Bases de connaissances pour asseoir la crédibilité des réponses d'un système de Q/R

L. Gillard, P. Bellot, M. El-Bèze

Laboratoire d'Informatique d'Avignon (LIA)
339 ch. des Meinajaries, BP 1228 ; F-84911 Avignon Cedex 9 (France)
{ laurent.gillard, patrice.bellot, marc.elbeze }@lia.univ-avignon.fr

Résumé – Abstract

Cet article présente un prototype de Question/Réponse (Q/R) impliquant un ensemble de bases de connaissances (BC) dont l'objectif est d'apporter un crédit supplémentaire aux réponses candidates trouvées. Ces BC et leur influence sur la stratégie d'ordonnancement mise en œuvre sont décrites dans le cadre de la participation du système à la campagne Q/R de TREC-2002.

This paper presents a Question-Answering system using Knowledge Databases (KDB) to validate answers candidates.

Mots Clés – Keywords

Système de Question/Réponse, Bases de Connaissances

Question Answering system, Knowledge Databases.

1 Introduction

Avec l'explosion de la quantité d'information disponible, les systèmes de recherche d'information sont devenus les principaux moyens d'accès à l'information. Pourtant, nombre d'entre eux présentent deux faiblesses : d'une part, le nombre de documents retournés en réponse à une requête utilisateur est souvent trop important ; d'autre part, c'est toujours à l'utilisateur de localiser l'information dont il a besoin dans les documents renvoyés. Les systèmes de question/réponse (Q/R) fournissent une alternative intéressante à ces faiblesses : ils proposent de renvoyer directement à l'utilisateur *LA* réponse à *LA* question posée.

Ainsi, depuis 1998, la campagne internationale « Text Retrieval Conference » (TREC) propose un sous-volet thématique Q/R (*« QA Track »*) dont le but est de faire avancer les techniques employées dans ces systèmes, notamment en proposant un référentiel et une méthodologie permettant leur évaluation et leur comparaison.C'est dans le cadre de TREC-2002 (Voorhees E., 2002), que le LIA s'est intéressé pour la première fois à la problématique Q/R.

Cet article décrit le prototype Q/R du LIA à travers une présentation chronologique de ses divers constituant (en section 2). Enfin, l'originalité de ce papier tient en la présentation d'un composant supplémentaire : un module de bases de connaissances (BC, section 3) et son influence (section 4) sur les résultats obtenus lors de notre participation (Bellot *et al.*, à paraître) à TREC-2002. En effet, pour certaines questions de « culture générale », il peut paraître intéressant de doter un système Q/R d'un ensemble de réponses calculées à l'avance afin d'en faciliter la recherche ou, comme dans le cas de notre prototype, d'en asseoir la confiance.

2 Prototype QR du LIA

Étiquetage des questions : les systèmes Q/R précurseurs comme celui de (Kupiec, 1993) ou les différents systèmes participants à TREC-QA intègrent tous un module chargé de classer les

questions (les différences résident dans les moyens d'effectuer la classification et dans sa granularité). À partir de la forme interrogative d'une question, il est généralement possible de déterminer un ou plusieurs types de réponses attendues et, par conséquent, d'en faciliter la recherche en établissant une correspondance directe entre la nature de la réponse attendue et celle d'une entité nommée détectée. En plus de l'étiquetage, certains systèmes comme (Ferret *et al.*, 2002) vont plus loin dans le traitement de la question : ils identifient au sein d'une question un « focus » (l'objet sur lequel porte la question), des modificateurs du « focus » (les informations qui le qualifient) ou encore un ensemble de mots clés à trouver dans le voisinage de la réponse.

De même, le prototype Q/R du LIA affecte à chaque question une ou plusieurs étiquettes (classées par ordre de probabilité parmi une hiérarchie d'étiquettes déduites des questions de TREC-9 et 10) ou même *« unknown »* en cas d'indécision. L'étiquetage de la question se fait de préférence à partir d'un étiqueteur à base de règles (reposant sur des motifs constitués par des mots de la question ou leur étiquette syntaxique), et sinon à partir d'un étiqueteur probabiliste (Béchet *et al.*, 2000).

Recherche des documents susceptibles de contenir la réponse : le prototype Q/R du LIA ne contient pas de tel module mais dépend de la liste des 1000 premiers documents proposés par les organisateurs de la campagne en considérant la question brute comme requête. Il est à noter qu'il n'y a aucune garantie que la réponse à la question soit effectivement comprise dans le jeu de documents fournis même si elle est comprise dans le reste du corpus de référence.

