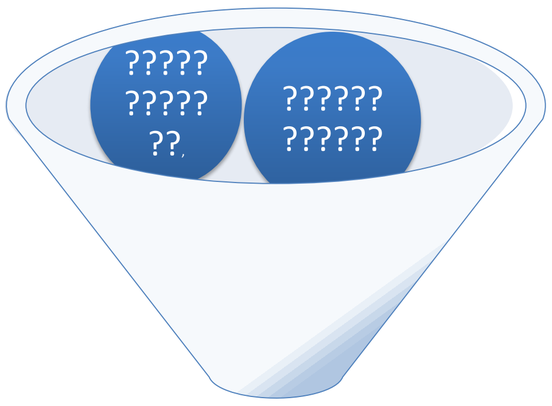
Moteur de question réponse pour les sciences biomédicales

Alexandre Racine, Philippe Cosandey et Romilda Chalard
Master en Sciences de l'information
M8 – Projet de recherche – 2015 / sous la dir. du Prof. Patrick Ruch

Les défis du biomédical

- Déluge de publications
- Manque de temps pour lire les articles
- Caractère urgent des questionnements
- Exigence de garantie de qualité
- Fiabilité
- Actualité



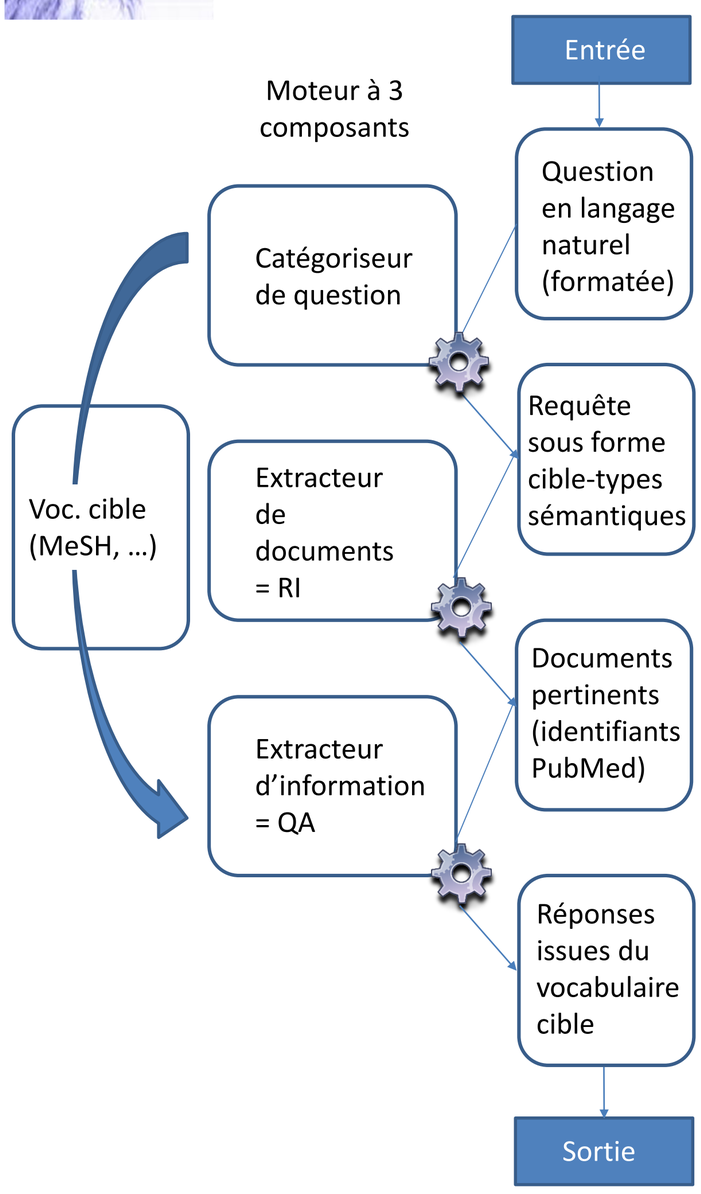
Un moteur de question réponse (QA pour Question Answering) retourne une réponse à une question posée en langage naturel (par opposition à un moteur de recherche d'information (RI), qui retourne des documents susceptibles de contenir la réponse).

