Deux pistes complémentaires pour améliorer l'appariement Question Réponse

Karine Lavenus (1), Jens Grivolla (2), Laurent Gillard (3), Patrice Bellot (4)
Laboratoire d'Informatique d'Avignon (LIA)
339 ch. des Meinajaries, BP 1228
F-84911 Avignon Cedex 9 (France)
{karine.lavenus, jens.grivolla, laurent.gillard, patrice.bellot}@lia.univ-avignon.fr

Résumé – Abstract

Cet article porte sur l'appariement au sein des systèmes de Question Réponse (QR). Nous présentons deux approches, l'une linguistique, l'autre statistique, utilisées à deux étapes du processus, afin de réduire le fossé entre question et réponse. Nous décrivons tout d'abord notre catégorisation, étape incontournable pour orienter la recherche de la réponse à partir de la question. Basée sur des patrons lexico-syntaxiques, elle permet entre autres de tenir compte de nuances sémantiques, et de faciliter ainsi l'appariement. Puis nous nous proposons de modifier le modèle vectoriel classique de Recherche d'Information (RI), afin de l'adapter aux spécificités du QR et d'améliorer l'appariement. Nos premiers résultats nous servirons à attribuer des poids aux mots-clefs des questions et à leurs expansions.

This paper presents two approaches at different steps of Question Answering to improve the match between question and answer. First, as categorization is unavoidable to guide the pairing, we describe a question pattern-based categorization, which processes the linguistic criteria. Then, to enhance the pairing probability, we propose a statistical method. This method aims to modify the weights of keywords and expansions within the classical Information Retrieval vector space model.

Mots Clés – Keywords

Systèmes de question réponse, QR, catégorisation des questions, appariement, mise en correspondance de patrons, pondération des mots-clefs et expansions.
Question answering systems, Q&A, categorization, pairing, pattern-matching, keyword and expansion weight.

Introduction

Une des difficultés majeures en Question Réponse (QR) consiste à réduire le fossé entre la question et la réponse pour les apparier. La catégorisation apparaît comme un module important dans la mesure où elle permet d'indiquer le type de *réponse* attendu d'après la sémantique de la *question*. La première section de notre article présente les particularités de notre catégorisation, basée sur des critères linguistiques. Nous verrons ensuite, dans la seconde section, quels sont les premiers résultats d'une étude statistique en cours, portant sur la fréquence des mots-clefs et de leurs expansions dans les réponses. Nous souhaitons en effet nous baser sur ces résultats afin d'adapter le modèle vectoriel classique au QR.

1 Catégorisation en vue de l'appariement question - réponse

1.1 Rôle et importance de la catégorisation

Les systèmes de question réponse utilisent des techniques propres à la recherche d'information. Cela suppose que la question posée par l'utilisateur soit transformée en requête dès le début du processus. Dès lors, les nuances les plus fines sont ignorées par les moteurs de recherche, qui, habituellement :
-transforment la question en un sac de mots, perdant ainsi des informations hiérarchico-syntaxiques contenant des informations sémantiques (notamment en ce qui concerne la voix passive) ;
-peuvent lemmatiser ou raciniser, ce qui supprime les informations de temps et mode, de genre (en français) et de nombre (singulier *vs* pluriel) ;
-éliminent des termes pourtant significatifs (mots-outils).
Or, si l'on offre à l'utilisateur l'opportunité de poser une question, ce n'est pas seulement pour lui donner une réponse concise, c'est aussi pour qu'il puisse exprimer la question qu'il pose de manière fine et complète. Lorsque la question est transformée en requête, de nombreuses informations sont perdues, puisque l'on obtient un « sac de mots ». De plus, les mots-outils jouent un rôle important au sein des systèmes de question réponse, soit parce que leur sens doit être pris en compte lors de la catégorisation de la question, soit parce qu'ils aident à localiser la réponse lors de son extraction.
Les subtilités qui ne peuvent être traitées par un moteur de recherche lors du passage de la question à la requête peuvent être prises en charge lors de la catégorisation de la question. Cette étape permet en effet de regrouper un maximum d'informations sur le type de réponse attendu d'après le contenu de la question avant l'« élagage » que constitue le passage à la requête. La catégorisation sémantique des questions consiste à regrouper des questions possédant des traits surfaciques communs dans des catégories parfois distinctes. En effet, le type sémantique ne tient pas exclusivement compte du pronom interrogatif ou de la seule syntaxe de la question. Certaines questions commencent par le même pronom interrogatif mais ne visent pas le même type de réponse (les questions commençant par Who[1], visant soit un nom de personne, soit certains éléments de réponse que l'on peut inférer : en général, quelqu'un est connu pour le poste, la fonction qu'il occupe ou pour des actions qu'il a menées, ou les événements auxquels il a été mêlé). Un apprentissage basé sur la reconnaissance de patrons lexico-syntaxiques associés à des catégories de questions devrait permettre d'orienter de façon adéquate des questions visant des types de réponse différents. (Voir aussi : Hermjakob, 2001 et Li, 2002).

