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

rapport de progrès 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
30 juin 2014

TABLE DES MATIÈRES

[Introduction 2](#_Toc391934554)

[Description des étapes complétées 3](#_Toc391934555)

[Familiarisation avec UIMA, cTAKES, OpenNLP, UMLS 3](#_Toc391934556)

[Lecture de la documentation et choix des composantes de cTAKES 3](#_Toc391934557)

[Description des étapes en cours 4](#_Toc391934558)

[Recherche et analyse de questions cliniques 4](#_Toc391934559)

[Construction de l’annotateur 5](#_Toc391934560)

[Test de l’annotateur avec des questions cliniques 7](#_Toc391934561)

[Test de l’annotateur avec des articles médicaux 7](#_Toc391934562)

[Description des prochaines étapes à entamer 8](#_Toc391934563)

[Utilisation de l’outil de recherche sémantique (composante de l’UIMA) 8](#_Toc391934564)

[Création de service web SOAP, réalisation d’une interface web et installation du prototype sur le serveur de l’UQO 8](#_Toc391934565)

[Calendrier d’exécution des tâches 9](#_Toc391934566)

[Bibliographie 11](#_Toc391934567)

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

Le projet qui est réalisé dans le cadre de ce cours a pour but d’améliorer la recherche d’articles en tentant d’établir si un article trouvé correspond à la question d’un clinicien. Pour ce faire, la création d’un annotateur qui permettra d’analyser une question clinique et d’identifier des concepts clés dans des articles est en cours de réalisation et constitue l’objectif principal de ce projet.

Ce rapport de progrès décrit les étapes nécessaires à sa réalisation en relatant d’abord les tâches complétées, puis celles en cours, avant de donner un aperçu du travail qui reste à accomplir pour réaliser le projet. Une mise à jour du calendrier des tâches est aussi incluse.

# Description des étapes complétées

## Familiarisation avec UIMA, cTAKES, OpenNLP, UMLS

Cette première étape consistait 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) par l’entremise de leur documentation. Afin de mieux comprendre l’architecture UIMA, ainsi que le fonctionnement d’annotateurs simples et agrégés, le tutoriel d’UIMA fut utilisé.

Cette étape comprenait aussi l’installation locale de l’infrastructure d’application et la trousse de développement logiciel Java UIMA ainsi que l’outil cTAKES. Quelques problèmes furent rencontrés lors de l’installation de l’outil cTAKES, particulièrement avec l’utilisation de l’outil Maven qui ne m’était pas familier : cette tâche pris donc plus de temps que prévu.

## Lecture de la documentation et choix des composantes de cTAKES

Une lecture plus approfondie sur les différentes composantes de cTAKES permit d’identifier les plus utiles dans le cadre de ce projet. Les composants *Core*, *POS Tagger*, *Chunker* et *LVG* seront soit compris dans l’annotateur agrégé développé, soit utilisés par l’entremise de l’utilisation d’un annotateur cTAKES les comprenant déjà. C’est que ces composantes essentielles servent entre autres à détecter les phrases, les mots, et les symboles, à départager les verbes des noms et à transformer les mots sous forme canonique [7]. La composante *Dictionary Lookup*, quant à elle, permet de chercher dans les dictionnaires SNOMED CT et RxNorm de UMLS et sera fort utile dans le cadre de travail, tous comme les annotateurs *Clinical Document Pipeline* et *Drug Named Entity recognition.* En effet, tel que son nom l’indique, les annotateurs compris dans la composante *Drug Named Entity recognition* permettent de détecter des médicaments et certains de leurs attributs, tels leur dosage et leur fréquence d’administration. Le *Clinical Document Pipeline*, qui constitue l’élément principal de cTAKES, permet en plus de détecter des traitements, des symptômes et des maladies.

