h e g

Moteur de question-réponse pour les sciences biomédicales

Mémoire de recherche réalisé par :

Alexandre RACINE

Philippe COSANDEY

Romilda CHALARD

Sous la direction de :

Patrick RUCH, Professeur HES

Genève, 18 janvier 2016

Master en Sciences de l'information

Haute École de Gestion de Genève (HEG-GE)

Hes·so GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

Déclaration

Ce mémoire de recherche est réalisé dans le cadre du Master en Sciences de l'information de la Haute école de gestion de Genève. Les étudiants acceptent, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans ce travail, sans préjuger de leur valeur, n'engage ni la responsabilité des auteurs, ni celle de l'encadrant.

« Nous attestons avoir réalisé le présent travail sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 18 janvier 2016

Alexandre Racine Philippe Cosandey Romilda Chalard

Remerciements

Plusieurs personnes nous ont apporté leur aide, leur soutien moral voir un apport physique ou intellectuel à des degrés variés, mais toutes nous ont motivés dans le cadre de ce travail. Nous tenons à les remercier sincèrement ;

Pour commencer, Patrick Ruch pour nous avoir proposé ce sujet de recherche et pour son encadrement, sa disponibilité constante et sa capacité à se mettre à notre niveau. Nous remercions ensuite, tout particulièrement, Julien Gobeill pour sa patience à toute épreuve et son appui sans faille durant ces mois où nous avons pu le solliciter à de nombreuses reprises, ainsi que pour la mise à disposition d'espace de travail et de solutions informatiques.

Egalement toutes les personnes croisées en interne qui nous ont encouragé et donné leur avis ;

Stéphanie Pouchot pour ses conseils divers, Michel Gorin pour son coup d'œil avisé sur notre poster scientifique ainsi que Arnaud Gaudinat pour son intérêt et ses remarques, avec une dernière pensée pour les assistants de la HEG et leurs encouragements.

Nous remercions également des intervenants externes tels que Zhiyong Lu du National Center for Biotechology Information (NCBI) que nous avons pu interroger par courriels ainsi que l'équipe BioASQ qui a répondu à nos nombreuses questions.

Enfin, nous tenons aussi à remercier chaleureusement nos familles et amis qui nous ont soutenus et épaulés dans les moments de doute comme de réussite et sans qui ce parcours d'une année pleine de rebondissements n'aurait jamais été possible.

Résumé

La littérature concernant le domaine médical atteint un volume dépassant l'entendement humain et ne cesse d'augmenter. Si nous nous concentrons sur les documents numériques qui permettent la recherche en ligne d'information quelconque, l'exploitation de cette quantité rend la précision complexe et chronophage. Ce problème a motivé le développement d'outils plus évolués comme les systèmes de question-réponse. Ces derniers autorisent l'utilisateur à poser des questions en langage dit naturel.

L'objectif de notre travail est d'augmenter la performance du mode question-réponse du moteur EAGLi. Pour mesurer les performances de notre apport, une analyse a été réalisée à partir de questions et de scores d'autres moteurs, le tout issu de la campagne d'évaluation internationale BioASQ.

Notre revue de la littérature, pour commencer, synthétise des données sur la prolifération des sources numériques, en particulier dans le domaine biomédical. Nous développons également les solutions de prospection grâce aux moteurs de recherches, plus particulièrement ceux dit de question-réponse. Cet environnement décrit nous aide à aborder les challenges liés au développement et à l'amélioration de ces outils avec la mise en exergue du concours en ligne BioASQ.

La focalisation sur ce challenge, nous permet de faire ressortir une des phases qui correspond à notre projet. Ce dernier consiste à reformuler des questions manuellement qui sont issues du challenge évoqué ci-dessus. Il s'agit aussi dans une certaine mesure d'améliorer la couverture du système en augmentant les données à sa disposition. Ce procédé nous a permis d'évaluer nos performances.

Le moteur EAGLi nécessite principalement une typologie de phrase précise ainsi que des patrons de questions pour interagir avec la base de données MEDLINE. Son architecture, notre méthodologie ainsi que l'évaluation de nos résultats sont développés dans ce rapport. Ces derniers sont satisfaisants et nous laissent à penser que malgré l'ampleur de la tâche associée au Question Answering, ce domaine particulier de la recherche d'information va sans nul doute se perfectionner.

Mots-clefs : Moteur de recherche ; Moteur de Question-Réponse ; Sciences biomédicales ; Données structurées

Table des matières

Liste des tableaux

Liste des figures

1. Introduction

La prédominance du Web dans nos vies fait de la recherche d'information une activité presque banale voir quotidienne. Cependant la masse documentaire ne permet pas forcément un retour d'éléments toujours pertinents. Il est estimé que 3000 articles sont publiés chaque jour dans le domaine des sciences biomédicales (Paliouras, Krithara 2015). Cette croissance exponentielle de données mises en ligne est trop importante pour être traitée ou assimilée à l'échelle humaine. Une telle multitude d'informations a besoin d'un soutien automatisé à la recherche, qui permet de faire un premier tri pour obtenir une sélection de documents pertinents. Dans ce contexte, la branche du biomédical nécessite de trouver des données qui peuvent s'avérer vitales et surtout qui demandent une réponse rapide en cas d'urgence.

C'est à ce niveau qu'émerge une nouvelle problématique, trouver plus rapidement des réponses à des besoins qui se veulent toujours plus précis. Les derniers éléments évoqués ci-dessus déterminent le besoin d'outils plus évolués que de simples moteurs de recherches classiques. En effet, une liste de références demande encore une étape pour faire ressortir un élément déterminé. Pour ce type de requêtes avancées il existe des moteurs de question-réponse (QA pour Question/Answering ci-après) dont les caractéristiques sont de permettre de poser des questions en langage naturel afin de trouver les informations en fonction de données structurées. Ces systèmes, inhérents à l'intelligence artificielle, analysent la question en langage naturel, reformulent la requête avant de rechercher dans des bases de données conséquentes. L'importance de ce procédé vient du fait que le système de QA ne fournit pas seulement une liste de documents pouvant contenir des informations pour l'utilisateur, mais opère une analyse pour extraire les réponses des documents sélectionnés. En ce sens, les systèmes de QA sont la suite évoluée des systèmes de recherches documentaires. Dès lors, l'objectif est que l'utilisateur n'ait plus à explorer l'entier des documents retournés par une recherche mais obtienne l'élément exact qu'il attend.

Dans le cadre de notre projet, nous avons travaillé sur le moteur de QA EAGLi qui exploite les informations de la collection MEDLINE (Gobeill 2012, p. 30). Ce moteur a été créé par le groupe BiTeM (Ruch [ca. 2015]), spécialisé dans la fouille de texte et la bibliométrie, dans le domaine biomédical entres autres. Il réunit des chercheurs issus de domaines différents, tels que l'informatique, la bioinformatique ou la médecine. Ces personnes sont affiliées à divers instituts de recherche sur Genève. La filière IS de la Haute Ecole de Gestion (HEG) et le service des sciences de l'information médicale (SIM) du département d'imagerie et des sciences de l'information médicale (DISIM) au sein des hôpitaux universitaires de Genève (HUG) sont particulièrement représentés. BiTeM conçoit, développe et maintient des instruments et des infrastructures de services pour aider les biocurateurs à tirer parti des données et les exploiter, par exemple grâce à des plateformes come GOCat (the Gene Ontology Categorizer) ou EAGLi (Engine for question-Answering in Genomic Literature). Comme nous l'indiquions plus haut, notre travail s'est focalisé sur EAGLi.

Celui-ci possède trois niveaux de fonction : recherche de publications, recherche sémantique et une fonction plus avancée de moteur de QA. Dans ce dernier mode, trois composants logiciels sont successivement appelés : un catégoriseur de question, un moteur de recherche de document (RI) et un extracteur de réponses (QA).

Les documents du domaine médical sont indexés à l'aide de thésaurus. Pour les articles scientifiques c'est le Medical Subject Headings (MeSH) qui est utilisé. En s'appuyant sur la structuration plus ou moins complexe de ces données, la puissance de calcul de l'informatique permet de rechercher les concepts attendus au-delà des capacités humaines. Il va sans dire que la notion de moteur de QA n'est pas exclusivement liée au domaine du biomédical, ce que nous développerons plus en détails ci-dessous, mais ce secteur de la recherche se trouve être extrêmement intéressant dans ce contexte et ceci pour plusieurs raisons.

Tout d'abord la redondance de l'information permet à un moteur d'exploiter la grande quantité de données disponibles pour améliorer ses performances. De manière basique, plus un terme qui a du sens est présent dans des publications retrouvées en réponse à une question précise, meilleures sont les chances que ce concept soit pertinent par rapport à cette question. La structure d'EAGLi s'appuie justement sur le fait qu'un concept répété à de nombreuses reprises dans la littérature biomédicale lui permet de ressortir la réponse correcte.

Ensuite, la structuration des données au moyens de vocabulaires contrôlés permet à la communauté de s'entendre sur les règles d'écriture et de sens des différents concepts, et à la machine de savoir exactement, de manière univoque, de quel concept il est question. Le fait que les données soient aussi pour la plupart concentrées dans des points de collecte centralisés et unanimement connus représente également un avantage. Ces atouts conduisent à la construction de systèmes complexes, qui tentent au quotidien de relever les nombreux défis du biomédical : nécessité de qualité telle que la fiabilité ou l'actualité et aspects fonctionnels cruciaux (intégration dans les outils de la pratique médicale) pour ne citer que deux exemples.

Les enjeux actuels sont immenses, le biomédical tend vers des réponses toujours plus précises. Si nous considérons que l'avenir se présente avec une prolifération de données, notamment sur l'individu, il est nécessaire d'affiner la précision au sein d'une masse en expansion constante. L'objectif de notre travail vise à augmenter la performance du mode de question-réponse du moteur EAGLi, à l'aide de la reformulation des questions et de l'écriture de règles manuelles. L'analyse de la performance se faisant à partir des questions d'une des tâches de la campagne d'évaluation internationale BioASQ (Paliouras, kakadiaris, Krithara 2015). Cette compétition a lieu chaque année, au printemps, sous la forme de soumission de 100 questions toutes les deux semaines. Le challenge BioASQ émane d'un projet transnational du septième programme-cadre de l'Union européenne. Il organise depuis 2013 des concours d'indexation sémantique et de réponse à des questions dans le domaine biomédical. Notre méthodologie est expliquée en deuxième partie de ce mémoire, suivie des résultats et d'une analyse. Au terme de celle-ci, nous pouvons étudier les questions traitées ainsi que les patrons de questions[1] afin de partir sur une discussion permettant de clarifier les pistes issues de nos résultats.

[1] La notion de patron de questions sera développée plus en détails dans la méthodologie de ce travail.

2. Etat de l'art

Les chiffres liés au numérique commencent à donner le vertige depuis quelques années, avec un développement de capacités et d'usages spectaculaire. Le phénomène se ressent autant au niveau de la production, du stockage que de l'usage. Nous constatons que la mise en ligne a accentué ce fait de base. Dans sa thèse de 2008 Mehdi Embarek évoque, l'expansion continue des documents électroniques depuis que l'accès à ceux-ci est facilité par l'intermédiaire d'Internet. A partir de ce constat, il met en évidence que la recherche d'information fait dorénavant partie de nos activités quotidiennes. Cette évolution provoque un accroissement du nombre d'utilisateurs de moteurs de recherche dont certains sont devenus leaders comme Google.

> *« Selon les chiffres de la société de mesure d'audience Comscore Networks (*http://comscore.com*) le moteur de recherche Google a ainsi traité en novembre 2006, 5,6 milliards de requêtes (+9,1% par rapport à novembre 2005). » (Embarek 2008, p. 23)*

Ces chiffres datent un peu, mais l'essentiel est de voir que, pour la recherche en ligne, l'acquisition par le grand public s'est accrue rapidement. Ce que nous tenons à soulever est que cette habitude qui s'ancre dans nos vies de tous les jours va développer des exigences qualitatives, la quantité étant déjà bien présente. En effet, une recherche finit toujours par développer un besoin de précision.

Mehdi Embarek évoque la difficulté, au final, à trouver de l'information pertinente selon des besoins précis.

> *« Deux facteurs en sont essentiellement responsables : le nombre de documents retournés par les moteurs de recherche d'une part ; l'hétérogénéité des informations disponibles sur le Web d'autre part. De plus, parmi tous les documents retournés par les moteurs, la plupart d'entre eux ne sont pas pertinents. » (Embarek 2008, p. 23)*

Il définit ainsi l'émergence d'un besoin de précision en déclarant en introduction que les systèmes de recherche d'information doivent s'améliorer pour pouvoir répondre toujours plus rapidement à des besoins mieux déterminés.

Cette nécessité qui se met en place évolue également en fonction de la masse documentaire en constant développement. Comme l'indique Gobeill (2012) dans sa thèse nous sommes de plus en plus dépendants de méthodes automatiques efficaces pour accéder à l'information contenue dans la littérature. La gestion à ce niveau dépasse effectivement les capacités humaines.

Les enjeux actuels sont immenses et vont s'élargir avec des données toujours plus importantes et plus précises en fonction de nos attentes, et du domaine considéré, comme le biomédical par exemple.

2.1 Les données en ligne

Nous l'avons évoqué ci-dessus, l'apparition d'Internet a bouleversé les comportements de la recherche mais également de la production. Ce second élément tend vers un but d'exhaustivité en reproduisant également l'existant :

> *« Une production numérique native de plus en plus importante s'est développée à partir des années 1990, complétée dans la foulée par la numérisation en cours de la production antérieurement existante sous diverses formes physiques ». (El Haldi 2010, p. 43)*

Nous retrouvons l'utopie d'obtenir un jour une bibliothèque, fusse-t-elle virtuelle, contenant l'ensemble du savoir humain associée à la volonté de rendre le tout accessible en ligne. Ce livre évoque également le fait que la publication est un élément de mise en valeur des données scientifiques. Ce dernier terme est un point important pour notre travail, car nous verrons par la suite que données et information sont interdépendantes pour effectuer une requête. C'est la donnée qui est cherchée pour permettre de construire une réponse contenant l'information requise. C'est ainsi que la notion d'information était liée à la connaissance dans un premier temps, puis avec les processus d'informatisation, à celles de données :

> *« En sciences de l'information, Charles T. Meadow, mathématicien … définissait en 1984 l'information comme des données assemblées et mises en forme de manière signifiante. » (Simonnot 2012, p. 23)*

Nous comprenons ainsi qu'il est nécessaire de les encoder (structurer) pour permettre leur enregistrement ou leur transmission. Il est communément évoqué pour Internet de parler du Web des données.

2.1.1 Les données structurées

Ce que nous appelons fréquemment une collection est un ensemble structuré. Une base de données est définie comme une collection organisée de données structurées. Dans le livre de El Haldi (2010, p. 78) nous avons une définition :

> *« Une donnée structurée est un triplet (i, d, v) composé des éléments suivants : un intitulé (i), renvoyant à un concept (une catégorie d'activité, par exemple), un domaine de définition (d), composé d'assertions formelles (contraintes d'intégrité) spécifiant l'ensemble des valeurs admises dans une base de données pour ce concept (une liste contrôlée de valeurs alphabétiques, par exemple) et, enfin, une valeur (v) à un instant t (le secteur de la chimie , par exemple). »*

Ce système dit contrôlé dépend d'une manière de faire ou d'un savoir qui a été formalisé et reconnu de manière institutionnelle. C'est un procédé permettant de centraliser de l'information. Pour conserver un lien entre les données et leur publication en ligne il est nécessaire de leur apporter une structure.

> *« Cette mise en relation, créatrice de sens, est facilitée par les liens hypertexte, par les métadonnées comme par l'indexation dans les moteurs de recherche, qui multiplient considérablement les possibilités de mettre les textes en relation, sans commune mesure avec les outils dont disposaient les bibliothécaires auparavant quand ils recouraient à la classification et à l'indexation selon les méthodes traditionnelles. » (El Haldi 2010, p.45)*

Il existe un grand nombre d'exemples de données structurées sur le web, utilisées en fonction de différents domaines d'application. Au niveau du domaine biomédical qui nous intéresse ici, le Medical Subject Headings (MeSH) en est un parfait exemple. Le MeSH est un thésaurus (vocabulaire contrôlé) produit par la bibliothèque de médecine américaine (National Library of Medicine, NLM ci-après). Il est utilisé pour indexer des informations relatives à la santé. Ce vocabulaire inclut dans ses concepts de nombreux descripteurs ainsi que leurs synonymes pour permettre de retrouver le descripteur le plus pertinent qui caractérisera un concept dans un contexte donné.

