Création d'un annotateur à l'aide de composantes de cTAKES pour le traitement de questions cliniques et la recherche sémantique dans des articles médicaux

plan de travail présenté à
M. Michal Iglewski,
superviseur et coordonnateur

par
Valérie Levasseur (LEVV10608406)

dans le cadre du cours
INF4173- Projet Synthèse

Département d'informatique et d'ingénierie
Université du Québec en Outaouais
16 mai 2014

TABLE DES MATIÈRES

Création d'un annotateur à l'aide de composantes de cTAKES pour le traitement de questions cliniques et la recherche sémantique dans des articles médicaux

Introduction

Lors de la formation des étudiants en sciences de la santé, plusieurs sont introduits à la médecine factuelle (*evidence-based medecine*). La médecine factuelle, qui est un concept datant du milieu du 19^{e} siècle [1], vise l'utilisation des données probantes par le clinicien, en plus de son jugement et de son expertise, lorsqu'il prend une décision à propos des soins d'un patient [1].

Un des problèmes reliés à la médecine factuelle est la recherche de ces données probantes; c'est-à-dire la recherche des meilleurs articles médicaux présentant le meilleur niveau de preuve. Une application, EBMPICO, visant, entre autres, à aider le clinicien à effectuer cette recherche des articles médicaux les plus pertinents, est d'ailleurs en cours de développement par le département d'informatique et d'ingénierie de l'Université du Québec en Outaouais, en collaboration avec l'Université McGill, l'Agence de la santé et des services sociaux de l'Outaouais et le Centre de santé et de services sociaux de Gatineau [2].

Dans le cadre d'une application telle que EBMPICO, il s'agit non seulement de trouver les meilleures sources d'information, mais aussi de les trouver le plus rapidement possible. Or, selon les lignes directrices de la médecine factuelle, l'utilisation du cadre PICO facilite la recherche de données par le clinicien [3]. Ce cadre propose la formulation d'une question clinique en termes de problème et/ou de population (élément P), d'intervention visée (élément I), de comparaison (éléments C) et de résultats (élément O – pour *outcome*). Non seulement il rend la recherche plus aisée pour le clinicien, mais il permet aussi de dégager certains concepts dans les questions cliniques. Dans le cadre d'une application visant à assister la recherche de données probantes, le dégagement de concepts sémantiques dans les questions cliniques permet d'améliorer la qualité de la recherche en ne retournant au clinicien que les articles correspondants aux concepts clés

de la question. Le dégagement de ces concepts peut être réalisé de diverses façons, et quelques outils sont disponibles pour ce faire. L'un d'eux, cTAKES, qui sera employé dans le cadre de ce projet, contient un ensemble d'annotateurs spécialisés dans le domaine médical et est basé sur l'infrastructure d'application UIMA [4]. C'est que les annotateurs permettent d'identifier des entités sémantiques dans des documents.

Le projet qui sera réalisé dans le cadre de ce cours a donc pour but d'améliorer la recherche d'articles en tentant d'établir si un article trouvé correspond à la question du clinicien. Ce plan de travail en décrit donc les objectifs, de même que les étapes nécessaires à sa réalisation et leur durée prévue.

Objectifs du projet

Le projet consiste à créer un annotateur à l'aide de composantes de cTAKES pour le traitement de questions cliniques et la recherche sémantique dans des articles médicaux afin d'améliorer la recherche d'articles, et ce, en tentant d'établir si un article trouvé correspond à la question du clinicien, tel que mentionné précédemment. C'est qu'avant que cette correspondance ne soit ou non établie, la question et l'article (ou du moins son résumé) seront analysés sémantiquement, en tirant profit des capacités de cTAKES.

L'objectif principal du projet est donc la création d'un annotateur qui permettra d'analyser la question clinique et d'identifier des concepts clés dans des articles.

Les autres objectifs du travail, découlant de l'objectif principal, comprennent l'identification de concepts fréquemment présents dans les questions cliniques, l'analyse des composantes de cTAKES et l'identification correcte des éléments communs à la question et à l'article choisi en utilisant l'outil de recherche sémantique compris dans la trousse de développement logiciel Java UIMA.

Description des étapes et méthodes de travail associées

1. Familiarisation avec UIMA, cTAKES, OpenNLP, UMLS

Cette première étape consistera à se familiariser avec la terminologie propre au traitement du langage naturel, l'architecture UIMA, la librairie OpenNLP et le système unifié de langage médical (UMLS). Cette étape comprendra aussi l'installation locale de l'infrastructure d'application et la trousse de développement logiciel Java UIMA ainsi que l'outil cTAKES.

2. Lecture de la documentation et choix des composantes de cTAKES

À ce stade, une lecture plus approfondie sur les différentes composantes de cTAKES sera effectuée afin de choisir les plus pertinentes à l'analyse de questions cliniques et d'articles médicaux (pouvant correspondre à différents niveaux de preuve : revue systématique, série de cas, étude, etc.)