# Description des étapes en cours

## Recherche et analyse de questions cliniques

Différentes sources de questions cliniques ont été identifiées et évaluées au cours de cette étape. D’abord, deux banques de questions potentielles ont été identifiées lors de recherches bibliographiques [8, 9] : *The Journal of Family Practice* [10] et *Parkhurst Exchange* [11]. Bien que les deux banques contiennent de nombreuses questions cliniques dans différents domaines, l’avantage de la première banque est qu’elle contient, dans les réponses à ces questions, les références vers les articles qui ont conduit les auteurs à leurs conclusions. Dans le cadre de ce projet, ces références peuvent être utilisées pour cibler des articles qui devraient contenir les éléments sémantiques identifiés dans les questions. Malheureusement, pour ces deux banques de questions, ni le type de question, ni l’identification des éléments P, I, C et O n’est fourni. C’est qu’idéalement, afin d’avoir une approximation de la qualité de l’annotateur développé, les questions cliniques devraient être annotées manuellement par des experts. C’est donc pourquoi la banque de *Cochrane Clinical Answers* pourrait être intéressante [12], puisqu’elle fournit pour chaque question des références, le type de question et les éléments P,I,C et O. Le seul inconvénient est que cette banque n’est pas accessible gratuitement (contrairement aux deux autres). Une demande d’accès gratuit pour trente jours est d’ailleurs en cours. L’évaluation de cette banque de questions est donc plutôt limitée en attendant.

Si l’accès à *Cochrane Clinical Answers* n’est pas obtenu la semaine prochaine, un choix d’une cinquantaine de questions cliniques de différents types sera fait dans *The Journal of Family Practice.* Les questions choisies seront simplement insérées dans un document texte qui sera analysé par l’annotateur développé.

Afin de compléter cette étape, une fois la banque de questions choisie, elle sera aussi étudiée afin de déceler des concepts fréquemment présents dans les questions cliniques. Il sera entre autres vérifié si les patrons de concepts qui ont été identifiés comme étant caractéristiques de certains types de questions par Xiaoli Huang [6] sont présents dans les questions choisies. De plus amples recherches bibliographiques sur de tels patrons seront aussi réalisées. Les patrons identifiés au cours de cette étape serviront à améliorer l’annotateur développé.

## Construction de l’annotateur

L’annotateur développé est de type agrégé et comprend plusieurs annotateurs simples représentant, entre autres, différentes classes sémantiques identifiées par Xiaoli Huang [9]. Les classes sémantiques représentant l’âge et le genre de la population ciblée par l’article ou la question ont été regroupées sous forme d’attributs de la classe sémantique «groupe» décrivant cette population. La classe sémantique «groupe», quant à elle, est directement tirée des travaux de Demner-Fushman et Lin [8] et de Cohen et Demner-Fushman [13]. Comme toutes ces classes sémantiques n’appartiennent pas toujours aux mêmes éléments P,I,C et O, dépendamment du type de question posée [9], des annotateurs représentants chacun des éléments P, I, C et O sont aussi développés indépendamment des annotateurs représentant les classes sémantiques mentionnées précédemment. Un annotateur déterminant le type de question pourrait aussi être inclus. Faute de temps, une annotation manuelle du type de question pourrait suffire dans le cadre de ce travail ou le choix d’un seul type de question.

L’annotateur développé comprend aussi des annotateurs directement tirés de cTAKES. Pour l’instant, le *Clinical Document Pipeline*, qui comprend entre autres les composantes *Core*, *POS Tagger*, *Chunker* et *LVG* est utilisé, bien que cet agrégat sera éventuellement modifié.

La figure 1 ci-dessous représente la structure visée de l’annotateur développé. Cette structure a été choisie afin de pouvoir indépendamment modifier chaque annotateur et d’ajouter facilement des classes sémantiques. Une séparation claire entre les éléments de cTAKES et ceux faits sur mesure permet aussi une mise à jour plus aisée de cTAKES.

Figure  : Représentation graphique de l'annotateur développpé



Bien que ça ne soit pas visible sur la figure précédente, il est à noter que l’annotateur agrégé cTAKES processor comprend les composantes essentielles au traitement du langage naturel, tel que mentionné précédemment.