La première liste officielle de références de données dans le domaine a été mise à disposition par la NLM en 1954 sous le titre « Subject Heading Authority List » (NLM 2014). Dans sa version actuelle, de 2016, elle contient plus de 27'000 descripteurs et 87'000 entrées qui permettent de trouver le sujet MeSH le plus approprié à une information dans un contexte précis. Rappelons aussi que le MeSH est étroitement lié à la base de données MEDLINE/PubMED qui permet de rechercher des articles scientifiques publiés par les revues phares du secteur biomédical. Il est enfin intéressant de souligner que même si le MeSH et MEDLINE/PubMED sont des outils structurés, certaines des informations contenues en leur sein peuvent, elles, ne pas l'être.

2.1.2 Les données non-structurées

Dans le cas de données sous forme non- ou semi-structurée, l'information est donnée sous forme de langage librement choisi par l'auteur. *« C'est le cas dans la littérature contenue dans MEDLINE – même si MEDLINE est une base de données structurée* ». *(Gobeill 2012, p.8).*

Comme évoqué ci-dessus, le MeSH définit des descripteurs (contrôlés et donc structurés) mais renvoie également à la plupart de leurs synonymes. Le problème est, si nous prenons l'exemple d'un utilisateur qui effectuera une recherche au moyen des termes MeSH, qu'il existe la possibilité que le ou les termes utilisés lors de cette recherche ne soient pas tous compris sous cette forme explicite dans le vocabulaire utilisé librement par les auteurs pour rédiger leurs articles.

Si l'on se réfère à la thèse de Gobeill (2012, p.10) :

> *« Le nombre par défaut d'articles affichés sur une page par PubMed est de 20. Seulement un tiers des résultats des recherches sont donc affichés sur une seule page. Or, 80% des clics pour lire le détail d'une citation concernent un article de la première page. »*

Ceci implique la possibilité, ou plutôt le risque, pour l'utilisateur de passer à côté d'une grande quantité d'informations pertinentes alors même que celles-ci sont pour la plupart sûrement comprises dans la collection en question. Partant de ce postulat, nous pouvons comprendre dès lors aisément pourquoi la structuration de ces données peut jouer un rôle crucial au niveau de la recherche d'information, a fortiori dans le domaine du biomédical.

2.2 La recherche d'information (RI)

Dans un premier temps, nous avons parlé de l'expansion du Web, composé de toujours plus de données qui nécessitent d'être mises en forme pour permettre de retrouver de l'information pertinente et adéquate. Nous pouvons concevoir ce monde virtuel avec l'idée d'organiser ces données à l'instar de l'indexation d'une bibliothèque mais sous une forme dématérialisée.

> *« Les systèmes de recherche d'information doivent inclure des fonctionnalités qui permettent d'organiser et indexer les documents d'une collection, de traiter leur langage de représentation et les requêtes des utilisateurs, ainsi qu'une interface d'interrogation et de présentation des résultats ». (Simonnot 2012)*

La définition d'un système de recherche d'information (SRI) contient un système d'information et l'accessibilité à un ensemble de documents à l'aide de leur contenu sémantique. (Ihadjadene 2004). C'est pourquoi nous distinguons plusieurs types de systèmes. D'une part nous avons des SRI dit plein texte, généralement à vocabulaire non contrôlé ou libre et d'autre part les SRI référentiel, qui contiennent une description bibliographique ainsi qu'une description du contenu avec en général un vocabulaire contrôlé.

En ce qui concerne la recherche à proprement parlé nous connaissons deux modèles principaux. Le premier s'intitule booléen :

> *« Une requête R est une expression logique – d'où la référence à la logique de Boole - composée de termes assemblés par les opérateurs ET, OU et SAUF. Le modèle booléen utilise le mode d'appariement exact (égalité ou similarité des mots). » (Ihadjadene 2004, p. 20)*

Ce procédé ne donne donc lieu qu'à une restitution de documents qui répondent de manière stricte aux termes de la requête. On a vu au point 2.1.2 que cet état de fait pouvait conduire à ne pas retrouver l'information recherchée dans le sens où *« L'efficacité de la recherche dépend d'une stricte égalité entre les termes de la requête et ceux des documents. » (Ihadjadene 2004, p. 30)*

Au niveau de la mise en œuvre, le modèle booléen s'appuie sur la technique dite de fichiers inversés :

> *« Après avoir enregistré, pour chaque document, la liste des termes qu'il contient, on crée un fichier inversé qui dresse, pour chaque terme, la liste des documents qui le contiennent. Cette facilité explique l'énorme succès de ce modèle » (Ihadjadene 2004, p. 21)*

Pour opérer une fonction de tri dans le modèle booléen il existe le RANK qui permet de repérer le rang pour classer des mots-clés en rapport à leur fréquence d'utilisation. En effet, malgré sa forte popularité le modèle booléen peut parfois avoir une performance médiocre au vu des éléments abordés ci-dessus. Il devient alors nécessaire de pouvoir filtrer les résultats et ceci nécessite la connaissance de techniques d'interrogation parfois complexes (exclusion de termes, troncatures), et d'une méthodologie rigoureuse quant au choix des termes utilisés (emploi de synonymes, variation du singulier/pluriel).

La logique booléenne a de ce fait *« sans doute [été] longtemps réservée à des spécialistes de la recherche documentaire [pour ces raisons]. A l'heure où elle est mise à la disposition du grand public à travers les moteurs de recherche, il convient de poursuivre les recherches pour en améliorer les performances » (Ihadjadene 2004, p. 32)*

L'amélioration se profile pourtant avec la possibilité de classement des résultats par ordre chronologique par exemple, ce qui est le cas dans PubMED, l'utilisation du langage naturel comme introduite dans les systèmes de QA ou le recours à des liens hypertexte.

Un second modèle a fortement influencé la RI. Il s'agit du modèle vectoriel, créé au début des années 1970 par Gérard Salton et son équipe dans le système de recherche d'informations SMART (Ihadjadene 2004)

Le concept de base est l'utilisation d'une représentation d'inspiration géométrique permettant de classer les documents par ordre de pertinence, avec l'utilisation de la fréquence des termes de la requête dans le document et non uniquement la présence ou l'absence du terme comme pour le modèle booléen.

Les termes de l'indexation sont considérés comme les dimensions d'un espace multidimensionnel. A l'intérieur de ce dernier, les documents et les requêtes sont représentés par des vecteurs. La pertinence d'un document par rapport à une requête est relative aux positions respectives du document et de la requête, et est estimée par une mesure de similarité (au sens mathématique) définie dans cet espace. La requête est ainsi considérée comme un texte particulier exprimé sous forme de langage naturel.

Cette représentation vectorielle nécessite de faire des choix sur un certain nombre de paramètres ; dimensions, termes d'indexation qui les supportent, la valeur de la composante du vecteur pour cette dimension, autrement dit le poids du terme d'indexation dans le document ainsi que la similarité choisie. (Ihadjadene 2004)

Concernant spécifiquement notre travail, ces deux modèles ont leur importance puisqu'EAGLi est capable dans sa version actuelle d'alterner entre les représentations booléennes et vectorielles pour effectuer ses requêtes. Nous en donnerons un détail un peu plus précis dans la suite de ce rapport.

2.2.1 Les méthodes d'évaluation

L'évaluation des systèmes de la recherche d'information a son origine dans les projets d'évaluation des systèmes d'indexation menés à Cranfield au Royaume-Uni. Les premières expérimentations ont été supposées fiables pour l'évaluation de la recherche d'information.

> *« En montrant de manière rigoureuse l'influence du type d'indexation sur la mesure de performance de recherche, Cranfield II a prolongé de manière significative les travaux initiés dans Cranfield I et consolidé un protocole d'évaluation qui allait être repris plusieurs années plus tard, en particulier dans les campagnes TREC. Ce protocole comprend deux aspects ; tout d'abord la collection de tests qui comprend :*
>
> - *Un ensemble de documents (la base documentaire) ;*
> - *Un ensemble de requêtes*
> - *Un ensemble de documents jugés pertinents qui constitue le référentiel*
>
> *et des techniques de mesures de la performance ». (Ihadjadene 2004, p. 190)*

La méthodologie TREC poursuit celle qui a été mise en place avec les tests Cranfield II, mais en apposant des modifications importantes. La première modification concerne la taille de la collection de documents définie pour les tests.

Ensuite, les systèmes qui participent disposent d'un groupe de thèmes, chacun étant une description d'un besoin d'information.

> *« Depuis TREC 4, les thèmes ont conservé une structure identique qui comporte quatre champs : un identifiant, un titre, une description d'une phrase qui explicite le domaine de recherche et un commentaire qui fournit les éléments permettant de juger qu'un document est pertinent ou non. » (Ihadjadene 2004, p. 201)*

Un autre point concerne la construction du référentiel qui est construit automatiquement et non plus manuellement comme c'était le cas pour les campagnes précédentes. Ce référentiel se construit en fonction de l'ensemble des documents jugés pertinents par les différents systèmes. Il ne retient que les n documents les plus fréquemment cités. Les systèmes concurrents verront ensuite leurs réponses comparées avec ce référentiel.

Un dernier point concerne le déroulement, avec une première étape dite d'entraînement qui précède l'évaluation réelle. Ce procédé permet aux différents systèmes de s'améliorer. Les campagnes TREC utilisent l'outil trec_eval pour évaluer les performances dans la recherche d'information. Notre choix concernant l'évaluation de ce travail s'est porté sur trec_eval. Des commentaires sur son utilisation seront donnés au point 3. Méthodologie.

2.3 Le Question Answering (QA)

Au vu de ce qui a déjà été défini plus en avant dans ce rapport, le domaine de la recherche d'information a connu l'émergence du QA. Les systèmes de QA sont une extension améliorée des systèmes de recherche d'information. Ils permettent de poser une question en langage dit naturel afin d'obtenir en retour une réponse à la place d'un ensemble de documents pertinents comme le propose les moteurs de recherche. Ces derniers ne font que la première étape du travail mais il demeure la tâche de passer en revue l'ensemble des documents pour trouver l'élément de réponse précis. La tâche principale des systèmes de QA est donc dans ce contexte d'extraire les réponses issues des documents préalablement retournés par un moteur de recherche. Pour arriver directement à cette étape il faut une analyse plus profonde, nécessitant une architecture plus complexe.

2.3.1 Général

Dans le QA, le fait de permettre à l'utilisateur de poser une question en langage naturel permet aussi dans une certaine mesure de s'affranchir des contraintes évoquées au point 2.2 quant à l'utilisation des techniques d'interrogations complexes sur le modèle booléen. Ceci offre aussi la possibilité d'occulter le fait de réfléchir en langage dit de machine. Des avantages conséquents donc, même s'il est intéressant de constater que ce langage dit naturel évolue. En effet la langue est codifiée dans nos dictionnaires par les institutions académiques, mais elle évolue avec l'usage social, ce qui influencera les définitions ultérieures. (El Haldi 2010).

La mouvance et l'ambigüité du langage humain restent donc les principales caractéristiques à prendre en compte concernant le domaine de la recherche d'information, à plus forte raison du Question Answering.

2.3.1.1 Watson

Un élément intéressant de l'évolution du QA que nous tenons à soulever ici est Watson (Ferruci 2012). L'évolution dans le concept de réponse induite par le passage de la RI au QA comme défini plus haut permet d'améliorer le dialogue, voir la collaboration entre l'humain et l'ordinateur. Dans cette optique, en se plaçant dans une démarche compétitive, un chercheur d'IBM a suggéré, en 2006, le développement d'un ordinateur pour participer au jeu télévisé Jeopardy[2]. Dans ce programme télévisuel, le candidat doit formuler une question à partir d'un petit texte comprenant un indice. Une équipe de 20 chercheurs et développeurs s'est vu confier cette mission et la première participation au jeu a eu lieu en 2011. Ce fût un succès pour « Watson » (nom de l'ordinateur en question, rendant ainsi hommage au fondateur d'IBM), ce qui mit en lumière les possibilités et le potentiel du QA.

Le succès de Watson, qui a été suivi notamment par une interaction avec les sciences biomédicales a permis différentes applications pour le domaine de la santé. Le Watson Discovery Advisor a, par exemple, permis aux chercheurs d'obtenir des réponses précises à leurs questions sans devoir fouiller une multitude d'articles. Ce gain de temps à également été validé par des compagnies pharmaceutiques pour analyser de la documentation permettant de découvrir que des médicaments existants pouvaient être utiles pour le traitement de maladies autres que celles visées initialement (Foster 2015).

Une analyse de données a servi à déterminer la cible à atteindre : que Watson puisse en environ trois secondes s'estimer capable de produire une réponse pour au moins 70% des questions, avec une réponse effectivement correcte dans 85% de ces cas. Deux techniques mises en place ont servi de fondation :

- l'architecture logicielle DeepQA (favorisant la recherche massivement parallèle de propositions de réponse ainsi que l'évaluation de l'indice de confiance).
- la méthodologie AdaptWatson (pour accélérer la recherche, le développement et l'intégration d'un très grand nombre de composants algorithmiques).

Ferrucci (2012) évoque les principaux résultats. La progression de la performance de Watson à la fois en précision et en confiance, la victoire à 71% de 55 jeux en conditions réelles avant le jeu télévisé et la robustesse du système attestée par la résistance de sa performance à la dégradation d'une partie de ses composants. Les projets pour la suite : l'application comme aide à la décision en résolution de problèmes et le développement vers l'interactivité, en particulier dans le domaine de la santé. La stratégie habituelle de recherche (recherche de passages) a été complétée par d'autres stratégies de recherche (recherche de document, de passages title-in-clue). Des pseudo-documents « orientés titre » ont été générés à partir de documents sans titre (par exemple pour les œuvres de Shakespeare avec le titre de l'histoire mais aussi avec l'auteur, les protagonistes de l'histoire, etc…).

[2] https://fr.wikipedia.org/wiki/Watson_%28intelligence_artificielle%29

Enfin, les documents « orientés titre » ont été réorganisés (par exemple en rassemblant les entrées de wiktionary pour un même terme écrit de différentes manières, en complétant les définitions de dictionnaires afin que chaque définition inclue le terme défini pour former une phrase complète). L'expansion de sources a servi à combler les lacunes du corpus. La complétion a visé les sujets les plus populaires, repérés par exemple par le nombre de citations d'un article Wikipédia par d'autres articles Wikipédia.

A partir du repérage de ces « graines », des pseudo-documents ont été générés sur la base d'extraits de textes du web (en quatre étapes : « retrieval, extraction, scoring, merging »). La croissance à partir de ces « graines » augmente non seulement le fond (les informations), mais aussi la forme (par la paraphrase). Les mesures attestent que la performance de Watson a été sensiblement améliorée par la transformation ainsi que l'expansion des sources (Ferrucci 2012). Il est intéressant de constater ici que les stratégies utilisées (synonymie, expansion des sources par exemple) se rapprochent de ce que nous avons pu définir plus haut dans le point sur les données structurées établissant ainsi un lien entre le QA général et le QA biomédical.

2.3.2 Biomédical

Le texte est le medium principal pour l'échange d'information scientifique entre les experts (Spasic et al. 2005, cité dans Gobeill 2012, p.1). Les chiffres concernant les sciences biomédicales et la littérature qui s'y rattachent donnent rapidement le vertige. Si nous nous concentrons sur MEDLINE abordée plus haut, nous pouvons avoir une idée de l'ampleur.

Comme évoqué à plusieurs reprises, il est difficile dans cet océan de données de trouver des réponses avec des moteurs de recherche classiques. Il faut tenir compte des synonymes pour les gènes et les maladies et le renvoi de documents nous placent encore devant des listes conséquentes pour un traitement rapide. Si nous nous contenons au cancer par exemple, un médecin spécialiste du domaine qui souhaite se tenir au courant en continu devrait consacrer 160 heures par semaine pour suivre la documentation sur les nouveaux traitements, les nouveaux médicaments voir les nouveaux essais cliniques (Foster 2015). Il est aussi estimé actuellement que seulement 20% des données probantes sont utilisées par les médecins dans le traitement de leurs patients. Les enjeux peuvent être ainsi cruciaux, pour ne pas dire vitaux. Le cas reporté en 2001 d'une patiente décédée dans une étude clinique, alors que l'information sur la toxicité de l'agent qu'elle avait reçue était documentée mais n'avait pas été trouvée par les médecins en charge de l'étude en est une illustration parfaite (McLellan 2001 cité dans Gobeill 2012, p. 1).

D'autres complications viennent s'ajouter. Selon Neves et Leser (2015), si la restriction d'accès au contenu des articles scientifiques paraît globalement encore problématique le domaine de la biologie semble toujours peu couvert par rapport au domaine médical car elle est affectée par la quantité supérieure de ses concepts et la sous-utilisation de ses ontologies par les solutions actuelles de QA. Ce constat demande une plus grande collaboration entre les développeurs et les biologistes pour améliorer la compréhension des besoins des uns et des autres. Dans ce contexte, il est important de souligner aussi que la santé a un coût toujours plus élevé et que la gestion de l'information est un élément non négligeable pour atténuer ce dernier. Surtout si nous posons le problème à l'envers ; la prolifération d'informations accentue le temps de recherche et rajoute un coût supplémentaire.

