Le système STIM/LIPN à EQueR 2004, tâche médicale

Thierry Delbecque[1, 2], Pierre Zweigenbaum[1, 2, 3], Jean-François Berroyer[4, 5], Thierry Poibeau[4, 5]
(1) INSERM, U729, 75006 Paris
(2) INALCO, CRIM, 75343 Paris Cedex 07
(3) Assistance Publique - Hôpitaux de Paris, STIM/DSI, 75674 Paris Cedex 14
(4) CNRS, UMR 7030, 93430 Villetaneuse
(5) Université Paris 13, LIPN, 93430 Villetaneuse

Mots-clefs : Systèmes de questions-réponses, médecine, projet EQueR

Keywords: Question-answering systems, medicine, EQueR project

Résumé Nous présentons les principes du système de questions-réponses STIM/LIPN, fruit d'une collaboration entre le STIM (AP-HP & INSERM U729) et le LIPN (CNRS & Université Paris 13), qui a pris part à la tâche médicale de l'évaluation EQueR 2004 (Ayache, 2005).

Abstract We present the principles of the question-answering system STIM/LIPN, the result of joint work by STIM (AP-HP & INSERM U729) and LIPN (CNRS & Université Paris 13), which participated in the medical track of the EQueR 2004 evaluation (Ayache, 2005).

1 Segmentation et indexation du corpus

1.1 Segmentation

Nous avons fait l'hypothèse que la réponse à une question pouvait être trouvée au sein d'une même phrase. Nous avons donc décomposé le corpus en phrases, plus précisément en unités que nous avons appelées *prédicats*, dont l'organisation comporte (figure 1) (i) une tête verbale ; (ii) un modérateur, pour exprimer les négations, intensité, etc. (iii) un sujet ; (ix) un ou plusieurs compléments. Pour cela, le corpus dans sa totalité à été soumis à TreeTagger[1] ; la détermination et la construction des prédicats se fait sur la base de patrons de parties du discours. Le résultat intermédiaire est un document XML.

Parallèlement à cette analyse, une détection des définitions d'abréviations est effectuée ; nous disposons ainsi d'un dictionnaires d'acronymes, avec pour chaque entrée les définitions possibles, les fréquences et localisations des définitions dans le corpus.

[1] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`.

```
<form id="57">
        <predicat> être </predicat>
        <moderator> souvent </moderator>
        <subject from="1339" to="1343">
                <SUI>1:S0226498</SUI> <CUI>C0000726</CUI> <TUI>T029</TUI>
                L examen de l abdomen </subject>
        <arguments> <item from="1346" to="1361">
                        <SUI>1:S0234187</SUI> <CUI>C0022828</CUI> <TUI>T007</TUI>
                        normal ( formes basses ) ou montre
                        parfois une sensibilité de la fosse iliaque gauche . </item> </arguments>
</form>
```

FIG. 1 – Exemple de prédicat

1.2 Repérage d'entités nommées générales et médicales

Entités nommées médicales Nous avons choisi, comme source d'EN médicales, d'utiliser les types sémantiques et les relations sémantiques fournies par l'UMLS (`http://www.nlm.nih.gov/research/umls/`). Ainsi, nous espérions faire ressortir aussi bien les prédicats portant par exemple sur une pathologie (type sémantique *Pathologic Function* :T046), que ceux pouvant exprimer un traitement en action (relation sémantique *treats* :T154)[2].

La projection des types sémantiques sur les prédicats nécessite la détermination des termes simples et des termes composés, dans les sujets et compléments des prédicats ; puis la recherche de ces termes, ou de termes hyperonymes, au sein de la partie francophone de l'UMLS[3]. On remonte des termes aux concepts, puis aux types sémantiques correspondants. Les prédicats sont ainsi enrichis avec les types sémantiques dont la présence est soupçonnée, ainsi qu'on le voit sur un extrait (figure 1).