1.2 Critères linguistiques pour la catégorisation

La catégorisation des questions que nous avons mise au point est sémantique ; elle s'appuie essentiellement sur le type de réponse attendu, autrement dit, la cible de la question. Nous définissons en effet la cible comme un pronom interrogatif et/ou un mot-clef qui *vaut pour* la réponse (se « substitue » à elle en quelque sorte[2]). Il est donc indispensable de reconnaître la cible afin de catégoriser une question. La cible est indiquée en gras dans les exemples suivants :

[1] Exemples de questions extraites de TREC-9, suivies de leur identifiant (comme toutes celles citées dans cet article) : Who is the richest person in the world? (294) *vs* Who is Desmond Tutu? (287). Questions et réponses sont distinguées du reste du texte grâce à une fonte particulière.
[2] Dans une équation mathématique, on pourrait représenter la cible par l'inconnue *x*.

1) **Name** a Salt Lake city **newspaper**. (745)
2) **Where** is Trinidad? (368)

Name indique que nous cherchons un nom et *vaut* pour le nom du journal (newspaper) que nous cherchons, tout comme Where indique que nous cherchons un lieu et *vaut* pour le lieu que nous cherchons. Le focus, en revanche, est constitué des mots-clefs qui devraient permettre de localiser la réponse (ici : Salt Lake city ; Trinidad) (Ferret *et al*., 2002). A partir de l'analyse de la cible, et en fonction de notre échantillon comprenant 693 questions tirées de TREC-9, nous avons déterminé six catégories de questions – plus ou moins conséquentes - basées sur le type sémantique de la réponse attendue : Entités Nommées (EN, 459 questions), Entités (105), Définitions (63), Explications (61), Actions (3), Autres (2). Par « Entités », nous désignons des réponses qui peuvent être extraites à l'aide de patrons, comme les Entités Nommées, mais qui ne sont ni des noms propres, ni des identifiants uniques (animaux, végétaux, armes, etc.). (Sekine, 2002) les inclut dans sa représentation hiérarchique des réponses possibles.

Dans le tableau ci-dessous, on voit que les quelques questions extraites de la catégorie Entités correspondent toutes au même patron lexico-syntaxique (colonne 4). La cible de la question (en gras) correspond au GN2 (groupe nominal complément d'objet) introduit par le pronom interrogatif What. Le focus, en italique, correspond au GN1 (groupe nominal sujet).

Type sém	Cible	Lien Q-R	Patrons des questions	Questions
Entité	Sport GN2	spec	What **GN2** aux *GN1* V?	What **sport** do the *Cleveland Cavaliers* play?
Entité	Animal GN2	spec	What **GN2** aux *GN1* V?	What **animal** do *buffalo wings* come from?
Entité	Instrument GN2	spec	What **GN2** aux *GN1* V?	What **instrument** does *Ray Charles* play?

Tableau 1 : Catégorie des questions, lien question-réponse et patrons des questions

Dans la colonne « lien Q-R », on constate qu'à chaque fois la réponse attendue est un spécifique (hyponyme) de la cible. Pour la première question par exemple, si l'on trouve un spécifique de sport près du focus Cleveland Cavaliers dans les documents fournis par le moteur de recherche, il y a de fortes probabilités pour que ce spécifique constitue la réponse. Les liens sémantiques unissant un des mots-clefs de certaines questions à la réponse sont exploitables sous WordNet. Le cas échéant, ils peuvent être intégrés dans la catégorisation (*i.e*. enrichir la requête) afin d'orienter la recherche de la réponse. Cependant, les liens que l'on peut détecter afin de retrouver la réponse ne sont pas toujours fournis par WordNet. Par exemple, la réponse à Which type of soda has the greatest amount of caffeine? (756), *Jolt*, n'est pas présente en tant qu'hyponyme de soda dans WordNet.