Le MeSH défini plus haut ou d'autres terminologies deviennent alors importantes quant à l'efficacité de la gestion et de la recherche documentaire pour nommer de manière univoque les concepts, d'où l'utilité d'un référentiel commun. En unifiant la sémantique, nous pouvons indexer et classer de manière unique, ce qui permet ensuite de comparer les informations pour identifier les concepts selon ce principe d'univocité. La communauté de chercheurs en sciences de la vie représente ces concepts de la même manière, que ce soit au niveau de l'orthographe, des variants singuliers ou pluriels ou de leur signification.

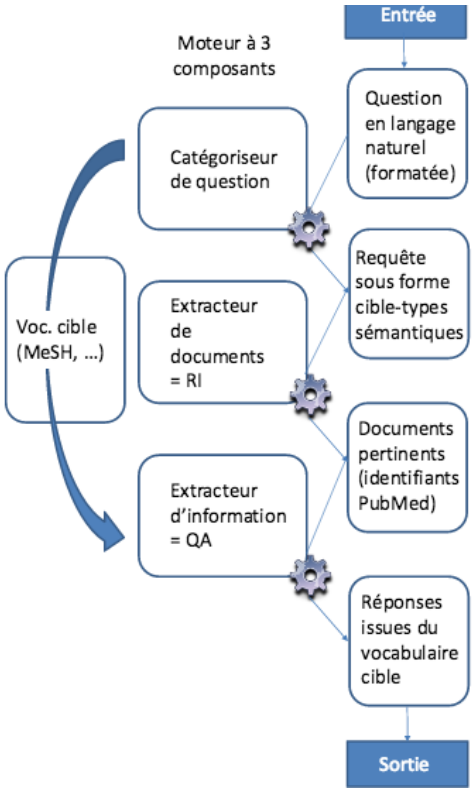
Au niveau des questions les plus communes, les chercheurs les classent généralement selon trois types : oui/non, factoïdes / liste ainsi qu'une définition ou un résumé. Les questions oui / non sont les plus simples, du fait que les deux seules réponses possibles sont connues à l'avance: «oui» ou «non». Les factoïdes sont des questions qui attendent un seul élément ou une liste de faits courte en retour. Une seule peut être par exemple, un gène, une maladie ou un chiffre (par exemple, le nombre de mutations pour un certain gène). Des types de questions différents pour des réponses hétérogènes, autant dire que la tâche est conséquente. Toujours selon Neves et Leser (2015), dans le médical, la restriction à une espèce et l'utilisation des terminologies de l'UMLS (Unified Medical Language System) facilitent la tâche, tout comme les portails rassemblant des questions. Il reste un problème soulevé ici, la fiabilité ne les convainc pas.

Les processus biologiques sont en effet extrêmement enchevêtrés et peuvent, par exemple impliquer l'interaction d'une multitude de molécules apparentées (Andrade-Navarro, Perez-Iratxeta 2015). Cet article définit un échantillon de méthodes d'extraction de texte. Il est possible d'annoter les fonctions des gènes et des protéines, de prédire les interactions des protéines avec des médicaments ou avec d'autres protéines. Ces constats imposent un enjeu de taille, auquel nous avons tenté de répondre par notre travail sur EAGLi.

2.3.2.1 EAGLi

Comme défini dans notre introduction, EAGLi possède trois niveaux de fonction distincts : recherche de publications, recherche sémantique et une fonction plus avancée de moteur de QA qui nous intéresse plus particulièrement. Dans ce dernier mode, trois composants logiciels sont successivement appelés : un catégoriseur de question, un moteur de recherche d'information qui renvoie des documents (RI) et un extracteur de réponses (QA).

Figure 1 : Architecture d'EAGLi



Source : adapté de (Gobeill 2012, p.3)

Des informations plus détaillées sur le fonctionnement du système seront abordées dans notre point 3. Méthodologie mais il convient ici d'expliquer brièvement les bases de son fonctionnement. En premier lieu la question en langage naturel doit être adaptée à la structure acceptée en entrée par le moteur. Si la syntaxe est correcte, le catégoriseur reconnaît qu'il s'agit d'une question et formule la requête. Le moteur de recherche renvoie ensuite les documents contenus dans Pubmed et considérés comme pertinents par rapport à la question posée. Pour finir, le moteur de QA extrait ensuite les réponses potentiellement pertinentes issues de ces documents sélectionnés.

Nous l'avons déjà évoqué au point 2.2 La recherche d'information mais un élément important à préciser à nouveau ici est le fait qu'EAGLi soit capable d'alterner entre les modes de recherche booléen et vectoriel. En mode booléen le moteur renvoie les documents issus de Pubmed pour lesquelles les termes de la question apparaissent explicitement dans le texte, le tout classé par ordre chronologique. En mode vectoriel, il applique les spécificités définies plus haut et inhérentes à ce mode de recherche. C'est-à-dire qu'il ne va plus seulement prendre en compte le fait que les termes apparaissent ou non dans l'article, mais surtout leur attribuer une pondération en fonction de leur fréquence d'apparition. Précisons aussi que dans ce mode les documents ne sont plus classés par ordre chronologique.

Le point 2.2.1 Les méthodes d'évaluation nous a montré que pour tout type de recherche d'information il fallait ensuite pouvoir évaluer les résultats obtenus par le moteur. Dans notre cas, cette partie s'est faite au moyen de la campagne d'évaluation BioASQ.

2.4 BioASQ

L'évolution des systèmes de QA a donné lieu à la mise en place de compétitions afin de déterminer les bons systèmes. BioASQ est un challenge dans ce domaine. Il permet d'obtenir une vue des techniques de classification de textes, d'indexation sémantique, de récupération de passage et de résumés. Il est intéressant de noter que le but sous-jacent est de promouvoir les systèmes et approches qui sont en mesure de traiter l'ensemble de la diversité du web. Le challenge est particulièrement axé sur le biomédical mais ce n'est pas l'unique cible (Balikas 2014)

BioASQ se définit par différentes tâches, que nous passerons en revue. Cependant il faut déjà préciser que nous n'avons participé qu'à une seule d'entre elles. Relevons aussi qu'en recherche d'information, *« cette notion de tâche peut aller de la simple définition d'un but à une modélisation plus fine des actions à accomplir » (Grivel 2011, p. 72).*

Chaque expert biomédical travaillant sur l'élaboration du challenge doit formuler au moins 30 questions en anglais. Pour ce faire il s'inspire des besoins rencontrés dans sa vie professionnelle, que ce soit au niveau de la recherche ou pour un diagnostic. Chaque question doit être autonome et ne pas avoir de lien avec une autre.

Cette somme de 30 questions se répartit au travers de quatre catégories. La première contient des interrogations qui attendent des réponses oui/non. La seconde concerne des questions factoïdes, ce qui veut dire qu'en retour sont attendus des termes particuliers (une maladie, un médicament ou un gène par exemple). La troisième catégorie nécessite de fournir une liste attendue (par exemple une liste de gènes). La dernière comporte des interrogations qui requièrent de sortir un résumé à partir des informations les plus pertinentes qui ont été trouvées (Malakasiotis 2013).

A partir de cet ensemble de questions, un ensemble de termes pertinents est formé. Il est composé d'éléments attendus par les experts mais également de synonymes. Afin de récupérer l'information, la requête peut être enrichie avec des balises de recherche avancée sur PubMed avec des concepts pertinents (MeSH).

Divers modules sont testés au travers de différentes tâches qui comportent elles-mêmes différentes phases. L'idée principale de la tâche a est de classer des documents de la bibliothèque numérique PubMed vers des concepts de la hiérarchie PubMed (indexation automatique). Il faut préciser que les articles utilisés pour la compétition ne sont pas encore annotés (Balikas 2014). Notre travail ne concerne pas cette tâche mais s'intéresse en priorité à la tâche b.

La tâche b comprend une phase A qui teste la RI et une phase B qui comprend l'extraction de réponse. Pour la phase A BioASQ attend des systèmes participants qu'ils répondent par des concepts pertinents (de terminologies et d'ontologies désignées), des articles et extraits d'articles pertinents. Pour la Phase B, les systèmes participants devront répondre avec des réponses exactes (par exemple des termes précis suite à des questions factoïdes). L'évaluation de la phase B est réalisée manuellement par des experts biomédicaux (Balikas 2014).

Les mesures utilisées pour l'évaluation de la tâche b phase A (RI) par l'équipe BioASQ sont les suivantes : mean precision, mean recall, mean F-measure, mean average precision (MAP) et geometric mean average precision (GMAP). Pour la phase B (QA), il s'agit d'une part des mesures d'évaluation des réponses exactes pour chaque type de question : accuracy pour les oui/non ; strict accuracy, lenient accuracy et mean reciprocal rank (MRR) pour les factoïdes ; (mean average) precision, (mean average) recall et (mean average) F-measure pour les listes.

Pour la phase B, il s'agit d'autre part de l'évaluation des réponses idéales pour chaque question par le biais de ROUGE pour l'évaluation automatique et par le biais de scores de 1 à 5 sur quatre dimensions (recall, precision, repetition et readability) pour l'évaluation manuelle. Le détail de ses mesures est issu du texte (Malakasiotis, Pavlopoulos, Androutsopoulos 2015) et nous aide particulièrement à interpréter les résultats des différents systèmes. Des informations plus détaillées quant aux mesures qui nous intéressent spécifiquement dans le cadre de notre travail seront données au point ci-dessous sur les méthodes d'évaluation pour BioASQ.

2.4.1 Méthodes d'évaluation pour BioASQ

Nous avons évoqué brièvement plus haut l'évaluation propre à la RI. Dans ce sous-chapitre nous allons reprendre les informations définissant les mesures et calculs utilisés par les experts de BioASQ pour juger les participants au challenge. Un accent particulier sera mis sur ce qui nous intéresse directement dans le cadre de notre projet. L'aspect d'un développement en plusieurs tâches a déjà été défini, nous avons ci-dessous divisé les explications en fonction.

Tout d'abord concernant la RI :

Mean precision, mean recall et mean F-measure sont les mesures utilisées pour les tâches qui ne tiennent pas compte de l'ordre des réponses

Mean average precision MAP et geometric mean average precision GMAP pour les mesures qui tiennent compte de l'ordre des réponses (Malakasiotis, Pavlopoulos, Androutsopoulos 2015).

Ces différentes mesures ont toutes pour base les notions de précision (P) et rappel (R) inhérentes à l'évaluation des moteurs de recherche d'information. En prenant en compte l'ensemble des éléments classés selon le gold file[3] (par exemples les articles, et un ensemble d'éléments renvoyé par un des systèmes concurrents, pour une question particulière dans notre cas, P et R sont définis comme suit :

$P = TP / (TP + FP)$

$R = TP / (TP + FN)$

Les termes indiquent que : TP (vrais positifs) correspond au nombre d'articles retournés également présents dans le gold file ; FP (faux positifs) correspond au nombre d'articles retournés qui ne sont pas présents dans ce gold file et ; FN (faux négatifs) est le nombre des éléments de l'ensemble des documents issus du gold file qui ne sont pas retournés par le système.

[3] Le gold file est le fichier de référence fourni par les experts du domaine.

La F-mesure est la moyenne pondérée entre P et R. Afin de déterminer cette moyenne entre la précision et le rappel la formule suivante est utilisée :

$F = 2 \cdot (P \cdot R) / (P + R)$

Si nous prenons un ensemble de requêtes, dans notre cas les questions Q1, ….Qn, la précision moyenne, le rappel et la F-Mesure de chaque système est obtenue en calculant la précision, le rappel et la F-Mesure pour toutes les requêtes.

Pour prendre en compte une liste de retour particulière, il est commun dans la RI de calculer la précision (non-interpolée) de cette liste avec la formule suivante :

Équation 1 : Average Precision

$$AP = \frac{\sum_{r=1}^{|L|} P(r) \cdot rel(r)}{|L_R|}$$

(Malakasiotis, Pavlopoulos, Androutsopoulos 2015, p.4)

Elle se définit comme suit. |L| est le nombre d'éléments dans la liste, |LR| correspond au nombre d'éléments pertinents (pour BioASQ le nombre maximum d'éléments pertinents que les systèmes sont autorisés à retourner est 10), P(r) est la précision lorsque la liste est retournée et ne contient que le premier article ; rel (r) est égal à 1 si l'élément r-ème de la liste est dans l'ensemble "gold", par exemple s'il est pertinent autrement il correspond à 0. AP est transposé sous la forme d'une courbe rappel-précision.

Pour calculer la moyenne de AP sur un ensemble de questions, Q1, …Qn la formule suivante est utilisée:

Équation 2 : Mean Average Precision

$$MAP = \frac{1}{n} \cdot \sum_{i=1}^{n} AP_i$$

(Malakasiotis, Pavlopoulos, Androutsopoulos 2015, p.5)

Pour le QA :

Dans notre contexte nous évaluons le système seulement sous forme de réponse exacte. EAGLi ne fournit pas de réponse idéale (paragraphe). Pour les réponses exactes, on se réfère à la Strict Accuracy (SAcc ci-après qui place la bonne réponse en première position), Lenient Accuracy (LAcc ci-après qui place la bonne réponse dans une liste de 5 propositions) et Mean Reciprocal Rank MRR (avec considération du rang de la bonne réponse dans la liste).

Nous passons maintenant à l'évaluation des réponses "exactes" des questions factoïdes :

Pour chaque question factoïde, un système en compétition doit retourner une liste de cinq noms maximum, par ordre hiérarchique décroissant. L'équipe d'experts biomédicaux de BioASQ aura associé à chacune de ces questions un seul terme exact, ainsi que des synonymes possibles de ce nom. Une fois que les réponses des systèmes participants ont été soumises, ces experts vont analyser les noms des entités retournés par les systèmes

participants afin d'ajouter une nouvelle série de synonymes auxquels ils n'auraient pas pensé lors de leurs préparations des fameuses questions « gold ». Ceci permet de mesurer la stricte exactitude (SACC) et l'exactitude dite clémente (LACC) de chacun des systèmes. La stricte exactitude est valable si un terme issu de la liste d'or et des ses synonymes est le premier nom retourné par le système. En revanche la précision est dite clémente si ce terme est inclus dans la liste retournée sans être à la première place.

Dans les formules ci-dessous, n est le nombre de questions factoïdes, c1 est le nombre de questions factoïdes qui ont obtenu une réponse correcte avec la bonne réponse en première place et C5 est le nombre de bonnes réponses qui ne sont pas arrivées en première place

Équation 3 : Stric et Lenient Accuracy

$$SAcc = \frac{c_1}{n}$$
$$LAcc = \frac{c_5}{n}$$

(Malakasiotis, Pavlopoulos, Androutsopoulos 2015, p.7)

Ces précisions dites stricte et clémente seront mesurées pour l'exhaustivité. La mesure officielle pour les réponses « exactes » aux questions factoïdes considère le rang réciproque moyen (MRR) qui est souvent utilisé pour évaluer ce type de questions durant les défis de QA. Dans la formule ci-dessous, à partir de la liste renvoyée, la position du terme attendu ou un synonyme est déterminée. Si la position la plus haute est la j-ième une, alors r (i) = j ; sinon r (i) !+ 1, à savoir 1 r(i) =0

Équation 4 : Mean Reciprocal Rank

$$MRR = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{r(i)}$$

(ibidem)

MRR récompense en effet les systèmes qui ont les bonnes réponses (ou leurs synonymes) à la place la plus élevée dans les listes renvoyées.

2.4.2 Les bons systèmes au challenge BioASQ

A travers certains articles, nous découvrons quelques moteurs performants au challenge. Cependant nous bénéficions de peu d'information sur les rouages complexes de leur structure. Voici les principales informations que nous avons pu ressortir sur trois moteurs particulièrement bien classés pour la tâche b phase B de ce challenge.

Polysearch2 est le système nommé « Wishart system » (en référence au nom d'un de ses auteurs) dans la documentation trouvée sur le site de BioASQ. Pour ce système nous obtenons quelques explications de la part des auteurs. Ils décrivent le fonctionnement sur une architecture d'association. Ses performances sont expliquées notamment par l'augmentation des synonymes pris en compte et l'utilisation de nombreuses bases de données différentes (i.e. UniProt, DrugBank, HMDB, etc.) pour rechercher l'information ainsi que l'intégration de sources comme Wikipédia ou US Patent application abstracts.