Les autres entités nommées Les autres entités nommées sont reconnues grâce à un outil appelé TagEN, développé au LIPN. L'analyse effectuée est très classique : elle se fonde sur un ensemble de données lexicales définies dans des dictionnaires et sur des automates permettant de regrouper les unités pertinentes en fonction des informations contenues dans les dictionnaires. Les types d'entités classiquement définis ont été repris — noms de personnes, de lieux, dates, durées —, et d'autres ont été affinés — pour les entités chiffrées, on distingue des dosages, des posologies, etc. L'analyse fait alors appel aux algorithmes classiques sur les automates, tels qu'ils sont développés au sein de la boîte à outils Unitex (`http://www-igm.univ-mlv.fr/~unitex/`). En cas d'ambiguïté dans l'analyse, seule la séquence la plus longue est retenue.

1.3 Indexation

Nous gérons directement l'ensemble des prédicats dans une base de données MySQL ; l'indexation remplace le recours à un moteur de recherche. Quatre tables d'index sont créées : sur les têtes verbales (lemmatisées), sur les mots pleins (formes originales et lemmatisées), sur les EN générales ; sur les types et relations sémantiques projetés depuis l'UMLS. Un attribut du schéma relationnel est consacré à la position dans le corpus du prédicat, pour pouvoir extraire a posteriori le passage correspondant.

[2]Le réseau sémantique de l'UMLS contient 134 types sémantiques et 54 relations sémantiques, hiérarchisées par un lien *is-a*.

[3]Ce qui constitue en soi une limitation : le français ne couvre qu'à peine plus de 2% des concepts de l'UMLS (version 2002-AA).

2 Exécution d'un *run*

2.1 Analyse et transformation de la question

L'analyse d'une question doit permettre de décider de sa catégorie (en particulier, si l'on cherche une définition), du type d'entité nommée s'il y a lieu, et de l'ensemble des mots clés rencontrés (« mots d'ancrage »). Lorsqu'une question concerne la définition d'un acronyme, elle est traitée de manière particulière (voir la section 2.2)[4].

Le typage de la question est effectué en appliquant une série d'automates sur les questions, suivant la même stratégie que pour le repérage des entités nommées. L'analyse de la question produit un objet XML.

Dans le cas où la question ne porte pas sur la définition d'un acronyme, l'objet XML est transformé en requête dans un langage intermédiaire. Lors de cette transformation, la requête peut être étendue, en fonction du type d'entité nommée cherché, en particulier s'il s'agit d'une entité médicale. Par exemple si la question porte sur une pathologie, on adjoindra à la requête « *syndrome* » comme mot clé pertinent ; de telles extensions portent également sur la tête verbale. On associe également à la requête les poids à affecter aux réponses, en fonction de la présence ou non d'indices (EN, mots clés, etc.).

Enfin, une analyse complémentaire de la question y recherche des mots clés du thésaurus MeSH. Ces mots clés sont utilisés plus tard (voir la section 2.2) pour obtenir un indice de confiance supplémentaire dans certains documents.

2.2 Requête et tri des réponses

Lorsque l'analyse de la question a établi qu'il s'agit de trouver la définition d'un acronyme, la procédure de recherche consiste simplement à consulter le dictionnaire d'acronymes (1.1).

Dans le cas général, la requête en langage intermédiaire est traduite en SQL, puis soumise à MySQL ; les résultats sont utilisés pour isoler les fragments de texte adéquats dans le corpus, qui constitueront les résultats finals[5]. Ceux-ci sont alors pondérés (notés) selon différents critères.

L'un de ces critères tient compte du fait que la thématique de la question est fortement représentée dans le document où la réponse a été trouvée. Pour cela, nous nous aidons de l'indexation thématique faite par le portail CISMeF (`http://www.chu-rouen.fr/cismef/`) à l'aide de termes MeSH sur certains documents du corpus EQueR médical. Si la question contient des mots clés du thésaurus MeSH, une requête est envoyée au moteur de recherche du catalogue CISMeF. Les documents de CISMeF indexés par les mots clés MeSH de la question, lorsqu'il y en a, sont ainsi recensés. Si une réponse proposée à la question est trouvée dans l'un de ces documents, elle reçoit un bonus : sa note globale est augmentée d'un point.

