TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Sélection de réponses à des questions dans un corpus Web par validation

A. Grappy[1, 2], B. Grau[1, 3], M.-H. Falco[1, 2], A.-L. Ligozat[1, 3], I. Robba[1, 4], A. Vilnat[1, 2]
(1) LIMSI-CNRS
(2) Université Paris 11
(3) ENSIIE
(4) UVSQ
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. Les collections Web diffèrent des articles de journaux de par leurs structures et leur style. Pour tenir compte de ces spécificités nous avons développé un système fondé sur une approche robuste de validation où des réponses candidates sont extraites à partir de courts passages textuels puis ordonnées par apprentissage. Les résultats montrent une amélioration du MRR (Mean Reciprocal Rank) de 48% par rapport à la baseline.

Abstract. Question answering systems look for the answer of a question given in natural language in a large collection of documents. Web documents have a structure and a style different from those of newspaper articles. We developed a QA system based on an answer validation process able to handle Web specificity. Large number of candidate answers are extracted from short passages in order to be validated according to question and passage characteristics. The validation module is based on a machine learning approach. We show that our system outperforms a baseline by up to 48% in MRR (Mean Reciprocal Rank).

Mots-clés : systèmes de questions réponses ; validation de réponses ; analyse de documents Web.

Keywords: question-answering system ; answer validation ; Web document analysis .

1 Introduction

La recherche d'informations précises dans des textes, en réponse à des questions posées en langue naturelle, constitue un domaine largement étudié depuis la première évaluation de systèmes de réponses à des questions (SQR dans la suite) lancée à TREC en 1998 (*Q&A track*). Les meilleurs systèmes (Hickl *et al.*, 2006; Bouma *et al.*, 2005; Laurent *et al.*, 2010) utilisent des connaissances et des processus avancés de TAL notamment des analyseurs syntaxiques. Ces connaissances et processus interviennent notamment lors de la phase de sélection de passages pertinents et d'extraction de réponses, qui ont fait l'objet d'études spécifiques.

L'ordonnancement de réponses ou de passages consiste à ordonner les différentes réponses extraites afin d'obtenir la meilleure réponse en première position. Là aussi, les meilleures approches se fondent sur des correspondances syntaxiques ou sémantiques entre les passages (souvent constitués d'une phrase) et la question, correspondances obtenues par calcul de similarité entre arbres syntaxiques (Kouylekov *et al.*, 2006) ou en tenant compte de chemins de dépendances communs (Cui *et al.*, 2005).

Ces différents systèmes obtiennent de bons résultats sur des documents issus d'articles de journaux, mais ne peuvent être appliqués en l'état sur des collections provenant du Web, comme celle constituée dans le cadre du projet Quæro[1] pour évaluer les SQR. Pour le français, les SQR participants ont trouvé entre 27 % et 50 % des réponses en 2009 (Quintard *et al.*, 2010) après adaptation alors que le meilleur système en obtenait 69% lors de l'évaluation CLEF 2006 sur des articles de journaux (Laurent *et al.*, 2010). Ces difficultés sont dues entre autres aux spécificités des documents Web très souvent composés de tableaux, de listes ou de menus qui mettent en défaut les analyses syntaxiques une fois le texte extrait des pages.

[1] http ://www.quaero.org - Quæro est un programme financé par OSEO

Afin de tenir compte des problèmes dus au style des documents, nous avons conçu un système sur le français, QAVAL (Question Answering by VALidation) pouvant s'appliquer sur tout type de documents. Alors que nos précédentes approches appliquaient des filtres successifs visant à sélectionner des passages, puis des phrases, puis extraire des réponses, QAVAL extrait directement de nombreuses réponses à partir de passages de 300 caractères extraits des documents. Ces réponses candidates sont ensuite ordonnées par un module de validation de réponses.

