Analyse des échecs d'une approche pour traiter les questions définitoires soumises à un système de questions/réponses

Laurent GILLARD, Patrice BELLOT, Marc EL-BÈZE
Laboratoire d'Informatique d'Avignon (LIA)
Université d'Avignon et des Pays de Vaucluse
339 ch. des Meinajaries, BP 1228
F-84911 Avignon Cedex 9 (France)
{laurent.gillard,patrice.bellot,marc.elbeze}@univ-avignon.fr

Résumé. Cet article revient sur le type particulier des questions définitoires étudiées dans le cadre des campagnes d'évaluation des systèmes de Questions/Réponses. Nous présentons l'approche développée suite à notre participation à la campagne EQueR et son évaluation lors de QA@CLEF 2006. La réponse proposée est la plus représentative des expressions présentes en apposition avec l'objet à définir, sa sélection est faite depuis des indices dérivés de ces appositions. Environ 80% de bonnes réponses sont trouvées sur les questions définitoires des volets francophones de CLEF. Les cas d'erreurs rencontrés sont analysés et discutés en détail.

Abstract. This paper proposes an approach to deal with definitional question answering. Our system extracts answers to these questions from appositives appearing closed to the subject to define. Results are presented for CLEF campaigns. Next, failures are discussed.

Mots-clés : système de questions/réponses, questions définitoires.

Keywords: question answering, definitional question answering.

1 Introduction

Les systèmes de Questions/Réponses (sQR) se proposent d'aller au delà de la recherche de documents pertinents afin de répondre, précisément et avec concision, à une question directement formulée en langue naturelle. L'étude de ces systèmes est encouragée par des campagnes d'évaluation qui spécifient des axes de recherche comme la nature des questions à considérer. Cet article s'intéresse tout particulièrement aux questions « définitoires » (QD), en domaine ouvert, telles que *Qui était Alfred Nobel ?* (CLEF06/28), qui interroge sur les aspects biographiques d'un individu, ou les questions *Qu'est-ce que la RKA ?* (CLEF06/95) et *Qu'est-ce que Hubble ?* (CLEF06/02), qui attendent une forme étendue ou encore une (voire *La*) caractéristique remarquable de l'objet à définir. Ces questions ont été introduites lors du volet Questions/Réponses (QR) de la campagne TREC-9. Il est à noter que dans ce qui suit, nous nous limitons au contexte des campagnes EQueR (Ayache *et al.*, 2006) et plus spécifiquement CLEF (Vallin *et al.,* 2006), où une seule et unique réponse est à produire. En effet, depuis 2003 (Voorhees*,* 2003) et dans les campagnes TREC (Voorhees*,* 2005), les réponses attendues aux QD sont constituées de l'intégralité des faits pertinents connus sur le sujet à définir ; cela, au travers d'un découpage en *« pépites »* d'informations vitales (à maximiser), non vitales (indifférentes) et inintéressantes (à minimiser et pénalisantes). *A Contrario*, dans

ce travail, notre objectif est d'extraire *LA* meilleure des réponses pour une QD ; et nous souhaitons y parvenir depuis la détection de mises en apposition et l'emploi d'un minimum de ressources. De plus, ce travail, préliminaire du point de vue de la tâche, nous permet d'explorer l'utilisation de la proximité immédiate des objets à définir avec leur définition. Une autre motivation provenait des faibles performances obtenues par notre système lors d'EQueR sur ces QD : seulement 7% de réponses courtes correctes pour les QD concernant des personnes et 42% pour les autres. Cela est d'autant peu qu'une partie était obtenue grâce à des bases de connaissances et par conséquent la projection de ressources exogènes (couteuses à maintenir). La méthode employée était basée sur un appariement entre un type de réponse attendu et la détection au sein des documents d'Entités du type adéquat. La principale difficulté rencontrée était liée à la l'identification des limites précises de l'énoncé correspondant à la définition. Pour illustrer cette détection parfois mal aisée, considérons une fonction (ou profession), qui débute pourtant par *chef* (une telle construction est assez fréquente en QR) :

- Bouraima Koné, [$_{Fonction}$chef] des opérations techniques] de lutte d'urgence] contre les criquets] au ministère malien] de l'agriculture], au micro de Jean Paul Ade.
(Lu en ligne, http://www2.dw-world.de/french/Politik_Afrika/1.173728.1.html)

