il a été montré que la fusion des réponses était plus pertinente que la fusion des passages.

(Chalendar *et al.,* 2002) combinent les résultats obtenus par le même système, QALC, qui recherche la réponse à une question dans deux ensemble de documents : le Web et les articles de journaux. Une des méthodes consiste alors à modifier le score de confiance étant donné le rang de la réponse si elle a été obtenue dans les deux collections.

Par la suite nous présenterons différentes méthodes de fusion de réponses. L'une d'entre elle est proche de l'approche de (Tellez-Valero *et al.,* 2010) et utilise un module de validation de réponses. D'autres sont des méthodes de vote qui tiennent compte du rang des réponses extraites ainsi que de leur score de confiance.

Cette tâche peut également se rapprocher de celle qui consiste à fusionner les résultats provenant de systèmes de recherche d'information. Parmi les différentes approches, (Shaw *et al.,* 1994) présentent différentes combinaisons tenant compte du score de confiance des différents systèmes. Ainsi, l'une d'entre elles tient compte du score le plus élevé et d'autres de la somme pouvant être pondérée des différents scores. Ces méthodes sont également utilisées pour notre travail.

3. Les systèmes de questions réponses

3.1. *Le système QAVAL*

3.1.1. *Présentation globale*

Dans le système QAVAL (Grappy *et al.,* 2011) , les questions sont tout d'abord analysées afin d'en obtenir leurs informations pertinentes telles que le type de réponse attendu ou les mots clés. Puis de courts passages de textes, de 300 caractères environ,

1. Mean Reciprocal Rank

sont recherchés en utilisant le moteur de recherche Lucene. Les passages obtenus sont ensuite annotés afin de marquer les termes de la question et leurs variantes (morphologiques et de synonymie) ainsi que certaines réponses potentielles (entités nommées, application de patrons d'extraction). De nombreuses réponses sont ensuite extraites. Si la question attend une réponse relevant d'un certain type d'entité nommée (« Quand a eu lieu la chute du mur de Berlin ? ») alors toutes les entités du bon type présentes dans les passages retournés seront extraites. Dans le cas contraire (« Quel film reçu la palme d'Or en 1998 ? ») tous les groupes nominaux des passages sont extraits. Les réponses sont ensuite ordonnées par un module de validation de réponses.

3.1.2. *L'ordonnancement des réponses par validation*

Le module d'ordonnancement de réponses a pour but de fournir un score de confiance aux différents candidats. A cette fin, nous avons développé une approche par apprentissage qui combine des critères de différentes natures fondé sur une combinaison d'arbres de décision par la méthode bagging.

Les premiers critères portent sur la présence dans le passage des termes de la question. Les seconds (calcul de la plus longue chaîne de mots de la question et de la réponse dans le passage, calcul de la distance entre la réponse et les termes de la question) permettent de vérifier que la réponse est reliée aux termes de la question.

Certaines questions attendent une réponse d'un type particulier, sans que celui-ci soit forcément un type d'entité nommée. Par exemple « Quel film remporta la palme d'or en 1998 ? » attend un nom de film en réponse. Un critère évalue la compatibilité entre la réponse et le type attendu calculé par l'utilisation d'un système de vérification (Grappy *et al.,* 2010) .

Les autres critères sont relativement fréquents dans les systèmes de questions réponses. Ainsi, on retrouve le rang du passage duquel la réponse a été extraite, la catégorie de la question et la redondance de la réponse.

3.2. *Le système RITEL*

Le système Ritel, complètement décrit dans (Bernard *et al.,* 2009), a été conçu dans l'optique d'un système de dialogue (Toney *et al.,* 2008). Il s'appuie sur une analyse multi-niveaux qui est appliquée sur les questions et les documents. Les documents sont totalement analysés puis indexés. La recherche est effectuée dans cet index complet. Cette analyse permet de repérer et typer des éléments pertinents d'information comme des entités nommées, des chunks morpho-syntaxiques ou des actes de dialogue (Galibert, 2009).

La première étape de ce système est de construire un *descripteur de recherche* (DDR) qui contient toutes les informations utiles pour la recherche de documents, l'extraction de passages pertinents et l'extraction des candidats réponses : les éléments de la question avec leurs transformations possibles (dérivations morphologiques, sy-

nonymes) et leurs poids ainsi que les types de réponse attendus. Ces types, le plus souvent des types d'entités (*personne*, *lieu* ...), correspondent à la taxonomie utilisée par le système d'analyse.