3. Recherche et analyse de questions cliniques

Pour simplifier le projet, l'annotation se fera uniquement sur des questions cliniques et des articles anglophones. Afin de tester l'annotateur, il faudra donc trouver une banque de questions cliniques anglophones à analyser. Pour pouvoir avoir une approximation de la qualité de l'annotateur développé, les questions cliniques devraient être annotées manuellement par des experts. Différents types d'annotations pourraient être utiles, tels l'identification des éléments P, I, C et O et la détermination du type de question (intervention, diagnostic, pronostic ou exposition).

En plus d'être annotée lors de la phase de test de l'annotateur développé, cette banque de questions sera aussi étudiée afin de déceler des concepts fréquemment présents dans les questions cliniques pour lesquels il serait intéressant de développer des annotateurs lorsque ceux-ci ne sont pas déjà présents dans cTAKES.

Afin de réaliser l'analyse de concepts présents dans les questions cliniques, certaines recherches bibliographiques seront effectuées. C'est que certains patrons de concepts ont été identifiés comme étant caractéristiques de certains types de questions par Xiaoli Huang [3]. La recherche de nouveaux concepts et/ou de patrons se fera donc à partir des résultats publiés dans la littérature.

4. Construction de l'annotateur

Cette étape consiste à combiner les différentes composantes cTAKES choisies précédemment avec les nouveaux annotateurs, s'il y lieu, en un seul annotateur à l'aide de la trousse de développement logiciel Java UIMA. L'assemblage des différentes composantes, comme tel, ne nécessite pas de programmation. Il est cependant à noter que la création des nouveaux annotateurs nécessite, quant à elle, de la programmation Java.

Si le temps le permet et que l'analyse des composantes aux étapes précédentes le justifie, différentes versions de l'annotateur à des fins de comparaisons pourraient être réalisées. Il pourrait être intéressant, par exemple, de comparer une version de l'annotateur comprenant seulement des composantes cTAKES avec une version dans laquelle ont été ajoutés de nouveaux annotateurs créés dans le cadre de ce projet.

5. Test de l'annotateur avec des questions cliniques

Les questions cliniques identifiées lors de la phase 3 (recherche et analyse de questions cliniques) seront annotées par l'annotateur proposé. Les entités identifiées seront comparées aux annotations manuelles des questions, si de telles annotations sont disponibles. Il est possible qu'à cette étape l'annotateur soit modifié en fonction des résultats obtenus (retour à l'étape précédente de construction).

6. Test de l'annotateur avec des articles médicaux

Dans cette partie, l'annotateur sera testé avec différents articles correspondants à différents niveaux de preuve, provenant de différentes sources et étant présentés dans différents formats (texte, page web, XML). Des résumés structurés et non structurés

seront annotés. Cette étape est essentielle pour que la comparaison entre la question et les articles puisse être effectuée par la suite.

7. Utilisation de l'outil de recherche sémantique (composante de l'UIMA)

Le but de cette étape est d'être capable d'identifier des entités présentes à la fois dans la question clinique et dans un résumé d'article fourni. Après avoir annoté la question et un article, l'outil de recherche sémantique présent dans la trousse de développement logiciel Java UIMA sera utilisé pour chercher chaque terme important identifié dans la question clinique dans l'article, de même que sa position. Un peu de programmation Java sera nécessaire à ce stade afin de faire une recherche automatique des entités.

8. Création de service web SOAP

Le service web créé permettra d'annoter une question soumise de même qu'un ou plusieurs articles. Il devra être capable de retourner les éléments identifiés dans la question de même que leurs positions dans l'article ou la série d'articles. Bien que la documentation de l'UIMA comprenne une partie sur le déploiement de composantes en tant que service SOAP, plus de recherches sur la mise en œuvre de services web devront être faites. Les détails du service, tels les formats acceptés, la quantité d'articles annotés et le format de la réponse du service seront déterminés après ces lectures. La création comme telle du service nécessitera de la programmation.

Il est cependant à noter que cette étape, ainsi que les tâches subséquentes, pourraient ne pas être réalisées si les autres tâches nécessitent plus de temps que prévu.

9. Réalisation d'une interface web

L'interface web réalisée servira pour la démonstration du projet synthèse. Elle permettra d'afficher visuellement les résultats des annotations (comme le font les outils graphiques fournis avec cTAKES et avec la trousse de développement logiciel Java UIMA). La création d'une interface propre à ce projet permettra d'afficher à la fois la question

clinique annotée et les éléments communs présents dans l'article choisi. Cette interface servira aussi de prototype pour l'implémentation éventuelle dans EBMPICO.