Ils indiquent également des améliorations algorithmiques permettant de pénaliser les co-occurrences distantes tout en donnant plus de poids aux co-occurrences proches. (Liu, Liang, Wishart 2015)

Il faut noter, en plus, que dans sa première version, Polysearch accédait aux différentes bases de données directement via le web (API) alors que dans cette deuxième version, l'outil cherche au travers de bases de données maintenues localement. Ce procédé permet d'accélérer le processus de recherche et d'extraction de réponses.

Un autre article présente les méthodes utilisées par l'équipe du National Center for Biotechnology Information (NCBI) ainsi que leurs bons résultats. (Mao, Wei, Lu 2014). Les auteurs expliquent qu'ils cherchent à déterminer automatiquement le type de questions attendues (nombre, choix multiples, bio-concepts). Dans un second temps, ils créent des patrons d'expressions régulières afin de déterminer différents types et sous-types de réponses en fonction des catégories de questions. Au préalable, les auteurs ont développé des outils de reconnaissance automatique de concepts biomédicaux. Ces derniers sont aussi utilisés pour générer des candidats de réponses. Il s'agit, par exemple de GenNorm pour les gènes, tmChem pour les éléments chimiques, DNorm pour les maladies, SR4GN pour les espèces et tmVar pour les mutations.

Un autre moteur qui se classe assez bien est le SNUMedinfo (Seoul National University). A travers nos lectures, nous avons pu voir qu'il était toujours premier en RI documents en 2014 et très bon en réponses idéales en 2014 (trois fois 1er et deux fois 2e).

Il conserve de bonnes performances sur les mêmes dimensions en 2015. Nous donnons ici les informations disponibles à son sujet. Pour la phase B de la tâche b, les auteurs ne visent que la génération de réponses idéales (et non exactes) à partir des extraits fournis dans le gold file. Ils suivent trois méthodes heuristiques pour sélectionner n passages et les combiner pour former la réponse idéale.

- Sélection des passages les plus courts : sélection des n passages dont le nombre de tokens est minimum, classés par ordre croissant.
- Sélection des passages clefs : identification des mots clefs comme étant les mots présents dans une certaine proportion des passages (par exemple 10 sur 20), puis classement des passages en fonction du nombre de mots clefs uniques apparaissant dans chacun, et enfin sélection des n passages comportant le plus de ces mots clefs uniques.
- Sélection de passages complémentaires : méthode basée sur le classement de la méthode précédente, avec rétention du passage le mieux classé, puis vérification de la proportion de tokens qui apparaissent dans le prochain passage et pas dans le(s) passage(s) retenu(s), avec rétention ou écartement de ce prochain passage en fonction d'un seuil, et répétition du processus jusqu'à la sélection de n passages.

A la rédaction de notre document, leur équipe ne peut pas encore se prononcer sur les résultats de chacune de ces méthodes, ceux-ci n'ayant pas fini d'être évalués. Nous ne sommes malheureusement pas en mesure de déterminer quelle était la meilleure des trois stratégies. La phase A de la tâche b, bornée au document retrieval, est basée sur le moteur de recherche Indri. Le détail des paramètres testés est détaillé et les résultats sont analysés.

Le Sequential Dependence Model (SDM) et le Semantic Concept-enriched Dependence Model (SCDM) montrent de meilleures performances que la baseline, et l'avantage est plus net pour le SCDM. La référence qui suit détaille le SCDM (Choi, Choi 2014). Le Semantic Concept-enriched Dependence Model (SCDM) évoqué ci-dessus se définit comme suit :

Pour la phase A de la tâche b. Les concepts sémantiques sont identifiés dans la requête originale par le biais de MetaMap. Chaque groupe de termes de la requête appartenant à un même concept sémantique est considéré avoir des dépendances implicites avec les autres.

L'algorithme de classement est révisé pour favoriser les documents qui préservent ces dépendances implicites. Il s'agit d'incorporer ces caractéristiques de dépendance de concept sémantique enrichi dans le cadre d'un modèle de langage formel (SCDM), dont plusieurs variants sont élaborés. Les résultats montrent l'efficacité et la robustesse de la démarche. Le gain de performance est net, au contraire du modèle basé sur l'idée de dépendance séquentielle entre les termes de la requête, et il est indépendant de l'expansion de requête ainsi que de l'approche par pondération des concepts.

3. Méthodologie

Notre état de l'art a permis de situer la RI et les moteurs de QA spécifiquement par rapport au domaine du biomédical. Nous avons également cerné la structure d'EAGLi parallèlement à celui du challenge BioASQ. Nous allons maintenant décrire et développer notre approche pour mener à bien notre objectif de départ.

Il a tout d'abord fallu passer par une phase importante de dépouillement d'un grand nombre de publications du domaine et définir une bibliographie concise nous permettant d'analyser les besoins du benchmark BioASQ. Parallèlement à cette étude nous avons pris connaissance d'EAGLi, de son fonctionnement et de sa structure (Gobeill et al. 2009 ; Gobeill 2012) pour voir comment il nous était possible de mener à bien notre objectif dans le cadre du challenge BioASQ.

3.1 BioASQ

Comme nous cherchons dans ce travail à évaluer et améliorer la performance d'un moteur de QA en regard de cette compétition internationale, il a fallu prendre connaissance de toutes les informations disponibles quant à ce challenge. Elles sont nombreuses, parfois complexes, et usent régulièrement d'un vocabulaire technique. Dans un premier temps, nous avons étudié un nombre important de rapports et d'articles concernant la compétition. Ces éléments englobent les deux premières années du challenge et nous ont permis de comprendre l'ensemble tout en cernant mieux la partie QA qui nous concernait spécifiquement. Les exemples concrets donnés dans cette littérature nous ont aussi permis d'appréhender, en vue de la partie résultats, les techniques d'évaluation utilisées par BioASQ.

Comme mentionné au point 2.3. BioASQ est une compétition qui se déroule de mi-mars à mi-mai et comprend plusieurs tâches et types de questions distincts. Durant ce laps de temps, les participants ont un temps défini pour répondre aux différentes questions des différentes tâches. Nous pouvons dès lors dire que la compétition se déroule en temps réel avec une chronologie très stricte à respecter. A cette époque nous en étions aux balbutiements de notre projet et il ne nous a donc pas été possible de participer en direct. Les runs dont nous avons tiré nos résultats portent de ce fait uniquement sur des données historiques, couvrant les campagnes des années précédentes. Pour obtenir ces données, il a été nécessaire de nous inscrire sur le site internet du challenge pour procéder à leur téléchargement concernant la tâche b. Ces données comprenaient les questions et le qrel (ou gold file)[4] au format JSON. A partir du moment où nous avions le qrel des deux premières années, nous avons pu démarrer en les retravaillant. Nous avons attendu celui de la 3ème année, celle en cours, mais il est arrivé trop tard en fonction de l'avancée de notre rapport.

En mentionnant la tâche b dans le paragraphe ci-dessus, il est également important de rappeler que le périmètre de notre mandat ne couvrait pas l'ensemble des tâches des campagnes BioASQ. En effet, l'indexation sémantique (tâche a), qui vise à attribuer automatiquement plutôt que manuellement des termes MeSH aux nouveaux articles, n'est pas couverte par EAGLi. Nous nous sommes donc concentrés uniquement sur la seconde tâche, le biomedical question answering (QA). Il faut encore préciser que dans sa fonction de

[4] Fichiers donnés aux participants par les experts et comprenant les réponses correctes (attendues) aux questions d'origine.

question-réponse, EAGLi est actuellement paramétré pour le traitement des questions factoïdes uniquement. Notre mandat s'est donc limité à ce type de questions, qui représente environ un quart des 100 questions de chaque lot. Comme explicité plus haut dans ce rapport, les autres types de questions attendent une réponse par oui ou non, sous forme de liste ou sous forme de résumé.

Il a fallu ensuite procéder au tri des questions pour isoler les factoïdes (n=328 sur les 3 ans) puis procéder enfin à un second tri des questions factoïdes pour déterminer celles qui pouvaient être traitées par EAGLi. En effet, il était nécessaire de mettre à l'écart les questions numériques, de définition, d'explication, ainsi que celles dont la réponse attendue était hors du vocabulaire contrôlé car trop précise ou trop vague. C'est-à-dire ici les réponses relevant du MeSH (n=30 Q sur 2013-2014 + n=5 Q en considérant aussi les concepts supplémentaires). Cette étape a été faite de manière automatique par Julien Gobeill. A cet instant nous cherchions à savoir le nombre de questions à traiter pour obtenir des résultats statistiquement significatifs. Il a été convenu avec nos superviseurs qu'au-delà d'une trentaine de questions les résultats obtenus pourraient nous apporter du sens.

Pour finir, dans un second temps, nous avons procédé à un ultime nettoyage du qrel pour récupérer des réponses supplémentaires relevant du MeSH. En effet, nous avons constaté que comme BioASQ n'attendait pas toujours des réponses provenant d'un vocabulaire unique, il était possible que les concepts attendus par les experts ne soient pas du MeSH mais puissent y être associés sans trop de difficultés. Un exemple précis peut être donné ici avec « Fibroblast Growth Factor Receptor 3 (FRGR3) » attendu en sortie. Nous avons choisi de conserver la forme développée « Fibroblast Growth Factor Receptor 3 » sans l'acronyme, qui devient alors un terme associé au MeSH. Au final nous avons pu récupérer n=13 Q supplémentaires de 2013-2014 ce qui a porté notre échantillon total à 48 questions.

3.2 UMLS (Unified Medical Language System)

UMLS est défini sur le site de la US National Library of Medicine (NLM) (2015) comme

> *« Un ensemble de fichiers et logiciels rassemblant une grande quantité de vocabulaires et de standards dans les domaines de la santé et du biomédical permettant l'interopérabilité entre les systèmes informatiques. »*

La version d'UMLS est distribuée deux fois par an. Elle contient tous les termes de l'ensemble des terminologies ainsi que leurs types sémantiques. L'UMLS peut être obtenu gratuitement en signant une convention avec la NLM. Dans notre cas, il ne nous a pas été nécessaire de prendre directement contact avec cette institution puisque nous avons pu obtenir, par l'intermédiaire de Julien Gobeill, le MeSH avec sa mise-à-jour récente. Celle-ci, sous la forme d'un ensemble de fichiers, se nommant mshall. A partir de là, nous avons récupéré la structure arborescente des types sémantiques de mshall pour profiter d'une meilleure visibilité et naviguer plus aisément entre les différents types sémantiques et leurs concepts.

Il est important de préciser ici le choix du vocabulaire utilisé qui n'est pas anodin. En effet, il existe plusieurs vocabulaires différents dans le contexte des données biomédicales et chacun d'entre eux peut avoir son avantage. Pour n'en nommer que quelques-uns, nous pouvons donner les exemples de Gene Ontology (GO), Uniprot, DrugBank ou le MeSH qui nous intéresse ici.

L'utilisation du vocabulaire MeSH dans notre travail s'est imposée d'elle-même pour plusieurs raisons. Tout d'abord, et il s'agit de la raison principale, EAGLi dans sa version actuelle est construit pour travailler sur la base de différents vocabulaires liés au domaine mais il s'avère que le MeSH se trouve être le plus adapté au type de requêtes prises en compte par le système. Les autres terminologies telles que décrites plus haut ne sont donc, en l'état, pas exploitées par le moteur. Ceci a pu nous poser quelques difficultés dans le travail que nous avions à effectuer. Nous y reviendrons plus tard.

Ensuite, à la différence de Gene Ontology ou Uniprot, le MeSH dispose d'une bonne couverture générale des concepts biomédicaux et d'un niveau de granularité intermédiaire (pas trop fin, ce qui relèverait du travail de spécialiste). Ceci permettant de construire les questions, ou requêtes, au plus proche de ce que peut représenter la réalité. En effet, même si à priori cet outil est à destination d'un public de professionnels de la santé, il paraît peu probable que tous les utilisateurs formulent des questions en usant d'un vocabulaire extrêmement pointu. Pour finir, n'étant pas nous-même, ou en tout cas pour deux d'entre nous, des experts du domaine biomédical, le MeSH représentait une terminologie relativement facile à appréhender et à comprendre. Il reste à préciser que la prise en main d'une terminologie médicale nous a tout de même demandé un certain temps.

Comme défini dans l'état de l'art, le MeSH est un vocabulaire contrôlé, faisant ainsi partie de la catégorie des données structurées. Pourtant, nous l'avons abordé dans le point 3.1 ci-dessus, BioASQ attend parfois comme réponses des concepts qui sont rattachés à un vocabulaire libre. En ce sens, l'utilisation du MeSH dans notre travail nous permet d'établir des correspondances précises entre les réponses attendues et les réponses renvoyées par EAGLi, tout en utilisant un langage qui nous soit accessible dans sa compréhension et qui corresponde à l'architecture du système que nous cherchons à tester. Ainsi, nous avons pu débuter ce qui allait être une des pièces maîtresses de notre travail, à savoir la reformulation des questions et la mise en place des éléments nécessaires pour atteindre notre objectif.

3.3 EAGLi

Concernant EAGLi en lui-même, il a fallu étudier son fonctionnement à la lumière de lectures telles que Gobeill (2012) ; Gobeill et al. (2009) ainsi que par les quelques exemples disponibles en ligne. En effet, la page d'accueil du site[5]nous propose une liste déroulante comprenant une dizaine d'exemples explicitant les formulations de questions acceptées par le système. Ensuite, lorsque la recherche est lancée les résultats sont retournés par ordre de confiance. Avoir un visuel sur les formulations correctes et sur les résultats renvoyés nous a, il est vrai, beaucoup aidé afin de comprendre comment il nous était possible de travailler à partir de là. Nos superviseurs nous ont mis à disposition un espace sur leur serveur « casimir » d'où sont tirées toutes les captures d'écran ci-après.

3.3.1 Patrons de questions et « negatives »

EAGLi fonctionne sur une architecture qui peut paraître complexe au premier abord et pour laquelle un certain temps nous a été nécessaire pour en appréhender les bases. En effet, outre les trois composants principaux évoqués plus haut (catégoriseur de questions, moteur de RI et moteur de QA), le moteur nécessite des informations supplémentaires pour pouvoir fonctionner. Il s'agit ici des fichiers que l'on nommera

[5] http://eagl.unige.ch/eagli/

« patrons de questions » et ***« negatives »*** principalement. Les ***« patrons de questions »*** sont les fichiers qui indiquent au moteur dans quelle partie du vocabulaire MeSH il doit aller chercher la réponse à la question qu'on lui pose. Les ***« negatives »*** indiquent pour leur part les termes à exclure lors de cette recherche pour différentes raisons. Par analogie, les ***« patrons de questions »*** pourraient s'apparenter aux coordonnées géographiques du concept que nous cherchons à obtenir comme réponse (sa localisation) et les ***« negatives »*** à l'opérateur **« NOT »** dans une recherche en mode booléen pour les familiers de ce type de langage.

Dans leur construction, les ***« patrons de questions »*** se présentent de la sorte :

- Ils doivent commencer par un pronom interrogatif matché *« What | Which »* voir *« Where »*.
- Le pronom interrogatif doit être suivi par une cible (un mot en langage naturel)
- Ils doivent se terminer par le numéro d'identification du type sémantique MeSH.

Les types sémantiques du MeSH utilisés dans ces ***« patrons »*** contiennent les concepts potentiellement pertinents pour répondre à une question. Le numéro du type sémantique sert de numéro d'identification unique. Par exemple, dans la figure 2 ci-dessous, le type sémantique ***« T007 »*** rattaché à une question concernant les bactéries contiendra des termes comme ***« Escherichia coli »*** ou ***« Herellea »***.