Reconnaissance des entités nommées : ce travail a été effectué par la société Sinequa (http://www.sinequa.com). Les entités sont détectées à l'aide d'une cascade de transducteurs utilisant le contexte avoisinant (mots et étiquettes syntaxiques) ainsi que des traits sémantiques présents dans des lexiques. Une normalisation des noms propres ainsi qu'une résolution élémentaire d'anaphore (pour les pronoms personnels « il » et « elle ») a également été faite.

Sélection de la réponse : l'approche du système QR du LIA pour la sélection des réponses candidates repose sur l'utilisation conjointe du moteur de recherche vectoriel SIAC (Bellot & El-Bèze, 2000) et sur la localisation d'une entité du type recherché. Ainsi, dans un premier temps, l'ensemble des phrases de tous les documents est ordonné suivant le calcul d'une proximité avec la question. Ensuite, un filtrage est effectué afin de conserver seulement les phrases contenant au moins une entité correspondant au type attendu et les noms propres apparaissant éventuellement dans la question. La réponse du système est alors la première entité du type attendu apparaissant dans la phrase. Néanmoins, cette façon de procéder a au moins deux faiblesses : l'étape de filtrage, en écartant toutes les phrases de manière trop draconienne, a parfois rendu impossible l'extraction d'une réponse ; enfin, la sélection de la « première » entité pose problème dans le cas ou plusieurs entités du même type sont présentes dans la même phrase.

3 BC pour crédibiliser et ordonner les réponses

3.1 Bases de connaissances

En raison des nouvelles contraintes introduites lors de la campagne TREC-11 : celle pour un système Q/R d'évaluer la « confiance » qu'il a en une réponse, mais aussi celle d'ordonner l'ensemble de celles-ci par ordre de « certitude » ; il nous a paru intéressant d'associer à notre prototype Q/R une source fiable de réponses déjà connues. Cette source a été modélisée au travers d'un module de bases de connaissances.

En effet, la présence de telles bases permet de crédibiliser une réponse candidate. Ainsi, pour notre participation à TREC-11, ce module a simplement eu une fonction de validation des réponses suggérées par le reste du système.

De plus, le choix d'inclure un module de BC a été conforté par un examen des questions des précédentes campagne TREC-QA, puisqu'il est aisément possible de constater qu'un certain nombre d'entre elles relève de sujets assimilables à des thématiques de « culture générale » (géographie, dates et faits historiques, noms d'écrivains et d'œuvres, *etc.*) mais également que ces sujets sont fréquents (au sein d'une même campagne ou sein de l'ensemble de celles-ci).

Il est à noter que de telles BC peuvent être utilisées comme des listes de réponses préenregistrées qui serviraient alors de support à une fouille dans un corpus de référence. L'un des buts du critère de réponse supportée était d'ailleurs d'éviter dans la mesure du possible un système Q/R complètement architecturé sur ce principe. Cependant, dans un tel cas, il est tout de même intéressant de se demander, notamment au niveau applicatif, quel est l'intérêt de chercher à nouveau une réponse (ou une autre formulation d'une réponse) qui est déjà connue.

À plus long terme, l'objectif visé est de disposer d'une source de connaissances de qualité pour servir d'amorces à des itérations dont le but est d'extraire des connaissances similaires (motifs d'extraction, motifs de réponses ; Soubbotin & Soubbotin, 2001 ; Ravichandran & Hovy, 2002) soit à l'intérieur du corpus, soit à partir d'Internet ; mais surtout d'évaluer et comparer la crédibilité de différentes réponses candidates (notamment au travers des relations permettant de l'établir ou liant ses différents constituants).

3.1.1 Construction des BC

Les « connaissances » ont été extraites de tableaux et de listes provenant de différents sites Internet de nature encyclopédique. En fait, l'un des principaux critères était que ces informations soient présentées sous une forme fortement structurée pour en faciliter l'extraction (manuelle) et qu'elles soient en nombre suffisant. En outre, puisque ces connaissances sont destinées à corroborer une réponse candidate extraite par ailleurs d'un document du corpus, le critère de la fiabilité de leur provenance qui pourrait paraître primordial de prime abord ne l'est pas. La fiabilité provient plutôt de la convergence entre une réponse candidate et sa présence dans les BC. Il est assez improbable qu'un document donne une même réponse fausse que celle contenue dans des BC de large couverture (tant du point de vue des questions que des réponses). Dans un tel cas, et si cette réponse est considérée comme incorrecte, il est alors possible de se demander si les capacités des systèmes à évaluer ne sont pas surestimées.