La validation de réponses vise à valider des réponses extraites par des SQR en vérifiant qu'elles sont correctes et justifiées par le passage de texte extrait. La plupart des approches (Herrera *et al.*, 2006) utilisent des critères lexicaux et syntaxiques, tels que la présence des termes de la question dans le passage et leur proximité, pour mesurer la similarité entre la question et le passage et évaluer la pertinence de la réponse. Quelques SQR ont intégré la validation de réponses : pour ordonner les passages et les réponses (Harabagiu & Hickl, 2006) ou pour choisir la réponse à partir de plusieurs ensembles proposés (Téllez-Valero *et al.*, 2010). Le système d'IBM (Martin *et al.*, 2001) s'améliore ainsi (MRR 0,496 vs 0,458) en utilisant une approche par apprentissage.

Dans QAVAL, nous avons mis en œuvre une approche par apprentissage afin de pouvoir se fonder sur des critères locaux et robustes pour caractériser les réponses à valider. Cette approche permet de s'affranchir de l'absence de phrases complètes bien formées et de gérer la dispersion éventuelle des informations utiles à la validation. Nous appliquons QAVAL sur des questions factuelles, qui attendent la précision d'un fait en réponse comme par exemple sa date dans « Quand le pont de Normandie a-t-il été inauguré ? », et dont les informations sont souvent réparties sur plus d'une phrase.

L'article présente d'abord les prétraitements effectués sur les documents Web. Puis il s'intéresse aux différents modules du système : l'analyse de la question qui extrait de celle-ci les informations utiles à la recherche de la réponse, la recherche et le traitement des passages, l'extraction des réponses depuis ces passages et l'ordonnancement des réponses. Il se termine par une partie expérimentation qui présente le corpus et l'évaluation de QAVAL et montre que notre approche obtient de bons résultats avec un MRR[2] supérieur de 48% à celui de la baseline.

2 Le système QAVAL

Le système QAVAL est constitué de modules séquentiels, qui peuvent être regroupés selon quatre grandes étapes : 1) L'analyse des questions ; 2) La recherche de passages et leur annotation à partir des documents Web prétraités ; 3) L'extraction de réponses candidates ; 4) La validation de réponses. Les deux premières étapes font l'objet de cette section.

2.1 Prétraitement des documents

Nous avons prétraité l'ensemble des documents HTML de la collection Quæro avec notre logiciel Kitten[3] afin de les rendre homogènes et utilisables (notamment pour le traitement syntaxique). Les documents HTML sont tout d'abord formatés et convertis au format XHTML en appliquant *HTMLCleaner*[4] et *jTIDY*[5]. Leur contenu textuel est ensuite extrait selon un filtre sur les types de balises (script, paragraphe) puis selon des expressions régulières afin de délimiter les phrases par ajout de ponctuation. En effet, de par la disposition visuelle des pages HTML (titres, sections, menus), les phrases sont visuellement séparées (une fois le HTML interprété) bien que ne se terminant pas par un point. Enfin, une extraction non-linéaire est effectuée pour les structures visuelles spécifiques telles que les tableaux en répétant les en-têtes pour chaque valeur afin de faciliter l'extraction des réponses puisqu'une extraction linéaire éloignerait de l'en-tête la valeur d'une case d'un tableau.

2.2 Analyse des questions

L'analyse des questions vise à extraire les informations utiles à la recherche de passages et à l'extraction de réponses. Outre le type de réponse attendu, qui correspond à un type d'entité nommée que l'on sait reconnaître

[2] Mean Reciprocal Rank : moyenne sur l'inverse du rang de la bonne réponse
[3] Kitten Is A Textual Treatment for Extraction and Normalization
[4] `http://htmlcleaner.sourceforge.net/`
[5] `http://jtidy.sourceforge.net/`

dans les textes, et les termes de la question, nous reconnaissons le type spécifique de la réponse, s'il est explicite, le focus et la catégorie de la question. Le focus désigne l'élément à propos duquel on demande une information (une entité ou un événement) et peut donc être représenté par un nom ou un verbe. Par exemple, la question « Quel président succéda à Jacques Chirac ? » a « succéder » comme focus, « personne » comme type d'entité nommée et « président » comme type spécifique. Selon l'existence du focus, son type, entité ou événement, et le type de réponse attendue, nous assignons une catégorie à la question qui représente le type de relation qui devra exister entre la réponse et le focus ou le type dans les documents : modifieur du nom, sujet ou complément d'objet du verbe, complément circonstanciel ... A la question précédente, on attend une réponse sujet du verbe en focus.