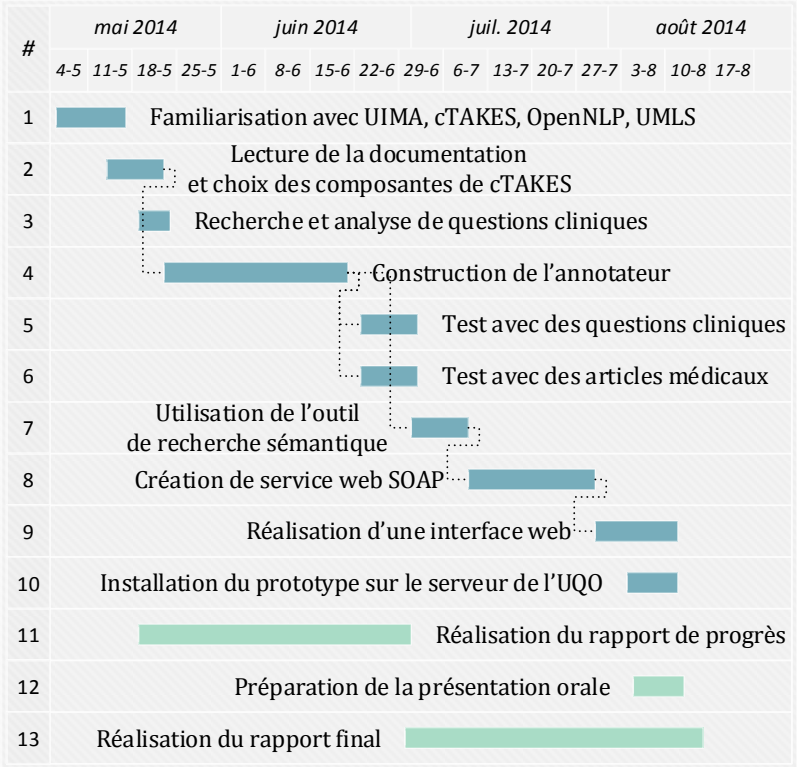
10. Installation du prototype sur le serveur de l'UQO

En dernier lieu, le service web de même que l'interface seront installés sur les serveurs de l'UQO afin que le jury et les autres étudiants puissent tester le projet.

Calendrier d'exécution des tâches

Ci-dessous, à la figure 1, est une représentation graphique des étapes du travail décrites précédemment à laquelle les différents rapports à réaliser ont été ajoutés. Les lignes en pointillé indiquent les dépendances : une tâche reliée à une précédente par une ligne pointillée ne peut débuter avant que la première ne soit terminée.

FIGURE 1 : REPRÉSENTATION GRAPHIQUE DES TÂCHES À EFFECTUER DURANT LA SESSION



Le tableau 1 présente pour sa part la durée et la date de fin prévues pour la réalisation de chacune des tâches composant le projet.

TABLEAU 1 : DURÉE ET DATE DE FIN PRÉVUES POUR CHACUNE DES TÂCHES

TÂCHE	DURÉE PRÉVUE (JOURS)	DATE DE FIN PRÉVUE
Familiarisation avec UIMA, cTAKES, OpenNLP, UMLS	9	16/05/14
Lecture de la documentation et choix des composantes de cTAKES	7	22/05/14
Recherche et analyse de questions cliniques	5	23/05/14
Construction de l'annotateur	21	20/06/14
Test de l'annotateur avec des questions cliniques	7	01/07/14
Test de l'annotateur avec des articles médicaux	7	01/07/14
Utilisation de l'outil de recherche sémantique (composante de l'UIMA)	7	09/07/14
Création de service web SOAP	14	29/07/14
Réalisation d'une interface web	9	11/08/14
Installation du prototype sur le serveur de l'UQO	6	11/08/14

Estimation des coûts

Aucun coût n'est estimé pour ce projet. L'ensemble des logiciels et des outils utilisés est gratuit et à code source ouvert. De plus, aucun déplacement n'est prévu, outre les rencontres hebdomadaires avec le superviseur.

Bibliographie

[1] D. L. Sackett, W. M. Rosenberg, J. A. Gray, R. B. Haynes, and W. S. Richardson, "Evidence based medicine: what it is and what it isn't," *BMJ,* vol. 312, pp. 71-2, Jan 13 1996.

[2] "PICO", 2014. Disponible : http://larip.uqo.ca/ebmpico/index.php. [Consulté le : 14 mai 2014].

[3] M. Xiaoli Huang, Jimmy Lin, Ph.D., and Dina Demner-Fushman, M.D., Ph.D., "Evaluation of PICO as a Knowledge Representation for Clinical Questions," in *AMIA 2006 Symposium Proceedings*, 2006, p. 359.

[4] "cTAKES", 2013. Disponible : https://ctakes.apache.org/index.html. [Consulté le : 9 mai 2014].