Il apparaît, comme le signale les crochets fermants, que la frontière droite de l'expression est délicate à apprécier, et donc sujette à erreurs lors d'un étiquetage en Entités Nommées (ou même une extraction à l'aide de patrons). Par conséquent, le risque existe de faire glisser la réponse extraite vers une réponse incorrecte ou inexacte. Pourtant, cet exemple reste un cas simple : pour les QD qui commencent par *Qui*, la définition à trouver n'est pas systématiquement un rôle social mais peut être n'importe quelle raison pour laquelle une personnalité est connue (de telles QD sont parfois étiquetées *« WhyFamous »*). Et pour les *Qu'est-ce que*, les possibilités sont encore plus vastes. Aussi, notre approche a été de partir des expressions mises en apposition avec les objets à définir pour sélectionner celle qui semble la meilleure (selon différents critères et indices, qui sont présentés en section 3) et de la proposer comme réponse. L'utilisation d'une apposition permet de s'affranchir d'une détection plus hasardeuse, et puisqu'il s'agit d'un extrait du document, cela nous permet de supposer une réponse mieux construite. L'approche que nous présentons est évaluée dans le cadre de notre participation à la campagne QA@CLEF-2006 (Gillard *et al.*, 2006) et obtient autour de 80% de bonnes réponses. Ensuite (section 4), les cas d'échecs que nous avons rencontrés sur CLEF-2006 sont analysés et discutés en détails afin de mettre en évidence des améliorations à mettre en œuvre ou des points importants à considérer pour traiter de telles questions.

2 Autres travaux sur les questions définitoires

Un travail similaire à celui-ci a été fait par (Malaisé *et al.*, 2005) mais il porte sur la détection d'énoncé définitoires dans le domaine médical de la tâche spécialisée de la campagne EQueR. Étrangement, les constructions faisant intervenir des appositions ne sont pas listées dans les nombreux patrons d'extraction lexico-syntaxiques utilisés pour répondre aux questions.

De nombreuses approches ont été envisagées, et pratiquement chaque système emploie une stratégie différente pour traiter les questions définitoires. Par exemple, (Greenwood, Saggion, 2004) utilisent une étape préalable d'acquisition de termes secondaires depuis des ressources exogènes (WordNet, l'encyclopédie Britannica et le Web, ce dernier contribue d'ailleurs a 78%) pour aider à sélectionner des définitions d'abord extraites avec l'aide de patrons. (Fleischman *et al.,* 2003) centrent également leur travaux sur des patrons pour la collecte des réponses candidates, mais se limitent à deux : la succession *Nom Commun Nom Propre* et la

mise en apposition ; ensuite un apprentissage automatique permet un filtrage. L'apposition est également l'un des patrons de (Hildebrandt *et al.,* 2004) parmi 11 autres constructions pour extraire *a priori* des connaissances du corpus ; ils s'aident également d'une projection des définitions de dictionnaires en ligne. (Prager *et al.*, 2001) utilisent les liens d'hyperonymie de WordNet pour localiser un passage contenant une réponse. D'autres, comme (Cui *et al.*, 2005) proposent d'utiliser des patrons lexico-syntaxique probabilistes aux travers de deux modèles (bigrammes et PHMM). (Han *et al.*, 2006) proposent un modèle purement probabiliste basé sur une séparation entre les modèles pour la question et la définition.

3 Une approche et son évaluation pour les questions définitoires

Comme dans tous les systèmes de Questions/Réponses, les questions définitoires sont d'abord identifiées comme telles puis classées suivant quatre catégories (au moyen de motifs d'expressions régulières) : *D+Personne*, pour les questions telles que *Qui est Neil Armstrong ?* (CLEF06/51) ; *D+Acronyme* pour les questions telles que *Qu'est-ce que l'OUA ?* (CLEF06/48) ; *D+minuscules* pour celles comme *Qu'est-ce que l'effet de serre ?* (CLEF06/189) ; et enfin, ce qui constitue le choix par défaut, la catégorie *D* pour les questions définitoires qui n'entrent dans aucune des précédentes catégories comme par exemple *Qu'est-ce que Challenger ?* (CLEF06/81), *Qu'est-ce qu'Euro Disney ?* (CLEF06/29) ou *Qu'appelle-t-on le Knesset ?* (CLEF06/103). Ensuite, l'objet à définir est obtenu en filtrant, après un étiquetage morphosyntaxique (TreeTagger), les différents pronoms interrogatifs et mots vides de sens comme *être*, *appeler*, *acronyme*, *sigle*, *etc.* Puis toutes les phrases du corpus contenant l'objet à définir sont conservées et étiquetées morphosyntaxiquement. Si aucune phrase n'est trouvée la réponse *NIL* est retournée.

Ce n'est qu'après ces différents prétraitements que les différentes expressions candidates sont extraites. Chacune d'entre elles est accompagnée de différents critères qu'il est possible de percevoir comme des juges dont les votes pondérés permettent de faire préférer *in fine* l'une plutôt que l'autre. La fusion de ces différents jugements est variable suivant l'appartenance de la question à l'une des quatre catégories initiales. Les réglages utilisés ont été obtenus de manière empirique sur les QR des campagnes CLEF-2004, 2005 et EQueR.