Deux méthodes différentes sont utilisées pour l'extraction et l'évaluation des candidats réponses. La première s'appuie sur la redondance de ces mêmes candidats et leur position dans les documents et les passages, la seconde sur des modèles bayésiens. Les deux approches exploitent la même extraction de passages et la même analyse (le même DDR est utilisé).

Extraction de candidats fondée sur des distances *(Ritel Standard)*

On considère que tous les éléments de tous les passages correspondant à un type possible de réponse sont des candidats réponses. A chacun de ces candidats réponses, un score est associé. Pour le calculer, chaque élément du DDR s'additionne au score du candidat réponse proportionnellement à son poids et inversement proportionnelle à sa distance au candidat. Le score est ensuite pondéré par le score du passage et par la fréquence du candidat réponse dans tous les documents et les passages. Les différents scores d'une même réponses sont enfin additionnés.

Modèles bayésiens pour l'extraction des candidats *(Ritel Probabiliste)*

Il s'agit de modéliser le processus d'estimation de la qualité d'une réponse par un modèle bayésien. Cette approche s'appuie sur de nombreux modèles élémentaires dont les probabilités de co-occurrence d'éléments, la probabilité d'apparition d'un élément de question dans le contexte proche du candidat réponse et en dehors de ce contexte.

3.3. *Combinaison des systèmes*

Les systèmes QAVAL et RITEL appliquent des stratégies différentes pour sélectionner des passages et pondérer les réponses. C'est pourquoi il nous a paru intéressant de combiner ces stratégies de manière interne aux systèmes : des passages de QAVAL sont donnés en entrée de RITEL et des réponses de RITEL sont données pour validation à QAVAL, exploitant ainsi des résultats intermédiaires.

Par ailleurs, de manière plus classique, nous avons testé et proposé plusieurs méthodes de fusion de réponses opérant à la sortie des systèmes et reposant sur les informations fournies par ces systèmes : le score des réponses, leur rang, leur redondance. Nous présentons dans les sections suivantes les différentes méthodes proposées afin de combiner les résultats provenant des trois systèmes de questions réponses, à savoir QAVAL, RITEL probabiliste et RITEL Standard.

4. Combinaison interne

4.1. *Combinaison de la sélection de passages*

RITEL applique une indexation fine des documents pour extraire des passages et obtient 80,3 % de questions ayant au moins un passage contenant la bonne réponse. QAVAL sélectionne d'abord de courts passages (150) par application de Lucene sur des documents ayant subi un stemming et obtient un score de 88 %. Nous avons voulu faire collaborer ces deux modes de recherche en appliquant la recherche de réponses de RITEL dans les passages retrouvés par QAVAL.

Tout d'abord l'analyseur de RITEL est passé sur l'ensemble des passages et le résultat de cette analyse est indexé. Ensuite la procédure décrite dans la section 3.2 est appliquée : analyse des questions, construction du DDR, extraction des documents puis des passages et enfin extraction et évaluation des candidats réponses. Les deux méthodes d'extraction et scoring des réponses ont été appliquées.

	Tous les documents		Passages QAVAL	
	Ritel-S	Ritel-P	Ritel-S	Ritel-P
top-1	32,0%	22,4%	29,9%	22,4%
MRR	0,41	0,29	0,38	0,32
top-20	61,2%	48,7%	54,4%	49,7%

Tableau 1. *Résultats de la combinaison de QAVAL dans RITEL*

On constate (cf tableau 1) une dégradation de tous les résultats sauf pour le MRR de Ritel Probabiliste. Cette approche n'est donc pas concluante, sans doute due au fait que les passages de QAVAL sont trop courts et ne répondent pas bien aux critères de RITEL utilisés pour évaluer leur pertinence.

4.2. *Validation de réponses*

L'ordonnancement des réponses du système QAVAL est effectué par un module de validation de réponses. L'idée propre à cette méthode, semblable à celle présentée par (Tellez-Valero *et al.,* 2010), consiste à utiliser ce module (cf 3.1) afin de traiter les réponses des différents systèmes. La validation de réponses a pour but de décider si la réponse est valide en examinant le passage de texte l'accompagnant. Comme les différentes réponses sont accompagnées d'un passage de texte permettant de les justifier dans les trois systèmes, il nous est possible d'utiliser cette méthode sur les réponses provenant de RITEL. De plus, ceci permet d'évaluer l'applicabilité du module de validation de réponses à des systèmes différents. Malheureusement, les tests ont montré que cette combinaison ne permet pas de dépasser les scores du meilleur système. Cela peut être dû au fait que la taille des passages renvoyés par RITEL est différente de celles des passages renvoyés par QAVAL alors que le système n'a pas été entrainé sur de tels passages.

