TALN 2005, Dourdan, 6–10 juin 2005

Recherche en corpus de réponses à des questions définitoires

Véronique Malaisé[1, 2] Thierry Delbecque[2, 3] Pierre Zweigenbaum[2, 3, 4]
(1) DRE de l'Institut National de l'Audiovisuel
4, avenue de l'Europe, 94366 Bry-sur-Marne Cedex
(2) INSERM, U729, 75006 Paris
(3) INALCO, CRIM, 75343 Paris Cedex 07
(4) Assistance Publique - Hôpitaux de Paris, STIM/DSI, 75674 Paris Cedex 14
vmalaise@ina.fr, thd@biomath.jussieu.fr, pz@biomath.jussieu.fr

Mots-clefs : Systèmes de questions-réponses, repérage d'énoncés définitoires, patrons lexico-syntaxiques, médecine

Keywords: Question-answering systems, mining definitions, lexico-syntactic patterns, medicine

Résumé Les systèmes de questions-réponses, essentiellement focalisés sur des questions factuelles en domaine ouvert, testent également d'autres tâches, comme le travail en domaine contraint ou la recherche de définitions. Nous nous intéressons ici à la recherche de réponses à des questions « définitoires » portant sur le domaine médical. La recherche de réponses de type définitoire se fait généralement en utilisant deux types de méthodes : celles s'appuyant essentiellement sur le contenu du corpus cible, et celles faisant appel à des connaissances externes. Nous avons choisi de nous limiter au premier de ces deux types de méthodes. Nous présentons une expérience dans laquelle nous réutilisons des patrons de repérage d'énoncés définitoires, conçus pour une autre tâche, pour localiser les réponses potentielles aux questions posées. Nous avons intégré ces patrons dans une chaîne de traitement que nous évaluons sur les questions définitoires et le corpus médical du projet EQueR sur l'évaluation de systèmes de questions-réponses. Cette évaluation montre que, si le rappel reste à améliorer, la « précision » des réponses obtenue (mesurée par la moyenne des inverses de rangs) est honorable. Nous discutons ces résultats et proposons des pistes d'amélioration.

Abstract Question-answering systems mostly focus on open-domain, factoid questions, but also test other tasks such as restricted-domain and « definitional » questions. We address here the search for definitional questions in the medical domain. Searching for answers to definitional questions generally resorts to two kinds of methods : those which mostly rely on the contents of the target corpus, and those which call on external resources. We have chosen to limit ourselves to the first kind. We present an experiment in which we reuse lexico-syntactic patterns, formerly designed for another task, to locate answers to definitional questions. We have integrated these patterns in a processing chain which we evaluate on the medical definitional questions and corpus of project EQueR (evaluation of French QA systems). This evaluation shows that, while recall still needs to be increased, the « precision » of the obtained answers (as measured through the mean reciprocal rank) is honorable. We discuss these results and propose directions for improvement.

1 Introduction

Les systèmes de questions-réponses ont été jusqu'ici principalement évalués sur des questions de type « factuel », dont la réponse attendue est un fait (Voorhees & Tice, 2000). Des questions recherchant des définitions ont cependant été introduites lors de la campagne d'évaluation TREC-12 QA de 2003 (Voorhees, 2003). Sur les 500 questions de TREC-12, 50 portaient sur des définitions : 30 concernaient un personnage (*« Who is Andrea Bocceli ? »*), 10 une organisation (*« What is ETA in Spain ? »*), et 10 d'autres « choses » (*« What is feng shui ? »*) [1].

La campagne EQueR[2] pour l'évaluation de systèmes de questions-réponses en français a également inclus des questions « définitoires ». Une particularité de cette campagne est d'avoir mis en place, à côté d'une traditionnelle tâche de questions-réponses en domaine ouvert, une tâche de questions-réponses à domaine restreint (corpus et questions de domaine médical). Dans la tâche médicale, sur 200 questions, 70 (soit plus d'un tiers) portaient sur des définitions, dont aucune ne concernait un personnage ou une organisation.