Les expressions apposées à celle à définir et séparées d'elle par des virgules sont extraites et constituent l'ensemble préférentiel dans lequel la réponse devrait être extraite. Un critère correspondant à leur fréquence d'apparition dans l'ensemble des phrases leur est associé. Un autre ensemble plus particulièrement adapté aux acronymes et abréviations est défini et correspond à une construction *Expression (CAPITALES)* ou *CAPITALES (Expression)*, par l'intermédiaire de deux extractions de ces *Expressions* : l'une est obtenue depuis un alignement dynamique entre la forme *CAPITALES* et l'*Expression*, l'autre depuis la partie gauche (respectivement droite) la plus redondante ; là encore, un critère de fréquence est associé avec chacune d'elles. D'autres juges sont définis : la présence en tête de l'expression du nom le plus fréquent à la position immédiatement à gauche ; de même, et après application d'un motif morphosyntaxique minimal pour détecter des groupes nominaux, la présence en tête de ce groupe nominal du nom principal déterminé le plus fréquent ; un taux de couverture avec le centroïde des noms les plus fréquents au sein d'une fenêtre de 10 mots autour de l'objet à définir ; un fonction de la longueur de l'expression ; un fonction du nombre de noms ; et deux autres juges binaires pour la présence des noms *président* ou *société* en tête d'expression. Enfin, depuis ces expressions, et une stratégie de fusion des provenances, catégories et juges, les meilleures sont proposées comme réponse avec un comportement par défaut qui consiste à répondre par : l'expression apposée qui est à la fois la plus longue et en adéquation avec le motif de détection des groupes nominaux, si elle existe, ou le nom le plus

fréquent précédant l'objet à définir. Cette réponse par défaut est systématiquement proposée en position 5.

Contexte et évaluation de la méthode : Le tableau 1 propose un référentiel pour l'évaluation de notre méthode au travers d'un bilan sur les résultats obtenus par l'ensemble des participants aux questions définitoires en français des 3 campagnes QA@CLEF passées, mais selon les ventilations que nous avons retenues. En 1ière colonne (*#Q*) est présenté le nombre de QD pour chacune des campagnes et catégories : 20 en 2004, 50 en 2005 et 42 en 2006. Le 2ième groupe de colonnes correspond aux nombres et pourcentages de ces questions pour lesquelles une réponse correcte a été trouvée (RC) par au moins l'un des participants. Il ressort de cette colonne que l'ensemble des stratégies mises en œuvre par tous les systèmes permet, à partir de 2005, de s'approcher de la totalité des réponses à obtenir (94%). Enfin le dernier groupe (*RC par soumission*) permet de situer les performances des systèmes puisque figure le nombre de réponses correctes obtenues par : la/les moins bonnes des soumissions (*Min.*), la/les meilleures (*Max.*), ainsi que leur moyenne arithmétique (*Moy.*). Il est à noter que l'année 2004 est moins représentative puisqu'un seul système a participé. Également, si le (ou les) meilleur des systèmes dépasse 80% de bonnes réponses, la moyenne de ceux-ci se situe en deçà (34% et 50%). Les questions portant sur les mots/concepts les plus généraux (*D+minuscules*), et dont la réponse devrait être proche d'une définition du type dictionnaire, rencontrent une réussite moindre, il est possible d'envisager deux raisons à cela : le fait que les corpus journalistiques employées se prêtent peu à ce type d'extraction (contrairement à des d'informations plus biographiques), ou tout simplement leur relative nouveauté en QR.

		#Q	*Réponses Correctes trouvées (RC)*		*RC par soumission*				
					Min.	*Max.*		*Moy.*	
CLEF-2004	D+Personne	12	7	58%	1	6	50%	3	25%
	D+Acronyme	5	0	0%	-	-	-	-	-
1 participant, 16 soumissions	D	3	0	0%	-	-	-	-	-
	Total	***20***	***7***	***35%***	***1***	***6***	***30%***	***3***	***15%***
CLEF-2005	D+Personne	25	24	96%	2	22	88%	11,50	46%
	D+Acronyme	22	21	95%	0	20	91%	10,75	49%
6 participants, 12 soumissions	D	3	2	67%	0	1	33%	0,33	11%
	Total	***50***	***47***	***94%***	***2***	***43***	***86%***	***16,94***	***34%***
CLEF-2006	D+Personne	17	16	94%	1	15	88%	9,67	57%
	D+Acronyme	5	5	100%	0	5	100%	2,33	47%
7 participants, 15 soumissions	D	16	15	94%	1	13	81%	8	50%
	D+minuscules	4	3	75%	0	2	50%	1	25%
	Total	***42***	***39***	***93%***	***2***	***35***	***83%***	***21***	***50%***

Tableau 1: Bilan sur les réponses correctes(RC) proposées par les participants aux questions définitoires des campagnes QA@CLEF.

Le tableau 2 présente les résultats obtenus par notre méthode sur les mêmes jeux de QD que le tableau 1. Les questions des campagnes CLEF 2004, 2005 et EQueR (non présentées) ont été utilisées pour raffiner la méthode mise au point après notre participation à EQueR. L'évaluation de ces résultats a été faite manuellement. Il en est de même pour la dernière colonne (*Au moins une réponse correcte dans les 5 premières réponses*). En revanche pour les autres données de CLEF 2006, les résultats présentés sont ceux obtenus lors de notre participation à la campagne au volet Français-Français (les résultats en Anglais-Français sont inférieurs en raison d'erreurs de traduction ; 67% de réponses correctes sont trouvées au rang 1 au lieu de 79%). Les *(+1)* et *(+2)* qui figurent dans le tableau correspondent à des réponses qui auraient dû être extraites mais qui ont été perdues à cause de problèmes d'ingénierie, ou pour l'une de la ligne *D+Personne* en raison d'une incertitude que nous avons : est-il

acceptable de définir *Boris Becker* comme une *tête de série n°7* ? (*cf.* discussion en section suivante). Il apparaît de ce tableau que les taux de bonnes réponses sont constants et aux alentours de 80% au premier rang et dépassent 83% pour une réponse correcte placée parmi les 5 premières. Cependant, et comme précédemment souligné par le tableau 1, les questions définitoires portant sur des concepts génériques (*D+minuscules*) ou qui ne concerne ni les personnes ni les acronymes, rencontrent également moins de succès avec notre méthode.