Afin de mieux repérer et circonscrire la réponse, d'autres éléments d'informations peuvent être utilisés. Il s'agit de nuances généralement ignorées par un moteur de recherche lors de la transformation de la question en requête. Ces nuances concernent le nombre attendu de réponses (nombre demandé ; nombre possible), les ordinaux et les superlatifs et enfin les modalités. Pour répondre de manière satisfaisante à certains types de questions, la réponse

doit contenir plusieurs renseignements. Par exemple, pour être valide, la réponse à la question suivante doit être composée de *trois* termes, correspondant aux noms des trois navires sur lesquels Christophe Colomb voyagea : What were the names of the three ships used by Columbus? (388). D'autres questions utilisant un déterminant indéfini permettent de proposer plusieurs réponses différentes et malgré tout valides : Name a female figure skater. (567). Certaines questions circonscrivent cependant la réponse à un petit sous-ensemble de réponses possibles (c'est en particulier le cas des questions contenant un superlatif) : Name one of the *major* gods of Hinduism? (237). Pour des questions du style Where do lobsters *like* to live? (258), on cherche à savoir où les homards *préfèrent* vivre - endroit qui ne correspond pas systématiquement à l'endroit où ils vivent effectivement. Pour répondre correctement aux questions, le système doit donc détecter et prendre en compte ces nuances.

Ce sont essentiellement les patrons lexico-syntaxiques des questions qui vont permettre d'orienter chaque question vers la catégorie adéquate tout en en saisissant les nuances. Nous avons représenté les questions sous forme de patrons, en essayant de factoriser un maximum d'éléments (en ne développant pas, par exemple, les éléments de type GN, qui peuvent être développés par ailleurs à l'aide de patrons de réécriture). Cependant, il est nécessaire de garder certains traits pertinents car particuliers et distinctifs, si l'on veut, à partir des patrons, rediriger la question vers la catégorie adéquate. Par exemple, le patron What be GN? n'est pas assez fin, car il convient à la fois aux questions de type Définition : What is a nematode? (354), Entité : What is California's state bird? (254) ; Entité Nommée : What is California's capital? (324). Afin de distinguer des questions appartenant à une catégorie dont la syntaxe est similaire à d'autres questions appartenant à une autre catégorie, nous avons besoin d'inclure dans nos patrons certains lemmes ou mots de la question. On aura ainsi, pour les exemples précédents, les patrons suivants : What be (**a**|the) N? (354) ; What be GNprep **GN**? (254) ; What be GNprep **capital**? (324). Le traitement de l'antonymie, des restrictions et des modaux se fait de manière identique, en incluant dans les patrons des termes distinctifs.

Ainsi, certains termes ne sont pas étiquetés mais conservés tels quels : What be the population of ? Le fait de garder le terme population, qui fait partie intégrante de la cible et renvoie à une réponse de type (EN, nombre) permet de catégoriser la question de manière adéquate afin d'orienter correctement la recherche de la réponse. De la même manière, les traits spécifiques sont indiqués par un S lorsqu'il s'agit de superlatifs (en gras) : What state have **S** GNp2? pour What state has **the most** Indians? (208). Pour repérer ces traits spécifiques, on peut générer des règles grammaticales qui relèvent des marqueurs comme most devant un adjectif, ou encore les suffixes « er » ou « est » ajoutés à la racine de l'adjectif, ou créer une liste de termes pour les exceptions.

Lorsque les termes sont interchangeables avec d'autres termes issus du même paradigme, cela limite au contraire le nombre de patrons. Par exemple, pour un ensemble donné de question appartenant à la catégorie Définition, on peut appliquer le patron suivant : What aux [the initials| acronym| abbreviation] X [stand for|mean]? Il y a donc un équilibre à trouver entre une représentation globale et trop abstraite de la question et une représentation trop fine qui ne permettrait pas de réemployer les patrons afin de catégoriser automatiquement de nouvelles questions.