Figure 2 : Patrons de questions

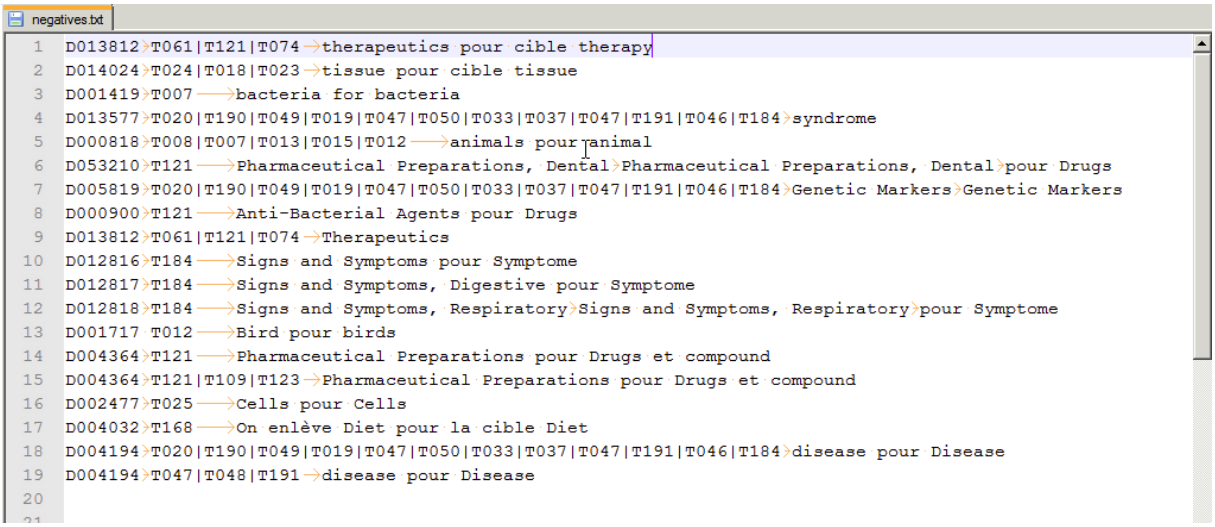
BiomedQuestionsPattems4students.txt

```
// le format est "[pronom|*]\t[nom|*]\type[\|type]*"
// les pronoms interrogatifs matchés sont what|which, et where
what|which→abnormality→T190|T018|T019|T047
what|which→abnormalities→T190|T018|T019|T047
what|which→activities→T052|T053|T038|T039|
what|which→activity→T052|T053|T038|T039|
what|which→administrative procedure→T061|T059
what|which→administrative procedures→T061|T059
what|which→alteration→T046|T047|T048|T191|T049
what|which→alterations→T046|T047|T048|T191|T049
what|which→amino acid→T116
what|which→amino acids→T116
what|which→amphibian→T011
what|which→animal→T008|T009|T010|T011|T012|T014|T013|T015|T012
what|which→anorexia→T047|T048→free comment encore pour essayer
what|which→antibodies→T129
what|which→antibody→T129
what|which→arrhythmia→T046|T047
what|which→bacteria→T007
what|which→bacterium→T007
what|which→behavior→T053|T054|T055
what|which→behaviors→T053|T054|T055
what|which→behaviour→T053|T054|T055
what|which→behaviours→T053|T054|T055
what|which→binding site→T085|T086|T087
what|which→binding sites→T085|T086|T087
what|which→biologic function →T038|T039|T040|T041|T042|T043|T044|T045
what|which→biologic functions →T038|T039|T040|T041|T042|T043|T044|T045
what|which→biomarker→T116|T129
what|which→biomarkers→T116|T129
```

Comme évoqué ci-dessus, le fichier ***« negatives »*** indique au système quels termes exclure lors de la recherche. Prenons un exemple simple. Imaginons que nous recherchons un type de cellules (ex. Which cells …?) Le type sémantique ***« T025 »*** contient des termes comme ***« Erythroblast »*** ou ***« Red Blood Cells »*** mais il contient aussi tout simplement le terme

« Cells » lui-même. Il paraît évident qu'à une question commençant par Which cells, nous souhaiterions obtenir une réponse autre que le terme ***« Cells »***. Ceci sera défini dans le fichier ***« negatives »*** donné en exemple à la figure 3. A la ligne 16 de ce fichier, nous pouvons voir à la première colonne l'identifiant unique MeSH D002477 qui identifie le concept ***« Cells »*** de manière univoque. Nous pouvons constater à la colonne deux que ce concept fait lui-même partie du type sémantique ***« T025 »***. Par ce moyen nous indiquons au système que pour une question impliquant le type sémantique T025 il faudra exclure des réponses possibles le concept *« Cells »*.

Figure 3 : Fichier des negatives



Après avoir compris un peu plus en détails sur quelles bases fonctionnait EAGLi, il nous a été possible de démarrer la construction et la reformulation des questions fournies par BioASQ.

3.3.2 Reformulation et complétion des patrons de questions

Concrètement, si nous prenons un exemple parmi les 330 questions factoïdes du challenge, cela nous donne la composition suivante :

"Where is the histone variant CENPA preferentially localized ?"

Il s'agit ici de la question originale fournie par les experts de BioASQ.

A partir de ce type d'élément, nous devions reformuler la phrase afin de l'adapter à la structure d'EAGLi qui, même s'il est théoriquement capable de prendre en compte les pronoms interrogatifs *« Where ? »,* a tout de même beaucoup de difficultés à ressortir une réponse correcte, pour ne pas dire qu'il n'y arrive pas. Le procédé de reformulation découle d'un schéma type qui se définit comme suit :

"What | Which +1 cible + 1 verbe + 1 complément + 1 ?"

Ces éléments définissent les termes de la recherche. Il est important de préciser également que pour avoir une reformulation correcte qui apporte des résultats, un des éléments à respecter est le fait que la cible définie dans la question doit impérativement respecter la cible donnée dans le *« patron »* correspondant dans le fichier *« patron de questions »* ci-dessus. Reformulations et patrons de questions étant ainsi étroitement liés, nous avons aussi créé et ajouté de nouveaux patrons pour augmenter la couverture du système et ainsi tenter

d'améliorer ses performances. Cette partie sera développée plus en détails dans le point 4 Résultats.

Par exemple, en sachant que le patron suivant *« what | which cell component T026*[6] *»* est défini dans le fichier, si nous reprenons la phrase issue du challenge BioASQ indiquée plus haut, après notre intervention manuelle elle aura alors, par exemple, la forme suivante :

"Which cell component has the histone variant CENPA?"

3.3.3 Test par soumission au catégoriseur de questions

A partir de ce travail intermédiaire que nous avons opéré sur une sélection de questions, l'étape suivante a consisté à soumettre certaines d'entre elles à EAGLi. Julien Gobeill nous a ainsi fourni l'accès au catégoriseur de questions en ligne avec lequel il était possible de passer les arguments directement dans l'URL. Ceci pour définir si, dans un deuxième temps, notre reformulation correspondait bien à la structure acceptée par le système. C'est seulement lors des étapes suivantes que nous avons pu soumettre notre lot complet à EAGLi pour obtenir les résultats finaux.

Figure 4 : Résultat du catégoriseur de questions

```
− <output>
    <isQuestion>true</isQuestion>
    <understandQuestion>true</understandQuestion>
    <typeQuestion>factoid</typeQuestion>
    <QAtarget>T026</QAtarget>
    <canonicalQuery>histone variant CENPA</canonicalQuery>
    <whatWord>which</whatWord>
    <subject>cell component</subject>
  − <taggedPhrase>
      which/WDT cell/NN component/NN has/VBZ the/DET histone/NN variant/NN CENPA/NNP ?/PP
    </taggedPhrase>
  − <brut>
      <wdt>which</wdt>
      <nn>cell</nn>
      <nn>component</nn>
      <vbz>has</vbz>
      <det>the</det>
      <nn>histone</nn>
      <nn>variant</nn>
      <nnp>CENPA</nnp>
      <pp>?</pp>
    </brut>
  </output>
```

Les principaux éléments à commenter dans la figure 4 ci-dessus sont les suivants :

- Tout d'abord le système reconnaît qu'il s'agit bien d'une question et qu'il la comprend dans les parties <isQuestion>true</isQuestion> et <understandQuestion>true<understandQuestion/>
- Il reconnait le type sémantique et le sujet de la question dans les balises <QAtarget>T026</QAtarget> et <subject>cell component<subject/>

[6] Le type sémantique T026 compte environ 1'500 termes MeSH

- Entre les balises <brut>…</brut> dans le bas de la figure, le système analyse et comprend la structure syntaxique de la question qui correspond au schéma type de reformulation explicité plus haut.

Après avoir validé la structure de la plupart de nos reformulations au moyen du catégoriseur de questions en ligne, il nous était désormais possible de soumettre l'ensemble de celles-ci, sous forme de lot, à EAGLi. La partie suivante de ce rapport explique les différentes étapes de cette soumission en lot pour finalement conclure sur l'évaluation des résultats au moyen de la méthodologie trec_eval.

3.3.4 Soumission par lot au système EAGLi

Pour effectuer cette soumission, nous avons pu nous appuyer sur l'aide de Julien Gobeill qui nous avait développé un script automatisé. Le fonctionnement, assez simple, prend en compte les éléments suivants :

- Il faut donner en entrée au système le fichier « input.txt » comprenant l'ensemble des questions auxquelles répondre
- Il faut préciser le nombre de documents avec lesquels travailler (le nombre de documents dans lesquels le système ira chercher une réponse)
- Préciser le mode de recherche (booléen ou vectoriel)

Figure 5 : Extrait du fichier input.txt

A la figure 5 ci-dessus, nous pouvons voir sur la même ligne, en première position la question originale qui sera suivie par son numéro d'identification unique, tous deux fournis par BioASQ. Par la suite on retrouvera en troisième position notre reformulation ainsi que le patron correspondant. Précisons également que ces différents éléments doivent obligatoirement être séparés par des tabulations, sans quoi les questions ne seront pas reconnues et tout

simplement inexistantes du fichier final de résultats. Une fois la vérification faite sur l'intégrité du fichier, il est dès lors possible de faire tourner EAGLi selon le mode choisi.

A partir de cette opération, la RI, le composant d'extraction de document, retourne les identifiants PubMed (PMID) des documents jugés pertinents, c'est à dire ceux dont le résumé contient les termes qui sont attendus par notre requête (figure 6). Pour terminer, le composant d'extraction d'information, QA, retourne les réponses, les termes MeSH des résumés sélectionnés (figure 7).

Figure 6 : Fichier RI.res

RI_bool_48.res

```
5118dd1305c10fae75000001 dummy 26630011 1 999 dummy
5118dd1305c10fae75000001 dummy 26629984 2 998 dummy
5118dd1305c10fae75000001 dummy 26629845 3 997 dummy
5118dd1305c10fae75000001 dummy 26629366 4 996 dummy
5118dd1305c10fae75000001 dummy 26629364 5 995 dummy
5118dd1305c10fae75000001 dummy 26629363 6 994 dummy
5118dd1305c10fae75000001 dummy 26629181 7 993 dummy
5118dd1305c10fae75000001 dummy 26629129 8 992 dummy
5118dd1305c10fae75000001 dummy 26628988 9 991 dummy
5118dd1305c10fae75000001 dummy 26628706 10 990 dummy
5118dd1305c10fae75000001 dummy 26628608 11 989 dummy
5118dd1305c10fae75000001 dummy 26628607 12 988 dummy
5118dd1305c10fae75000001 dummy 26628605 13 987 dummy
5118dd1305c10fae75000001 dummy 26628599 14 986 dummy
5118dd1305c10fae75000001 dummy 26628597 15 985 dummy
5118dd1305c10fae75000001 dummy 26628596 16 984 dummy
5118dd1305c10fae75000001 dummy 26628594 17 983 dummy
5118dd1305c10fae75000001 dummy 26628312 18 982 dummy
5118dd1305c10fae75000001 dummy 26627986 19 981 dummy
5118dd1305c10fae75000001 dummy 26627497 20 980 dummy
5118dd1305c10fae75000001 dummy 26625200 21 979 dummy
5118dd1305c10fae75000001 dummy 26624970 22 978 dummy
5118dd1305c10fae75000001 dummy 26623013 23 977 dummy
5118dd1305c10fae75000001 dummy 26622511 24 976 dummy
5118dd1305c10fae75000001 dummy 26622470 25 975 dummy
```

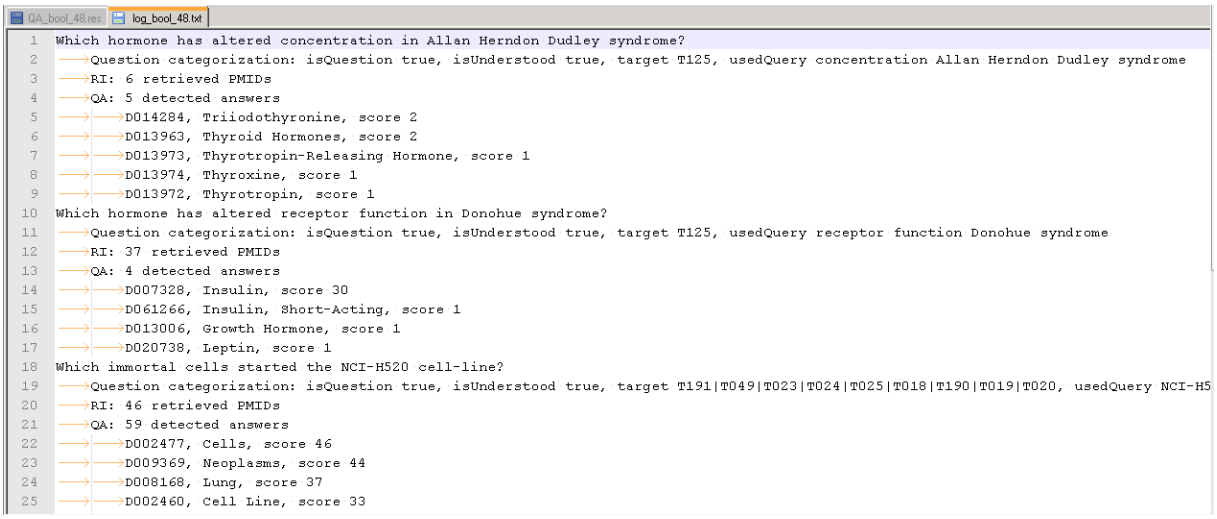
Figure 7 : Fichier QA*.res

QA_bool_48.res

```
512d0e635274a5fb07000005 dummy D008727 1 999 dummy
512d0e635274a5fb07000005 dummy C108577 2 998 dummy
512d0e635274a5fb07000005 dummy D018501 3 997 dummy
512d0e635274a5fb07000005 dummy D001685 4 996 dummy
512d0e635274a5fb07000005 dummy D009569 5 995 dummy
5139b31dbee46bd34c000004 dummy D012598 1 999 dummy
5139b31dbee46bd34c000004 dummy D009103 2 998 dummy
5139b31dbee46bd34c000004 dummy D004194 3 997 dummy
5139b31dbee46bd34c000004 dummy D007249 4 996 dummy
5139b31dbee46bd34c000004 dummy D001244 5 995 dummy
51405cd123fec90375000005 dummy D000519 1 999 dummy
51405cd123fec90375000005 dummy D005696 2 998 dummy
51405cd123fec90375000005 dummy D011506 3 997 dummy
51405cd123fec90375000005 dummy C018589 4 996 dummy
51405cd123fec90375000005 dummy D002787 5 995 dummy
5147c088d24251bc05000026 dummy D055752 1 999 dummy
5147c088d24251bc05000026 dummy D018288 2 998 dummy
5147c088d24251bc05000026 dummy D010257 3 997 dummy
5147c088d24251bc05000026 dummy D008175 4 996 dummy
5147c088d24251bc05000026 dummy D006689 5 995 dummy
5149f494d24251bc0500004c dummy D009270 1 999 dummy
5149f494d24251bc0500004c dummy D001569 2 998 dummy
5149f494d24251bc0500004c dummy D003932 3 997 dummy
5149f494d24251bc0500004c dummy D009020 4 996 dummy
5149f494d24251bc0500004c dummy D003061 5 995 dummy
```

Ces deux fichiers étant de prime abord difficilement lisible par un être humain, nous avons pu aussi nous appuyer sur un fichier de log (figure 8), plus clair, pour nous aider à comprendre ce qui fonctionnait plus ou moins bien, et ainsi assurer le suivi de nos résultats.

Figure 8 : Fichier log*.txt



3.3.5 Suivi des premiers résultats

Lors des premières tentatives, nous avions une liste de 31 questions reformulées ainsi qu'un dossier comportant encore 17 phrases susceptibles d'être retravaillées en fonction de termes MeSH supplémentaires.

Nous avons lancé un premier run afin de tester notre première liste, ce qui a donné une première analyse. En mode Booléen, nous avons obtenu trois questions déclarées comme telles mais non comprises. Elles apparaissaient cependant dans notre fichier de log. Nous avions par contre cinq questions qui n'étaient absolument pas comprises et n'apparaissaient plus du tout. Ce premier test a donc donné un lot de 26 questions qui étaient acceptées, dont trois non comprises par le moteur.

A ce stade, un simple réglage des tabulations en trop a suffi pour que l'ensemble de ce lot soit accepté, avec cette fois deux questions ajoutées aux trois non comprises. Pour ces dernières, le moteur retourne toujours la même indication : « is question true, is understood false ». Nous étions quand même rassurés d'avoir un log complet, ce qui nous indiquait que la structure de nos reformulations étaient correctes dans une certaine mesure, même si des réponses ne pouvaient être trouvées à toutes les questions.

Pour ce qui est du test en vectoriel, lorsqu'une question est non comprise le système se bloque. Nous nous sommes rendu compte qu'il ressortait trois questions qui n'étaient pas reconnues comme telle. Nous avons fait une nouvelle copie de notre lot en supprimant ces dernières et avons obtenus la prise en compte des 28 autres, dont deux à nouveau non comprises comme explicité avant.