Cet article porte sur cette recherche de réponses à des questions « définitoires » médicales[3]. Dans les travaux antérieurs, deux types principaux de méthodes sont employées pour rechercher ce type de réponse : des méthodes endogènes, par application sur le corpus de patrons d'énoncés définitoires, et des méthodes exogènes, qui projettent sur ce corpus des définitions obtenues dans des ressources dictionnairiques externes ((Hildebrandt *et al.*, 2004), par exemple, combinent les deux). Nous avions à notre disposition les définitions d'une partie des termes du thesaurus MeSH, rédigées par l'équipe CISMeF du CHU de Rouen. Pour les questions portant sur des termes de ce thesaurus, la projection en corpus de ces définitions aurait pu être pertinente. Cependant, les questions ayant été préparées par l'équipe CISMeF, l'usage de ces définitions constituerait un biais[4]. Nous nous sommes donc focalisés sur une méthode endogène, et avons cherché à réutiliser des travaux réalisés sur le repérage d'énoncés définitoires à des fins de construction d'ontologie (Malaisé *et al.*, 2004). C'est cette méthode qui fait l'objet de cet article.

Après une revue de travaux existants (section 2), nous présentons le corpus médical EQueR (section 3), la méthode que nous avons mise en place (section 4) et ses résultats (section 5). Nous discutons ces résultats (section 6) et concluons (section 7).

2 Travaux antérieurs

Les travaux réalisés autour de la tâche TREC-QA de 2003 constituent une source naturelle de bibliographie. Les réponses aux questions définitoires ont été considérées comme un ensemble de « pépites » d'information (Voorhees, 2003) correspondant à des éléments de définition à retrouver : éléments « vitaux », « non-vitaux » ou non pertinents. Les systèmes doivent ramener un maximum d'éléments vitaux et un minimum d'éléments non pertinents. Nous considérons

[1] La notion de « définition » d'un personnage peut sembler étrange, mais c'est le typage qui a été choisi dans cette campagne d'évaluation.

[2] EQueR fait partie du projet Technolangue EVALDA coordonné par ELDA ; contacts : Christelle Ayache (ELDA), Brigitte Grau (LIMSI) (`http://www.technolangue.net/article61.html`). La tâche médicale a été gérée par Magaly Douyère (CISMeF). L'évaluation a été menée en juillet 2004.

[3] La méthode présentée ici n'a pas été mise en place, faute de temps, dans le prototype que nous avons présenté à EQueR. Les connaissances utilisées ont néanmoins été préparées indépendamment des questions EQueR.

[4] Un examen *a posteriori* montre effectivement que 15 des réponses aux questions définitoires sont constituées par ces définitions, qui figurent dans certains documents du corpus.

également, suivant (Meyer, 2001), qu'un énoncé est « définitoire » s'il contient au moins un élément susceptible de servir de base à la construction d'une définition lexicographique. Cet élément peut être, par exemple, l'hyperonyme du terme ou une caractéristique qui permet de le distinguer d'autres termes proches dans le domaine. Idéalement, les énoncés que nous recherchons combinent ces deux types d'éléments. Parmi les participants à TREC 2003, (Hildebrandt *et al.*, 2004) s'appuient sur plusieurs sources de connaissances pour composer une réponse à une question définitoire. La première est obtenue en appliquant sur le corpus cible (AQUAINT) onze patrons recherchant des « pépites » de définition. La deuxième utilise le dictionnaire Merriam-Webster en ligne. La troisième consiste, en dernier recours, à collecter toutes les phrases du corpus contenant le terme cible. L'évaluation des réponses aux questions définitoires privilégiait largement le rappel par rapport à la précision. De ce fait, un système reposant essentiellement sur de simples techniques de recherche d'information, soumis par BBN après l'évaluation (Xu *et al.*, 2003), a obtenu de meilleurs résultats que les systèmes des participants. L'évaluation des définitions dans EQueR n'a pas mis en place une telle prime au rappel. Il était donc raisonnable de privilégier une approche plus précise.

Indépendamment des systèmes de questions-réponses, (Klavans & Muresan, 2001) se sont intéressées au repérage d'énoncés définitoires dans le domaine médical. Elles indiquent que 75 % des énoncés définitoires peuvent être retrouvés à l'aide des patrons qu'elles ont mis au point, et que ce rappel peut être augmenté par *bootstrapping*. Les différents travaux de recherche de définitions, que ce soit au moyen de patrons lexico-syntaxiques (Rebeyrolle, 2000), d'exploration contextuelle (Cartier, 1997) ou de règles (Klavans & Muresan, 2001), se basent schématiquement sur des indices ou *marqueurs* de la définition, associés à des contraintes concernant le voisinage lexical et/ou syntaxique du marqueur. Les contextes de ces marqueurs doivent vérifier en corpus l'ensemble des contraintes définies pour que l'énoncé correspondant soit considéré comme de type définitoire. La méthode retenue par (Malaisé *et al.*, 2004) reprend cette approche fondée sur des patrons lexico-syntaxiques ancrés sur des marqueurs.