Les GN représentant des personnes sont signalés par GNp, qui correspond souvent à une fonction, une nationalité ou une profession (en gras) : What state have S **GNp**2? pour What state has the most **Indians**? (208). Cette étiquette sert dans certains cas à savoir que l'on recherche une Entité Nommée de type Nom de personne : dans le cas de la question What was the name of the first Russian *astronaut*?, astronaut désignant une personne,

nous pouvons en déduire que nous recherchons le nom d'une personne (vs What was the name of the first *car*?). L'étiquetage des entités nommées, en particulier des noms de lieux, peut être utile lorsque la question vise à connaître la situation géographique d'un lieu. Une fois le terme correspondant à l'EN lieu repéré, en fonction de la syntaxe de la question, on recherchera dans WordNet un holonyme ou un méronyme de ce terme. Les questions contenant what kind|type|sort permettent aussi d'inférer que la réponse peut être un hyponyme du terme introduit par kind|type|sort.

IDQ	Q	patronQ	Typel	type sém	rel QR
201	What was the name of the first Russian astrona	What be the name of S GNp1 Vinf GN2'	ENS	NP	
202	Where is Belize located?	Where aux GNlieuPays ppé ?	EN	lieu	holo
203	How much folic acid should an expectant moth	How much GN2 aux GNp1 V adv ?	EN	num-qte	
204	What type of bridge is the Golden Gate Bridge?	What kind|type of GN be GN ?	DEF		spec
205	What is the population of the Bahamas?	What be the population of GNlieuPays?	EN	num-nbre-pop	
206	How far away is the moon?	How far away be GN?	EN	num-distance	
207	What is Francis Scott Key best known for?	What aux GNP known for?	EXP	WFP	
209	Who invented the paper clip?	Who V GN2 ?	EN	NP	
210	How many dogs pull a sled in the Iditarod?	How many GN1 V GN2 CCL?	EN	num-nbre	
211	Where did bocci originate?	Where aux GN1 V ?	EN	lieu	
212	Who invented the electric guitar?	Who V GN2 ?	EN	NP	
213	Name a flying mammal.	Name a GN	E	animal	spec
214	How many hexagons are on a soccer ball?	How many GN be (CCL|there)?	EN	num-nbre	
215	Who is the leader of India?	Who be GNp GNprep?	EN	NP	

Tableau 2 : Extrait de la base de données contenant la catégorisation des questions et leur patron

*Légende : **GN** : groupe nominal ; **GN1** : sujet ; **GN2** : complément d'objet direct ; **GNprep** : introduit par une préposition ; **GNp** : représentant une personne ; **GNP** : composé d'un nom de personne (EN) ; **V** : verbe*[3] *; **Vinf** : à l'infinitif ; **Aux** : auxiliaire ; **Ppé** : participe passé ; **S** : trait spécifique (superlatif, ordinal…) ; **CCL** : complément circonstanciel de lieu ; **Adv** : adverbe.*

2 Critères statistiques et appariement question – réponse

Nous avons vu que la catégorisation, basée sur des critères linguistiques, constitue une étape incontournable au sein du QR. Une autre façon d'améliorer l'appariement repose sur des données statistiques. Dans cette section, nous nous pencherons sur les réponses afin d'analyser le nombre d'occurrences des mots-clefs issus des questions ainsi que leurs expansions.

2.1 Mots-clefs à sélectionner

Comme les modèles en RI ont été créés afin de retrouver des documents traitant d'un sujet ou d'un thème – ce qui diffère du fait de retrouver une réponse concise à une question précise – nous envisageons de modifier le modèle vectoriel classique afin de l'adapter au QR (voir aussi Bellot *et al*., 2003 et Clarke *et al*., 2003). En prenant en compte les étiquettes morpho-syntaxiques des mots-clefs, les types d'expansion et la catégorie de la question, nous pourrions d'une part sélectionner les mots-clefs et les expansions utiles au repérage de la réponse et d'autre part leur attribuer un poids. Nous avons vu, en effet, que la simple présence

3 Tous les verbes sont lemmatisés lorsqu'ils ne sont pas représentés par l'étiquette V. Ainsi, pour What was the name, on a What be the name.