		#Q	*Au rang 1, Réponses*			*Au moins une Réponse Correcte dans les 5ières*	
			Correctes (RC)		*ineXactes*		
CLEF-2004	D+Personne	12	9	75%	2	12	100%
	D+Acronyme	5	5	100%	-	5	100%
	D	3	2	67%	1	3	100%
	Total	***20***	***16***	***80%***	***3***	***20***	***100%***
CLEF-2005	D+Personne	25	20	80%	3	21	84%
	D+Acronyme	22	20	91%	2	22	100%
	D	3	1	33%	1	2	67%
	Total	***50***	***41***	***82%***	***6***	***45***	***90%***
CLEF-2006	D+Personne	17	15	88%	1	15 (+2)	88%
	D+Acronyme	5	4 (+1)	80%	-	4	80%
	D	16	13	81%	2	15	94%
	D+minuscules	4	1	25%	-	1	25%
	Total	***42***	***33***	***79%***	***3***	***38***	***83%***

Tableau 2: Résultats obtenus par notre méthode sur les questions définitoires des campagnes QA@CLEF ; évaluation officielle pour 2006 (soumission FR-FR).

4 Analyse et discussion des cas d'échecs sur CLEF-2006

Cette partie propose une discussion sur les cas d'échecs que nous avons rencontrés. Aussi, elle s'articule autour de quelques questions qui illustrent des difficultés représentatives pour notre système et parfois la tâche elle-même. Il faut également rappeler que notre méthode n'utilise pas d'analyse syntaxique et repose essentiellement sur une recherche d'expressions apposées depuis un découpage en phrases, et, par conséquent, avec un contexte limité.

Qui est Boris Becker ? (CLEF2006/90) Au delà de la simplicité évidente de cette question pour un « lecteur moyen de journaux », la difficulté est réelle puisqu'aucun des systèmes n'a trouvé une réponse correcte. En effet, la réponse retournée par notre système est une nationalité au travers du nom *Allemand* soit le nom qui qualifie le plus fréquemment *Boris Becker* dans le corpus (*l'Allemand Boris Becker* est présent 21 fois sur les 104 occurrences de *Boris Becker*). Cette réponse n'a pas été jugée correcte, la raison étant qu'une nationalité n'est pas suffisante, à elle seule, pour qualifier convenablement une personne. Cette règle connue, il apparaît possible, notamment dans ce cas simple, de filtrer les « mauvais » candidats à l'aide de règles de rejet. Aussi, il faut poursuivre l'investigation. L'une des premières définitions qui viendrait à l'esprit pour définir *Boris Becker* serait très probablement sa qualité de *joueur de tennis*. Mais, au sein d'un découpage en phrase du corpus, *Boris Becker* entre que très rarement en cooccurrence avec un motif comprenant *tennis* (qu'il soit issu de l'expression *joueur de tennis* ou même *tennism(a|e)n*). Et, sur les neuf fois où cela se produit sur les 104 phrases contenant *Boris Becker*, une seule permet effectivement de le définir directement au travers de ce trait :

- *Sa fortune , il l' a bâtie sur les courts de tennis aux côtés de son partenaire de double , le fantasque Ilie Nastase , dans les années 60 ; puis en gérant les affaires des meilleurs tennismen , tels que l' Allemand Boris Becker hier ou le Croate Goran Ivanisevic aujourd'hui .* (LEMONDE94-000881-19940108)

Cependant, malgré la présence du *tels* introductif à une explicitation, il faut souligner ici à quel point le processus de réponse apparaît délicat : il est nécessaire de ne pas tenir compte de

la présence du nom *Allemand* pour revenir en arrière jusqu'à celui de *tennismen*. En outre, une inversion de la position des deux joueurs *Goran Ivanisevic* et *Boris Becker* dans la phrase aurait encore complexifié son analyse et aurait nécessité la mise en balance des deux groupes nominaux par la conjonction *ou*. Et dans ce dernier cas, la tâche aurait été compliquée par la présence des mots *hier* et *aujourd'hui* puisqu'ils apparaissent comme des éléments perturbateurs difficiles à prévoir (notamment dans le cas d'une extraction à partir de patrons morphosyntaxiques) mais pourtant à ignorer. Enfin, il faut se remémorer qu'il s'agit de l'unique cooccurrence entre *tennismen* et *Boris Becker* dans le corpus et par conséquent une prise en compte fréquentielle n'est pas envisageable dans ce cas (cependant une parenthèse mérite d'être ouverte : des procédés de résolutions d'anaphores pourraient augmenter le nombre de candidats mais notre système en est actuellement dépourvu).