TREC / Thématique	8	9	10	11	Total
Devise	0+0	0+0	1+0	1+0	**2+ 0**
Fleur	0+0	0+0	2+1	0+0	**2+1**
Hymne	0+0	0+0	0+0	1+0	**1+ 0**
Arbre	0+0	1+0	0+0	0+0	**1+ 0**
Oiseau	0+0	1+0	3+0	1+0	**5+ 0**
Gouverneur	0+0	0+0	1+0	3+0	**4+ 0**
Creation	0+0	1+0	3+0	2+0	**6+ 0**
Capitale	1+5	2+1	0+5	1+6	**4+17**
Population	0+4	4+5	1+4	1+3	**6+16**
Président	1+1	2+1	4+0	5+0	**12+ 2**
Total	**2+10**	**11+7**	**15+10**	**15+9**	**43+36**

Tableau 1 : Couverture des BC « USA » + Autres

3.1.2 Contenu et couverture des BC

Le choix du contenu des BC a été établi d'après un examen rapide de l'ensemble des questions des campagnes TREC, en fonction :

- de la constance des thématiques d'interrogation, cela afin d'obtenir un taux de couverture suffisant sur l'ensemble des précédentes campagnes TREC-QA (un pari a été fait sur le fait que ce taux reste stable sur le corpus TREC-2002) ;
- du fait que la réponse attendue met en jeu des *n-uples* d'entités nommées (par exemple : pays/capitale, nom de l'inventeur/date/invention, *etc.*).

Ainsi, une partie des BC porte sur des lieux géographiques comme : des cours d'eau et les pays traversés, les capitales des pays ; mais aussi des personnages célèbres comme : les prix Nobels par année et disciplines, des écrivains et leurs œuvres, les dieux des panthéons grecs, romains ; ou encore des couples de mots : abréviations et définitions, *etc.*

De même, dans le cadre des TREC-QA, 43 questions ont pour principal objet les États-Unis d'Amérique et portent notamment sur les attributs de ces états : capitales, nom des gouverneurs,

emblèmes (oiseau, arbre, fleur), *etc.* Ces sujets ont également donné lieu à la création de BC et, autant que possible, ces bases ont été enrichies par la version « extra-américaine » (*cf. Tableau 1 ; pour une thématique, le chiffre avant le + représente le nombre de questions centrées sur les USA et le chiffre après le + correspond au nombre portant sur d'autres pays).*

Enfin, il est apparu que certaines questions revenaient sous une forme identique lors des campagnes TREC suivantes. Aussi chacun des jeux des couples Q/R des TREC-QA précédents peut être considéré comme une base d'archives pour la campagne suivante. Il est à noter qu'il s'agit là d'une approche sujette à caution : la réponse à une même question peut varier d'année en année, tout comme son degré d'exactitude qu'il s'agisse d'exigences provenant de l'évaluation ou même de variation dans le jugement de la notion d'exactitude (Lavenus & Lapalme, à paraître).

Au final, le module de BC comprend 36 bases de connaissances thématiques et assure une couverture de l'ordre de 10% à 12% des réponses par campagne TREC (*cf. Tableau 2*).

TREC	8	9	10	11	Total
# Réponses incluses dans les BC	22	73	64	61	220
# Questions	200	694	500	500	1 894
%des Q couvertes	11.0	10.5	12.8	12.2	11.6

Tableau 2 : Couverture globale des BC

4 Stratégie d'ordonnancement des réponses

Dans le prototype Q/R du LIA, les informations disponibles qui permettent d'établir une préférence dans les réponses sont les suivantes :

- une étiquette de réponse attendue ainsi que sa probabilité associée. Cette affectation a lieu au niveau du module d'étiquetage. En cas d'échec ou d'incapacité, cette étiquette peut être *« unknown »* et dans ce cas le système ne « sait plus » répondre (car le processus de sélection repose sur la correspondance « étiquette attendue↔entité nommée ») ;
- le fait que les bases de connaissances corroborent une réponse trouvée par SIAC. Cette réponse peut être jugée comme fiable puisque proposée par deux systèmes indépendants ;
- si aucune réponse n'a été trouvée, la réponse donnée par le système est « NIL ».

Groupe	Modules impliqués dans le processus de réponse	Ventilation sur TREC-QA 2002	
		Position	Commentaires
1	SIAC & bases de connaissances	de 1 à 30	Accord entre deux sous modules donc réponses jugées fiables
2a	SIAC	de 31 à 424	
2b		de 425 à 473	Réponse « NIL » potentiellement abusive car filtrage draconien.
3	*Questions non étiquetées (« unknown »)*	de 474 à 500	Réponse « NIL » car système incapable de choisir une réponse.

