Bien représenter pour mieux raisonner : deux approches pour le dossier patient

Frédérique Segond[1], Aleksandra Ponomareva[1], Domoina Rabarijaona[1], André Bittar[2], Luca Dini[2], Ivan Kergourlay[3], Stefan Darmoni[3], Quentin Gicquel[4], et Marie Hélène Metzger[4]

[1] Centre de Recherche et Développement, Viseo,
{fsegond, aponomareva, drabarijaona}@objetdirect.com
[2] Holmes Semantic Solutions,
{bittar, dini}@ho2s.com
[3] CISMeF & TIBS, LITIS EA 4108 Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes, Centre Hospitalier Universitaire de Rouen,
{Ivan.Kergourlay, Stefan.Darmoni}@chu-rouen.fr
[4] UMR CNRS UCBL 5558, Université Lyon 1,
{quentin.gicquel, marie-helene.metzger}@chu-lyon.fr

Résumé : le Dossier Patient Informatisé (DPI) constitue une source d'information pour des applications aussi variées que la recherche médicale, l'aide à la décision, la médecine factuelle ou la surveillance épidémiologique. Les publications utilisant l'analyse textuelle pour traiter du dossier patient sont en progression constante. Bien que donnant de très bons résultats, une telle approche fait apparaitre un certain nombre de défis, tels que la nécessité d'intégrer connaissances linguistiques et connaissances métier. Se basant sur une expérience précédente, cet article montre comment lier l'analyse linguistique, la représentation des connaissances, les ontologies médicales et le raisonnement afin de fournir un système générique permettant d'extraire, de structurer et d'exploiter l'information provenant des DPIs.

Mots-clés : Traitement automatique de la langue, Terminologies, Extraction de données, Web sémantique, BRMS, Dossier patient informatisé, Raisonnement, Système d'aide à la décision

1 Introduction

L'intérêt pour l'analyse linguistique des données textuelles dans le domaine médical a augmenté progressivement ces dernières années. Cependant, les résultats obtenus varient en fonction de la granularité du traitement linguistique effectué. Par exemple, dans le système MedLEE, le traitement des textes médicaux est basé uniquement sur l'extraction des entités nommées et ne prend pas en compte les relations entre elles (Friedman et al, 1996). D'autres systèmes utilisent une approche plus sophistiquée liant l'analyse sémantique et syntaxique. Parmi ces systèmes, certains utilisent une approche basée sur les règles, (Wang, 2007) (Ben Abacha & Zweigenbaum, 2011), tandis que d'autres implémentent des méthodes statistiques (Ehrentraut et al, 2012).

Le projet *ALADIN* a montré l'efficacité du traitement linguistique à base de règles pour la détection des infections associées aux soins (Proux et al, 2011). Ce projet a également fait apparaitre un certain nombre de défis scientifiques : la nécessité d'une analyse plus en profondeur des événements médicaux, la nécessité d'intégrer les terminologies existantes avec des ontologies dans le processus du traitement linguistique et la nécessité de bien séparer les règles linguistiques des règles métier.

Cet article présente une solution développée dans le cadre du projet SYNODOS (url : http://www.synodos.fr). Ce projet consiste à développer une solution générique d'extraction des données médicales et de les organiser pour les rendre exploitables à des fins épidémiologiques ou d'aide à la décision médicale. La solution permettra au personnel médical de construire ses propres règles métier en utilisant de façon transparente l'intelligence d'un système linguistique. Le projet réunit 2 structures académiques, l'une experte dans le domaine de la recherche en Informatique Médicale (CISMeF) et l'autre dans le domaine de l'épidémiologie (LBBE, UMR UCBL CNRS 5558), et 2 partenaires industriels, l'un spécialisé dans le développement de logiciels et ressources linguistiques (Holmes Semantic Solutions) et l'autre dans l'intégration de solutions Web et Objet (Viseo/Objet Direct).