2.3 Recherche, sélection et annotation des passages

L'approche généralement utilisée dans les SQR consiste à retenir des passages de tailles variables (de 1 à 3 phrases). Nous avons choisi d'utiliser le moteur de recherche Lucene[6] pour procéder à l'indexation des documents et à la recherche de passages. Lucene peut renvoyer des extraits de documents et permet de paramétrer leur taille. Comme les passages renvoyés ne contiennent pas toujours des phrases complètes, la première et la dernière phrase sont donc complétées afin de faciliter l'analyse syntaxique des passages. Après expérimentations, nous avons décidé d'extraire un passage par document, d'environ 300 caractères, soit environ 3 phrases. Les passages retournés par Lucene sont ensuite analysés par Fastr (Jacquemin, 1996), qui repère les termes simples ou complexes de la question et leurs variantes. Ces variations peuvent être morphologiques, syntaxiques ou sémantiques, et à chacune d'elle est associé un poids, d'autant plus fort que la variation est fiable. Les passages sont ordonnés grâce à ces scores et les 50 meilleurs sont gardés.

Les passages sont ensuite annotés afin de faciliter l'extraction des réponses candidates. L'analyseur XIP (Aït-Mokhtar *et al.*, 2002) identifie pour chaque phrase du passage, ses syntagmes, calcule l'ensemble des relations de dépendances, et les entités nommées. Comme l'analyse syntaxique est moins fiable sur les documents Web, nous ne retenons que les syntagmes et les entités nommées. Les informations données par l'analyse de la question sont aussi annotées dans les passages : le focus, le type spécifique, le verbe principal et les noms propres.

3 Validation de réponses

3.1 Extraction des réponses candidates

Des réponses candidates sont extraites des passages retenus afin d'être ordonnées. La sélection des candidats est volontairement peu contrainte, dans le but de ne pas omettre la réponse correcte quitte à avoir davantage de réponses candidates à valider. Potentiellement tous les groupes nominaux pourraient constituer des candidats puisque les questions considérées sont d'ordre factuel et attendent en réponse un modifieur du nom ou un complément du verbe. Toutefois, afin de limiter le nombre de réponses extraites, un filtre est appliqué, pour ne conserver que les entités nommées correspondant au type de l'entité attendue par la question lorsqu'il existe. Ainsi la question « Qui est le président des États Unis ? » attend une personne en réponse et, dans ce cas, seuls les syntagmes marqués personne ou nom propre sont extraits.

Comme de très nombreuses réponses sont extraites des documents, une heuristique permettant de déclasser les réponses qui ont très peu de chances d'être correctes a été appliquée. Cela correspond aux cas où la réponse est constituée uniquement de mots contenus dans la question et à ceux où les entités nommées de la question ne se trouvent pas dans le passage justificatif. Les réponses restantes sont ensuite ordonnées par le module de validation de réponses suivant une approche par apprentissage : étant donnés des couples de réponses et passages dont elles sont extraites, annotés par le système, il s'agit de décider si les réponses sont valides et de leur degré de validité.

3.2 Ordonnancement de réponses

Pour l'ordonnancement de réponses,nous avons adapté l'approche développée pour la validation de réponses fournies par des SQR (Grappy *et al.*, 2008) en ajoutant des critères d'apprentissage, notamment pour tenir compte du

[6] http ://lucene.apache.org/

fait que la validité de la réponse ne soit pas la seule vérification à effectuer, et que les réponses n'ont pas été filtrées auparavant. Ces critères permettent d'évaluer la pertinence des passages (critère 2 ci-dessous) et des réponses par rapport au passage (critères 3b et 5). Nous avons aussi amélioré la vérification du type de la réponse. Globalement, les critères 1 et 2 traitent du passage afin d'évaluer une proximité avec la question. Les suivants portent sur la réponse et permettent notamment de distinguer plusieurs réponses issues d'un même passage.