de mots-clefs à proximité du type de réponse recherché ne constitue pas un critère suffisant pour repérer la bonne réponse (Lavenus, Lapalme, 2002).
Afin de mener à bien notre étude, nous avons transformé automatiquement chaque question de TREC-9 préalablement étiquetée en requête. Nous n'avons gardé que les noms, les noms propres, les adjectifs, les adverbes et les verbes. Ensuite, nous avons extrait les mots-clefs ainsi que leurs expansions (donnés par WordNet 2.0) automatiquement dans les réponses valides de TREC-9, constituées de 250 caractères maximum[4]. Tout d'abord cela nous a permis de voir quels sont les mots-clefs les plus fréquents et les plus proches de la réponse au sens strict. Nous envisageons ensuite de mener une étude complémentaire afin de voir si le nombre d'occurrences a un rapport avec le rôle syntaxique du mot-clef dans la question et avec la catégorie sémantique de la question (voir aussi Li, 2003).
On peut voir dans le tableau 3 que nous avons obtenu 2425 mots-clefs à partir des 693 questions de TREC-9, soit une moyenne de 3,49 mot-clefs par question. Dans la mesure où nous avons considéré les verbes to be et to have comme des mots-outils[5], il y a seulement 307 verbes pour 693 questions (ce qui représente 13,48 % des mots-clefs). La suppression des mots-outils intervenant après la catégorisation sémantique des questions, cette mesure ne devrait pas avoir d'incidence sur le processus de recherche de la réponse.
Les mots-clefs des questions sont essentiellement composés de noms (39,83 %), de noms propres (33,65 %) et d'adjectifs (9,65 %), ce qui n'est pas surprenant. Mais si l'on regarde la répartition des mots-clefs dans les réponses, on constate que le nombre de noms propres augmente (58,32 %) tandis que le nombre de noms (30,41 %), d'adjectifs (6,09 %) et de verbes (4,45 %) chutent. Ces nombres confirment l'intuition selon laquelle les noms propres constituent de bons critères pour trouver la réponse. Dans le cadre d'un modèle vectoriel adapté au QR, on pourrait ainsi leur attribuer un poids élevé. Cela signifie aussi que des questions contenant des noms propres, indépendamment de leur catégorie sémantique, devraient être plus faciles à traiter, et pourraient même être traitées en suivant le même processus.

Répartition des mots-clefs dans les Q			**Répartition des mots-clefs dans les R**		
étiquette	nombre	pourcentage	étiquette	nombre	pourcentage
cardinal	20	0,82%	CD	79	0.44%
adjectif	**208**	**8,58%**	**JJ**	**934**	**5.22%**
superlatif	**26**	1,07%	**JJS**	**156**	**0.87%**
nom	**844**	**34,80%**	**NN**	**4685**	**26.19%**
Nom, pluriel	**122**	5,03%	**NNS**	**755**	**4.22%**
Nom propre	**808**	**33,32%**	**NP**	**10395**	**58.11%**
Nom propre, pluriel	**8**	0,33%	**NPS**	**37**	**0.21%**
adverbe	36	1,48%	RB	130	0.73%
Adverbe, superlatif	2	0,08%	RBS	1	0.01%
verbe	72	2,97%	VV	154	0.86%
Verbe au passé	96	3,96%	VVD	187	1.05%
Verbe en -ING	16	0,66%	VVG	28	0.16%
Participe passé	94	3,88%	VVN	277	1.55%
Verbe, autre que 3eme pers sing	45	1,86%	VVP	91	0.51%
Verbe, 3eme pers sing	28	1,15%	VVZ	57	0.32%
	2425	100.00%		**17887**	100.00%

Tableau 3 : Répartition des étiquettes des mots-clefs dans les questions (Q) et les réponses (R)

[4] Les réponses valides *stricto sensu* fournies par les concurrents de TREC-9 sont encadrées par les balises ouvrantes <REP> et fermantes </REP> ont été disposées automatiquement par l'équipe du groupe LIR au sein des bribes de 250 caractères maximum.
[5] Ces termes, fréquents et peu significatifs en soi, ne constituent pas un indice pertinent pour localiser la réponse exacte. Il ne nous a donc pas semblé opportun de repérer leur présence dans les bribes de réponses.