3 Le corpus médical EQueR

3.1 Les documents

Le corpus médical EQueR est un sous-ensemble des documents indexés par le Catalogue et Index des Sites Médicaux Francophones (CISMeF, `http://www.chu-rouen.fr/cismef/`) : ceux de neuf « sites éditeurs »[5], auxquels viennent se joindre les documents référencés un lien plus loin sur le même site. L'ensemble comporte 5621 documents originellement au format HTML ou PDF (convertis en texte brut), pour un total d'environ 19 millions de mots.

3.2 Les questions définitoires

Les questions d'EQueR, comme celles de TREC, étaient typées au préalable : l'identifiant de la question indiquait s'il s'agissait d'une question factuelle, booléenne, définitoire ou à réponse

[5]Fédération Nationale des Centres de Lutte Contre le Cancer ; La Documentation Française ; Agence Française de Sécurité Sanitaire des Produits de Santé ; Agence Nationale d'Accréditation et d'Évaluation en Santé ; Orphanet, serveur d'informations sur les maladies rares ; site officiel du Sénat ; le CHU de Rouen ; Université de Rouen, restreinte à sa branche médicale ; site bilingue Santé Canada (ministère fédéral de la santé).

sous forme de liste. Parmi les 70 questions définitoires médicales d'EQueR, cinq portaient sur des acronymes : *« Comment l'IPS peut-il être défini ? »*, etc. Dans le système que nous avons présenté à EQueR (système STIM-LIPN, voir (Delbecque *et al.*, 2005)), les acronymes étaient repérés par le classique patron *« expression (ACRONYME) »* ou sa variante *« ACRONYME (expression) »*. Ces patrons se sont déclenchés par exemple sur le passage *« L'index de pression systolique (IPS) [...] »*. Ils ont été appliqués sur l'ensemble du corpus et les résultats stockés dans une base de données, avant de recevoir les questions. Les questions repérées comme portant sur un acronyme ont été envoyées sur un traitement spécifique qui accède à cette base.

Les 65 questions restantes portent sur des termes à définir : *« Quelle est la définition de la désinfection ? »*, *« Qu'est-ce que le syndrome du décalage horaire ? »*. Une chaîne spécifique a été conçue pour traiter ces questions. C'est sur ces traitements que nous nous concentrons ici.

4 Recherche d'énoncés définitoires pour un terme spécifique

La méthode que nous avons mise en œuvre peut se décrire en trois grandes parties :
- comme pour les « entités nommées » des questions factuelles (Delbecque *et al.*, 2005), nous cherchons à repérer et indexer au préalable tous les énoncés définitoires ;
- une question définitoire étant donnée, il faut en extraire le terme dont on cherche la définition ;
- il faut enfin sélectionner et classer les énoncés définitoires préindexés concernant ce terme.

4.1 Repérage d'énoncés définitoires en corpus

Nous nous sommes appuyés sur des travaux antérieurs, tant théoriques ((Fuchs, 1994), par exemple) qu'appliqués à la recherche en corpus (Rebeyrolle, 2000), et de nos propres corpus de test pour compiler une liste de marqueurs d'énoncés à intérêt définitoire. Nous avons mis au point des patrons lexico-syntaxiques à partir de ces marqueurs et les avons testés lors d'expérimentations antérieures visant à repérer des relations sémantiques entre termes (Malaisé *et al.*, 2004). Ces patrons sont appliqués sur un corpus préalablement analysé par Cordial Analyseur [6] : Cordial segmente en phrases, lemmatise, étiquette les mots et indique leurs relations syntaxiques. Nos patrons portent sur les lemmes et les catégories des mots.

Pour le présent travail, nous n'avons conservé que les patrons qui avaient donné les meilleurs résultats dans ces travaux antérieurs (voir le tableau 1), à savoir ceux modélisés autour de :
- verbes métalinguistiques : *« appeler »*, *« nommer »*, *« référer »*, *« dénommer »*, *« désigner »*, *« dénoter »*, *« signifier »*, *« définir »* ;
- noms métalinguistiques associés à un ensemble de verbes supports : *« nom »*, *« terme »*, *« mot »*, *« expression »*, *« vocable »*, *« appellation »*, *« désignation »*, *« dénomination »*, *« concept »*, *« notion »*, *« acception »*, associés à *« porter »*, *« appliquer »*, *« employer »*, *« réserver »*, *« utiliser »*, *« donner »*, *« renvoyer »*, *« référer »*, *« être »* ;
- indices de reformulation et d'hyperonymie : *« vouloir dire »*, *« entendre par »*, *« à savoir »*, *« sorte de »*, *« est un »*, *« par exemple »* ;
- la parenthèse.

