

Extraction d'information sur la maladie de Crohn par visualisation interactive

Ahmed Amrani¹ et Oriane Matte-Tailliez²

¹ESIEA Recherche, 9 rue Vésale, 75005 Paris, France
amrani@esiea.fr

²LRI, UMR CNRS 8623, Bât. 490, Université de Paris-Sud 11, 91405 Orsay, France
oriane@lri.fr

Abstract: *La maladie de Crohn est une affection inflammatoire chronique, provoquant des lésions sur différents segments du tube digestif. Dans le cadre d'une approche générale de fouille de textes de spécialités, nous proposons une approche interactive pour l'extraction d'informations relatives à cette maladie. L'expert écrit, de manière interactive, des patrons flexibles en utilisant une interface dédiée. Ces patrons d'extraction sont fondés sur une grammaire expressive permettant d'utiliser les ressources collectées pendant les phases précédentes de la chaîne de traitements des textes comme les termes du domaine et les concepts associés. Le système permet de visualiser les résultats obtenus.*

Keywords: Extraction d'Information, maladie de Crohn, visualisation.

1 Introduction

La maladie de Crohn est une affection inflammatoire chronique, provoquant des lésions sur différents segments du tube digestif. Une prédisposition génétique pourrait provoquer une réaction inappropriée du système de défense de l'intestin vis-à-vis d'un agent présent dans le tube digestif. L'agent en question pourrait être de nature infectieuse ou alimentaire. La maladie a une origine multifactorielle. Les liens en particulier les liens causaux entre la maladie de Crohn et d'autres phénomènes biologiques ne sont pas encore élucidés. De plus c'est l'une des deux affections inflammatoires majeures du tube digestif. Elle est donc l'objet de nombreuses recherches génétiques, épidémiologiques et cliniques. Notre objectif est de pouvoir extraire des informations liées à cette maladie. Comme les maladies génétiques orphelines ne sont pas très connues, il est nécessaire de pouvoir utiliser les informations se trouvant dans les grandes bases de littérature scientifique qui abritent des résumés et des textes complets comme par exemple Medline et PubMed Central, Highwire Press. Dans le cadre d'une approche générale de fouille de textes de spécialités, nous proposons une méthode interactive pour l'extraction d'informations. L'expert écrit, de manière interactive, des patrons flexibles en utilisant une interface dédiée. Ces patrons d'extraction sont fondés sur une grammaire expressive permettant d'utiliser les ressources collectées pendant les phases précédentes de la chaîne de traitements des textes comme les termes du domaine et les concepts associés. Le système permet de visualiser les résultats obtenus, les patrons peuvent être améliorés progressivement. De cette manière, on gagne en efficacité par une extraction guidée par l'expert.

Concernant la fouille de textes portant sur la maladie de Crohn, on peut citer le travail de [1] qui utilise les méta données de terminologie de la base MeSH, incorporés dans les résumés d'articles de la base Medline (NCBI), pour retrouver l'information contenue dans les textes. Le système « MeSHmaps » a été appliqué sur des textes portant sur la maladie de Crohn. A notre connaissance, aucun autre travail d'extraction d'information n'a été fait pour le domaine considéré.

2 Traitements préalables des textes par une chaîne logicielle

L'extraction d'informations à partir de textes bruts et spécialisés est une tâche difficile. Afin de traiter ce problème, plusieurs étapes de traitements doivent être effectuées par la chaîne de présentée ci-dessous :

Acquisition → *Normalisation* – (*Acronymes, synonymes*) → *Etiquetage grammatical* → *Etiquetage sémantique* : *Repérage des entités nommés* → *Extraction de la Terminologie nominale et verbale* → *Résolution des corréférences* → *Classification conceptuelle des termes*

La normalisation consiste à supprimer les éléments susceptibles d'engendrer des erreurs avec les outils

d'analyse des textes (étiqueteurs, analyseurs syntaxiques, etc.). Avec un expert du domaine, nous avons établi des règles afin d'uniformiser le vocabulaire employé. L'expert du domaine a également construit une règle lexicale afin de repérer les formules dans le corpus de biologie (par exemple, « GAL4 », « RAP1 », « FK506 », « Mcm1 », etc.).