Cette (probable) bonne réponse étant écartée, il est possible de s'intéresser à une autre réponse candidate ou plutôt un autre lot de réponses envisageables. En effet, *Boris Becker* est à plusieurs reprises qualifié de *tête de série* comme explicité dans les exemples ci-dessous :

- Victime d' une blessure au dos à l' échauffement , l' Allemand Boris Becker , tête de série numéro 10 , a déclaré forfait avant le match contre Stark (LEMONDE94-002991-19940525)
- Pete Sampras a gagné , dimanche 15 mai , les Internationaux d' Italie de tennis en battant en finale l' Allemand Boris Becker , tête de série n 8 (6-1 , 6-2 , 6-2) . (LEMONDE94-001994-19940517)
- Sur le court n° 1 , dos-à-dos à deux sets partout , l' Allemand Boris Becker , tête de série n° 7 , et l' Ukrainien Andrei Medvedev (n° 9) ont vu leur rencontre interrompue [...] (LEMONDE94-003405-19940629)

Ainsi, une première difficulté survient dans le cas d'une éventuelle factorisation fréquentielle sur *tête de série :* l'expression n'est pas suffisante si le qualifié n'est pas *tête de série numéro 1*. Aussi, il peut être nécessaire de la complémenter. Dans ce cas, cela suppose d'être en mesure de prendre en compte les différentes écritures de *numéro*. Pourtant cela serait une erreur de penser qu'il s'agit d'un même classement comparable et qu'il est possible de suivre sa variation dans le temps au travers des différents documents (auquel cas, une décision aurait pu être de ne considérer que le plus récent). En effet, une *tête de série* correspond à un classement préalable à un tournoi, sorte d'estimation faite en fonction du niveau d'un participant, pour faire en sorte que les meilleurs d'entre eux ne se rencontrent qu'à la fin de la compétition. Aussi la notion de *tête de série* n'a de sens que vis-à-vis d'une compétition sportive. Cependant, comme il est possible de le voir sur ces exemples, ces références ne sont pas toujours présentes dans la fenêtre de la phrase : seul le deuxième exemple propose à la fois la compétition et le domaine au travers des *Internationaux d'Italie de tennis*. Aussi, et finalement, l'ensemble de ces ambiguïtés liées à autant d'élisions peut diminuer considérablement l'intérêt d'une réponse extraite de ces phrases (sauf à pouvoir synthétiser une réponse telle que *tête de série n°8 aux Internationaux d'Italie de tennis* mais dans ce cas il serait probablement préférable d'aller jusqu'à *demi-finaliste aux Internationaux d'Italie de tennis en 1994*).

Il est d'ailleurs à noter qu'une partie de cette discussion eut été différente si plutôt que *tête de série n°X*, l'expression *n°X mondial* avait été présente, puisque alors il aurait fallu prendre justement en compte une variabilité dans le temps de ce classement parmi les meilleurs joueurs mondiaux. Et de s'interroger sur l'opportunité de qualifier *Boris Becker* avec autant de classements différents malgré un dénominateur commun d'être « l'un des 10 premiers joueurs mondiaux de tennis en 1994 » (d'ailleurs, n'est-ce pas la réponse, actuellement hors de portée, qu'il aurait fallu pouvoir inférer de ces différentes réalisations ?).

Après avoir considéré ces candidats de réponses propulsés en tête des possibles en raison de leur répétition, il apparaît parmi les appositions restantes quelques autres susceptibles de donner lieu à des réponses correctes. Cependant, dans trois de celles-ci, le rapprochement

avec le monde du *tennis* n'est pas présent, ce qui par conséquent amènera à présupposer cette connaissance (qui peut ne pas aller de soi). Il est aussi intéressant de noter que trois d'entres elles commencent par des adjectifs numéraux ou multiplicateurs (peut-être faut-il y voir une particularité de l'univers sportif ?). Enfin, et comme pour étayer la discussion passée, une erreur de typographie peut compliquer les rapprochements des *tête de sréie numéro 3*. Tout comme il apparaît délicat de décider si *Boris Becker* est *huitième joueur mondial* ou bien *no 3 mondial* (respectivement dans un document de 1994 et 1995, aussi a-t-il été successivement, l'un puis l'autre ; du moins au moment de l'écriture de chacun de ces documents).

- En déclarant , [...] sans toutefois apporter de preuves , l' Allemand Boris Becker , triple champion de Wimbledon , a relancé la rumeur autour du dopage dans le monde du tennis (LEMONDE94-000256-19940104)
- L' Allemand Boris Becker , huitième joueur mondial et tête de sréie numéro 3 , s' est imposé en finale du tournoi de New-Haven (Connecticut) en battant [...] (LEMONDE94-001895-19940823)
- La fédération accéderait ainsi aux exigences de Stich , qui réclame une somme identique à celle accordée à son compatriote Boris Becker , no 3 mondial , soit 1 million [...]. (LEMONDE95-010221)
- Boris Becker , multiple vainqueur dans les autres tournois du Grand Chelem , a fait une nouvelle fois son deuil des Internationaux de France . (LEMONDE95-022102)

Enfin, il est possible de remarquer que, depuis ce dernier exemple, et puisque la réponse est à fournir hors du contexte du document sans doute serait-il préférable d'être capable de perdre le *autres* pour ne conserver que *multiple vainqueur dans les tournois du Grand Chelem*.