5. Fusion des résultats par vote et combinaison des scores

Les méthodes de fusion de réponses reposent sur la détection que deux réponses sont égales. Comme les chaînes de caractères extraites peuvent être différentes et donner la même réponse , il est nécessaire de définir l'égalité de deux réponses.

Une réponse R_1 est considérée incluse dans une réponse R_2 si tous les mots non vides de sens de R_1 sont contenus dans R_2. Pour tester l'égalité de deux mots, nous avons utilisé l'égalité des lemmes. Deux réponses R_1 et R_2 sont considérées comme égales si R_1 est incluse dans R_2 et R_2 est incluse dans R_1.

5.1. *Mesures calculées sur les rangs des réponses*

La première information utilisée afin de combiner les systèmes tient compte du rang des réponses extraites. En effet, comme les systèmes fonctionnent relativement bien, la bonne réponse se trouve le plus souvent dans les premiers rangs.

5.1.1. *Interclassement des réponses*

La méthode la plus simple consiste à alterner les réponses différentes provenant de chaque système et à placer en première position la première réponse du premier système, en seconde position la première réponse du second système et ainsi de suite. L'ordre du jeu de réponses des systèmes a été défini en fonction des résultats obtenus. Le jeu de réponses de QAVAL est ainsi le premier, celui de RITEL standard le second et enfin celui de RITEL probabiliste. Notons que cette méthode ne peut pas placer d'avantage de bonnes réponses en première position que le meilleur système.

5.1.2. *Somme de l'inverse du rang*

La méthode que nous proposons ici calcule un poids pour chacune des réponses correspondant à la somme de l'inverse des rangs obtenus pour la même réponse par les différents systèmes (cf. formule 1). Ainsi si un système place une réponse en première position et que le suivant trouve la même réponse en seconde position le poids sera de 1,5 ($1 + \frac{1}{2}$). Cette méthode permet ainsi de placer davantage de bonnes réponses en première position en privilégiant notamment la première réponse d'un système si cette même réponse a été trouvée par un autre système.

$$poids = \sum \frac{1}{rang} \qquad [1]$$

5.2. *Mesures fondées sur les scores de confiance*

Lorsqu'une réponse est renvoyée, elle est accompagnée d'un score qui permet d'évaluer la confiance que le système a dans la réponse produite. Cette valeur est généralement utilisée par les systèmes de questions réponses afin d'ordonner leurs réponses. Comme les systèmes ne renvoient pas un score dans le même intervalle, les

scores provenant des différents systèmes de RITEL ont été normalisés afin de se placer dans l'intervalle correspondant aux scores de confiances de QAVAL ($[-1, 1]$) et une formule de régression linéaire a été appliquée (cf. formule 2).

$$valeur_{normalise} = \frac{2 * valeur_{origine}}{val_{Min} - val_{Max}} - 1 \quad [2]$$

5.2.1. *Somme des scores de confiances*

De manière à comparer nos méthodes de fusion aux méthodes classiques, nous avons mis en place deux des méthodes présentées dans (Shaw *et al.,* 1994) :

– **CombSum** qui somme les différents scores de confiance d'une réponse donnés par les différents systèmes ;

– **CombMNZ** qui somme les scores de confiance des différents systèmes et multiplie cette valeur par le nombre de systèmes ayant extrait la réponse.

5.2.2. *Méthode hybride*

Une méthode hybride, combinant des informations sur les rangs des réponses et les poids, a été définie. La mesure de pondération correspond à la somme de deux termes : le score de confiance le plus élevé pour la réponse et un terme qui tient compte du rang de la réponse dans les différents jeux de réponses.

$$poids(r) = score(r) + \prod bonus\_egalite * (|reponse(q)| - \sum rang(r)) \quad [3]$$

avec *r* une réponse et *q* une question. Le bonus d'égalité, *bonus_egalite*, est calculé pour chaque couple de systèmes et reçoit la valeur 3 si les deux réponses sont égales, 2 si une réponse est contenue dans l'autre et 1 sinon. Le premier score a donc une valeur comprise entre -1 et 1 et le second peut être très élevé. Ainsi la méthode va favoriser les réponses extraites par plusieurs systèmes et une réponse extraite par un seul système en première position sera probablement rétrogradée.