L'étape de l'étiquetage morphosyntaxique consiste à affecter une étiquette morphosyntaxique à chaque mot. Pour ce faire, nous avons adapté l'étiqueteur de Brill, qui est appris sur un corpus général, pour étiqueter les corpus de biologie [2]. Puis nous procédons à l'acquisition des termes : cette étape permet d'extraire les termes pertinents du domaine [3]. Nous avons également étudié l'influence de l'étiquetage morphosyntaxique sur la qualité de la terminologie en biologie moléculaire dans [4]. L'étape suivante est la construction de la classification conceptuelle du domaine. Cette étape consiste à associer chaque terme extrait à un concept du domaine. A partir d'un noyau de connaissances donné par l'expert et du corpus, l'algorithme Induction extensionnelle [5], permet de compéter automatiquement l'ontologie de l'expert. La fouille de textes à partir du corpus peut se faire par différentes approches comme par exemple la construction automatique d'une ontologie reliant les concepts découverts dans le corpus ou l'extraction de règles d'association entre les concepts du domaine [6].

3 Module EXTRACT

Le module EXTRACT permet de trouver dans les textes des informations sur un sujet à partir d'un corpus spécialisé. Il permet de visualiser, dans la phrase, la co-occurrence de concepts. Les termes, qui sont les traces linguistiques des concepts dans les textes, sont de nature nominale et verbale. Le module nécessite en entrée un corpus normalisé, étiqueté, et comportant une terminologie nominale et verbale et une classification conceptuelle (concepts et termes nominaux et verbaux associés). Une recherche dans les phrases de relations entre plusieurs mots, termes et/ou concepts est ensuite réalisée.

3.1 Module extraction d'informations

Ce module permet de d'extraire des informations à partir d'un corpus spécialisé d'une manière conviviale et interactive. L'expert établit des requêtes flexibles et expressives via une interface de saisie, les résultats sont par la suite visualisés.

Le but des requêtes est de trouver les cooccurrences pertinentes d'instances de concepts dans les phrases, et ainsi découvrir de nouvelles relations entre les différents concepts. Pour concevoir les requêtes, l'expert doit exprimer les conditions sur les instances potentielles de concept et sur les entités reliant les concepts (verbes, conjonction de coordination...etc.). Une fois les contraintes précisées, l'expert lance la recherche des cooccurrences des instances de ces concepts dans les phrases.

Les contraintes sur les instances de concepts peuvent être exprimées en fonction de plusieurs ressources externes. En effet, il est possible d'établir des contraintes sur :

- les étiquettes morphosyntaxiques.
- l'appartenance à un groupe d'étiquettes morphosyntaxique, par exemple : le groupe des étiquettes nominales, le groupe des étiquettes verbales...etc.).
- la chaîne de caractère exacte de l'instance du concept ou bien des expressions régulières sur cette chaîne. Les expressions régulières sont pré-codées et l'expert les insert facilement via une interface dédiée.
- l'appartenance de l'instance du concept à un concept particulier. Cette contrainte est établie en utilisant une classification conceptuelle existante.

Toutes ces contraintes peuvent être combinées en utilisant des opérateurs logiques : AND, OR et NOT. Une fois les contraintes sur chaque instance de concept ont été établies, l'expert peut lancer la recherche de la cooccurrence des instances de ces concepts dans les phrases.

Les instances de ces concepts sont mises en évidence dans le contexte de la phrase. En réaction à cette visualisation, l'expert peut affiner ces requêtes.

3.2 Exemples d'extraction d'information

Pour la démonstration, nous avons utilisé un corpus portant sur « la maladie de Crohn » constitué à partir de la requête « Crohn's disease » dans Medline. Nous avons utilisé une classification conceptuelle de 26 concepts construite par un expert du domaine à partir d'une terminologie extraite automatiquement par EXIT. La figure 1 représente un extrait de la classification conceptuelle de l'expert. Chaque terme est associé à son concept.

abnormalities-in-chromosomes-1-and-10,GENET
active-phase,PHASE-DIS
adenocarcinoma,CANCER
advanced-stage-of-cancer,CANCER
Anal-cancer,CANCER
annual-surveillance,EPIDEMIO
anti-Fhit-antibody,METHOD
any-medication,DRUG
appear-with,VERB-ASSOC
archival-material,METHOD
atypical-immunoreactivity,DIS
azathioprine,DRUG
azathioprine-therapy,DRUG
biomarkers-of-carcinoma,DIAGNOS
biopsies,CLINIC
biopsy,CLINIC
CARD15/NOD2,GENPROT
chordoma,CANCER
chronically-inflamed-tissues,CANCER
chronic-healed-phase,PHASE-DIS
chronic-inflammatory-process-in-IBD,DIS
chronic-severe-anorectal-disease,DIS
ciprofloxacin,DRUG
clinicopathological-data,EPIDEMIO
Colorectal-cancer,CANCER
colorectal-carcinoma,CANCER

Figure 1. Extrait de la classification conceptuelle de l'expert. Chaque terme (à gauche) est associé à un concept (à droite, en majuscules).