1. la proportion des termes de la question présents dans le passage. Quatre calculs sont effectués :
– les mots de la question pris à l'identique ou sous forme de variantes,
– les mots répartis par catégorie morphosyntaxique (noms propres, noms communs, verbes, adjectifs),
– les éléments remarqués lors de l'analyse de la question (focus, type spécifique et verbe principal),
– les multi-termes : ensemble de mots consécutifs reconnus comme liés, comme « Prix Nobel » ;

2. rang du passage obtenu lors de la sélection des passages ;

3. proximité des termes. Si la réponse est proche en surface des mots de la question, elle a plus de chance d'être liée à ceux-ci, et donc valide. Pour évaluer cette proximité, deux critères sont utilisés :
a) la longueur de la plus longue chaîne de mots consécutifs présents dans le passage et constituée des mots de la question et de la réponse. Deux mots sont dits consécutifs s'ils sont adjacents, séparés par des éléments autorisés (virgule, déterminant...) ou séparés par un unique mot,
b) la distance entre la réponse et les mots de la question. La moyenne des distances séparant la réponse de chacun des mots de la question est calculée ;

4. redondance Plus la même réponse est extraite de différents documents, plus elle a de chances d'être correcte ;

5. la catégorie de la question, critère caractérisant la relation de dépendance avec la réponse ;

6. la vérification du type de la réponse. Certaines questions précisent le type de la réponse attendue, comme « Quel président succéda à Jacques Chirac ? » qui attend un nom de *président* en réponse. Ce critère vérifie que la réponse est du type spécifique attendu par la question. Cette vérification se fait aussi par apprentissage sur différents critères :
– la fréquence d'apparition commune de la réponse et du type dans les documents,
– l'utilisation des entités nommées du système de questions réponses RITEL (Rosset *et al.*, 2006) qui permet de reconnaître 70 types différents,
– la recherche du type dans les pages Wikipédia associées à la réponse,
– la recherche de structures de phrases indiquant une correspondance entre la réponse et le type (« Albert Einstein est un physicien ») dans les pages Wikipédia ;
La vérification du type de la réponse est présentée plus en détail dans (Grappy & Grau, 2010). Une évaluation consistant à détecter les cas où une réponse correspond au type a été menée et a obtenu 80 % de bons résultats. Notons cependant que ce critère ne peut être appliqué qu'à certaines questions, toutes n'ayant pas un type spécifique.

L'apprentissage de la validation de réponse, utilisant l'ensemble des critères ci-dessus, est effectué par une combinaison d'arbres de décision grâce à la méthode *bagging* fournie par WEKA[7]. Ce classifieur fournit, pour chaque réponse, un score compris entre -1 et 1 indiquant sa confiance dans le fait que la réponse soit valide. La valeur 1 indique que la réponse est correcte et -1 qu'elle est non valide. Ce score nous permet d'ordonner les réponses.

Afin de constituer la base d'apprentissage, nous avons sélectionné les questions factuelles de QA@CLEF05 et QA@CLEF06 et en avons cherché des réponses dans la collection avec QAVAL. Les réponses correspondant à un patron de réponse connu sont ensuite validées de manière manuelle. La base d'apprentissage contient 349 réponses valides et 698 non valides.

4 Expérimentation

La collection sur laquelle nous avons évalué notre système correspond à un corpus Web constitué par la société Exalead [8] à partir des requêtes d'utilisateurs sur leur moteur de recherche : deux millions de documents ont ainsi

[7] WEKA : http ://sourceforge.net/projects/weka/
[8] http://www.exalead.com/software/

été collectés. Un sous ensemble de 500 000 documents a été sélectionné pour les évaluations.