Nos patrons lexico-syntaxiques permettent d'extraire des « définitions candidates » et deux groupes syntaxiques dans ces énoncés, qui sont susceptibles de contenir l'élément défini : le *definiendum*. Ces groupes syntaxiques sont extraits selon deux modalités :

[6] http://www.synapse-fr.com/

Type de marqueur	*Patrons lexico-syntaxiques*
Verbes métalinguistiques	*VerbeMeta* NonPrécédéDe *« se »* ET NonSuiviDe [pas/.]
Noms métalinguistiques	*NomMeta* {0-6}MOTS *VerbeSupport* ; *VerbeSupport* {0-6}MOTS *NomMeta* ; *« être »* {0-1}MOT [le/ce] *NomMeta* ; sous [le/ce] NomMeta
Indices de reformulation ou d'hyperonymie	*« vouloir »* {0-n}MOTS *« dire »* NonSuiviDe *« que »* ; *« entendre »* {0-6}MOTS *« par »* ; *« par »* {0-6}MOTS *« entendre »* ; *« à savoir »* PrécédéDe *Ponctuation* ; *« à savoir »* NonPrécédéDe *Ponctuation* ; *Determinant « sorte de »* ; *« être »* {0-3}MOTS [le/la/les/un/une/des] *Nom* ; *« par exemple »* PrécédéDe 1MOT ET SuiviDe 1MOT
Parenthèse	*NomCommun* (1MOT) ; *NomCommun* (*« ou » Nom* ; *NomCommun* (*« qui »* [est/se] ; *Verbe* (*VerbeInfinitif*

TAB. 1 – Patrons lexico-syntaxiques pour le repérage d'énoncés définitoires.

– si le marqueur est un verbe, nous extrayons son sujet et son objet direct dans l'énoncé, s'il en contient, et sinon, nous extrayons respectivement :
  – le groupe syntaxique ayant la même fonction que le nom précédant le marqueur ;
  – le groupe syntaxique ayant la fonction du premier mot plein suivant le marqueur ;
– si le marqueur n'est pas un verbe, nous extrayons les groupes syntaxiques précédant et suivant le marqueur de la manière décrite ci-dessus.

Dans les cas où deux marqueurs doivent être présents dans la phrase (*définir* associé à *comme*,...), nous ne spécifions pas la position relative des deux marqueurs dans la phrase, et extrayons les sujets et objets ou les contextes droits et gauches du verbe. Ce procédé rudimentaire donne toutefois des résultats de l'ordre de 55 % de précision (Malaisé *et al.*, 2004) et permet de factoriser les patrons. La qualité de cette extraction dépend également de la qualité de la segmentation initiale des phrases et de leur analyse. Par exemple, dans l'énoncé (5598-1) (qui comporte un titre mal segmenté) *« RECOMMANDATIONS ET RÉFÉRENCES Les patients dyslipidémiques sont définis par une augmentation des taux1 sériques du cholestérol et ou des triglycérides. [...] »*, le marqueur *« défini »* a permis de repérer les deux groupes *« RÉFÉRENCES Les patients dyslipidémiques sont »* et *« par une augmentation des taux1 sériques du cholestérol et ou des triglycérides. Ils ont de ce fait »*.

Selon la phrase, l'un des deux groupes peut également être vide : pour l'énoncé (5601-6) *« [...] La SFHH recommande le terme de pré-désinfection pour l' étape préalable à la désinfection ou à la stérilisation Opération utilisant des détergents contenant au moins un principe actif reconnu pour ses propriétés bactéricides , fongicides , sporicides ou virucides (SFHH) [...] »*, le système n'a extrait qu'un groupe « droit » : *« de pré-désinfection [...] Opération utilisant des détergents contenant au moins un principe actif reconnu pour ses propriétés bactéricides »*.

Les énoncés définitoires candidats ainsi trouvés, avec le ou les deux groupes extraits correspondant aux positions hypothétiques du definiendum, sont notés dans une table, qui est indexée par les mots (leurs formes graphiques et lemmes) présents dans chacun des groupes. C'est cet index des definienda hypothétiques qui servira lors de la recherche de réponses.