La requête sur l'existante d'instances de concepts de façon concomitante dans les phrases du corpus « Maladie inflammatoire du tube digestif » (DIS) et tumeurs bénignes ou malignes (CANCER) (figure 2) nous permet de visualiser 14 phrases vérifiant la condition. Deux possibilités sont offertes à l'expert, soit voir les instances de concepts dans le contexte de la phrase, ou voir uniquement les instances des concepts.

La figure 2 nous montre un exemple d'utilisation d'expression régulière pré-codée (l'existence de la séquence « Crohn » dans une chaîne de caractères). Cette figure montre également deux exemples de requête :

- existence d'une chaîne « Crohn » et le concept « épidémiologie » (EPIDEMIO) dans la même phrase ;
- existence dans la même phrase d'une chaîne « Crohn », d'une instance du verbe « associer » (VERB-ASSOC) et une instance du concept CANCER.

ETIQ

Lexical stage Contextual stage Contextual annotation Tags Concept classification Evaluation Help

Lexical rules Contextual rules Contextual annotation Tags Taxonomy Knowledge extraction Evaluation Induction

Save

The number of sentences : 14

word 0	word 1	word 2	word 3
Colorectal-cancer	inflammatory-bowel-disease	ulcerative-colitis	colorectal-carcinoma
inflammatory-bowel-dis...	Crohn-disease	severe-Crohn's-dis...	
Human-papilloma-virus-i...	recurrent-squamous-cell-carci...		
Crohn's-disease	chordoma		
Crohn's-disease	chordoma		
Crohn's-disease	chordoma		
Inflammatory-bowel-dis...	increased-risk	development-of-col...	
chronic-inflammatory-pr...	CRC		
Metastatic-carcinoma-of...	Crohn's-disease		
histologic-type-of-stoma...	atypical-immunoreactivity	primary-colon-carci...	
Epstein-Barr-virus-asso...	Crohn's-disease		
Anal-cancer	rectal-cancer	Crohn's-disease	
increased-risk	rectal-cancer	severe-proctitis	severe-chronic-perianal...
sporadic-cancer	Crohn's-disease		

word 0	word 1	word 2	word 3
Colorectal-cancer	in	inflammatory-bowel-disease	.
Patients	with	inflammatory-bowel-disease	,
Human-papilloma-virus-infection	in	a	recurrent-squamo
Crohn's-disease	associated-to	chordoma	:
The	purpose-of-this-...	is	to-present
A	review-of-literat...	revealed-that	Crohn's-disease
Inflammatory-bowel-disease-[IBD]	is-a	typical-example	since
Therefore	,	genetic-factors	predisposing-to
Metastatic-carcinoma-of-the-colon	similar-to	Crohn's-disease	:
The	ileocecal-lesion	was-diagnosed	as
Epstein-Barr-virus-associated-lymphoma	in	Crohn's-disease	.
Anal-cancer	and	rectal-cancer	in
Present-knowledge	from	the	literature
Multimodal-treatment	is-similar-to	that	in

EXTRACT

Word form
Regular expression
PoS tag
General PoS tag
Concept

Add Simple Condition

Taxo DIS
Taxo CANCER

AND OR NOT

Add Final Condition

Taxo CANCER
Taxo DIS

Display sentences Display concepts

Figure 2. Ecran du module EXTRACT présentant un exemple d'informations extraites sur les maladies inflammatoires du tube digestif et cancer.