6. Fusion par apprentissage

La méthode la plus élaborée que nous avons mise en place consiste à combiner par apprentissage différents critères permettant de caractériser les réponses . Le premier type de critère correspond au rang et au score de confiance obtenu par les différents systèmes. Ces deux informations sont très pertinentes et ont été utilisées par les méthodes précédentes. Si une réponse n'est pas reconnue par un système alors son poids et son rang seront de -2 ce qui permet de la sortir de l'intervalle des valeurs classiques. Comme précédemment nous utilisons la normalisation des scores de confiance du système RITEL et notre détection d'égalité. Notons que certaines réponses ont tout d'abord été filtrées si le passage associé ne contient pas une entité nommé de la question ou que la réponse ne correspond pas au type d'entité nommé attendu.

Une autre information fournie sous forme de critère, indique quels sont le ou les systèmes ayant renvoyé la réponse. Cela permet entre autre de distinguer les cas où la même réponse a été extraite par QAVAL et RITEL standard donc deux systèmes totalement différents des cas où la réponse a été extraite par les deux systèmes RITEL qui comportent des modules communs.

Un autre critère encode le fait que certaines questions attendent une réponse étant d'un type spécifique, plus précis que les types d'entités nommées classiquement utilisés. La valeur du critère indique alors si la réponse est compatible avec le type de la question (une instance ou un hyponyme) et ce critère devrait permettre de déclasser des réponses erronées. La méthode, déjà utilisée dans QAVAL pour évaluer cette compatibilité (Grappy *et al.,* 2010), est une combinaison par apprentissage des résultats de différentes méthodes de vérification comme l'utilisation des entités nommées, l'exploitation des pages Wikipédia ou des recherches de coocurrences en corpus.

Le dernier critère que nous avons mis en place concerne la redondance de la réponse. Dans les systèmes de questions réponses, le fait qu'une réponse a été extraite de nombreuses fois est un indice pertinent et généralement utilisé pour ordonner les réponses. Il permet de détecter une relation fréquente entre la réponse et les mots de la question. Le critère utilisé correspond au nombre de documents duquel la réponse a été extraite. En pratique, la valeur du critère est égale à la somme des redondances.

Après avoir défini l'ensemble des critères, il reste à les combiner, ce qui est fait grâce à SVMRank (Joachims, 2006). C'est une méthode d'ordonnancement fondée sur les SVM. Nous l'avons utilisée car elle obtient de bons résultats et est pertinente pour notre tâche d'ordonnancement. La base d'apprentissage est calculée à partir des réponses calculées par les trois systèmes sur les questions issues de la campagne d'évaluation proposée dans le cadre du projet QUAERO en 2009. Notons que seules les questions ayant au moins une bonne réponse associée ont été conservées. Ainsi la base d'apprentissage est constituée des réponses à 104 questions pour lesquelles seules les 10 premières réponses provenant de chaque systèmes ont été conservées.

7. Expérimentations et résultats

7.1. *Base de test*

Les tests ont été effectués sur les sorties des systèmes QAVAL, RITEL standard et RITEL probabiliste appliqués à 147 questions factuelles provenant de la campagne d'évaluation effectuée dans le cadre du projet QUAERO en 2010. Les systèmes recherchent les réponses dans un ensemble de 500 000 documents extraits du Web [2]. Les systèmes ont fourni jusqu'à 20 réponses différentes pour les différentes questions.

Les méthodes que nous avons mises en place sont évaluées par le MRR sur les cinq premières réponses qui est une mesure classique en QR fondée sur l'inverse du

2. Crawling réalisé par Exalead, http ://www.exalead.com/search/

rang de la meilleure réponse, ainsi que sur le nombre de bonnes réponses en première position. Le tableau 2 présente les résultats obtenus par les trois systèmes. Ces résultats nous serviront ensuite de baseline. Nous pouvons voir que les résultats obtenus par QAVAL sont légérement supérieurs à ceux obtenus par RITEL standard, tous deux bien supérieurs à ceux de RITEL probabiliste. Parmi les résultats 37 % des questions ont une réponse renvoyée par tous les systèmes et 20 % n'ont aucune bonne réponse renvoyée quelque soit le système. Le tableau indique également les résultats pouvant être obtenus par une méthode de fusion parfaite.