4.2 Analyse de la question

Les questions posées sont traitées en fonction de leur type. Le traitement général appliqué par défaut a été réalisé par l'équipe du LIPN (Thierry Poibeau), et s'appuie sur une série de transducteurs mis en œuvre avec Unitex (`http://www-igm.univ-mlv.fr/~unitex/`). Pour les questions de type « définition » (par exemple, *« Quelle est la définition de "chimiothéra-*

pie" ? »), un traitement spécifique vise à extraire de la question le terme dont on cherche la définition[7]. Il procède par élimination, en supprimant de la question tous les mots considérés comme « vides » : principalement la copule, les déterminants, les particules interrogatives et les verbes de parole. Les mots des deux premières catégories (*« est »*, *« la »*, *« de »*) sont puisés dans plusieurs listes collectées dans des travaux antérieurs, qui ont été augmentées par une liste de particules interrogatives (*« quel »*, *« quelle »*, *« quelles »*, *« quels »*, *« quoi »*, *« comment »*,...) et d'autres mots (*« façon »*, etc.). Les mots qui désignent une définition (*« définition »*, *« définir »*, *« appeler »*...) ont été pris dans les listes des principaux marqueurs employés dans les patrons : ceux qui se sont appliqués sur le corpus. Les listes employées contiennent directement les formes fléchies des mots. Ainsi, pour la question ci-dessus, le terme restant est *« chimiothérapie »*.

4.3 Recherche de définitions en réponse à une question

Il s'agit ici de proposer des définitions pour le terme extrait d'une question. On va pour cela le chercher parmi les definienda hypothétiques relevés précédemment (section 4.1). Ils sont classés en fonction du nombre de mots du terme de la question qu'ils contiennnent. Dans les expériences présentées ici, nous avons imposé que tous les mots du terme recherché soient présents. Les réponses ont alors été classées dans l'ordre des documents du corpus. Dans l'évaluation EQueR, un système peut renvoyer jusqu'à cinq réponses ordonnées. Nous conservons donc les cinq premiers candidats. Une réponse se compose d'un passage (l'énoncé, tronqué à 250 caractères si nécessaire) et d'une réponse courte (la définition). Nous proposons comme définition celui des deux groupes qui ne contient pas le terme de la question (ou celui qui le contient si l'autre est vide). Si l'énoncé est tronqué, il est centré sur ce groupe.

Ainsi, à la question *« Qu'est-ce qu'une aniridie ? »* , l'énoncé (5590-2) donnera comme réponse « courte » *« comme l' absence totale d' iris »*, et comme passage *« Aniridie sporadique TITLE L' aniridie est une absence clinique d' iris [...] L' aniridie se définit comme l' absence totale d' iris . »*. Pour la question *« Quelle est la définition de l'asthme ? »*, l'énoncé (5586-2) donne la réponse courte erronée *« Définition Le clinicien »*, mais le passage correct *« National des Prescriptions et Consommations des Médicaments [...] Définition Le clinicien définit l' asthme comme un accès de dyspnée , de toux et de sifflement paroxystique , dont l' expression peut »*[8].

4.4 Évaluation

L'objectif de ce travail était de répondre aux questions de type définitoire posées dans la campagne EQueR d'évaluation de systèmes de questions-réponses, tâche médicale. Pour ne pas biaiser le système, la mise au point du module de recherche de réponse à des questions définitoires s'est faite indépendamment des questions EQueR. Nous avons utilisé pour cela un ensemble de 735 termes médicaux tirés du thesaurus MeSH pour lesquels nous disposions de définitions en français[9]. Ces termes nous ont servi à tester notre chaîne de traitement.

L'évaluation proprement dite s'est faite sur les questions de la campagne EQueR. Ce module ayant été terminé après la fin de l'évaluation officielle, nous avons nous-mêmes calculé les

[7] Ce terme sera alors stocké sous la forme graphique selon laquelle il apparaît dans la question.

[8] Ce passage n'a hélas pas été classé parmi les cinq premiers par notre système.

[9] Ces définitions sont celles mises au point par l'équipe CISMeF, dans une version du printemps 2004.

scores, en reprenant les principes utilisés dans EQueR. Cette évaluation repose sur un jugement humain de pertinence des réponses courtes et passages produits par le système, avec une possibilité de variation non négligeable[10]. Nous l'avons effectuée nous-mêmes, les jugements individuels et les chiffres synthétiques qui en découlent ne sont donc pas directement comparables à ceux de l'évaluation EQueR officielle. Nous estimons cependant qu'ils devraient rester dans le même esprit.

Pour EQuER, une réponse courte est jugée correcte si elle est juste et précise ; inexacte si elle n'est pas assez précise, incorrecte si elle n'est pas juste, et non justifiée si elle est correcte mais que le document indiqué ne la corrobore pas. Un passage est correct s'il contient au moins une partie d'une réponse juste (le reste étant dans le document) ou incorrect s'il ne contient pas assez ou pas du tout d'éléments corrects (non complétés par le document).