Après une brève présentation de l'architecture générale du système, nous nous concentrons sur le module d'analyse terminologique et sémantique. Nous discutons ensuite de l'extraction et de la structuration des données. Deux approches sont étudiées pour cela: une « classique » basée sur un BRMS (Business Rules Management System), l'autre, plus orientée recherche, basée sur le Web sémantique. L'objectif, par ces deux approches et leur comparaison, est d'avoir une solution complète et aussi générique que possible.

2 Architecture générale du système

La solution SYNODOS combine analyse multi-terminologique, analyse sémantique, représentation des connaissances et raisonnement.

Afin d'interroger les services web de la solution (extracteur multi-terminologique et analyseur sémantique), SYNODOS comprend un module d'anonymisation qui masque toutes les informations pouvant servir à identifier un patient, un médecin, un lieu de domicile ou de soins dans les éléments textuels qui seront traités par ces services. Ce module permet aussi de désanonymiser ces éléments textuels s'ils s'avèrent nécessaires à l'activité intra-hospitalière.

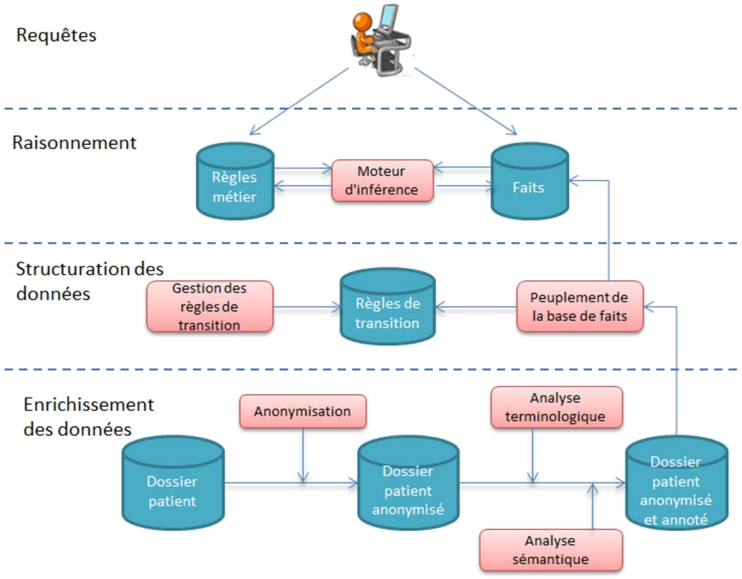

FIGURE 1 – Architecture générale du système

La section suivante détaille l'enrichissement des données des DPIs qui se fait à 2 niveaux :
- terminologique via l'association d'un terme avec son code médical
- linguistique via le traitement automatique de la langue

3 Analyse terminologique et sémantique

Chaque phrase est tout d'abord analysée par le serveur terminologique au sein de SYNODOS. Cette étape consiste à associer à chaque terme médical trouvé dans le texte, les codes terminologiques qui lui correspondent. Ces informations proviennent des différentes terminologies médicales et ontologiques : CIM-10, SNOMED CT et NCIt (la traduction française de ces deux dernières ontologies est partielle et complétée dans le cadre du projet), accessibles via les différents portails UMLS (Bodenreider, 2006), BioPortal (Whetzel et al. 2011) et HeTOP (Soualmia et al. 2011). Par exemple, pour la phrase "La masse volumineuse, détectée dans le duodénum, correspond à une récidive du gastrinome", ces ontologies permettent de savoir que le duodénum est une partie de l'anatomie et qu'un gastrinome est un trouble lié à un développement anormal de cellules.