A partir de ces éléments, nous avons regardé avec Julien Gobeill ce qui pouvait être source d'erreur et de blocage. Il ne s'agissait en fait que de fautes de rédaction, comme des tabulations en trop ou l'emplacement de majuscules inadéquates.

Dès lors que nous savions rédiger définitivement les bonnes phrases interrogatives nous avons décidé de reprendre également la liste de 17 questions potentiellement améliorables. Cet ajout nous a permis de déterminer qu'un lot d'une quarantaine de questions serait suffisant pour mener à bien notre phase de test complète. Cette proposition d'échantillon a été validée par notre mandant, Patrick Ruch.

Nous avons dissocié nos deux listes dans un premier temps, en lançant d'une part les 31 questions correctement rédigées, qui passaient toutes cette fois ainsi que notre lot de 17 questions que nous venions de retravailler. Le premier lot a obtenu de bons scores dès le premier lancement. Notre deuxième lot obtenait par contre de mauvais résultats que nous avons analysés. Il en découle que la faiblesse venait d'un problème de formulation et du sens de la phrase. Les autres éléments étaient corrects, y compris les patrons. Nous avons donc effectué un travail de vérification afin de déterminer si les questions que nous venions de reformuler correspondaient bien au sens biologiquement parlant de la question d'origine.

En passant, nous pouvons voir à ce stade que nous nous éloignons de la logique du challenge BioASQ. En effet, les résultats que nous cherchons à obtenir permettent de définir la voie à prendre pour élaborer nos questions. Cette action s'opère avec une série de ratés suivis de rectifications. Notre mode opératoire n'est donc pas compatible avec une compétition en directe comme l'impose BioASQ.

La suite du programme s'est poursuivie en faisant à nouveau deux lots. Un premier qui comporte l'ensemble des questions acceptées et donnant lieu à des réponses satisfaisantes et un second avec les questions qui n'étaient pas adéquates. Nous avons pris ce dernier lot, en recherchant les passages indiqués comme « false » afin de tester une par une les phrases problématiques avec le catégoriseur de questions. Ceci pour déterminer précisément l'origine des erreurs. Dans certains cas, il s'agissait du verbe qui ne passait pas, ou dans d'autres, les termes s'avéraient non compris. Nous avons également, à nouveau, retrouvé des problèmes de tabulations. Ces questions, une fois retravaillées correctement et testées à nouveau, ont donné de bons résultats. A ce stade nous avions donc un lot unique de 48 questions exploitables, que nous avons testé avec deux runs complets (1 booléen et 1 vectoriel).

Nous avons aussi fait le choix de conserver un fichier sans formulation. Ce dernier avait pour objectif d'évaluer les résultats sur les données brutes, selon le modèle avant-après comme nous en avions discuté avec nos superviseurs. Nous n'avons pas réussi à faire un run en vectoriel pour ce dossier « sans Formulation » car le système bloque sur des questions qu'il ne comprend pas et tourne sans fin dans le vide. Pour le booléen cela fonctionne mais, sans surprises, comme les questions ne sont pas reformulées il renvoie beaucoup de résultats indiquant que ce n'est pas une question ou qu'il ne la comprend pas. Par la suite nous avons transmis ces fichiers à Julien Gobeill afin d'obtenir des résultats en utilisant la méthodologie trec_eval.

3.4 trec_eval

Afin d'évaluer nos résultats par méthodologie trec_eval, plusieurs informations différentes sont nécessaires. En effet, comme évoqué précédemment, nous devons disposer d'une part du qrel (ou gold file) fourni par les experts de BioASQ et d'autre part, bien sûr, de nos fichiers de résultats retournés par EAGLi.

Pour les fichiers de qrel (figures 9 et 10), nous disposions des fichiers originaux de BioASQ mais ceux-ci étaient au format JSON et structurés différemment de nos fichiers de résultats au niveau de leur contenu. Pour les fichiers qrel du QA et de la RI, il a donc fallu récupérer les identifiants uniques de chaque question **(1)**. Ensuite, spécifiquement pour chacun des deux fichiers, il a fallu récupérer en plus les identifiants uniques des articles retrouvés dans Pubmed (PMID) pour la RI **(2)**. Les identifiants uniques MeSH des concepts attendus en réponse pour le QA **(3)**. Pour cette étape, comme nous l'avons déjà évoqué, il a été nécessaire d'associer manuellement des concepts MeSH lorsque les réponses attendues n'y correspondaient pas. Il faut aussi préciser que pour le fichier QA, il était nécessaire d'ajouter également l'identifiant unique établissant le lien entre le fichier qrel du QA et de la RI pour chaque question **(4)**.

Figure 9 : Qrel RI

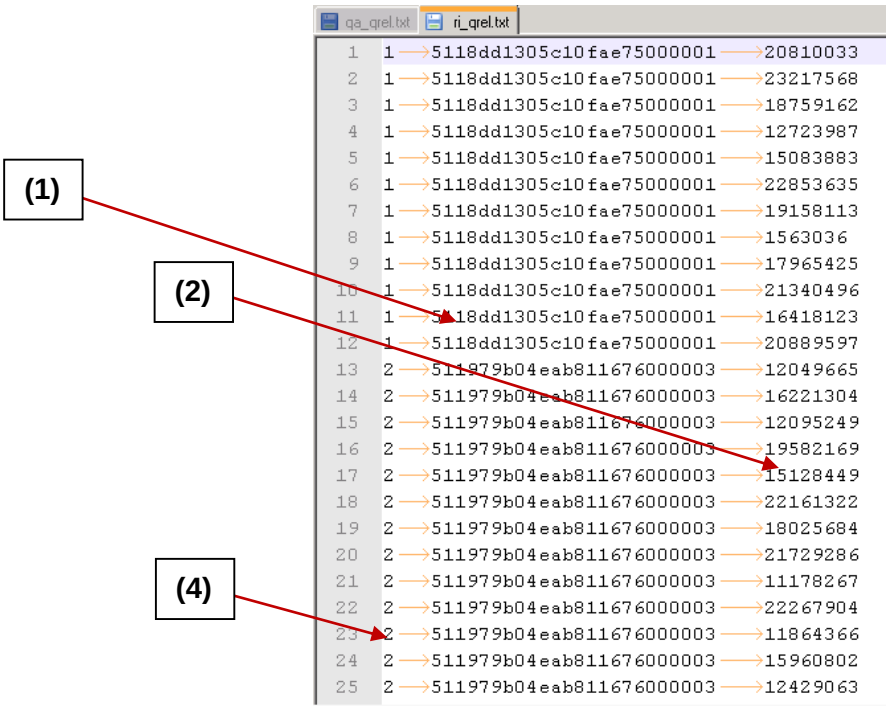
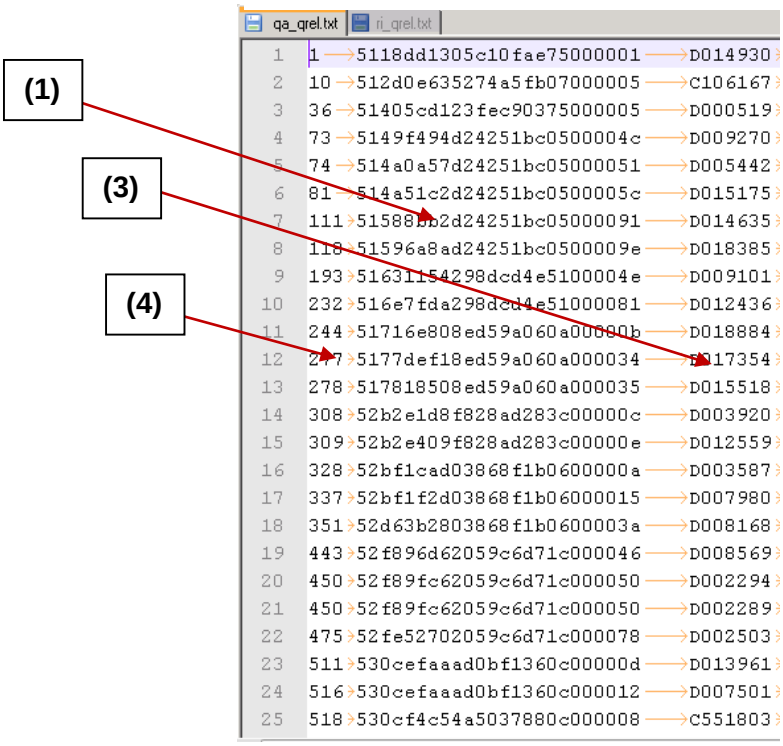


Figure 10 : Qrel QA



Il était dès lors possible de comparer les fichiers de qrel avec nos fichiers de résultats afin de mesurer finalement l'étendue de notre travail. Pour mener à bien cette comparaison, les fichiers doivent correspondre au niveau de leur nommage. Julien Gobeill a de ce fait écrit un script permettant en premier lieu de générer de manière automatique les fichiers de qrel pour la RI et le QA, en prenant comme sources les fichiers de qrel d'origine et nos fichiers de résultats. Ceci en y ajoutant une étape de renommage. Concrètement, si dans notre dossier nous disposons des fichiers suivants : qa_qrel.txt et ri_qrel.txt (les deux fichiers de qrel fournis par BioASQ), QA.res et RI.res (les deux fichiers de résultats du moteur EAGLi), alors après exécution du script nous aurons un dossier contenant les fichiers suivants : QA.qrel et RI.qrel ainsi que QA.res et RI.res. Par la suite, pour permettre la comparaison, les fichiers doivent également correspondre au niveau de la mise en forme de leur contenu. Deux exemples sont disponibles aux figures 11 et 12.

Figure 12 : Qrel QA réécrit automatiquement

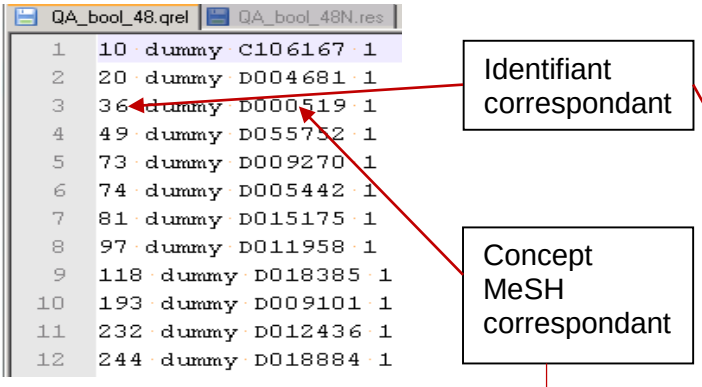
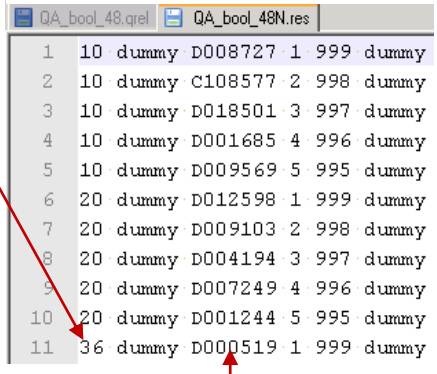


Figure 11 : Résultats QA réécrits automatiquement



Nous remarquons aux figures 11 et 12 ci-dessus une correspondance de concept MeSH entre le fichier de qrel et notre fichier de résultats. Cela indique que la réponse est comptabilisée comme correcte pour cette question. Trec_eval va maintenant pouvoir analyser de manière automatique l'ensemble du fichier, question par question, pour comptabiliser toutes les réponses et nous fournir les statistiques finales données à la figure 13.

Figure 13 : Exemple de statistiques trec_eval

```
Queryid (Num):          42
Total number of documents over all queries
    Retrieved:         208
    Relevant:           43
    Rel_ret:            27
Interpolated Recall - Precision Averages:
    at 0.00          0.4544
    at 0.10          0.4544
    at 0.20          0.4544
    at 0.30          0.4544
    at 0.40          0.4544
    at 0.50          0.4544
    at 0.60          0.4544
    at 0.70          0.4544
    at 0.80          0.4544
    at 0.90          0.4544
    at 1.00          0.4544
Average precision (non-interpolated) over all rel docs
                     0.4544
```

4. Résultats

Dans cette partie nous abordons les principaux résultats que nous avons obtenus, d'une part au regard des meilleurs systèmes actifs dans le challenge BioASQ pour la phase B de la tâche b, d'autre part sans comparaison avec ces meilleurs systèmes. Nous évoquons aussi le nombre de questions reformulées ainsi que l'ensemble des patrons de questions, ajoutés au fichier d'origine.

4.1 Reformulation des questions

La reformulation des questions et la création des patrons sont étroitement liées. En effet, comme ceci a été abordé dans la méthodologie, la cible comprise dans la question reformulée soumise au système doit impérativement se retrouver sous la même forme dans le patron correspondant. Aucune différence ne doit être perceptible, y compris en termes orthographique. Pour reprendre un exemple cité plus haut dans ce rapport, si nous souhaitons obtenir une réponse à une question au sujet de cellules en considérant la forme singulière et la forme plurielle, il faudra alors créer deux patrons distincts. Ceux-ci seront par exemple *« What | Which Cell et What | Which Cells »*. Les patrons doivent aussi se trouver sur deux lignes distinctes.

En retirant les questions autres que celles attendant des réponses exactes ainsi que celles dont les concepts paraissaient vraiment trop obscurs pour les non-professionnels de la médecine que nous sommes, sur un total de 330 questions factoïdes potentiellement traitables par EAGLi sur l'ensemble des données historiques, une reformulation a été faite pour un grand nombre d'entre elles mais testée effectivement pour seulement 48, soit 14.5%.

En mode booléen, 42 questions sur 48 ont été comprises par le système (88%) et en mode vectoriel 46 sur 48 (96%). Il s'agit ici uniquement des questions qu'EAGLi a réussi à traiter. Le fait qu'elles aient été comprises n'implique pas pour autant que les réponses étaient correctes dans tous les cas. Les résultats concernant la validité ou non des réponses seront abordés dans la partie sur les métriques de nos résultats. Nous remarquons au passage qu'en mode vectoriel 4 questions supplémentaires sont correctement comprises et traitées (8%). Encore une fois, cela n'indique pas forcément que les réponses étaient correctes.

4.2 Complétion du fichier de patrons de questions

Comme nous l'avons indiqué au point ci-dessus, après une prospection de toutes les questions factoïdes traitables, nous avons donc débuté les reformulations. Reformulation et création de patrons étant liées, nous avons ainsi augmenté les patrons lorsque ceux-ci n'étaient pas déjà présents dans le fichier d'origine. Cette démarche permettait de tenter autant que possible d'élargir la couverture du système et sa capacité de réponse. Comme nous imaginions au départ pouvoir tester plus de 48 questions au total, à chaque reformulation faite, nous avons rajouté un patron s'il n'était pas présent. Cette démarche implique que sur un total de 330 questions, 195 patrons ont été créés et rajoutés, en incluant les variations orthographiques (ex. singulier ou pluriel) pour une même question mais aussi des patrons supplémentaires pour les nouvelles questions. Une explication supplémentaire sur ces chiffres sera donnée dans notre partie discussion.

4.3 Évaluation des résultats et métriques

Nous l'avons évoqué au point 2.4.1, les méthodes d'évaluation pour BioASQ concernant les réponses exactes aux questions factoïdes sont les suivantes :

- Strict Accuracy (en considérant la réponse attendue en première position des résultats renvoyés)
- Lenient Accuracy (en considérant la réponse attendue comme faisant partie d'une liste de 5 réponses renvoyées mais sans considérer son rang dans la liste)
- Mean Reciprocal Rank (en considérant le rang de la réponse attendue dans la liste des propositions renvoyées)

Notre choix concernant l'évaluation de nos résultats s'est porté sur la méthodologie TREC en utilisant l'outil à disposition trec_eval. Ce choix a été motivé par plusieurs raisons dont la première est une méthodologie robuste et déjà éprouvée de la part de TREC en matière d'évaluation de la recherche d'information. Ensuite, l'outil trec_eval nous paraissait, dans une certaine mesure, le plus simple à appréhender en raison du fait que nous avions pu déjà étudier son fonctionnement durant nos cours sur la RI. Enfin, il semblait de prime abord mieux correspondre aux principales contraintes qui nous étaient imposées durant ce projet. Ne pas pouvoir participer en direct aux sessions d'évaluation de BioASQ d'une part et ne participer qu'à une seule phase d'une des deux tâches de la campagne d'autre part. Il nous paraissait ainsi plus simple de pouvoir obtenir uniquement les métriques qui nous intéressaient dans ce contexte mais aussi et surtout celles qui apporteraient du sens à notre travail en permettant de nous comparer aux meilleurs systèmes du challenge.