Les atouts du biomédical

- Redondance de l'information
- Vocabulaires contrôlés tels que le MeSH (Medical Subject Headings de la National Library of Medicine)
- Centralisation des données (PubMed, ...)
- Financement du fait des enjeux stratégiques
- Support de la bioinformatique

Amélioration et évaluation d'EAGLi à travers le challenge BioASQ

EAGLi (Engine for question-Answering in Genomics Literature) : moteur de QA pour les sciences biomédicales. Basé sur la redondance d'information et les vocabulaires contrôlés.



Exemple d'une des 330 questions factoïdes du challenge BioASQ :
Where is the histone variant CENPA preferentially localized?

Reformulation pour structure EAGLi :
What|Which + cible + verbe + complément (= termes de la recherche)
Which cell component has the histone variant CENPA?

Identification des types sémantiques de la cible parmi les n= ~200 du MeSH :
T026 (cell component) (n= 1'500 termes MeSH)

Soumission des questions reformulées à EAGLi.
Le catégoriseur de question reconnaît la question et formule la requête :
What|Which cell component T026 has the histone variant CENPA?

La RI retourne les identifiants PubMed (PMID) des documents jugés pertinents (= dont le résumé contient les termes attendus) :

52fe52702059c6d71c000078 dummy 25898113 1 999 dummy
52fe52702059c6d71c000078 dummy 21508988 2 998 dummy
52fe52702059c6d71c000078 dummy 24213134 3 997 dummy
52fe52702059c6d71c000078 dummy 20940262 4 996 dummy
52fe52702059c6d71c000078 dummy 20119530 5 995 dummy
52fe52702059c6d71c000078 dummy 12906131 6 994 dummy

Le QA retourne les réponses (= termes MeSH des résumés sélectionnés) :

D002503, Centromere, score 5
D002875, Chromosomes, score 5
D018386, Kinetochores, score 4
D008941, Spindle Appa

48 questions factoïdes reformulées

Tests effectués par runs : booléens + vectoriels

1. Résultats booléens sans comparaison :
Top précision : 0.45
Rappel : 0.64

2. Résultats vectoriels sans comparaison :
Top précision : 0.42
Rappel : 0.69

3. Résultats avec comparaison:
Booléen :
EAGLi : MRR 0.45
Vectoriel :
EAGLi : MRR 0.42
BioASQ 1ère année : **MRR 0.31**
BioASQ 2ème année : **MRR 0.46**

Nettoyage du qrel pour exploiter les réponses relevant du MeSH ou les y associer
Evaluation par méthodologie trec_eval

- Un bon score de rappel en booléen et vectoriel.
- En 1 et 2 top précision évaluée sans comparaison avec meilleurs systèmes BioASQ.
- En 3 top précision évaluée en comparaison avec meilleurs systèmes BioASQ sur deux ans.

Challenge international de traitement automatique du langage naturel pour le biomédical. Tâches d'indexation sémantique et de QA.
QA en 2 phases : RI et extraction d'information.

10 experts reconnus de 10 spécialités différentes sont à l'origine des questions et du «gold file» ou «qrel» (les réponses attendues).
Evaluation automatique, en exact match, et par experts, avec amélioration du qrel

4 types de questions : oui / non, factoïdes, listes et résumés.
Ressources à disposition : databases et ontologies (MeSH, GO, UniProt, …) (0.9 million de concepts), résumés PubMed (22 millions d'entrées), …

Depuis 2013, au printemps :
100 questions toutes les 2 semaines.
QA sur 48 heures : 24 h pour la RI et 24 h pour le QA
Total de 1'300 questions sur 3 ans dont ~ 330 factoïdes