Répartition des MC avant la réponse exacte			**Répartition des MC dans la réponse exacte**			**Répartition des MC après la réponse exacte**		
étiquette	nombre	pourcentage	étiquette	nombre	pourcentage	étiquette	nombre	pourcentage
CD	30	0.37%	CD	1	0.11%	CD	48	0.54%
JJ	**391**	**4.84%**	**JJ**	**54**	**5.82%**	**JJ**	**489**	**5.45%**
JJS	54	0.67%	JJS	6	0.65%	JJS	96	1.07%
NN	**2017**	**24.98%**	**NN**	**175**	**18.86%**	**NN**	**2493**	**27.81%**
NNS	245	3.03%	NNS	63	6.79%	NNS	447	4.99%
NP	**4942**	**61.22%**	**NP**	**602**	**64.87%**	**NP**	**4851**	**54.11%**
NPS	17	0.21%	NPS	0	0.00%	NPS	20	0.22%
RB	53	0.66%	RB	1	0.11%	RB	76	0.85%
RBS	0	0.00%	RBS	0	0.00%	RBS	1	0.01%
VV	69	0.85%	VV	16	1.72%	VV	69	0.77%
VVD	59	0.73%	VVD	0	0.00%	VVD	128	1.43%
VVG	9	0.11%	VVG	2	0.22%	VVG	17	0.19%
VVN	131	1.62%	VVN	5	0.54%	VVN	141	1.57%
VVP	33	0.41%	VVP	3	0.32%	VVP	55	0.61%
VVZ	23	0.28%	VVZ	0	0.00%	VVZ	34	0.38%
	8073	100.00%		**928**	100.00%		**8965**	100.00%

Tableau 4 : Répartition des étiquettes des mots-clefs (MC) avant, dans et après les balises <REP> qui indiquent la réponse exacte

Tout d'abord, le tableau 4 montre que la plupart des mots-clefs se trouvent avant (**8073**, soit 44,93 %) ou après (**8965**, soit 49,89 %) la réponse exacte qui ne contient que 5,16 % (**928**) des mots-clefs issus des questions. Cette répartition s'explique par la nature même des réponses, étant donné qu'une réponse peut compter jusqu'à 250 caractères, tandis que la réponse *stricto sensu* est limité à un ou quelques mots. Alors que le pourcentage d'adjectifs trouvés dans différentes positions dans les réponses est stable, il y a plus de noms avant et surtout après les réponses que dedans. En revanche, les noms propres sont plus nombreux dans et avant la réponse qu'après. Ces pourcentages représentent certainement les entités nommées.

étiquette	Répartition à –1 et +1	pourcentage	**Répartition dans REP**	pourcentage
NP	120	**36.59%**	73	**15.53%**
NN	117	**35.67%**	41	**8.72%**
NNS	22	**6.71%**	12	**2.55%**
JJ	21	6.40%	5	1.06%
VVD	13	3.96%	0	0.00%
VVN	10	3.05%	2	0.43%
VV	6	1.83%	1	0.21%
VVP	5	1.52%	2	0.43%
RB	5	1.52%	1	0.21%
VVG	4	1.22%	2	0.43%
VVZ	3	0.91%	0	0.00%
JJS	1	0.30%	2	0.43%
CD	1	0.30%	1	0.21%
	328	100.00%	142	**30.21%**

Tableau 5 : Répartition des étiquettes à proximité immédiate et dans la réponse exacte (REP)

Dans le tableau 5, nous avons considéré l'étiquette morpho-syntaxique du mot-clef se trouvant juste avant la réponse au sens strict (-1) et juste après (+1). Nous avons aussi, bien entendu, pris en compte l'étiquette morpho-syntaxique du mot-clef trouvé dans la réponse exacte. Les noms propres semblent précéder ou suivre la réponse exacte (36,59 %) mais en

fait les noms sont plus nombreux (NN + NNS = 42,38 %). Les verbes, avec un total de 12,49 %, sont bien plus présents avant et après la réponse exacte que dedans (4,45 %). On peut supposer que les réponses exactes contenant des verbes correspondent à la catégorie de question « Action », une des plus faibles (voir section 1.2).
30,21 % des mots-clefs qui sont proches des réponses sont contenus dans les balises <REP> et sont majoritairement composés de noms propres (15,53 %) et de noms (NN + NNS = 11,27 %). Nous pouvons avancer que ces critères morpho-syntaxiques correspondent aux réponses possibles des questions visant respectivement des Entités Nommées et des Entités (voir section 1.2).

étiquette	avant	dedans	après	Moyenne du nombre d'occurrences
CD	40.74%	1.67%	**57.59%**	6.58
JJ	47.26%	1.99%	50.75%	7.13
JJS	40.10%	4.97%	54.93%	8.21
NN	45.25%	4.22%	50.53%	8.25
NNS	32.41%	**12.63%**	**54.96%**	8.88
NP	50.27%	4.72%	45.01%	14.46
NPS	36.59%	0.00%	**63.41%**	9.25
RB	45.73%	1.04%	53.23%	5.42
RBS	0.00%	0.00%	100.00%	1.00
VV	50.17%	3.38%	46.46%	5.13
VVD	36.53%	0.00%	**63.47%**	3.82
VVG	27.99%	**11.11%**	**60.90%**	3.11
VVN	49.27%	3.13%	47.60%	5.54
VVP	44.59%	2.22%	53.18%	3.37
VVZ	35.93%	0.00%	**64.07%**	6.33