Il reste à préciser que sur les trois métriques principales définies plus haut concernant l'évaluation des réponses exactes aux questions factoïdes, trec_eval ne fournit pas la Strict Accuracy et la Lenient Accurary mais la Mean Reciprocal Rank oui. Trec_eval est en effet dans une certaine mesure un outil plus adapté aux exigences de l'évaluation de la RI (renvoi de documents) qu'à celles du QA (renvoi de réponses). Toutefois, et cela est précisé dans la documentation de BioASQ comme défini plus en avant dans notre rapport, la Strict Accuracy et la Lenient Accuracy s'avèrent utiles pour tendre à l'exhaustivité des résultats mais la mesure officielle utilisée pour l'évaluation est bien la Mean Reciprocal Rank. En ce sens, nous disposons de tout le nécessaire pour évaluer nos résultats.

4.3.1 Métriques et résultats avec comparaison

Nous présentons ici les métriques obtenues en comparaison aux deux meilleurs systèmes ayant participé à BioASQ sur les données historiques que nous avons traitées en travaillant avec 50 puis 100 et enfin 200 documents. Ces chiffres sont valables pour les réponses exactes aux questions factoïdes uniquement.

Les informations retrouvées dans la littérature nous indiquent que le système « Wishart » nommé dans les résultats BioASQ est en fait associé au moteur Polysearch2 décrit dans notre état de l'art. Concernant « Ming » il semblerait que ce soit l'équipe NCBI mais aucune indication précise à l'heure actuelle ne nous permet de l'affirmer avec certitude.

Tableau 1 : Résultats comparés aux meilleurs systèmes BioASQ (50 docs)

	EAGLi	Wishart		Ming	
Métrique : MRR	**Données historiques sur deux ans**	**1ère année**	**2ème année**	**1ère année**	**2ème année**
Booléen*	0.45	0.31	0.46	Pas participé	0.16
Vectoriel*	0.42	0.31	0.46	Pas participé	0.16

Tableau 2 : Résultats comparés aux meilleurs systèmes BioASQ (100 docs)

	EAGLi	Wishart		Ming	
Métrique : MRR	**Données historiques sur deux ans**	**1ère année**	**2ème année**	**1ère année**	**2ème année**
Booléen*	0.43	0.31	0.46	Pas participé	0.16
Vectoriel*	0.41	0.31	0.46	Pas participé	0.16

Tableau 3 : Résultats comparés aux meilleurs systèmes BioASQ (200 docs)

	EAGLi	Wishart		Ming	
Métrique : MRR	**Données historiques sur deux ans**	**1ère année**	**2ème année**	**1ère année**	**2ème année**
Booléen*	0.43	0.31	0.46	Pas participé	0.16
Vectoriel*	0.41	0.31	0.46	Pas participé	0.16

*Les résultats booléen et vectoriel ne sont valables que pour EAGLi. Les valeurs sont donc identiques concernant la première et la deuxième année de Wishart. Ming n'a pas participé la première année.

4.3.2 Métriques et résultats sans comparaison

Nous présentons ici les métriques obtenues sans comparaison aux deux meilleurs systèmes ayant participé à BioASQ sur les données historiques que nous avons traitées en travaillant avec 50, puis 100 et enfin 200 documents. Ces chiffres sont valables pour les réponses exactes aux questions factoïdes uniquement.

Les valeurs prises en compte dans ces résultats comprennent, en plus de la Mean Reciprocal Rank, les valeurs de rappel obtenues pour chacun des deux modes de recherche, booléen et vectoriel.

Tableau 4 : Résultats sans comparaison aux meilleurs systèmes BioASQ (50 docs)

	EAGLi	
Données historiques sur deux ans	**Booléen**	**Vectoriel**
MRR	0.45	0.42
Rappel	0.64	0.69

Tableau 5 : Résultats sans comparaison aux meilleurs systèmes BioASQ (100 docs)

	EAGLi	
Données historiques sur deux ans	**Booléen**	**Vectoriel**
MRR	0.43	0.41
Rappel	0.64	0.67

Tableau 6 : Résultats sans comparaison aux meilleurs systèmes BioASQ (200 docs)

	EAGLi	
Données historiques sur deux ans	**Booléen**	**Vectoriel**
MRR	0.43	0.41
Rappel	0.64	0.67

4.3.3 Commentaires sur les résultats

Dans ces tableaux nous remarquons d'une part qu'EAGLi, sur les deux ans, obtient des valeurs supérieures aux deux moteurs concurrents pour la première année. Il se trouve légèrement inférieur à Wishart qui obtient de bons résultats pour la seconde. EAGLi reste toutefois très proche de ce système la deuxième année. Nous observons toutefois une légère baisse (en booléen et en vectoriel) en augmentant le nombre de documents.

Une petite différence est à soulever en fonction du mode de recherche choisi. Le mode booléen nous montre en effet une valeur légèrement supérieure au mode vectoriel pour la Mean Reciprocal Rank alors même que, rappelons-le, c'est en mode vectoriel qu'un plus grand nombre de questions sont comprises par EAGLi (+8%). Cela nous laisse à penser que pour un plus grand nombre de questions interprétées en mode vectoriel, les réponses fausses renvoyées par le système sont plus importantes. Nous constatons par contre, dans le tableau des résultats sans comparaison, que la couverture de rappel reste plus importante en mode vectoriel, même si elle diminue légèrement en augmentant le nombre de documents.

5. Discussion

A ce point précis du travail et compte tenu de notre expérience pratique dans le domaine de l'évaluation de la recherche d'information, de la connaissance du domaine biomédical que nous possédons et d'autres points qui seront abordés dans cette partie discussion, nous pouvons dire que nous obtenons des résultats satisfaisants. Ces scores sont toutefois à relativiser d'une certaine manière, et pour plusieurs raisons.

5.1 Sur le QA

Tout d'abord, comme déjà abordé à maintes reprises dans ce travail, les questions factoïdes qui sont traitées pas les compétiteurs du challenge BioASQ sont au nombre de 330. Nous l'avons évoqué, notre travail a porté au final sur la reformulation de 48 de ces questions. Ceci en excluant les questions qu'EAGLi n'arriverait probablement pas à traiter. Le système éprouve effectivement des difficultés concernant certains types de questions. Les numériques par exemple (dimension, poids, nombre de, etc…) ou les définitions de type « What is… ? » (EAGLi peut donner une définition, mais uniquement d'un terme MeSH). Ensuite, les questions relatives à des explications comme « How does it work? » ou d'interdépendance (quelle est la relation entre X et Y) sont, avec celles dont la réponse attendue est trop précise (Serine 5) ou trop large (autosomal dominant) pour relever du MeSH, des exemples types qui ne peuvent être traités à l'heure actuelle. Il faudrait pouvoir entrainer le système. A ce niveau, pour être conséquent, nous ne pouvons ainsi nullement certifier que nous obtiendrions sensiblement les mêmes résultats en participant à l'ensemble de la tâche.

5.2 Sur la RI

Concernant spécifiquement la RI, la liste d'articles du qrel pourrait différer de la liste établie par une machine du fait qu'un expert d'un domaine des sciences biomédicales n'est pas en mesure de (et ne prétend pas) éplucher l'intégralité des 22 millions d'articles de Medline/PubMed. L'expertise du spécialiste devrait compenser en grande partie ce biais, et il est à espérer que la liste du qrel établie par l'expert et complétée en fonction des réponses des candidats au challenge est proche de celle qui serait établie sur la base de l'intégralité de PubMed. La restriction en 2015 à 10 documents retournés vise d'ailleurs à faciliter le travail de révision du qrel par les experts sur la base des propositions des candidats. Ensuite, la variation dans le qrel du nombre d'articles jugés pertinents en fonction de la question est un autre facteur qui peut empêcher la machine de coller au qrel.

La différence de pool d'articles du fait de l'évolution de PubMed dans le temps fait perdre au qrel son sens de valeur de référence du fait de sa perte d'actualité. BioASQ publie les dates de limite d'accès des experts aux ressources (freeze dates). Pour évaluer un système offline, il faudrait conserver des copies figées de MEDLINE/PubMed qui correspondent à ces dates. La National Library of Medicine propose d'ailleurs depuis 2002 une MEDLINE/PubMed Baseline annuelle, établie chaque fois sur la base des ressources de mi-novembre de l'année précédente et conservée ensuite intacte (sans révision ni mise à jour).

D'autres points importants peuvent aussi être abordés concernant la RI. Nous avons choisi de ne pas nous attarder sur les scores de cette dernière pour plusieurs raisons. D'une part il ne s'agissait pas du périmètre de notre mandat puisque celui-ci était de participer uniquement à la tâche b phase B, à savoir le biomédical question answering. La RI est comprise, elle, dans la phase A. D'autre part nous aurions trouvé intéressant d'y participer mais l'utilisation par

BioASQ de documents non encore annotés dans Pubmed implique certaines contraintes et nous assurent d'avoir de mauvais résultats en RI, du moins sur papier. En effet, les valeurs de précision et de rappel identifiées se montent respectivement à 0.10 et 0.19 en mode booléen et à 0.07 et 0.19 en mode vectoriel. Cependant, le décalage de scores bas en RI et plus généreux en QA « prouve » si besoin est que la RI n'est en fait pas si mauvaise puisque le système arrive tout de même à ressortir les bonnes réponses. La différence est sans doute liée au problème de pool d'articles évoqué plus haut.

5.3 Sur le qrel

Nous avons pu parfois constater certains problèmes qui nous ont retardés, ou tout du moins imposé un certain défi dans la réalisation de notre tâche. BioASQ n'attend pas que du MeSH et lorsqu'une association de concepts doit être faite manuellement de notre part, la tâche s'avère parfois plus rude que prévue. Des termes comme «thyroid» au lieu de «thyroid hormone» nous ont posé problème à deux reprises. Aussi, le « nettoyage » des concepts en fonction de nos capacités se limite à un changement de forme (pluriel (2x), trait d'union (1x), dissociation de deux termes MeSH accolés dont un en nom complet et l'autre en sigle abrégé (7x)). A relever encore que nous n'avons pas nettoyé les réponses du qrel sous forme de phrase englobant la véritable «réponse exacte», entendons par là les réponses qui s'orientent déjà vers le format de «réponse idéale», comme si un humain en était à l'origine.

5.4 Sur la méthodologie

Les derniers points à mettre en avant dans cette discussion concernent notre méthodologie, pour laquelle nous prenons conscience de certains biais. Le challenge BioASQ se déroule de mi-mars à mi-mai, et lors de cette période, à la réception des questions toutes les deux semaines, les compétiteurs ont 48h pour participer à l'ensemble de la tâche b (24h pour la RI et 24h pour le QA). Il s'agit donc d'une participation en direct, laissant peu de temps à la reformulation des questions et aux ajustements nécessaires à de bons scores. En effet, notre méthodologie, due aux contraintes inhérentes de ce travail à notre calendrier académique, ne nous a pas permis de nous tester en conditions réelles.

Comme déjà évoqué, nous avons passé par des phases de ratés, puis d'analyses et de reformulations diverses afin de faire accepter les questions. Nous avons, en quelque sorte, cherché à obtenir les bonnes réponses. Dans cette même optique, deux membres de notre groupe ne sont pas experts en terminologie biomédicale et de ce fait, pour effectuer correctement les reformulations et la création des patrons, nous avons souvent dû identifier au préalable dans quel type sémantique se trouvait le bon terme par rapport à la question. Ceci peut être sensiblement différent d'une situation où un utilisateur se retrouve devant le moteur, sans aide, pour formuler sa requête. Encore une fois, rien ne nous assure qu'en participant au challenge dans de vraies conditions nous aurions obtenu des résultats semblables à ceux d'aujourd'hui.

Un autre point important à aborder ici est la difficulté que nous avons pu ressentir à évaluer notre travail selon la méthode avant-après pour identifier la marge de progression du système. Il faudrait pouvoir ne faire que de la reformulation pour que les questions soient acceptées en entrée par EAGLi mais sans ajouter de nouveaux patrons à ceux déjà existants. Puis par la suite, tester avec nos questions reformulées en augmentant cette fois-ci le fichier en y insérant les patrons correspondants. Mais, reformulation de questions et construction des patrons étant étroitement liées, il paraît légitime de s'interroger sur la pertinence effective d'une telle

évaluation dans notre contexte. Au final, nous aurions aussi voulu pouvoir créer divers sous-types sémantiques du MeSH, pour réduire le bruit et affiner les résultats. Un exemple précis se rattache à une question où il était demandé chez quel type de population l'arthrite rhumatoïde était la plus fréquente. La réponse attendue était les femmes mais le type sémantique qui contenait ce terme était trop large en incluant par exemple des notions de provenance géographique comme les européens, asiatiques ou africains. Il aurait fallu pouvoir créer un sous-type « genre » n'incluant que les hommes et les femmes pour affiner la précision de la recherche. Ceci n'a pas été possible par manque de temps et en raison de considérations trop délicates quant à la modification du MeSH.

6. Conclusion et recommandations

Le biomédical poursuivant son expansion ne va définitivement plus stopper sa fulgurante production documentaire. Au contraire les possibilités de stockage et l'ère du big data vont même décupler les données sur le genre humain. Dans les années à venir les prévisions liées aux nombreux objets connectés nous permettent de penser effectivement que toute action humaine sera documentée et source d'archives. Nous parlions dans l'état de l'art de ce rapport de la dénomination de web des données communément utilisée aujourd'hui, cet élément ajouté à cela, nous entendons aujourd'hui plus fréquemment parler de web des objets. Dans le domaine du médical, à l'heure où la médecine personnalisée prend vraisemblablement de l'importance du fait de la précision de son diagnostic, il devient urgent de générer des outils de recherche d'information performants.

6.1 QA en général

La force du QA à l'heure de la prolifération de données en tout genre est la possibilité pour l'utilisateur d'interagir en langage naturel et de s'affranchir d'une certaine manière des techniques d'interrogation booléennes, réservées jusqu'ici principalement aux professionnels de l'information. Dans ce contexte, une évolution de nos métiers de l'information est à prévoir, si elle n'est pas déjà en cours. En effet, le spécialiste des sciences de l'information (bibliothécaire, documentaliste, archiviste, etc…) ne se contente plus uniquement d'être la personne mettant à disposition de l'information, il a aujourd'hui plus la tâche de la structurer puisqu'elle est de plus en plus présente en ligne sous forme native ou sous forme numérisée. Ceci pour permettre une plus vaste interaction hommes-machines.

En ce sens notre travail nous a démontré l'importance de l'utilisation des vocabulaires contrôlés (thésaurus, ontologies) qui permettent l'unification de l'indexation et de l'usage, facilitant ainsi l'accès au savoir.

Pourtant, nous avons pu le voir, les données sont tellement nombreuses qu'un grand chemin reste encore à faire pour que ces outils atteignent une précision inégalée et soient adoptés plus sérieusement par le grand public.

6.2 EAGLi

Au niveau des moteurs de QA spécifiques au biomédical, la tâche semble encore plus complexe bien qu'extrêmement intéressante. L'enjeu est ici de réussir à développer l'indispensable collaboration entre des professionnels de la médecine et des professionnels de l'information n'étant pas forcément experts des vocabulaires biomédicaux.

Concernant notre travail sur EAGLi, nous avons justement pu nous rendre compte des limites qui peuvent être imposées à un professionnel de l'information lorsque le vocabulaire devient extrêmement pointu pour représenter certains concepts. S'agissant également de l'évaluation de la performance des moteurs, même si des méthodologies solides et fiables ont déjà été mises en place concernant la recherche d'information (RI), nous nous sommes aperçus que tous les modèles ne pouvaient pas forcément être transposables d'office.

A travers le challenge BioASQ nous avons pu constater qu'il était parfois nécessaire de s'affranchir des techniques d'évaluation traditionnelles de la RI. Durant notre cursus, nous avons appris que normalement un moteur de recherche s'évalue selon deux dimensions

principales, la précision et le rappel. Pour ces deux mesures il est nécessaire de faire la moyenne communément appelée la F-Mesure et qui permet de faire un classement des listes de résultats. Dans notre cas, comme il s'agissait de concepts uniques renvoyés par le système (justes ou faux), il nous a fallu comme expliqué plus haut rentrer dans le détail et nous concentrer sur un aspect plus spécifique de ces métriques d'évaluation.

En définitive, nous considérons ce travail comme un élément formateur. Nous prenons cependant conscience que nous n'avons fait qu'un petit bout du chemin dans le vaste domaine du QA, et qu'EAGLi aura sans doute encore un rôle à y jouer.

6.3 Recommandations pour EAGLi

A cet instant, nous pouvons dire globalement que nous sommes satisfaits de nos résultats et qu'EAGLi se montre être un moteur relativement performant au regard des tâches spécifiques qu'il est capable d'accomplir.