Ensuite, le traitement sémantique est réalisé par un analyseur sémantique qui a été développé par Holmes Semantic Solutions. Il se fait par plusieurs étapes successives:

- lemmatisation et étiquetage morphosyntaxique de chacun des mots (Figure 2)

  Pour le mot *volumineuse*, on obtient :
  - Lemme : volumineux
  - Catégorie syntaxique : ADJ (Adjectif)
  - Morphologie : féminin singulier
- identification des nœuds et des dépendances syntaxiques:

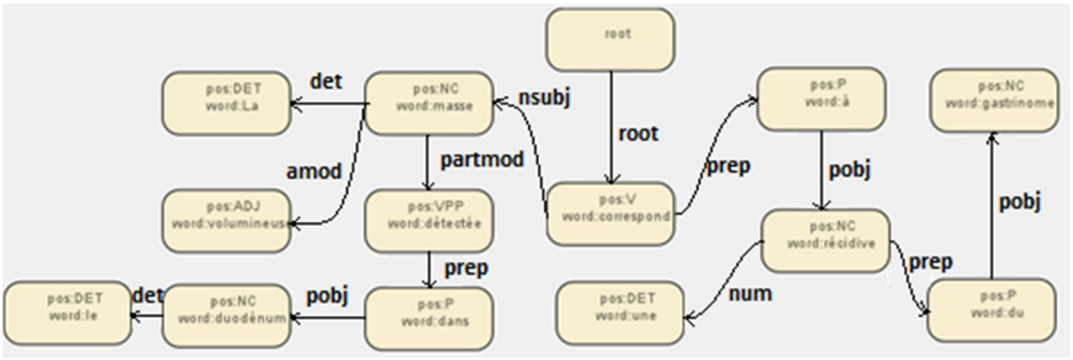

FIGURE 2- Exemple de dépendances syntaxiques

Toutes ces informations sont utilisées pour fournir une analyse sémantique liant les nœuds de la phrases via des relations sémantiques comme *agent, patient, but, thème, cause, instrument, bénéficiaire, lieu*, …

Les serveurs sémantique et terminologique donnent la sortie suivante (Figure 3):

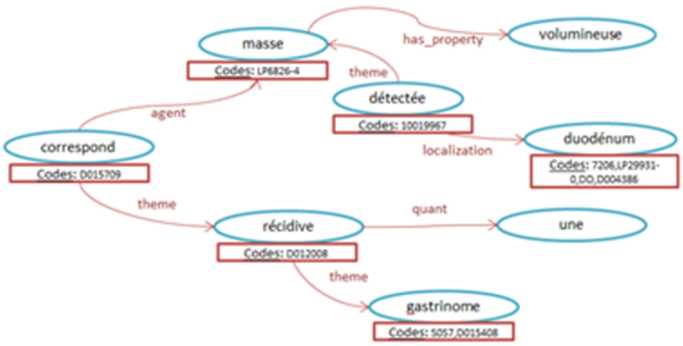

FIGURE 3- Exemple de graphe sémantique avec codes terminologiques

Pourtant ces données enrichies ne sont pas suffisantes pour répondre aux besoins des médecins. En effet, les médecins doivent avoir la possibilité de savoir, par exemple, que " gastrinome " est un antécédent, information non fournie par les analyses en amont. Le développement d'un ensemble de règles, dites « de transition », qui combinent à la fois les connaissances linguistiques et les connaissances métier, a donc été nécessaire pour exploiter au maximum ces données et les structurer. Dans le cadre du projet, deux approches sont développées et testées en parallèle : une basée sur un BRMS et l'autre basée sur le web sémantique.

4 Représentation des connaissances et raisonnement

Chaque approche, en fonction de ce qu'offrent les outils qui lui sont associés, propose une façon de représenter au mieux les connaissances extraites à l'aide des règles de transition.

Les deux approches seront comparées via un exemple dans le dernier paragraphe de cette section.

4.1. Approche classique : BRMS

Un BRMS est un outil composé d'un moteur de règles et de l'environnement nécessaire pour le manipuler. On peut ainsi définir une base de connaissances et raisonner sur des faits à partir d'un ensemble de règles métier. L'outil open source Drools a été retenu dans le cadre de ce projet. Comme Drools nécessite une approche objet, les faits sont stockés dans une base de données relationnelle.

Le schéma de la base de données a été conçu de manière à ce que le modèle de données soit extensible et permette de répondre aux besoins génériques de la solution (Figure 5).