Nous avons noté nos réponses (courtes ou passages) sur trois niveaux : sûrement correcte, possiblement correcte, incorrecte. Le niveau intermédiaire vise à prendre en compte l'écart qui peut exister entre notre jugement et celui qu'auraient rendu les évaluateurs d'EQueR. Nous donnons ainsi deux séries de résultats, un « score strict » qui considère les « possiblement correctes » comme incorrectes, et un « score laxiste » qui les compte comme correctes.

La note assignée à une question est l'inverse du rang de la première bonne réponse renvoyée par le système. La note globale est la moyenne de ces inverses du rang.

5 Résultats

Le repérage des énoncés définitoires a été appliqué à l'ensemble du corpus médical EQueR[11]. 17 792 énoncés définitoires potentiels ont été repérés. Les patrons les plus productifs sont ceux centrés sur les marqueurs : *« définir »* (4950), *« exemple »* (4347), *parenthèse* (1851), *« appeler »* (1336), *« être »* (1272). Les 65 questions définitoires d'EQueR hors acronymes ont été analysées, et les termes extraits ont été recherchés dans les groupes représentant les definienda hypothétiques[12]. Notre système a proposé des réponses à 22 des 65 questions (tableau 2). Selon le jugement porté, entre 5 et 10 des réponses courtes et entre 9 et 16 des passages étaient corrects. Le rang des bonnes réponses trouvées varie du premier au cinquième (et dernier) rang. Dans la version stricte, le bon passage est en moyenne trouvé au troisième rang ($MRR = 0,33$), et la moitié des 22 questions n'obtient aucune bonne réponse. Dans la version laxiste, le bon passage est en moyenne au second rang ($MRR = 0,53$), et seules 6 de ces questions n'obtiennent pas de bonne réponse.

6 Discussion

Pour les raisons expliquées plus haut, il est difficile de comparer nos résultats à ceux obtenus pour les définitions dans les compétitions TREC-QA (pour 2003, f-mesure médiane à 0,192,

[10] Dans le cadre de TREC-QA, la marge d'erreur liée à des différences d'appréciation sur la pertinence de certaines réponses semble globalement ne pas modifier le classement des systèmes.

[11] Il a été programmé en XSLT, et son application sur l'ensemble du corpus EQueR médical prend 25 minutes sur un biprocesseur Xeon 2,4 GHz avec 1,2 Go de mémoire.

[12] Les programmes pour l'analyse des questions et la recherche des réponses aux 65 questions définitoires EQueR, implémentés en Perl, mettent 10 secondes pour traiter l'ensemble des questions sur la même machine.

		Score strict		Score laxiste	
n° EQueR	*Question*	C	P	C	P
MD28	*Quelle est la définition de chimiothérapie ?*	0	0	0	0,25
MD52	*Qu'est-ce qu'une aniridie ?*	0,5	0,5	0,5	0,5
MD56	*Qu'est ce qu'un mésothéliome ?*	1	1	1	1
MD57	*Qu'est-ce qu'une anomalie congénitale ?*	0	0	0	0
MD61	*Qu'est-ce qu'une anorexie ?*	0,33	0,33	0,33	0,33
MD62	*Quelle est la définition de la désinfection ?*	0	0,33	0	0,33
MD64	*Qu'est-ce que la radiothérapie ?*	0	0	0	0,5
MD66	*Quelle est la définition de l'asthme ?*	0	0	0	0
MD69	*Qu'est-ce que le séquençage ?*	0	0	0	0
MD70	*Qu'est-ce que le syndrome du décalage horaire ?*	1	1	1	1
MD71	*Qu'est-ce qu'un trouble dépressif ?*	0	0	1	1
MD75	*Qu'est-ce que l'Index de Pression Systolique ?*	0,2	1	0,2	1
MD77	*Qu'est-ce que la schizophrénie ?*	0	0	1	1
MD79	*Qu'est-ce qu'une amblyopie ?*	0	0	0,33	0,33
MD82	*Qu'est-ce qu'un scanner ?*	0	0	0	0
MD87	*Quelle est la définition de la génomique ?*	0	0	0	0
MD89	*Quelle est la définition du neuroblastome ?*	0	1	0	1
MD91	*Qu'est-ce que la thérapie génique ?*	0	1	1	1
MD94	*Qu'est-ce que la virémie ?*	0	1	0	1
MD98	*Qu'est-ce qu'une sialographie ?*	0	0	0	0
MRD153	*Que signifie le terme chimiothérapie ?*	0	0	0	0,5
MRD189	*Comment peut-on définir un trouble dépressif ?*	0	0	1	1
Moyenne des inverses de rang (MRR)		0,138	0,326	0,335	0,534
Nombre de réponses trouvées		5	9	10	16