Les autres critères portent sur la présence ou non d'entités nommées, de verbes ou de mots clés précis[6] ; la note globale est réévaluée en conséquence, conformément à ce qui est spécifié dans la requête intermédiaire. Un score final est affecté à chaque réponse, ce qui permet de les trier.

[4] Après l'évaluation, nous avons aussi mis au point un traitement spécifique pour les questions « définitoires » (Malaisé *et al.*, 2005).

[5] Sauf dans le cas des réponses oui/non, qui réclament un traitement supplémentaire, présenté dans la section 2.3.

[6] Dans le prédicat, et non dans le passage.

2.3 Décisions finales pour les réponses

La compétition prévoyait trois formats de réponses : longues, courtes, et binaire, ce qui a nécessité un reformatage des réponses.

Le mécanisme de construction d'une réponse courte n'était pas encore finalisé lorsque le système a été présenté. Nous avons donc décidé de ne pas demander l'évaluation des réponses courtes, sauf pour les cas suivants qui étaient prêts :
- les réponses booléennes, qui n'auraient pas de sens sinon ;
- les définitions d'acronymes : questions de type définition où le passage trouvé contenait un sigle (au moins deux lettres majuscules consécutives).

Toutes les autres réponses courtes ont été mises à « NUL ».

Pour les questions booléennes, il faut déterminer, sur la base des passages trouvés, si la réponse est oui ou non. Nous utilisons pour cela une heuristique simple. Le système, tel que présenté à l'évaluation, renvoie pour chaque question booléenne au plus un passage, dans lequel il a trouvé un ou plusieurs « mots d'ancrage » (mots de la question). Si au moins trois mots d'ancrage (de plus de trois lettres) ont été trouvés, la réponse fournie est « oui », et « non » dans le cas inverse.

3 Conclusion

Ce travail a été pour nous l'occasion de proposer l'utilisation de ressources terminologiques médicales dans une optique de QR. L'un des enjeux était l'utilisation de l'UMLS comme source d'entités nommées spécifiques au domaine. Sur 200 questions, le système a proposé 112 passages, dont 16 corrects (3^{e} ex æquo, moyenne globale de l'inverse des rangs de 0,09), et 35 réponses courtes, dont 12 correctes (2^{e} ex æquo, MRR = 0,06). Les réponses correctes obtenues portaient sur des questions booléennes (MRR = 0,33) et définitions (en l'occurrence, des sigles).

Le système est encore très jeune ; tous les composants ont été développés pour l'occasion, et il a été assemblé juste à temps pour l'évaluation, sans phase de test préalable. Parmi ses points faibles, outre sans doute des bogues à éliminer, on trouve certainement la qualité insuffisante de la construction des prédicats, qu'une véritable analyse syntaxique devrait améliorer. D'autre part, seule une infime partie des apports possibles de l'indexation par « concepts » de l'UMLS réalisée ici a été exploitée. Enfin, une méthode simple, de type recherche de passages en texte intégral, pourrait sans doute aider à obtenir des réponses lorsque la méthode présentée ici reste silencieuse. L'analyse *post mortem* des résultats du système par rapport aux réponses attendues va nous permettre d'établir des priorités pour nos efforts futurs.

Références

AYACHE C. (2005). *Campagne EVALDA/EQueR – Évaluation en Question-Réponse, rapport final*. Rapport interne, ELDA, Paris. Disponible à`http://www.technolangue.net/IMG/pdf/rapport_EQUER_1.2.pdf`.

MALAISÉ V., DELBECQUE T. & ZWEIGENBAUM P. (2005). Recherche en corpus de réponses à des questions définitoires. In *Actes Traitement automatique des langues naturelles (Traitement automatique des langues naturelles)*, Dourdan. *À paraître*.