Tableau 6 : Pour chaque étiquette morpho-syntaxique : position préférentielle dans la réponse

Dans le tableau 6, pour chaque étiquette de mot-clef, nous avons estimé le nombre de fois où elles apparaissent avant, dans et après la réponse exacte. Rappelons que la longueur de la réponse exacte est inférieure à celle des bribes la suivant ou la précédant. C'est pourquoi les nombres de la colonne « dedans » sont peu élevés. Cependant, on remarque parfois un déséquilibre entre la répartition des étiquettes avant et après la réponse exacte. En effet, certaines étiquettes de mots-clefs apparaissent surtout après la réponse exacte (CD, NNS, NPS, VVD, VVG, VVZ). Ce tableau montre aussi la moyenne du nombre d'occurrences par étiquette morpho-syntaxique.

2.2 Expansions à sélectionner

Après avoir présenté la répartition des mots-clefs dans les réponses, voyons maintenant quels sont les expansions que l'on peut retrouver dans les bribes de 250 caractères fournies par les concurrents de TREC-9 (voir aussi : Prager, Chu-Carroll: 2001 ; Yang *et al*.: 2003). Nous sommes en effet partis du constat que la réponse à certains types de questions peut être constituée (voir section 1.2) ou introduite (voire suivie) par une relation trouvée sous WordNet. Les relations fournies par WordNet ont été simplifiées, de manière à ne rechercher que les synonymes, hyperonymes, hyponymes, holonymes et méronymes dans ou près de la réponse au sens strict.
Lors de nos prochaines expérimentations, nous chercherons, le cas échéant, à établir un lien entre une catégorie de question donnée et un type d'expansion. Nous ferons ainsi la différence entre les expansions qui introduisent la réponse exacte et celles qui la constituent. La

répartition globale[6] des expansions dans les réponses est la suivante : les hyponymes sont les expansions les plus fréquentes (28,76 %), suivies par les hyperonymes (21,19 %) et les synonymes (17,93 %). Viennent ensuite les méronymes (11,97 %) et les holonymes (10,41 %). Mais il est plus intéressant de considérer la répartition des expansions en fonction de leur position dans la réponse :

Étiquette des expansions	Nombre AVANT	pourcentage	Étiquette	Nombre DEDANS	Pourcent.	Étiquette	Nombre APRES	Pourcent.
Also see	9	0.18%	Also see	3	0.19%	Also see	7	0.13%
attribute	10	0.20%	attribute	2	0.12%	attribute	11	0.21%
cause	29	0.59%	cause	2	0.12%	cause	51	0.95%
entailment	106	2.14%	entailment	1	0.06%	entailment	97	1.81%
holonym	415	8.39%	holonym	356	**22.10%**	holonym	469	8.76%
hyperonym	1020	20.63%	hyperonym	297	18.44%	hyperonym	1207	22.54%
hyponym	1391	28.13%	hyponym	488	30.29%	hyponym	1547	28.89%
meronym	557	11.26%	meronym	223	13.84%	meronym	646	12.06%
Similar to	39	0.79%	Similar to	1	0.06%	Similar to	46	0.86%
synonym	1020	**20.63%**	synonym	217	13.47%	synonym	899	16.79%
Verb group	349	**7.06%**	Verb group	21	1.30%	Verb group	375	**7.00%**
	4945	**100.00%**		**1611**	**100.00%**		**5355**	**100.00%**

Tableau 7: Répartition des expansions avant, dans et après la réponse exacte

Au total, 41,51 % des expansions se trouvent avant la réponse exacte, 13,52 % se trouvent dedans, et 44,95 % après. Si l'on compare cette répartition avec la répartition globale (*cf. supra*), les différences ne sont pas significatives. Cependant, on remarque que les holonymes se trouvent surtout dans la réponse, les synonymes surtout avant et les groupes verbaux avant ou après la réponse exacte.