TAB. 2 – Les 22 questions auxquelles le système a répondu, et le score des réponses fournies. C = réponse courte, P = passage. Les scores sont des inverses de rangs (un score de 0,33 correspond à une réponse trouvée au rang 3). Le score « laxiste » accepte des réponses moins complètes.

meilleure à 0,555) : ils comportaient de nombreuses questions sur des personnes ou des organisations et n'étaient pas calculés de la même façon (« pépites » de connaissance). La comparaison aux résultats d'EQueR est elle aussi malaisée, du fait de la part de jugement humain impliquée dans l'évaluation des réponses individuelles[13].

On peut néanmoins noter qu'un nombre relativement faible (un tiers) de questions obtiennent une réponse, et qu'une partie seulement de ces questions reçoivent une réponse correcte parmi les cinq premières proposées par le système (entre 40 et 73 % pour les passages). D'après les données globales dont nous disposons, nous avons calculé que dans EQueR, 39 questions définitoires ont obtenu au moins un passage correct de la part d'un participant, dont 35 avec la réponse courte correcte. Il est cependant important d'analyser l'origine de ces pertes de façon à augmenter le rappel du système actuel. Ce système étant composé de modules enchaînés, chaque module est susceptible de participer à cette perte d'information.

Le repérage des énoncés définitoires et des definienda hypothétiques est une première origine : dans les travaux précédents, leur rappel a été évalué à environ 50 %, et leur précision de 10 à 69 % suivant la complexité syntaxique des énoncés. De plus les patrons n'ont pas été adaptés au domaine médical, ce qui entraîne à la fois du bruit (marqueurs polysémiques en médecine) et du silence (patrons de définitions de type « médical » non modélisées). L'analyse des questions,

[13]Cette part pourra être réduite si tous les passages corrects sont relevés pour chaque question, ce qui a été fait au moins partiellement par les organisateurs.

réalisée ici de façon simpliste, a laissé passer plusieurs mots « vides »[14] non prévus ou dont la forme employée n'était pas prévue (*« dire »*, *« sigle »*, *« définie »*, *« possible »*, ...). Les termes ainsi obtenus (MD15 *« possible ostéosynthèse »*, MD16 *« dire noyade sublétale »*, ...) ne peuvent être trouvés en position de definiendum. Huit termes ont ainsi été mal identifiés. On peut espérer qu'une analyse employant par exemple des transducteurs et une lemmatisation comme pour les autres questions, donnerait des résultats plus précis. Enfin, dans les cas où plus de cinq réponses ont été trouvées, un meilleur classement des réponses pourrait augmenter le rappel si la bonne réponse ne fait partie des premiers passages renvoyés (une dizaine de cas). Par exemple, la méthode consistant à collecter toutes les phrases où apparaît le terme à définir, puis les mots les plus fréquents dans ces phrases, et à privilégier les phrases qui contiennent le plus grand nombre de ces mots (Xu *et al.*, 2003) semble une piste intéressante.

Il semblerait pertinent de comparer ce que donnerait l'application brute de cette méthode à ce que nous obtenons en filtrant à l'aide du repérage d'énoncés définitoires. On peut espérer que cette focalisation sur les énoncés définitoires réduit le bruit (même si elle ne le supprime pas, en particulier du fait que les patrons employés sont eux-mêmes sources de bruit) : on a vu qu'il n'était pas excessif dans les réponses données par notre système (MRR entre 0,33 et 0,53).

Ensuite, on peut supposer que l'usage de connaissances extérieures (dictionnaires, terminologies avec définitions, locaux ou interrogeables en ligne), que nous nous sommes interdit ici, devrait aider à localiser des définitions candidates non contraintes par les patrons dont nous disposons. Par exemple, certaines définitions sont données en plusieurs phrases, voire en faisant implicitement référence au titre du document, avec une présentation du type *« <TITLE>Noyade sublétale</TITLE> Définition [MeSH Scope Note ; traduction CISMeF] : immersion non fatale dans l'eau. Le sujet peut être réanimé. »* ou *« <TITLE>Adénite</TITLE> Définition [MeSH Scope Note ; traduction CISMeF] : inflammation des ganglions lymphatiques. »*. Ce type de présentation ne semble pas possible à détecter par des patrons génériques. En revanche, le fait de savoir que l'*« Adénite »* est une *« Inflammation des ganglions lymphatiques »* permet d'associer cet énoncé au terme correspondant. Cette information peut se trouver dans un dictionnaire médical, comme celui disponible en ligne à l'URL `http://www.AtMedica.com`.