Tableau 3 : Stratégie d'ordonnancement

Cependant, dans ce dernier cas, une nuance doit être apportée pour notre prototype Q/R. En effet, aucun module spécifique n'a été créé pour vérifier l'absence d'une réponse. Il ne comprend pas non plus de processus itératif qui permettrait de relâcher des contraintes afin d'élargir la liste des réponses envisageables. Et, pire, il lui arrive parfois d'écarter abusivement les réponses proposées par SIAC, en particulier durant une étape de filtrage lorsque aucune entité nommée du type attendu n'est présente dans la phrase retournée. Ainsi, il apparaît que le fait qu'une réponse ne peut être trouvée ne doit pas être considérée comme une absence catégorique de réponse dans le corpus mais juste comme une présomption d'absence.

À partir de ces considérations, une stratégie *(cf. Tableau 3)* a été mise au point pour ordonner les réponses du système et notamment positionner en premier les réponses crédibilisées par les BC.

À l'intérieur de chacun des 1er et 2nd groupes, les réponses sont ordonnées d'abord suivant la difficulté de reconnaissance des entités nommées correspondant à l'étiquette attendue (ordre obtenu expérimentalement) et ensuite suivant la probabilité décroissante de ces étiquettes. Enfin, les réponses du 3^{e} groupe peuvent être considérées comme des réponses « NIL » par défaut.

5 Résultats obtenus et apport des BC

Nombre de réponses fausses (W)	440
Nombre de réponses non supportées (U)	4
Nombre de réponses inexactes (X)	4
Nombre de réponses correctes (R)	**52**
« Confidence-weighted score » (CWS)	**0.246**
Précision des réponses « NIL »	7 / 75 = 0.093
Rappel des réponses « NIL »	7 / 46 = 0.152

Tableau 4 : Résultats officiels

Le tableau 4 présente les résultats officiels obtenus par le prototype Q/R du LIA à la campagne TREC-QA de 2002. Notre soumission a obtenu un score de 0.246 en répondant correctement à 52 questions sur les 500 de la campagne.

Après analyse, les BC interviennent dans 30 des Q/R et permettent d'asseoir correctement la crédibilité de 24 réponses. Sur les six restantes et jugées non « correctes » :

- trois d'entre elles sont non supportées. C'est-à-dire que les BC permettent de corroborer une bonne réponse mais que le document dont SIAC l'extrait ne permet pas de l'établir. Il s'agit d'un effet pervers d'une sélection des réponses candidates basée seulement sur le type de l'étiquette ;
- deux sont incorrectes pour des problèmes d'ordre de grandeur. Il s'agit de questions portant sur les populations « actuelles » de Mexico et d'Afrique du Sud. Les réponses retournées *(soulignées)* bien que compatibles avec les données des BC, sont des populations à une « échéance » à savoir *« Mexico's population will reach 100 million in the year 2000. »* et *« it expected the total population to be more than 42 million people, based on the census in 1991, when South Africa's total population was estimated to be 38 million. ».* Les motifs des réponses correctes attendues étaient respectivement *« 98.1 million »* et *« (37.9|38|41.8|42) million »*.
- enfin, la dernière *« When did George W. Bush get elected as the governor of Texas? »* s'est vue répondre par la date de la prise de fonction (*« In January 1995 , when George W. Bush took office as governor »*) plutôt que celle de l'élection (les réponses correctes *« 1994 »* ou *« 1998 »*).

Après que les motifs des réponses ont été communiqués, nous avons évalué notre prototype sans le module de BC : il n'identifie alors plus que 34 réponses correctes. Cette différence s'explique par le fait que, dans ce cas, l'entité nommée extraite de la phrase retournée est la première du type attendue et non plus celle issue de la BC.

Enfin, le composant d'ordonnancement utilisé pour notre soumission comportait une erreur : la probabilité décroissante des étiquettes n'était pas prise en compte comme initialement prévu sur les groupes 2 et 3. Cette correction a entraîné une hausse d'environ 9% dans le score (de 0,246 à 0,268) et a une influence encore plus flagrante dans le cas d'une utilisation sans BC puisque la progression est alors de +65% (de 0,084 à 0,139).

La figure 1 (page suivante, d'après Voorhees, 2002) permet de visualiser à quel point, une bonne stratégie d'ordonnancement permet de s'approcher d'une courbe idéale (en noir) dans le cadre d'une évaluation TREC-2002. Ainsi, bien que notre prototype réponde à seulement 52 questions il parvient assez bien à classer ses bonnes réponses. En effet, un point situé plus haut et plus à gauche d'un autre est un système qui ordonne mieux ses réponses qu'il ne répond correctement.