Conclusion

Les systèmes de QR utilisent des méthodes issues de la RI et de l'extraction d'information pour retrouver des documents contenant une réponse valide. Comme les méthodes de recherche d'information ont été créées pour retrouver des documents – et non pas une information spécifique – les systèmes de QR ont besoin de méthodes complémentaires pour affiner leur recherche. Nous avons ainsi démontré l'importance de la catégorisation des questions afin de réduire le fossé question – réponse. Notre catégorisation, basée sur des patrons de questions, permet de délimiter la cible de la question afin de classer celle-ci sémantiquement. Elle permet aussi de tenir compte de nuances et, dans certains cas, d'inférer le lien sémantique qui permettra de retrouver la réponse.

Par ailleurs, nous projetons de modifier le modèle vectoriel classique. Si l'on attribue des poids aux mots-clefs de la question et à leurs expansions en fonction d'une catégorie de question donnée, on pourrait adapter le modèle vectoriel au QR. Il faudrait voir, en moyenne, de quoi est constituée une réponse (mots-clefs, expansions, étiquettes morpho-syntaxiques des autres mots), et vérifier à chaque fois si les informations précédentes peuvent être rattachées à une catégorie particulière. Cette étude n'est pas terminée mais nous avons déjà quelques résultats intéressants sur la répartition des mots-clefs et des expansions avant, dans et après la réponse exacte. Nos résultats pourraient en effet servir à adopter une stratégie de recherche différente de celle pratiquée jusqu'alors, en se basant en partie sur l'étiquetage morpho-

[6] Globale, c'est-à-dire quelle que soit la place des expansions au sein de la bribe de 250 caractères.

syntaxique. Nous envisageons aussi de mesurer la distance entre les mots-clefs, les expansions et la réponse au sens strict. Chaque mot se trouvant entre un mot-clef ou une expansion et la réponse comptera pour 1. Cette étape supplémentaire nous permettra d'effectuer une analyse croisée, entre :
- les mots-clefs ou expansions les plus nombreux trouvés dans les réponses à une question donnée ;
- les mots-clefs ou expansions les plus proches de la réponse *stricto sensu*.
Afin de valider notre hypothèse, qui consiste à attribuer des poids aux mots-clefs et à leurs relations sémantiques de façon à modifier le modèle vectoriel pour l'adapter aux spécificités du QR, nous procéderons aux mêmes traitements sur les réponses *non* valides de TREC-9. Pour compléter notre étude, nous utiliserons aussi les dérivés morphologiques, disponibles sous WordNet 2.0.

Remerciements

Nous tenons à remercier Emmanuel Cartier pour sa contribution.

Références

BELLOT P., CRESTAN E., EL-BÈZE M., GILLARD L., DE LOUPY C. (2003), Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC-11 Question Answering Track, Actes de *The Eleventh Text Retrieval Conference*, Gaithersburg, Maryland: NIST Special Publication.

CLARKE C.L.A., CORMACK G.V., KEMKES G., LASZLO M., LYNAM T.R., TERRA E.L., TILKER P.L. (2003), Statistical Selection of Exact Answers, Actes de *The Eleventh Text Retrieval Conference*, Gaithersburg, Maryland: NIST Special Publication.

FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G., MONCEAUX L., ROBBA I., VILNAT A. (2002), Finding An Answer Based on the Recognition of the Question Focus, Actes de *The Tenth Text REtrieval Conference,* Gaithersburg, Maryland: NIST Special Publication.

HERMJAKOB U. (2001), Parsing and Question classification for Question Answering, Actes de *Workshop on Open-Domain Question Answering at ACL-2001*, Toulouse, France.

LAVENUS K., LAPALME G. (2002), Évaluation des systèmes de question réponse - Aspects méthodologiques, *Traitement Automatique des Langues*, Vol. 43, n°3/2002, pp. 181-208.

LI X., Roth D. (2002), Learning Question Classifiers, Actes de *COLING'02*.

LI X. (2003), Syntactic Features in question Answering, Actes de *SIGIR 2003.*

PRAGER J., CHU-CARROLL J. (2001),Use of WordNet Hypernyms for answering What-Is Questions, Actes de *The Tenth Text Retrieval Conference* (TREC 2001), Gaithersburg, Maryland: NIST Special Publication.

SEKINE S., SUDO K., NOBATA C. (2002), Extended Named Entity Hierarchy, Actes de *LREC-2002*, pp. 1818-1824.

YANG H., CHUA T.-S. (2003), The Integration of Lexical Knowledge and External Resources for Question Answering, Actes de *The Eleventh Text Retrieval Conference,* Gaithersburg, Maryland: NIST Special Publication.