Pour terminer, soulignons qu'une source importante de bruit est constituée par le corpus lui-même, obtenu par conversion en texte de documents HTML et PDF, conversion dont on sait qu'elle reste difficile à réaliser proprement de façon automatique. Sur un corpus de taille importante, une révision manuelle complète reste hors de portée, et les passages bruités sont nombreux. Ces passages sont à l'origine d'énoncés mal segmentés ou incohérents, que nos programmes n'ont pas su prendre en compte correctement. Ce point constitue aussi probablement une différence importante par rapport au corpus AQUAINT utilisé dans TREC.

7 Conclusion

Le système assemblé ici permet de proposer des réponses à des questions de type définitoire à partir du corpus médical EQueR. Si son rappel doit être amélioré, la précision des réponses proposées est honorable. Notons qu'il a trouvé quelques réponses correctes qu'aucun participant à EQueR n'a trouvées, par exemple celle de *« Quelle est la définition de la désinfection ? »*.

Nous avons souligné qu'au-delà des améliorations à apporter à la préparation du corpus et aux

[14] Particules interrogatives, verbes de parole, etc. (cf. section 4.2).

méthodes présentées elles-mêmes, essentiellement endogènes, un recours à des connaissances extérieures (définitions existantes) devrait aider à renforcer la détection des définitions correctes en corpus. Ce sera le thème de nos prochains travaux.

Remerciements

Nous remercions Magaly Douyère pour ses conseils sur l'évaluation des définitions trouvées.

Références

CARTIER E. (1997). La définition dans les textes scientifiques et techniques : présentation d'un outil d'extraction automatique de relations définitoires. In *2e Rencontres Terminologie et Intelligence Artificielle*, p. 127–140, Toulouse : ERSS.

DELBECQUE T., JACQUEMART P. & ZWEIGENBAUM P. (2005). Utilisation du réseau sémantique de l'UMLS pour la définition de types d'entités nommées médicales dans un système de questions-réponses : impact de la source des documents explorés. In *CORIA (COnférence en Recherche d'Informations et Applications)*, p. 101–115, Grenoble : CLIPS.

FUCHS C. (1994). *Paraphrase et énonciation*. Paris, Ophrys.

HILDEBRANDT W., KATZ B. & LIN J. (2004). Answering definition questions using multiple knowledge sources. In S. DUMAIS, D. MARCU & S. ROUKOS, Eds., *Actes HLT-NAACL*, p. 49–56, Boston, Massachusetts, USA : ACL.

KLAVANS J. & MURESAN S. (2001). Evaluation of the DEFINDER system for fully automatic glossary construction. *Journal of the American Medical Informatics Association*, **8**(suppl), 324–328.

MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. In P. BLACHE, Ed., *Actes de TALN 2004 (Traitement automatique des langues naturelles)*, p. 269–278, Fès, Maroc : ATALA LPL.

MEYER I. (2001). Extracting knowledge-rich contexts for terminography. In D. BOURIGAULT, C. JACQUEMIN & M.-C. L'HOMME, Eds., *Recent Advances in Computational Terminology*, p. 279–302. Amsterdam : John Benjamins.

REBEYROLLE J. (2000). *Forme et fonction de la définition en discours*. Thèse de doctorat, Université de Toulouse II - Le Mirail.

VOORHEES E. M. (2003). Overview of the TREC 2003 question answering track. In E. M. VOORHEES & L. P. BUCKLAND, Eds., *Actes Twelfth Text Retrieval Conference (TREC 2003)*, p. 54–68, Washington DC : NIST.

VOORHEES E. M. & TICE D. M. (2000). The TREC-8 question answering track evaluation. In E. M. VOORHEES & D. K. HARMAN, Eds., *Eighth Text REtrieval Conference (TREC-8)*, NIST Special Publication 500-246, p. 83–105.

XU J., LICUANAN A. & WEISCHEDEL R. (2003). TREC2003 QA at BBN : Answering definitional questions. In E. M. VOORHEES & L. P. BUCKLAND, Eds., *Actes Twelfth Text Retrieval Conference (TREC 2003)*, p. 98–106, Washington DC : NIST.