Une réflexion sur les annotations finales est toujours en cours, à savoir, par exemple, si seuls les éléments P,I,C et O doivent être présents. Il faudrait aussi déterminer si les classes sémantiques doivent être représentées sous forme d’attributs de ces éléments.

La construction de l’annotateur demande encore beaucoup de travail puisque celui-ci a actuellement des capacités moindres au *Clinical Document Pipeline*. De plus, beaucoup plus de temps que prévu a été utilisé pour la compréhension du fonctionnement des dépendances entre les annotateurs de cTAKES et leur utilisation hors du *Clinical Document Pipeline*. C’est pourquoi le calendrier d’exécution des tâches a été passablement modifié pour accorder plus de temps à la construction de l’annotateur.

## Test de l’annotateur avec des questions cliniques

Pour le moment, pendant la construction de l’annotateur et en attendant la sélection de la banque de questions qui servira à l’évaluation finale de l’annotateur développé, des questions cliniques mentionnées dans certains articles rencontrés dans le cadre de recherches bibliographiques [8, 9], de même qu’une question exemple de *Cochrane Clinical Answers,* accessible gratuitement, sont utilisées pour tester l’annotateur.

## Test de l’annotateur avec des articles médicaux

Bien qu’il fût mentionné dans le plan de travail que l’exécution de l’annotateur se ferait sur des résumés d’articles, l’exécution sur les articles complets fut aussi considérée au cours de cette étape. Cependant, comme les articles complets ont une structure différente (par exemple, ils comportement généralement des phrases plus longues et contiennent plus souvent du texte entre parenthèses [14]), il est parfois plus difficile de les traiter. C’est donc pourquoi seule l’analyse des résumés sera faite dans le cadre de ce travail.

Pour cette partie du travail, une dizaine de questions cliniques, pour lesquelles l’annotateur semble bien identifier les éléments importants seront choisies. Pour chaque question, les deux ou trois premiers articles cités dans les références, lorsque disponibles, seront choisis et analysés, peu importe leur provenance et leur structure. Chaque résumé d’article sera présenté et analysé sous forme d’un fichier texte.

Pendant la construction de l’annotateur et en attendant la sélection de la banque de questions, seul un résumé d’article structuré est utilisé présentement pour le développement de l’annotateur. Ce résumé d’article, issu de "*Antipyretic efficacy of ibuprofen vs acetaminophen*" (Am J Dis Child 146(5): 622-625), a été choisi puisqu’il apparaît, annoté, dans un article deDemner-Fushman et Lin [8], et que les annotations produites par l’annotateur développé peuvent ainsi être directement comparées à celui développé par ces auteurs.

# Description des prochaines étapes à entamer

## 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 des résumés d’articles fournis.

La lecture de la documentation disponible sur l’outil de recherche sémantique présent dans la trousse de développement logiciel Java UIMA sera nécessaire avant d’entreprendre la programmation Java. Cette programmation sera nécessaire afin de faire une recherche automatique des entités trouvées dans les questions.

## Création de service web SOAP, réalisation d’une interface web et installation du prototype sur le serveur de l’UQO

Ces étapes servant à la création d’une interface propre au projet permettraient d’afficher à la fois la question clinique annotée et les éléments communs présents dans l’article choisi. Cependant, tel que mentionné dans le plan de travail, ces étapes pourraient ne pas être réalisées si les autres tâches nécessitent plus de temps que prévu ou si ces étapes s’avèrent plus difficiles à exécuter qu’anticipées. C’est qu’étant donné la mise à jour du calendrier des tâches, très peu de temps leur est accordé.

# Calendrier d’exécution des tâches

Ci-dessous, à la figure 2, est une représentation graphique des étapes du travail décrites précédemment. Cette figure, présente dans le plan de travail, a été modifiée afin de mieux refléter l’ampleur du travail que constitue la création de l’annotateur, l’état des tâches entreprises et la nouvelle estimation du travail requis pour les tâches à entreprendre.