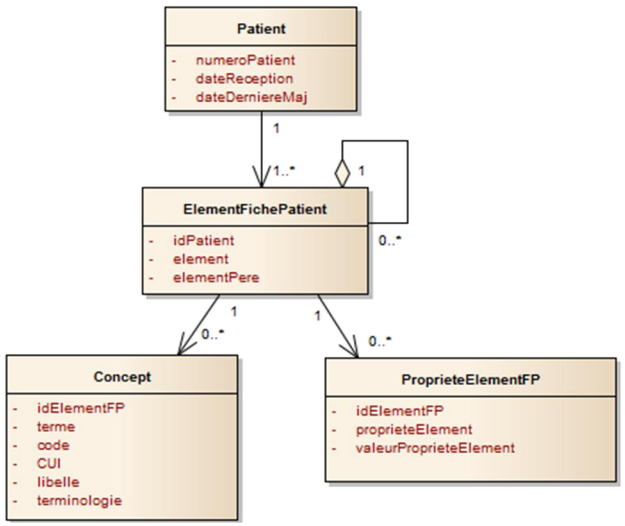

FIGURE 5 - Diagramme de classes simplifié du modèle de données

Un DPI structuré est constitué d'un ensemble d'éléments et de propriétés. Par exemple, un gastrinome est un élément "antécédent" qui a comme propriété "type d'antécédent = médical".

4.2. Approche web sémantique

Une 2ème approche de représentation des connaissances est basée sur OWL (*Web Ontology Language*), plus précisément sur OWL-DL (Description Logic), car il fournit un

vocabulaire riche pour décrire l'univers de modélisation (concepts et propriétés) et permet un haut niveau d'expressivité, tout en restant calculable. Les concepts et les propriétés sont basés sur un modèle conceptuel développé par le LBBE dans le cadre du projet SYNODOS (Gicquel et al, 2014). La figure 6 modélise l'exemple précédent.

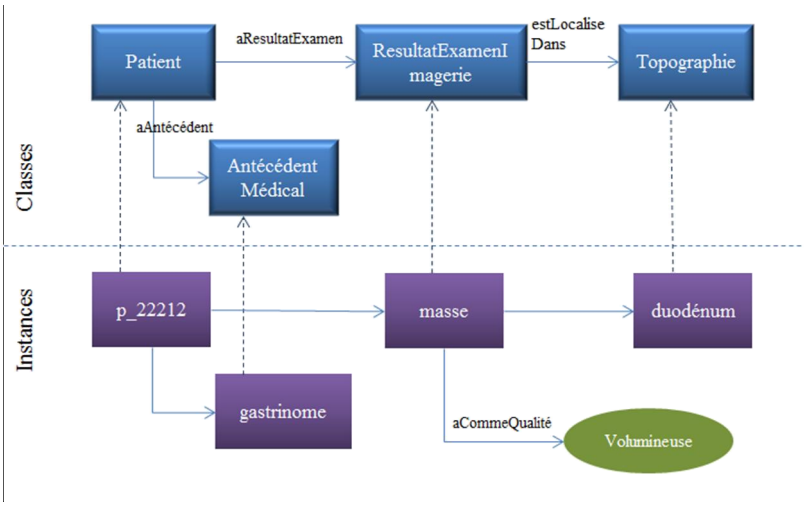

FIGURE 6 – Modélisation OWL

Une telle modélisation permet de représenter des connaissances plus explicitement que dans la base de données classique car elle ne définit pas seulement les entités, mais aussi les relations entre ces entités. L'exemple suivant permet de comparer l'approche ontologique et l'approche classique en termes de l'information extraite.