Qui était Alexander Graham Bell ? (CLEF2006/0050) C'est un problème d'ingénierie lié aux nombreuses ponctuations qui a empêché notre système de répondre à cette question depuis :

- Parmi les lieux historiques , le canal de St-Peters (entre le lac du Bras-d' Or et l' Atlantique) , le site dédié à l' inventeur du téléphone , Alexander Graham Bell (à Baddeck , à l' ouest de Sydney) , Louisbourg (lire notre reportage) , la citadelle de Halifax (fortifications du XIXe) , Port-Royal , à 210 km à l' ouest de Halifax [...] . (LEMONDE95-023507)

Mais cette question nous permet d'illustrer une autre difficulté qu'il ne faut pas oublier de considérer lors des étapes de recherche ou d'appariement, qu'il s'agisse de la question, des documents voire de l'un avec l'autre. En effet, il faut autoriser une certaine variabilité dans les noms propres afin de s'assurer qu'*Alexandre* et *Alexander* soit une même personne, qu'un éventuel nom intermédiaire, deuxième prénom, *etc.* puisse également être facultatif ou présent. Cela pour qu'*Arantxa Sanchez Vicario* (CLEF05/175) puisse s'écrire *Arantxa Sanchez-Vicario* mais n'être parfois qu'*Arantxa Sanchez*. S'il en doit en être de même pour *Bill*, *William Jefferson, W.J.* ou même *William J. Clinton*, force est de constater que dans ce cas la solution n'apparaît pas triviale (quelle forme canonique pour les noms propres ?). Enfin, dans certains cas, plus chanceux, si l'écriture fautive apparaît à la fois dans la question et les documents, le système peut établir qu'*Hil(l)ary Clinton* (CLEF06/120) est tout de même *l'épouse du président américain*. Néanmoins, il est aussi possible d'utiliser des algorithmes de tolérance aux fautes.

- Un compromis [...]sans délai en échange de la confirmation par Washington de la présence de Hilary Clinton , l' épouse du président américain , à la conférence [...]. (LEMONDE95-031398)

Qu'est-ce que le Crédit Suisse ? (CLEF2006/0107) Cette question est particulièrement intéressante : comment peut-on définir quelque chose qui apparaît comme déjà transparent ? Notre système n'y est pas parvenu et s'y est même trompé : le *groupe Crédit Suisse* est effectivement un *groupe* (et à de nombreuses reprises), mais ce n'est pas très satisfaisant et si c'est le *seul groupe*, c'est déjà trop. En effet, il faut prendre garde à certains adjectifs qui nuisent plus qu'ils n'apportent. Nous avions envisagé ce problème mais oublié de l'implémenter : s'il est intéressant de savoir que *Stephen Hawking* (CLEF06/0027) est un *célèbre physicien anglais*, ou que

Nick Leeson (CLEF06/0041) est un *ancien courtier de la banque Barings* ; d'autres adjectifs tels *nouveau* doivent être soigneusement évités (et à plus forte raison pour un corpus ancien). Ainsi notre système aurait pu filtrer *Windows* (CLEF06/107), le *nouveau système d' exploitation* jugé inexact (mais aussi parce qu'il était question de l'une des versions de *Windows* dans le document). En outre, et pour revenir au cas *Crédit Suisse*, les systèmes participants n'ont pu faire mieux que de répondre *groupe* qui a été la seule bonne réponse acceptée, à deux reprises, lors de l'évaluation (mais il reste possible de s'interroger sur l'intérêt de cette définition, surtout lorsque *Prix* de *Le Prix Crédit Suisse* est refusé, peut être parce qu'il s'agit d'une *« copie »* ou une *« imitation »*).

Il est à noter que notre système propose en 4ième position, *banque qui a placé les emprunts Biber*, depuis :

- Ces créanciers spéculent [...] hausse des cours , a expliqué Dieter Enkelmann , membre de la direction du Crédit Suisse , banque [est-ce à conserver ? qui a placé les emprunts Biber] . (ATS.940426.0097)

mais la fréquence *banque* (7) n'a pu prendre le pas sur celle de *groupe* (31). Un autre point est qu'il peut apparaître opportun (d'ailleurs plus par conformité avec les spécifications des réponses à produire dans le cadre des campagnes d'évaluations CLEF que du point de vue de la pertinence de la *« pépite »* d'information elle-même) d'éloigner la proposition subordonnée relative pour ne conserver que *banque*. Et encore une fois, dans le cas de *Windows* (CLEF06/107) :

- Microsoft Network [...] que la prochaine version de Windows , le système d' exploitation [est-ce à conserver ? qui a permis au groupe d' asseoir sa domination du marché des logiciels] (ATS.950516.0148)

Pour conclure avec l'institution financière, dans l'exemple ci-dessous, il ne faut pas extraire trop rapidement *banque bâloise* puisqu'il s'agit en fait de la *Société de Banque Suisse (SBS)* et non du *Crédit Suisse*, l'une des *grandes concurrentes* du *SBS*. Mais la compréhension nécessaire n'apparait pas dans la fenêtre de la phrase.