En ce sens nos recommandations ne sont pas un revirement d'ordre technique mais plutôt des encouragements à continuer sur la lancée de notre travail, à savoir :

- Reformulation de plus de 48 questions factoïdes, que ce soit au travers du challenge BioASQ ou d'autres compétitions existantes ou à venir.
- Concernant BioASQ, il serait extrêmement intéressant si le planning le permet de pouvoir participer en « live » à la compétition. Ceci apporterait des précisions supplémentaires sur les résultats en conditions réelles.
- Actuellement 195 patrons de questions ont été ajoutés au fichier. Il existe un tel degré de synonymie et d'ambiguïté propres au langage, à plus forte raison au langage biomédical, qu'il nous est difficile à l'heure actuelle de cerner l'ampleur de la tâche restante. Quoi qu'il en soit, il serait utile de continuer la complétion du fichier pour augmenter encore la couverture du système.

Au niveau de deux points que nous n'avons pas eu réellement le temps d'aborder et d'impacter, il resterait à :

- Compléter plus largement le fichier des « negatives » excluant certains termes lors de la recherche.
- Opter si possible pour la création de sous-types sémantiques du MeSH pour permettre de réduire le bruit et d'améliorer la précision.

Au final, nous rendons ce travail avec la ferme intention de nous tenir au courant de l'avancement futur du moteur EAGLi dans le domaine du QA. Les enjeux du biomédical imposent un challenge de taille, mais pas impossible.

Bibliographie

ANDRADE-NAVARRO, Miguel et PEREZ-IRATXETA, Carol, 2015. Text mining of biomedical literature: Doing well, but we could be doing better. Methods. mars 2015. Vol. 74, pp. 1‑2. DOI 10.1016/j.ymeth.2015.01.014.

BALIKAS, George et al., 2014. Results of the BioASQ Track of the Question Answering Lab at CLEF 2014. In : Proceedings of Question Answering Lab at CLEF [en ligne]. [Consulté le 4 juin 2015]. Disponible à l'adresse : http://ceur-ws.org/Vol-1180/CLEF2014wn-QA-BalikasEt2014.pdf

CHOI, Sungbin et CHOI, Jinwook, 2014. Classification and retrieval of biomedical literatures : SNUMedinfo at CLEF QA track BioASQ 2014. In : Proceedings of Question Answering Lab at CLEF [en ligne]. [Consulté le 12 juin 2015]. Disponible à l'adresse : http://ceur-ws.org/Vol-1180/CLEF2014wn-QA-ChoiEt2014.pdf

EL HALDI, Widad Mustafa [dir.], 2010. Organisation des connaissances et web 2.0. Paris : Lavoisier, septembre 2010. Les cahiers du numérique, Vol.6, No 3. ISBN 978-2-7462-3255-6.

EMBAREK, Mehdi, 2008. Un système de question-réponse dans le domaine médical: le système Esculape [en ligne]. Paris : Université Paris-Est. Thèse de doctorat [Consulté le 18 janvier 2016]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-00432052/

FERRUCCI, David A., 2012. Introduction to « This is Watson ». IBM Journal of Research and Development. Mai / Juillet 2012. Vol. 56, n° 3.4, pp. 1:1–15

FOSTER, Yvan, 2015. Watson d'IBM, le petit futé de l'informatique cognitive. Gestion. 2015. Vol. 40, n° 1, pp. 86. DOI 10.3917/riges.401.0086.

GOBEILL, Julien et al., 2009. Question answering for biology and medicine. In : Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine, ITAB 2009, Larnaca, Cyprus, 5-7 November 2009. IEEE. pp. 1–5.

GOBEILL, Julien, 2012. Modèles automatiques de questions/réponses pour les sciences biomédicales [en ligne]. Genève : Université de Genève. Thèse de doctorat. [Consulté le 4 juin 2015]. Disponible à l'adresse : http://archive-ouverte.unige.ch/unige:30032

GRIVEL, Luc [dir.], 2011. La recherche d'information en contexte : outils et usages applicatifs. Paris : Lavoisier, février 2011. Traité des sciences et techniques de l'information. ISBN 978-2-7462-2581-7.

IHADJADENE, Madjid [dir.], 2004. Les systèmes de recherche d'informations : modèles conceptuels. Paris : Lavoisier, mars 2004. Traité des sciences et techniques de l'information. ISBN 2-7462-0821-0.

LIU, Yifeng, LIANG, Yongjie et WISHART, David, 2015. PolySearch2 : a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic Acids Research [en ligne]. 29 avril 2015. pp. 1-8. [Consulté le 4 juin 2015]. Disponible à l'adresse : http://nar.oxfordjournals.org/content/early/2015/04/28/nar.gkv383.full.pdf+html

MALAKASIOTIS, Prodromos et al., 2013. Tutorials and Guidelines. Intelligent Information Management Targeted Competition Framework [en ligne]. Janvier 2013. FP7-318652 / BioASQ, D3.4. [Consulté le 12 juin 2015]. Disponible à l'adresse : http://bioasq.org/sites/default/files/PublicDocuments/2013-D3.4-TutorialsGuidelines.pdf

MALAKASIOTIS, Prodromos, PAVLOPOULOS, Ioannis, ANDROUTSOPOULOS, Ion, 2015. Evaluation Measures for Task 1B. Intelligent Information Management Targeted Competition Framework. Février 2015. FP7-318652 / BioASQ, excerpt from D4.1

MAO, Yuqing, WEI, Chih-Hsuan et LU, Zhiyong, 2014. NCBI at the 2014 BioASQ challenge task : large-scale biomedical semantic indexing and question answering. In : Proceedings of Question Answering Lab at CLEF [en ligne]. [Consulté le 4 juin 2015]. Disponible à l'adresse : http://ceur-ws.org/Vol-1180/CLEF2014wn-QA-MaoEt2014.pdf

NATIONAL LIBRARY OF MEDICINE (NLM), 2014. Medical Subject Headings : preface. National Library of Medicine [en ligne]. 6 août 2014. [Consulté le 18 janvier 2016]. Disponible à l'adresse : https://www.nlm.nih.gov/mesh/intro_preface.html

NATIONAL LIBRARY OF MEDICINE (NLM), 2015. Unified Medical Language System (UMLS). [en ligne]. 22 décembre 2015. [Consulté le 18 janvier 2016]. Disponible à l'adresse : https://www.nlm.nih.gov/research/umls/

NEVES, Mariana et LESER, Ulf, 2015. Question answering for Biology. Methods [en ligne]. Mars 2015. Vol. 74, pp. 36‑46. [Consulté le 4 juin 2015]. Disponible à l'adresse : http://www.sciencedirect.com/science/article/pii/S1046202314003491

PALIOURAS, Georgios et KRITHARA, Anastasia, 2015. Final Report. Intelligent Information Management Targeted Competition Framework [en ligne]. Mars 2015. FP7-318652 / BioASQ, D1.6. [Consulté le 4 juin 2015]. Disponible à l'adresse : http://www.bioasq.org/sites/default/files/BioASQ_publishable_summary_report.pdf

PALIOURAS, Georgios, KAKADIARIS, Ioannis et KRITHARA, Anastasia, 2016. BioASQ : a challenge on large-scale biomedical semantic indexing and question answering. [en ligne]. 2016. [Consulté le 18 janvier 2016]. Disponible à l'adresse : http://www.bioasq.org/

RUCH, Patrick, [ca. 2015]. BiTeM : Bibliomics and Text Mining Group [en ligne]. [Consulté le 18 janvier 2016]. Disponible à l'adresse : http://bitem.hesge.ch/

SIMONNOT, Brigitte, 2012. L'accès à l'information en ligne : moteurs, dispsotifs et médiations. Cachan : Lavoisier, avril 2012. Collection systèmes d'information et organisations documentaires. ISBN 978-2-7462-3829-9.

Annexe 1 : Extrait du qrel fourni par BioASQ


```
{"questions": [
    {
        "body": "Is Rheumatoid Arthritis more common in men or women?",
        "concepts": [
            "http://www.disease-ontology.org/api/metadata/DOID:7148",
            "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D001171",
            "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D012217",
            "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D013167",
            "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D015535"
        ],
        "documents": [
            "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
            "http://www.ncbi.nlm.nih.gov/pubmed/22853635",
            "http://www.ncbi.nlm.nih.gov/pubmed/21340496",
            "http://www.ncbi.nlm.nih.gov/pubmed/20889597",
            "http://www.ncbi.nlm.nih.gov/pubmed/20810033",
            "http://www.ncbi.nlm.nih.gov/pubmed/19158113",
            "http://www.ncbi.nlm.nih.gov/pubmed/18759162",
            "http://www.ncbi.nlm.nih.gov/pubmed/17965425",
            "http://www.ncbi.nlm.nih.gov/pubmed/16418123",
            "http://www.ncbi.nlm.nih.gov/pubmed/15083883",
            "http://www.ncbi.nlm.nih.gov/pubmed/12723987",
            "http://www.ncbi.nlm.nih.gov/pubmed/1563036"
        ],
        "exact_answer": [
            "Women"
        ],
        "id": "5118dd1305c10fae75000001",
        "ideal_answer": "Disease patterns in RA vary between the sexes; the condition is more commonly seen in women, who exhibit a more aggressive disease and a poorer long-term outcome.",
        "snippets": [
            {
                "beginSection": "sections.0",
                "document": "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
                "endSection": "sections.0",
                "offsetInBeginSection": 591,
                "offsetInEndSection": 678,
                "text": "Our results show a high prevalence of RA in LAC women with a ratio of 5.2 women per man"
            },
            {
                "beginSection": "sections.0",
                "document": "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
                "endSection": "sections.0",
                "offsetInBeginSection": 1140,
                "offsetInEndSection": 1394,
                "text": "RA in LAC women is not only more common but presents with some clinical characteristics that differ from RA presentation in men. Some of those characteristics could explain the high rates of disability and worse prognosis observed in wom
            },
            {
                "beginSection": "sections.0",
                "document": "http://www.ncbi.nlm.nih.gov/pubmed/22853635",
                "endSection": "sections.0",
                "offsetInBeginSection": 559,
                "offsetInEndSection": 718,
                "text": "The expression and clinical course of RA are influenced by gender. In developed countries the prevalence of RA is 0,5 to 1.0%, with a male:female ratio of 1:3."
            },
```

Annexe 2 : Résultats des systèmes participants

	Résultats des familles de systèmes par testset (de 0 à 1)							
	1e année			**2e année**				
	set1	**set2**	**set3**	**set1**	**set2**	**set3**	**set4**	**set5**
yes-no	0.92 Wishart	0.96 Wishart		0.94 Ming	0.93 Wishart	0.89 Wishart	0.94 Wishart	1 Ming
acc				0.94 System	0.82 Ming	0.83 Ming	0.88 Hippocrat	0.83 Asclepius
				0.84 Wishart	0.82 System	0.83 System	0.88 Asclepius	
						0.83 Asclepius	0.88 Ming	
	0.48 BioASQ	0.50 BioASQ	0.65 BioASQ			0.72 BioASQ		
	0.32 System	0.43 System	0.58 System	0.53 BioASQ	0.5 BioASQ		0.5 BioASQ	0.46 BioASQ
Facto				0.46 Wishart				
mrr	0.31 Wishart	0.30 Wishart	–					
				0.16 Ming	0.13 Wishart	0.08 Ming	0.28 Wishart	0.14 Ming
				0.1 System	0.09 System	0.06 Wishart	0.13 Asclepius	0.05 Asclepius
						0.04 Asclepius	0.11 Hippocrat	
							0.10 Ming	
List	0.23 Wishart	0.33 Wishart		0.36 Wishart	0.43 Wishart	0.39 Wishart	0.30 Wishart	
F								
	0.02 BioASQ	0.08 BioASQ	0.07 System	0.06 Ming	0.16 Ming	0.13 Ming	0.11 BioASQ	0.13 BioASQ
	0.007 System	0.07 System	0.03 BioASQ	0.05 BioASQ	0.10 BioASQ	0.11 BioASQ		

Annexe 3 : Arbre des types sémantiques du MeSH

ENTITY **enty T071**
Physical Object **phob T072**
Organism **orgm T001**
Plant **plnt T002**
Fungus **fngs**
T004 Virus **virs**
T005 Bacterium
bact T007
Archaeon **arch**
T194 Eukaryote
euka T204
Animal **anim T008**
Vertebrate **vtbt T010**
Amphibian **amph T011**
Bird **bird**
T012 Fish
fish T013
Reptile **rept**
T014
Mammal **mamm T015**
Human **humn T016**
Anatomical Structure **anst T017**
Embryonic Structure **emst T018**
Anatomical Abnormality **anab T190**
Congenital Abnormality **cgab**
T019 Acquired Abnormality
acab T020
Fully Formed Anatomical Structure **ffas T021**
Body Part, Organ, or Organ Component **bpoc T023**
Tissue **tisu T024**
Cell **cell T025**
Cell Component **celc T026**
Gene or Genome **gngm T028**
Manufactured Object **mnob T073**
Medical Device **medd T074**
Drug Delivery Device **drdd T203**
Research Device **resd T075**
Clinical Drug **clnd T200**
Substance **sbst T167**
Chemical **chem T103**
Chemical Viewed Functionally **chvf T120**
Pharmacologic Substance **phsu T121**
Antibiotic **antb T195**
Biomedical or Dental Material **bodm T122**
Biologically Active Substance **bacs T123**
Neuroreactive Substance or Biogenic Amine **nsba T124**
Hormone **horm**
T125 Enzyme
enzy T126
Vitamin **vita**
T127
Immunologic Factor **imft T129**
Receptor **rcpt T192**
Indicator, Reagent, or Diagnostic Acid **irda T130**
Hazardous or Poisonous Substance **hops T131**
Chemical Viewed Structurally **chvs T104**

Organic Chemical **orch T109**

Nucleic Acid, Nucleoside, or Nucleotide **nnon T114**

Organophosphorus Compound **opco T115** Amino Acid, Peptide, or Protein **aapp T116** Carbohydrate **carb T118**
Lipid **lipd T119**
Steroid **strd T110**
Eicosanoid **eico T111** Inorganic Chemical **inch T197** Element, Ion, or Isotope **elii T196**
Body Substance **bdsu T031**
Food **food T168**
Conceptual Entity **cnce T077**
Idea or Concept **idcn T078**
Temporal Concept **tmco T079** Qualitative Concept **qlco T080** Quantitative Concept **qnco T081**
Functional Concept **ftcn T169**
Body System **bdsy T022**
Spatial Concept **spco T082**
Body Space or Junction **bsoj T030** Body Location or Region **blor T029** Molecular Sequence **mosq T085**
Nucleotide Sequence **nusq T086**
Amino Acid Sequence **amas T087**
Carbohydrate Sequence **crbs T088**
Geographic Area **geoa T083**
Finding **fndg T033**
Laboratory or Test Result **lbtr T034**
Sign or Symptom **sosy T184**
Organism Attribute **orga T032**
Clinical Attribute **clna T201**
Intellectual Product **inpr T170**
Classification **clas T185**
Regulation or Law **rnlw T089**
Language **lang T171**
Occupation or Discipline **ocdi T090**
Biomedical Occupation or Discipline **bmod T091**
Organization **orgt T092**
Health Care Related Organization **hcro T093**
Professional Society **pros T094**
Self-help or Relief Organization **shro T095**
Group Attribute **grpa T102**
Group **grup T096**
Professional or Occupational Group **prog T097**
Population Group **popg T098** Family Group **famg T099** Age Group **aggp T100**
Patient or Disabled Group **podg T101**
EVENT **evnt T051**
Activity **acty T052**
Behavior **bhvr T053**

Social Behavior **socb T054**
Individual Behavior **inbe T055**
Daily or Recreational Activity **dora**
T056 Occupational Activity **ocac T057**
Health Care Activity **hlca T058**
Laboratory Procedure **lbpr**
T059 Diagnostic Procedure
diap T060

Therapeutic or Preventive Procedure **topp**
T061

Research Activity **resa T062**
Molecular Biology Research Technique **mbrt**
T063 Governmental or Regulatory Activity **gora T064**
Educational Activity **edac T065**
Machine Activity **mcha T066**
Phenomenon or Process **phpr T067**
Human-caused Phenomenon or Process **hcpp T068**
Environmental Effect of Humans **eehu T069**
Natural Phenomenon or Process **npop T070**
Biologic Function **biof T038**
Physiologic Function **phsf T039**
Organism Function **orgf T040**
Mental Process **menp**
T041 Organ or Tissue Function
ortf T042 Cell Function **celf T043**
Molecular Function **moft T044**
Genetic Function **genf T045**
Pathologic Function **patf T046**
Disease or Syndrome **dsyn T047**
Mental or Behavioral Dysfunction **mobd T048**
Neoplastic Process **neop T191**
Cell or Molecular Dysfunction **comd T049**
Experimental Model of Disease **emod**
T050
Injury or Poisoning **inpo T037**

http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

Last reviewed: 17 February 2015 Last updated: 19 May 2010 Extrait le 6 avril 2015