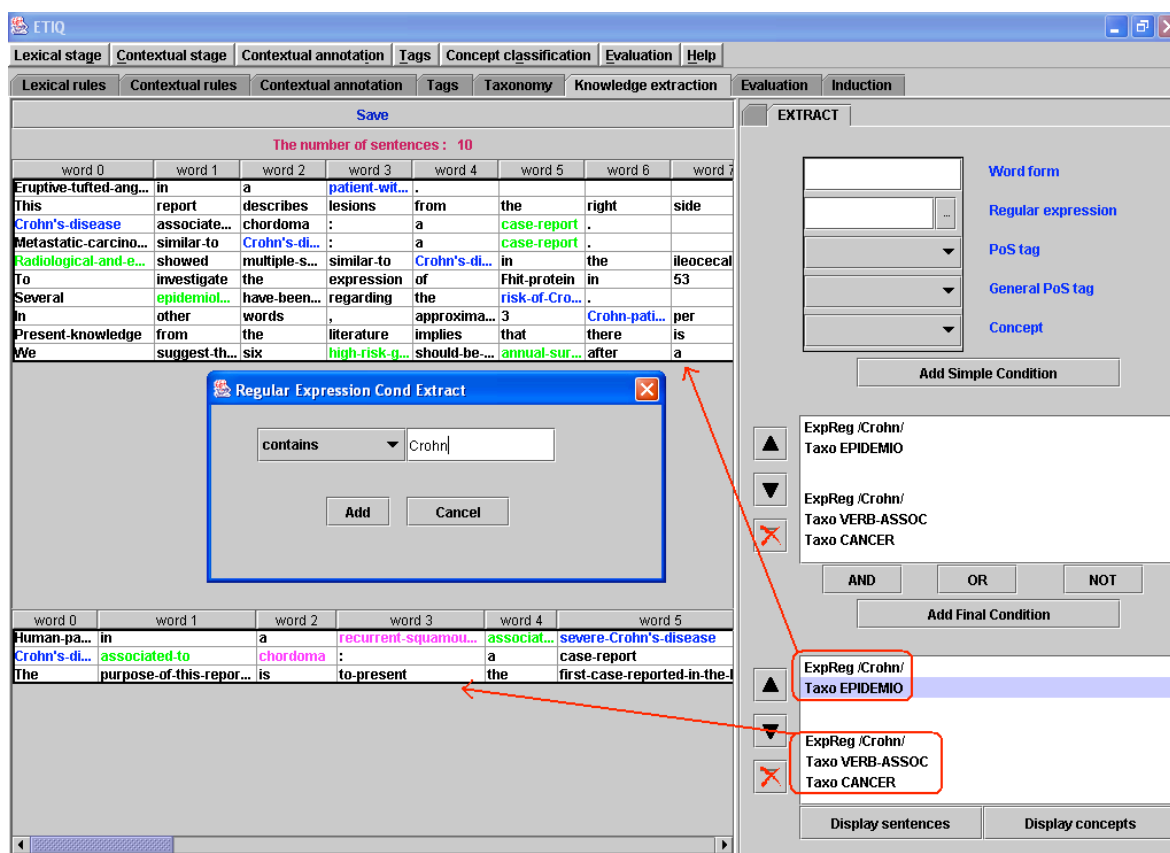


Figure 3. Ecran du module EXTRACT présentant un exemple d'informations extraites sur la maladie de Crohn associée à des cancers.

4 Conclusion et perspectives

Il est important de rechercher les liens, les facteurs de la maladie de Crohn, d'élucider les corrélations, de mesurer les poids des facteurs, de faire des inférences avec d'autres maladies inflammatoires afin de mieux comprendre cette maladie.

La capacité à traiter effectivement des textes de spécialité exige la coopération d'experts du domaine de spécialité et donc l'utilisation de logiciels conviviaux permettant un travail efficace des experts. La validation de la chaîne de traitement ainsi constituée ne peut se faire que par l'utilisation effective de l'information extraite dans le domaine de spécialité. Nous allons maintenant extraire des textes complets des informations pouvant élucider les causes et conséquences de la maladie de Crohn. Les textes devront porter sur les aspects de génétique, de biologie moléculaire et d'épidémiologie. Plus précisément, les textes portant sur toute maladie inflammatoire du tube digestif pourront être pris en considération car il est intéressant de faire une recherche de liens avec des phénomènes biologiques comme les phénomènes de cancérisation.

Références

- [1] P. Srinivasan, (2001). MeSHmap, a text-mining tool for Medline. Proceedings of the symposium AMIA, pp642-646.
- [2] A. Amrani, Y. Kodratoff, O. Matte-Tailliez, (2004). A Semi-automatic System for Tagging Specialized Corpora, H. DAI, R. SRIKANT, C. ZHANG (eds.), Advances in Knowledge Discovery and Data Mining, PAKDD, May, Sydney, LNAI, Vol. 3056, 670-681.
- [3] M. Roche, T. Heitz, O. Matte-Tailliez, Y. Kodratoff, (2004). EXIT: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés, International Conference on Statistical Analysis of Textual

Data (JADT'04) : 946–956.

- [4] A. Amrani, M. Roche, Y. Kodratoff, O. Matte-Tailliez, (2005). Inductive Improvement of Part-of-Speech Tagging and its Effect on a Terminology of Molecular Biology. In Proceedings of the 18th Conference Canadian AI 2005, Victoria, British-Columbia, Canada, 2005. LNAI, Vol. 3501,366-376.
- [5] Y. Kodratoff. (2004). Induction extensionnelle. Définition et application à l'acquisition de concepts à partir de textes. RNTI, Actes de EGC'04.
- [6] M. Roche, Azé. J, O. Matte-Tailliez, Y. Kodratoff, (2004). Mining texts by association rules discovery in a technical corpus. Dans Proceedings of IIPWM'04 (Intelligent Information Processing and Web Mining), Springer Verlag series "Advances in Soft Computing", pages 89-98.