- A l' instar de ses deux grandes concurrentes , l' UBS et le Crédit Suisse , la banque bâloise a subi le contre-coup des turbulences qui ont affecté l' an passé les marchés financiers . (ATS.950315.0084)

Qu'est-ce qu'un samovar ? (CLEF2006/0188) Cette question cristallise les difficultés. Dans le corpus, il n'est pas possible d'obtenir la définition à laquelle on s'attend. Pourtant, c'est justement grâce à une apposition qu'il est envisageable de répondre depuis l'une des 9 phrases utilisant le mot (les 8 autres sont susceptibles de brouiller les pistes). Mais l'analyse pour aboutir apparaît particulièrement complexe : il faut considérer la seconde occurrence du mot plutôt que la première (laquelle seconde est à ignorer sinon), passer outre la forme plurielle, les guillemets fermants, une partie de la première apposition, pour s'arrêter après la troisième, et non plus loin. Ensuite, un samovar peut être *ceux qui, au front, avaient perdu bras ou jambe*. Aucun des systèmes participants n'y est parvenu. Et finalement, cette question pose un autre problème sous-jacent à toute la tâche Questions/Réponses : est-il toujours possible d'extraire automatiquement une réponse concise depuis un passage qui la contient manifestement ?

- Près du [1samovar] , ou entourée de " [2samovars] " , comme on appelait ceux qui , au front , avaient perdu bras ou jambe , la voilà soudain loin de la littérature , se souvenant [...] (LEMONDE95-036764)

Qu'est-ce qu'un t-shirt ? (CLEF2006/0144) Cette question a donné lieu à une réponse vestimentaire sans intérêt liée à une énumération : *veste et cravate*. La réponse attendue était *NIL*, soit celle correspondant à une absence de réponse dans le corpus. Cette bonne absence de réponse a été proposée par 5 systèmes, dont le nôtre mais uniquement dans sa version anglais vers français et cela, seulement à cause d'une erreur de traduction. En effet, notre approche tend à toujours

proposer une réponse (par défaut l'expression apposée la plus longue ou le nom le plus fréquent précédant l'objet de la question). Le seul cas où un *NIL* est retourné survient lorsque l'expression à définir est absente du corpus (comme ce fut le cas à raison pour *Linux* (CLEF06/03), pas encore assez populaire en 1994 et 95, dates des documents de la campagne).

- " Casquette de base-ball , jeans , T-shirt , veste et cravate , tenue de jogging : il est impossible de dresser un portrait-robot de ces candidats à une arme . (LEMONDE95-028295)

5 Perspectives : vers une synthèse de définitions

Un tel processus de réponse à des questions définitoires ouvre une perspective intéressante : il permet d'améliorer la notion de satisfaction en s'essayant à une génération ou plutôt à une synthèse pour la réponse proposée. En effet, les réponses extraites depuis un contexte unique ne sont que très rarement exhaustives (surtout s'il est limité à quelques phrases). Pourtant, et à leur lecture, il apparaît évident qu'il est possible d'améliorer *LA* réponse grâce à l'ensemble des réponses candidates (mais il est alors nécessaire de s'assurer de la qualité de ces réponses candidates, par exemple, depuis des critères fréquentiels ou des coefficients d'association).

Ainsi, lorsque nous cherchons à définir *Airbus* (CLEF06/53), le nom le plus fréquemment associé est *Consortium* et parmi les expressions les plus fréquentes sont *Consortium européen*, *Consortium civil*, et *Consortium aéronautique (Consortium, Programme* et *Consortium européen* ont étés présentées par les systèmes et jugées correctes lors de l'évaluation). D'autre part *Consortium aéronautique européen* apparaît dans le corpus mais épisodiquement. Toutefois, force est de constater que ce dernier semble plus satisfaisant (même si sa fréquence moindre le rend peu sujet à extraction). En outre, il peut être synthétisé depuis les premiers. Il suffit de complémenter le nom le plus fréquemment apposé par tous les adjectifs épithètes qui l'accompagnent dans ses réalisations. Un simple étiquetage morphosyntaxique est suffisant. Enfin, une projection (sur le web ou dans le corpus) des expressions ainsi créées peut permettre de vérifier leur validité du point de vue de leur construction et, notamment, dans ce cas, fixer l'ordre des adjectifs (sauf à définir des règles : une nationalité apparaît souvent en dernier). Il en est de même pour le classique *Bill Clinton* (CLEF06/91) tour à tour *président américain* et *président démocrate* mais pourtant les deux à la fois ou des *navettes Atlantis* (CLEF06/01) et *Challenger* (CLEF06/81), avant tout *spatiales* et *américaines*. Par ailleurs, il peut être opportun de disposer d'une ressource, même limitée, afin, par exemple, de manipuler comme un même concept des noms tels que *Chef* et *Leader* (certaines fonctions reviennent particulièrement dans les QD de ces campagnes), et/ou éviter une éventuelle redite des adjectifs.