Pour tester notre système, nous avons utilisé un ensemble de 147 questions factuelles venant de la campagne Quæro 2010. Afin d'évaluer automatiquement notre SQR, nous avons recueilli un ensemble de réponses correctes pour chacune des questions. L'évaluation consiste donc à comparer chaque réponse extraite aux réponses attendues. Nous avons utilisé la mesure MRR, sur les cinq premières réponses, ainsi que la proportion de questions ayant une bonne réponse en première position ou dans les cinq premières positions pour évaluer les résultats.
Nous avons comparé notre méthode à une baseline portant sur l'extraction et la validation de réponses. Elle extrait les candidats les plus proches des mots de la question dans les cinq premiers passages ordonnés suivant leur rang après leur extraction. Le tableau 1 présente les résultats obtenus avec 50 passages retenus sur 150 ramenés par Lucene. 11 405 réponses candidates sont extraites, soit 77 en moyenne par question. Les passages de 300 caractères servant à extraire les réponses contiennent 6 phrases en moyenne, alors que sur un corpus d'articles de journaux ils en contiennent 3, et possèdent moitié moins de verbes, ce qui rend compte du fait qu'ils sont souvent formés de suites de syntagmes.

	MRR	premier rang %(#)	cinq premiers rangs %(#)
Quæro	0,43	34% (50)	55% (81)
baseline	0,29	21% (32)	43% (64)

TAB. 1 – Résultats QAVAL

QAVAL surpasse les résultats de la baseline de 48 % ce qui montre l'apport de notre méthode. Afin d'évaluer uniquement le module de validation de réponses, nous avons évalué notre système sur les 125 questions pour lesquelles la bonne réponse se trouve dans un passage sélectionné. 65% des questions ont une réponse correcte parmi les cinq meilleures et 40% en première position ce qui est meilleur que les résultats obtenus par (Cui *et al.*, 2005) qui trouvait 39% de réponses correctes en première position sur des passages extraits de journaux en anglais.

L'article (Quintard *et al.*, 2010) montre que les autres systèmes cherchant une réponse pour les questions factuelles sur le corpus Quæro obtiennent un MRR entre 0,284 et 0,54. Nos résultats sont donc de même ordre de grandeur que ceux obtenus par les meilleurs systèmes.

Afin de tester la robustesse de notre méthode, nous avons appliqué QAVAL sur une collection constituée d'articles du journal Le Monde et de dépêches ATS. Pour cela, nous avons pris 128 questions factuelles provenant de la campagne EQUER. Une nouvelle base d'apprentissage a été utilisée, dans laquelle les réponses sont extraites de cet ensemble de documents. Les résultats sont comparables à ceux obtenus sur les documents Web (MRR de 0,47) ce qui témoigne de la robustesse de la méthode et portent notre système aux résultats de l'état de l'art avec ce type d'approche. L'apport de la validation de réponses est encore important puisqu'il montre une amélioration de 38 % par rapport à la baseline.

Une analyse des arbres de décision a montré que tous les critères sont utilisés et que les trois critères les plus importants sont la fréquence de la réponse, le rang du passage et la proximité des termes. L'apport de la vérification du type de la réponse a également été mesuré en retirant ce critère de l'ensemble de départ. Le MRR passe alors de 0,43 à 0,41 ce qui montre bien que ce critère est important d'autant plus qu'il ne s'applique pas à toutes les questions. Les mêmes résultats sont obtenus avec la catégorie de la question.

Afin d'évaluer le prétraitement des documents, nous avons appliqué QAVAL avec trois prétraitements différents : le premier est Kitten, le second, la baseline, est une extraction complète du contenu textuel et le troisième applique le logiciel BoilerPipe (Kohlschütter *et al.*, 2010) qui utilise des traits textuels de surface. En plus du MRR, nous avons calculé la proportion de questions ayant au moins un document contenant la réponse peu importe son rang (cf. tableau 2). Les résultats montrent que notre traitement améliore les résultats de la baseline puisque le MRR lui est supérieur de 30%. Nous pouvons aussi voir que la méthode est supérieure aux résultats de BoilerPipe qui semble moins adapté à l'utilisation d'une collection Web dans le cadre des SQR.

	BoilerPipe	baseline	Kitten
% bons doc. (#)	77 %(114)	82 %(121)	88% (130)
MRR	0,28	0,33	0,43

TAB. 2 – Rôle du prétraitement des documents dans QAVAL

5 Conclusion

Alors que les SQR développés sur des collections d'articles de journaux rencontrent des difficultés sur les documents Web, le système QAVAL[9] applique une méthode robuste de validation de réponses permettant d'ordonner les nombreuses réponses extraites depuis des passages de 300 caractères environ à partir d'un apprentissage sur des critères locaux. Ce type de critères permet de tenir compte de la dispersion des informations sur plus d'une phrase dans les passages. QAVAL permet de surpasser la baseline de 48% et obtient des résultats analogues à ceux obtenus sur des documents issus d'articles de journaux.