Système	MRR	% Première position (#)
QAVAL	0,44	36 (53)
RITEL	0,41	32 (47)
RITEL probabiliste	0,26	18 (27)
Système parfait	0,79	79 (115)

Tableau 2. *Résultats de base*

Afin de mieux analyser les données, la figure 1 présente la répartition des réponses entre les rang 2 et 20, le nombre en rang 1 étant montré dans le tableau 2. Les systèmes placent le plus grand nombre de réponses parmi les premières positions ce qui témoigne de leur efficacité. Il y a très peu de bonnes réponses après le rang 10. Suite à ces observations, nos méthodes seront évaluées sur les 5 premières réponses.

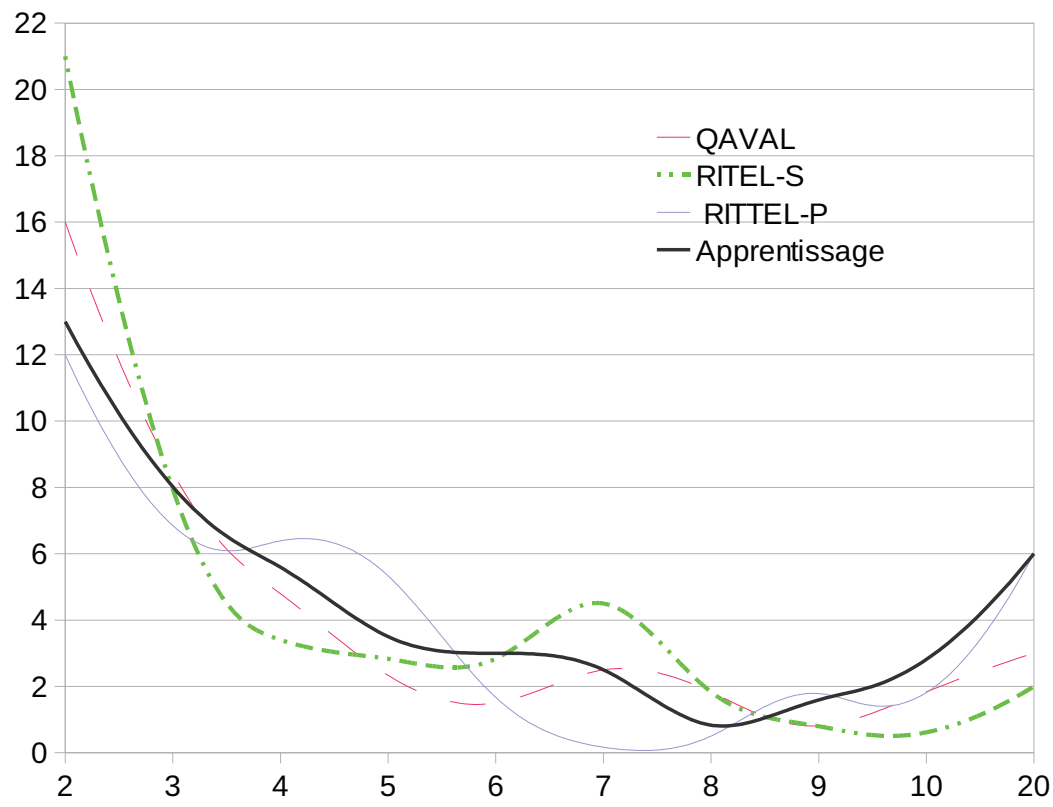

Figure 1. *Répartition des réponses*

7.2. *Évaluations*

Les méthodes de fusion ont été évaluées en considérant soit les trois systèmes soit les deux meilleurs, afin de définir l'intérêt du troisième système (RITEL probabiliste) qui a des résultats clairement inférieurs aux deux premiers. Les résultats (cf. tableau 3) ont ainsi montré que ce système est uniquement utile pour l'approche par apprentissage. Cela est dû à la forte ressemblance entre les deux systèmes RITEL qui peuvent plus fréquemment effectuer les mêmes erreurs.

