Fouille de texte et extraction d'informations dans les données cliniques

Clément Dalloux[1]
(1) IRISA - CNRS, Campus de Beaulieu, 35042 Rennes, France
clement.dalloux@irisa.fr

Résumé des travaux

Débutés le 1^{er} décembre 2016, mes travaux de thèse s'inscrivent dans le cadre du projet BigClin, dont l'objectif est de permettre la réutilisation des données cliniques non-structurées à une grande échelle. Intitulée *Fouille de texte et extraction d'informations dans les données cliniques*, cette thèse a pour objectifs de développer et tester des méthodes et outils de traitement automatique du langage naturel (TALN) destinés aux traitements robustes et efficaces des données cliniques non structurées en français. Parmi les aspects visés se trouvent par exemple l'indexation du contenu médical, la reconnaissance d'entités nommées (maladies, opérations, médicaments, quantité, date, etc.),la détection automatique de l'incertitude et de la négation. En outre, les méthodes doivent être suffisamment robustes pour traiter des données hétérogènes (provenant de différents hôpitaux) nécessaires pour différents cas d'usage.

L'étiquetage d'entités médicales à l'aide de terminologies est primordial afin de résoudre les problèmes de coréférence et de synonymie. L'UMLS (Unified Medical Language System), qui condense de nombreuses bases de connaissances en science biomédicale, indexe un concept, ses synonymes et son acronyme sous le même identifiant (CUI). La conception d'un outil d'étiquetage des concepts médicaux (en anglais et français) utilisant l'UMLS est en cours.

Illustration de nos travaux de recherche sur la détection de l'incertitude et de la négation, l'article intitulé *Détection de l'incertitude et de la négation : un état de l'art* a été accepté en tant que poster dans le cadre des 19ème Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL) qui auront lieu du 26 au 30 juin 2017, dans les locaux de l'Université d'Orléans.

Dans cet article, nous présentons une partie des approches proposées ces dernières années selon deux axes, la détection des marqueurs, c'est-à-dire des termes qui expriment une négation ou une incertitude, et la détection de la portée, l'influence sémantique que les marqueurs ont sur les parties voisines de la phrases. Nous présentons les différents problèmes liés à la détection, d'abord en évoquant les principaux marqueurs, puis les différents types de portées. Nous présentons ensuite les corpus annotés dont les chercheurs ont disposé pour entraîner des modèles par apprentissage automatique, pour extraire des descripteurs supplémentaires, ou bien pour la conception de systèmes experts. La revue de ces différents travaux nous a montré que ces systèmes experts, bien que rapide et performants sur les données pour lesquels ils sont conçus, sont désormais délaissés au profit des approches par apprentissage artificiel, plus performants et adaptables dans les évaluations dont nous avons rendu compte.

Nous serons vite amenés à traiter ces phénomènes pour des données cliniques et aucun corpus en français n'est, à ce jour, disponible pour l'apprentissage supervisé. Nous en avons donc débuté la conception et incorporons actuellement trois sources textuelles :

— Des essais cliniques de l'hôpital Gustave Roussy [1] (critères de sélection, détails des protocoles expérimentaux, etc.),
— Des résumés d'articles scientifiques récupérés sur HAL (discipline : Sciences du Vivant),
— Des articles de presse présentant, principalement, les conlusions d'études scientifiques dans le domaine de la santé [2].

Le corpus se présentera sous la forme suivante :

#Phrase	#Unité	Unité	Lemme	POS	Marqueur	Portée	Évènement médical
4	0	Dans	dans	PRP	_	_	_
4	1	les	le	DET :ART	_	_	_
4	2	formes	forme	NOM	_	_	_
4	3	peu	peu	ADV	_	_	_
4	4	sévères	sévère	ADJ	_	_	_
4	5	,	,	PUN	_	_	_
4	6	une	un	DET :ART	_	une	_
4	7	antibiothérapie	antibiothérapie	NOM	_	antibiothérapie	antibiothérapie
4	8	spécifique	spécifique	ADJ	_	spécifique	spécifique
4	9	n	n	VER :simp	n	_	_
4	10	’	’	PUN	’	_	_
4	11	est	être	VER :pres	_	est	_
4	12	habituellement	habituellement	ADV	_	habituellement	_
4	13	pas	pas	ADV	pas	_	_
4	14	nécessaire	nécessaire	ADJ	_	nécessaire	_
4	15	.	.	SENT	_	_	_

TABLEAU 1 – Extrait du corpus

En outre, il nous semble important d'envisager l'utilisation d'une approche semi-supervisé afin de limiter le nombre d'exemples à présenter aux systèmes de classification. Cela permettra de tirer parti plus efficacement des experts médicaux en charge d'étiqueter manuellement nos corpus d'entraînement. Une autre piste sera de produire un système d'indexation des patients sur la base des informations extraites de leurs dossiers médicaux (sexe, âge, maladies, médicaments prescrits et posologie, acte médical subit, etc.). Ces indexations permettront, par exemple, de comparer un patient et de multiples critères de sélection lors de recrutements pour essais cliniques

Une fois finalisés, les jeux de données annotés que nous aurons produit ainsi que Les logiciels d'étiquetage produits seront mis à disposition des chercheurs.

1. https://www.gustaveroussy.fr/fr/essais-cliniques
2. sources : https://www.pourquoidocteur.fr/, http://www.medisite.fr/, https://www.topsante.com/, http://www.doctissimo.fr/, https://www.sciencesetavenir.fr/, etc.