Idéalement une telle méthode pourrait être appliquée avec des subordonnées relatives mais ces « *pépites* » d'informations seraient probablement jugées comme inexactes dans le contexte actuel des campagnes, tout en étant susceptibles d'induire plus d'erreurs durant leur génération. Cependant, cela permettrait d'apporter une réponse concise adéquate à la question (fictive) *Qu'est-ce que Bell's Beach ?*, depuis le seul passage du corpus (EQueR) contenant l'expression *Bell's Beach*, puisqu'elle serait alors *(la|une) plage où éclatent les plus grosses vagues d'Australie*. Il est à noter qu'ici, le procédé de réponse implique une mise en relation entre *Beach* et *plage* puis d'utiliser ce qui joue le rôle d'un générique comme antécédent de la proposition relative. En d'autres termes, cela revient à complémenter une sorte de classe/hyperonyme par la proposition relative initialement apposée à l'expression à définir.

C' est du côté de Torquay , petite ville côtière du Victoria au sud -est du pays , située à quelques kilomètres de Bell' s Beach , où éclatent les plus grosses vagues d' Australie , que [...] (LEMONDE99-19935)

6 Conclusion

Dans cet article nous avons étudié l'utilisation de l'apposition pour répondre à des questions définitoires telles qu'elles sont proposées dans les campagnes d'évaluation des systèmes de Questions/Réponses. En effet, dans notre système, les réponses candidates pour ces questions sont extraites depuis leur mise en apposition (ou entre parenthèses) avec les objets à définir. Ensuite, afin de filtrer et retenir la meilleure des expressions apposées comme réponse, un choix est effectué depuis une stratégie impliquant différents indices (principalement fréquentiels) dérivés du voisinage des objets à définir. L'hypothèse forte de notre approche est qu'elle ne nécessite pas de connaissances externes ni même une analyse syntaxique. Évaluée dans le cadre d'une participation à la campagne QA@CLEF-2006, elle trouve environ 80% de bonnes réponses. Cependant une analyse détaillée de quelques cas d'échecs rencontrés nous a amené au constat qu'il n'est pas toujours possible d'aboutir à une réponse. Enfin, et parce que les réponses aux questions définitoires s'y prêtent particulièrement, nous avons esquissé une perspective quant à la synthèse de réponses (depuis les meilleures réponses extraites) pour aller plus avant vers des réponses plus satisfaisantes.

Références

AYACHE C., GRAU B., VILNAT A. (2006). EQueR: The French Evaluation campaign of Question Answering Systems. Actes de *LREC'2006*.

CUI H., KAN M.-Y., CHUA T.-S. (2005). Generic Soft Pattern Models for Definitional Question Answering. Actes de *The 28th ACM SIGIR Conference*, 384-391.

FLEISCHMAN M., HOVY E., ECHIHABI A. (2003). Offline Strategies for Online Question Answering: Answering Questions Before they Are Asked. Actes de *ACL-2003*, 1-7.

GILLARD L., SITBON L., BLAUDEZ E., BELLOT P., EL-BÈZE M. (2006). The LIA at QA@CLEF2006. *The Working Notes for the CLEF 2006 Workshop*.

GREENWOOD M.A., SAGGION H. (2004). A Pattern Based Approach to Answering Factoid, List and Definition Questions. Actes de *The 7th RIAO Conference*, 232-243.

HAN K.-S., SONG Y.-I., RIM H.-C. (2006). Probabilistic model for definitional question answering. Actes de *The 29th ACM SIGIR Conference*, 212-219.

HILDEBRANDT W., KATZ B., LIN J. (2004). Answering Definition Questions Using Multiple Knowledge Sources. Actes de *HLT-NAACL 2004,* 49-56.

MALAISÉ V., DELBECQUE T., ZWEIGENBAUM P. (2005). Recherche en corpus de réponses à des questions définitoires. Actes de la Conférence TALN 2005, 43-52.

PRAGER J., RADEV D., CZUBA K. (2001). Answering What-Is Questions by Virtual Annotation. Actes de *HLT-2001 Conference*, 26-30.

VALLIN A., MAGNINI B., GIAMPICCOLO D., AUNIMO L., AYACHE C., OSENOVA P., PEÑAS A., DE RIJKE M., SACALEANU B., SANTOS D., SUTCLIFFE R. (2006). Overview of the CLEF 2006 Multilingual Question Answering Track. *The Working Notes for the CLEF 2006 Workshop*.

VOORHEES E.M. (2003). Overview of the TREC 2003 Question Answering track, Actes de *The 12th TREC Conference*, 54-68.

VOORHEES E.M. (2005). Chapter 10: Question Answering in TREC. Dans VOORHEES E. M., HARMAN D. (éd.): *TREC Experiment and Evaluation in Information Retrieval.* 233-257.