Figure 2 : Représentation GRAPHIQUE des tâches à effectuer durant la session



Le tableau 1 présente pour sa part les nouvelles dates de fin prévues pour la réalisation de chacune des tâches composant le projet, de même que leur état actuel.

Tableau 1 : Durée et date de fin prévues pour chacune des tâches

|  |  |  |
| --- | --- | --- |
| Tâche | Date de fin prévue | État |
| Familiarisation avec UIMA, cTAKES, OpenNLP, UMLS | 16/05/14 | Terminé |
| Lecture de la documentation et choix des composantes de cTAKES | 22/05/14 | Terminé |
| Recherche et analyse de questions cliniques | 18/07/14 | En cours |
| Construction de l’annotateur | 25/07/14 | En cours |
| Test de l’annotateur avec des questions cliniques | 28/07/14 | En cours |
| Test de l’annotateur avec des articles médicaux | 28/07/14 | En cours |
| Utilisation de l’outil de recherche sémantique (composante de l’UIMA) | 01/08/14 | À faire |
| *Création de service web SOAP* | *05/08/14* | *À faire* |
| *Réalisation d’une interface web* | *08/08/14* | *À faire* |
| *Installation du prototype sur le serveur de l’UQO* | *08/08/14* | *À faire* |

# Bibliographie

[1] The Apache Software Foundation. (2014, 15 mai 2014). *Apache UIMA*. Disponible: https://uima.apache.org/index.html

[2] Apache OpenNLP Development Community. (2010, 15 mai 2014). *Apache OpenNLP Developer Documentation*. Disponible: https://opennlp.apache.org/documentation/1.5.3/manual/opennlp.html

[3] U.S. National Library of Medicine. (2013, 16 mai 2014). *Unified Medical Language System® (UMLS®) - Source Vocabularies*. Disponible: <http://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_002.html>

[4] U.S. National Library of Medicine. (2013, 16 mai 2014). *Unified Medical Language System® (UMLS®) - The Semantic Network*. Disponible: <http://www.nlm.nih.gov/research/umls/new_users/online_learning/SEM_001.html>

[5] The Apache UIMA Development Community. (2007, 14 mai 2014). *UIMA Tutorial and Developers' Guides*. Disponible: https://uima.apache.org/downloads/releaseDocs/2.1.0-incubating/docs/html/tutorials\_and\_users\_guides/tutorials\_and\_users\_guides.html

[6] The Apache Software Foundation. (2013, 9 mai 2014). *cTAKES*. Disponible: https://ctakes.apache.org/index.html

[7] J. Masanz. (2013, 22 mai 2014). *cTAKES 3.1 Component Use Guide*. Disponible: https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+3.1+Component+Use+Guide

[8] D. Demner-Fushman and J. Lin, "Answering Clinical Questions with Knowledge-Based and Statistical Techniques," *Computational Linguistics,* vol. 33, pp. 63-103, 2007/03/01 2007.

[9] 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.

[10] The Journal of Family Practice. (2014, 25 juin 2014). *Clinical Inquiries*. Disponible: <http://www.jfponline.com/articles/clinical-inquiries.html>

[11] Parkhurst Publishing. (2014, 25 juin 2014). *Parkhurst Exchange*. Disponible: <http://www.parkhurstexchange.com/searchQA>

[12] The Cochrane Library. (2014, 25 juin 2014). *Cochrane Clinical Answers*. Disponible: <http://cochraneclinicalanswers.com/>

[13] K. B. Cohen and D. Demner-Fushman, *Biomedical natural language processing*. Amsterdam: John Benjamins Publishing Company, 2014.

[14] K. B. Cohen, H. Johnson, K. Verspoor, C. Roeder, and L. Hunter, "The structural and content aspects of abstracts versus bodies of full text journal articles are different," *BMC Bioinformatics,* vol. 11, p. 492, 2010.

[15] 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.

[16] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?," *J Biomed Inform,* vol. 42, pp. 760-72, Oct 2009.