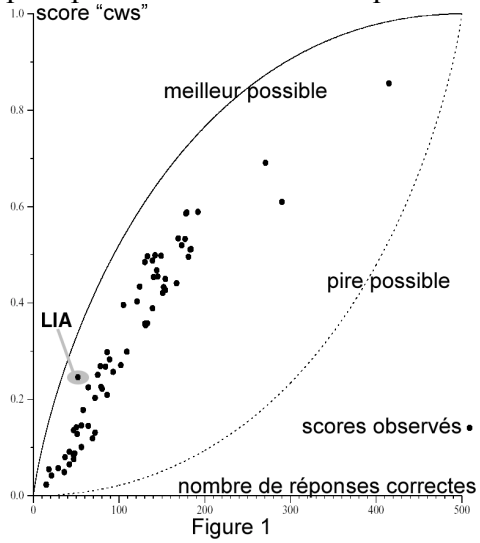

Figure 1

6 Conclusion

Pour notre première participation à la piste TREC-QA, notre objectif était de créer un prototype qui nous permette d'étudier la problématique des systèmes Q/R, mais également d'envisager les premières briques d'une stratégie mettant en avant la crédibilité des réponses. Certains modules de notre système Q/R doivent être

améliorés comme celui concernant l'étiquetage des questions mais cela n'était pas une priorité pour notre première participation à TREC.

Ainsi, les résultats obtenus dénotent une bonne orientation puisque les réponses trouvées le sont avec une confiance élevée et cela grâce à l'usage des BC comme moyen de certification. En outre, dans le cadre de TREC-QA, il est possible d'inclure aisément un sous-ensemble des Q/R d'au minimum 10% dans de telles bases (principalement les questions portant sur des *n–uples* d'entités nommées ou des faits établis).

Il faut cependant nuancer l'usage de ces BC. En effet, si elles remettent partiellement en cause le principe de la tâche de Q/R, il apparaît rapidement qu'il est difficile de constituer des bases exhaustives (toutes les questions n'autorisent pas des réponses pré-calculées). Cela est d'autant plus vrai dans le cadre des questions ouvertes puisque leur domaine n'est pas connu (pour notre prototype, une thématique donne lieu à une BC). De même, dans le cadre du problème de leur couverture l'intérêt d'utiliser de telles bases plutôt qu'une redondance issue d'Internet est à démontrer. Et, comme pour toutes les bases de données, se pose le problème de leur durée de vie, de leur mise à jour mais aussi de leur intégration/utilisation dans d'autres langues que celle de leur contenu.

Références

Béchet F., Nasr A., Genet F. (2000), "*Tagging Unknown Proper Names Using Decision Trees"*, in Proc. of ACL'2000, Hong-Kong, Chine, p.77-84.

Bellot P., Crestan E., El-Bèze M., Gillard L., de Loupy C. (à paraître), *"Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC-11 Question Answering Track"* in Proceedings of the 11th Text REtrieval Conference.

Bellot P. & El-Bèze M. (2000), "*Clustering by means of decision trees without learning or hierarchical and K-Means like algorithms"*, RIAO'2000, Paris, p.344-363.

Ferret O., Grau B., Hurault-Plantet M., Illouz G., Monceaux L., Robba I., Vilnat A. (2002), *« Recherche de la réponse fondée sur la reconnaissance du focus de la question »*, TALN2002, 24-27 juin Nancy, p.307-316.

Kupiec J. (1993), *"Murax: A robust linguistic approach for question answering using an on-line encyclopedia"*, in Proceedings of the 16th annual international ACM SIGIR conference, Pittsburgh, PA, USA, pp. 181-190.

Lavenus K. & Lapalme G. (à paraître), *« Évaluation des systèmes de question réponse - Aspects méthodologiques »*, dans la revue Traitement Automatique des Langues.

Ravichandran D. & Hovy E.H. (2002), *"Learning Surface Patterns for a Question Answering System."* In Proceedings of the ACL conference, Philadelphie, USA, p. 41-47.

Soubbotin M.M. & Soubbotin S.M. (2001), *"Patterns of Potential Answer Expressions as Clues to the Right Answers"* in Proceedings of the 10th Text Retrieval Conference (TREC-10), p.175-182, Gaithersburg, Maryland.

Voorhees E. (2002), "*Overview of the TREC 2002 Question Answering Track"*, http://trec.nist.gov/pubs/trec11/papers/QA11.pdf.