TALN 2008, Avignon, 9–13 juin 2008

Extraction d'informations en cardiologie dans le projet Akenaton

Cyril Grouin, Bernard Jacquemin, Patrick Paroubek,
Isabelle Robba, Xavier Tannier, Pierre Zweigenbaum (LIMSI, CNRS)
Anita Burgun, Arnaud Rosier (EA 3888, Université de Rennes)

Le projet Akenaton (ANR-07-TECSAN-001) a pour but de collecter des informations sur des patients suivis en cardiologie. Certaines de ces données vont être extraites de comptes rendus hospitaliers en français. Cela constitue donc une tâche d'extraction d'information à partir de textes. Les spécificités de la tâche proviennent du type de textes traités (des comptes rendus d'hospitalisation et opératoires) et des informations à en extraire. Le LIMSI est responsable de cette tâche, réalisée en collaboration avec l'EA 3888 (Université de Rennes 1, coordinateur du projet) et le CERIM (Université Droit et Santé de Lille 2).

Le but du projet dans sa globalité est d'améliorer la prise en charge des porteurs de défibrillateurs cardiaques automatiques en gérant de façon plus efficace les nombreuses alertes envoyées par ceux-ci (survenue d'arythmies par exemple). Dans cette optique, les informations captées par le pace-maker seront recoupées avec le dossier médical du patient, écrit en français par les médecins, pour déterminer avec une plus grande finesse la nature et la gravité du problème. Cette dernière phase mettra en jeu des techniques de data mining, et le présent résumé ne concerne que l'extraction d'information dans le dossier médical.

Le projet a démarré en décembre 2007 et doit durer trois ans. Le corpus n'étant pas encore collecté, nous nous limitons ici à décrire la méthodologie prévue pour mener à bien cette tâche.

Le corpus est en cours de collecte par nos collègues médecins. Une partie sera ensuite anonymisée pour permettre le développement de nos outils. Si la partie « administrative » des dossiers est souvent structurée, la partie « médicale », concernant par exemple les antécédents médicaux du patient, son traitement actuel ou le stade de sa maladie, est intégralement en texte libre.

Les types de données à extraire sont en cours de spécification par des experts du domaine dans une autre tâche du projet. Ces données sont sélectionnées en fonction du type de recherches qui devront être possibles pour les utilisateurs finals du projet (médecins cardiologues). Afin d'assurer d'emblée une bonne complétude, des thésaurus et glossaires existants sont examinés, ainsi qu'un premier corpus de comptes rendus et des manuels du domaine. Ces types de données sont formalisés sous la forme d'une ontologie du domaine. Un sous-corpus sera muni d'une annotation de référence selon ces types de données, préparée par nos partenaires médecins. Il sera divisé en deux parties, l'une utilisée pour le développement du système, l'autre réservée pour son évaluation finale.

Une étude des environnements existants pour ce type de travaux a été effectuée (GATE, LinguaStream, LingPipe, UIMA, etc.) et le choix devrait se porter sur la plateforme GATE.

L'extraction des données utiles présentes dans les comptes rendus utilisera un étiquetage en entités nommées très spécifiques au domaine, et essentiellement fournies par l'ontologie, ainsi que le développement de patrons d'extraction. Étant donné le caractère critique du résultat es-

compté, il sera utile de déterminer quel taux de rappel ou de précision minimum est nécessaire pour transmettre au mieux des alertes vers les acteurs concernés (médecin traitant, cardiologue en charge du suivi du défibrillateur, etc.).