Méthode	MRR (2 systèmes / 3 systèmes)	% Première position (#) (2 systèmes / 3 systèmes)
Interclassement	0,47 / 0,45	36 (53) / 36 (53)
$\sum \frac{1}{rang}$	0,48 / 0,46	38 (56)/ 36 (53)
CombSum	0,46 / 0,44	38 (56) / 34 (50)
CombMNZ	0,46/ 0,44	38 (56) / 35 (51)
Méthode hybride	0,49 /0,44	40 (59) / 34 (50)
Apprentissage	0,48 / **0,51**	39 (57) / **43** (63)

Tableau 3. *Résultats de l'ensemble des méthodes*

Nous pouvons également voir que les différentes méthodes permettent d'améliorer les résultats et que la méthode obtenant les meilleurs résultats est celle qui suit une approche par apprentissage avec une amélioration de 19 % du nombre de bonnes réponses en première position. Cette amélioration est très significative et montre l'efficacité de la méthode proposée. La figure 1 présente également la répartition des réponses suivant le rang pour cette méthode. De manière générale, nous pouvons remarquer que l'utilisation du rang des réponses permet d'obtenir d'avantage de meilleurs résultats que l'utilisation du poids.

Les résultats ont été calculés en considérant 10 réponses par système. Nous avons évalué l'impact de ce choix pour la méthode par apprentissage en considérant 20, 15, 10 et 5 réponses par système (cf. tableau 4). Le choix de 10 réponses est bien le meilleur car ainsi il n'y a ni trop ni trop peu de réponses considérées.

L'impact de la normalisation a été testé et on observe une nette diminution en ne l'utilisant pas ; pour la méthode par apprentissage, le nombre de bonnes réponses en première position passe ainsi de 63 à 58. De plus, les critères ne correspondant ni au rang ni au poids des réponses sont également utiles puisque, sans ces critères, le nombre de bonnes réponses en première position passe à 61.

Nombre de réponses	5	10	15	20
# Bonnes réponses en première position	55	63	58	58

Tableau 4. *Impact du nombre de réponses par système*

8. Conclusion

Nous avons étudié et expérimenté différentes méthodes permettant de combiner des systèmes de questions réponses de manière interne ou externe. Certaines suivent un mécanisme par vote, d'autres utilisent un module de validation de réponses. La méthode la plus pertinente combine différents critères par apprentissage. Les résultats obtenus sont bons puisque le nombre de questions ayant une bonne réponse en première position augmente de 20 %. La méthode mise en place est robuste puisqu'elle dépend de critères indépendants des systèmes et des stratégies appliquées, comme le type de réponse attendu, et de caractéristiques que les systèmes peuvent généralement produire, à savoir un score et la redondance de leurs réponses. Ainsi de nouveaux systèmes peuvent être ajoutés simplement et nous envisageons d'appliquer notre méthode à l'ensemble des réponses des systèmes participant à Quæro.

9. Bibliographie

Aceves-Pérez R. M., y Gómez M. M., Villaseñor-Pineda L., Ureña-López L. A., « Two Approaches for Multilingual Question Answering : Merging Passages vs. Merging Answers », *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 13, n° 1, p. 27-40, 2008.

Bernard G., Rosset S., Galibert O., Bilinski E., Adda G., « The LIMSI participation to the QAst 2009 track », *Working Notes of CLEF 2009 Workshop*, Corfu, Greece, October, 2009.

Chalendar G. D., Dalmas T., Elkateb-gara F., Ferret O., Grau B., Hurault-plantet M., Illouz G., Monceaux L., Robba I., Vilnat A., « The question answering system QALC at LIMSI : experiments in using Web and WordNet », *TREC 11*, 2002.

Galibert O., Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert, PhD thesis, Université Paris Sud, Orsay, 2009.

Grappy A., Grau B., « Answer type validation in question answering systems », *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO, 2010.

Grappy A., Grau B., Falco M.-H., Ligozat A.-L., Robba I., Vilnat A., « Selecting answers to questions from Web documents by a robust validation process », *WI*, 2011.

Jijkoun V., Rijke M. D., « Answer Selection in a Multi-Stream Open Domain Question Answering System », *ECIR*, 2004.

Joachims T., « Training linear SVMs in linear time », *ACM SIGKDD i*, 2006.

Shaw J. A., Fox E. A., « Combination of Multiple Searches », *TREC-2*, 1994.

Tellez-Valero A., Gomez M. M., Pineda L. V., Penas A., « Towards Multi-Stream Question Answering Using Answer Validation. », *Informatica (Slovenia)*, vol. 34, n° 1, p. 45-54, 2010.

Toney D., Rosset S., Max A., Galibert O., Bilinski E., « An Evaluation of Spoken and Textual Interaction in the RITEL Interactive Question Answering System », *LREC*, 2008.