Pour améliorer les résultats, nous envisageons d'introduire de nouveaux critères. Une possibilité est d'utiliser un module de paraphrases sous phrastique permettant de rapprocher les expressions contenues dans la question et les passages. QAVAL pourrait aussi être utilisé pour traiter des questions booléennes. Dans celles-ci la valeur à donner est OUI si un des passages trouvés justifie la forme affirmative de la question.

Références

AÏT-MOKHTAR S., CHANOD J.-P. & ROUX C. (2002). Robustness beyond shallowness : incremental deep parsing. *Nat. Lang. Eng.*, **8**.

BOUMA G., FAHMI I., MUR J., VAN NOORD G., VAN DER PLAS L. & TIEDEMANN J. (2005). Linguistic Knowledge and question answering. *Traitement automatique des langues spécial Répondre à des questions*, **46**(3).

CUI H., SUN R., LI K., YEN KAN M. & SENG CHUA T. (2005). Question answering passage retrieval using dependency relations. In *SIGIR 2005*.

GRAPPY A. & GRAU B. (2010). Answer type validation in question answering systems. In *Recherche d'Informations Assisté par Ordinateur*.

GRAPPY A., LIGOZAT A.-L. & GRAU B. (2008). Evaluation de la réponse d'un système de question-réponse et de sa justification. In *COnférence en Recherche d'Infomations et Applications*.

HARABAGIU S. & HICKL A. (2006). Methods for using textual entailment in open-domain question answering. In *Proceedings of the 44th annual meeting of the Association for Computational Linguistics*.

HERRERA J., RODRIGO A., PENAS A. & VERDEJO F. (2006). UNED submission to AVE 2006. In *Working Notes for the CLEF 2006 Workshop (AVE)*.

HICKL A., WILLIAMS J., BENSLEY J., ROBERTS K., SHI Y. & RINK B. (2006). Question answering with LCC's CHAUCER at TREC 2006. In *Proceedings of the Fifteenth Text REtrieval Conference*.

JACQUEMIN C. (1996). A Symbolic and Surgical Acquisition of Terms Through Variation. In *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, p. 425–438.

KOHLSCHÜTTER C., FANKHAUSER P. & NEJDL W. (2010). Boilerplate detection using shallow text features. In *WSDM*.

KOUYLEKOV M., NEGRI M., MAGNINI B. & COPPOLA B. (2006). Towards Entailment-based Question Answering : ITC-irst at CLEF 2006. In *7th Workshop of the Cross-Language Evaluation Forum*.

LAURENT D., SÉGUÉLA P. & NÈGRE S. (2010). Cross lingual question answering using qristal for clef 2006. *Evaluation of Multilingual and Multi-modal Information Retrieval*, p. 339–350.

MARTIN A. I., FRANZ M. & ROUKOS S. (2001). Ibm's statistical question answering system-trec-10. In *In Proceedings of TREC10*.

QUINTARD L., GALIBERT O., ADDA G., GRAU B., LAURENT D., MORICEAU V., ROSSET S., TANNIER X. & VILNAT A. (2010). Question Answering on web data : the QA evaluation in Quæro. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*.

ROSSET S., GALIBERT O., ILLOUZ G. & MAX A. (2006). Interaction et recherche d'information : le projet ritel. *Traitement Automatique des Langues (TAL), numéro spécial Répondre à des questions, volume 46 :3*, **46**(3).

TÉLLEZ-VALERO A., MONTES-Y GÓMEZ M., VILLASEÑOR-PINEDA L., DEL LENGUAJE L. & PEÑAS-PADILLA A. (2010). Towards Multi-Stream Question Answering Using Answer Validation. *Informatica*, **34**.

[9]Ces travaux ont été en partie réalisés dans le cadre du programme QUAERO, financé par OSEO, agence française pour l'innovation.