4.3. Exemple de règles de transition sur une phrase

Soit la phrase précédente (voir figure 3) *: La masse volumineuse, détectée dans le duodénum, correspond à une récidive du gastrinome*

Les règles de transition qui peuvent être appliquées sont :

- *Si THEME(x,y) et x=récidive alors y=antécédent.*
  ⇨ Ici, on obtient gastrinome comme antécédent.
- *Si THEME(x,y) et x=détecter alors y=résultat d'examen.*
  ⇨ Masse est donc un résultat d'examen.
- *Si THEME(x, y) et LOCALIZATION(x, z) et groupeSemantique(z)=ANAT alors z=topographie et (y-estLocaliseDans-z)*
  ⇨ Duodénum est une topographie dans laquelle est localisée la masse.
- *Si HAS PROPERTY (x, y) et y<> « gauche » et y<> « droit» alors (x-aCommeQualite-y)*
  ⇨ Masse a comme qualité "Volumineuse"

Les deux dernières règles impliquent des propriétés qui ne peuvent être représentées dans le modèle classique. Par exemple, la localisation n'est pas une propriété définie pour tous les résultats d'examen. Elle est très liée au contexte, contrairement à une propriété comme "Date de résultat d'examen" qui sera (ou devrait être) renseignée pour tous les résultats d'examen. Ainsi, grâce au modèle ontologique, l'approche web sémantique permet d'exprimer le type exact de relation entre les concepts « masse» et «duodénum» et pas seulement le fait que la relation existe. Il en est de même pour les propriétés qualificatives comme "volumineuse".

5 Conclusion

Dans cet article, nous avons présenté deux approches pour analyser le contenu textuel du DPI : une approche classique qui utilise un BRMS, et une approche orientée Web sémantique. Cette deuxième approche, par son expressivité, devrait permettre une meilleure description des informations et par conséquent offrir de meilleures possibilités de raisonnement que l'approche classique. Le travail futur consistera à comparer et à évaluer ces deux approches dans deux scénarios : surveillance hospitalière des infections liées aux soins et délais de prise en charge du cancer du côlon.

Remerciements

Ce projet a reçu une aide financière de l'Agence Nationale de Recherche – Programme TecSan (aide allouée : 785 000€, Projet SYNODOS ANR-12-TECS-0006)

Références

Ben Abacha A., Zweigenbaum P. (2011) Automatic extraction of semantic relations between medical entities: a rule based approach. Journal of Biomedical Semantics, 2(Suppl 5) : S4.

Coden A., Savova G., Sominsky I., Tanenblatt M., Masanz J., Schuler K., Cooper J., Guan W., de Groen PC. (2009). Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform; 42(5) : 937-49

Dupuch M., Segond F., Bittar A., Dini L.,Soualmia L., Darmoni S., Gicquel Q.,Metzger MH (2013) Separate the grain from the chaff : make the best use of language and knowledge technologies to model textual medical data extracted from electronic health records. In the Proceedings of the 6th Language & Technology Conference, (LTC2013) Poznan.

Ehrentraut C., Tanushi H.,Dalianis H. and Tiedemann J.. (2012) Detection of Hospital Acquired Infections in sparse and noisy Swedish patient records. A machine learning approach using Naïve Bayes, Support Vector Machines and C4.5. In the Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data, AND, held in conjunction with Coling 2012, Bombay.

Friedman C., Shagina L., Socratous SA,Zeng X. (1996) A WEB-based version of MedLEE : A medical language extraction and encoding system, In Cimino JJ, ed. Proceeding of the fall 1996 AMIA Conference, 938.

Gicquel Q., Dini L., Kergourlay I., Arnod-Prin P., Chariout S., Bittar A., Soualmia L., Guedez P., Segond F., Ruhlmann M., Darmoni S., Metzger MH (2013), SYNODOS SYstème de Normalisation et d'Organisation de Données médicales textuelles pour l'Observation en Santé, Medinfo, FRSIGIMIA, Copenhague - Danemark, August 2013 (http://www.synodos.fr/).

Jensen PB, Jensen LJ, Brunak S. (2012). Mining electronic health records : towards better research applications and clinical care, Nat Rev Genet; 13(6) : 395-405

Gicquel Q., Kergourlay I., Gerbier-Colomban S., Chariout S., Bittar A., Segond F., Darmoni S., Metzger MH. Annotation methods to develop and evaluate a medical expert system based on natural langage processing in electronic medical records, MIE, Istanbul, Turkey August 2014

Proux D., Hagège C., Gicquel Q., Pereira S., Darmoni S., Segond F., Metzger MH. (2011) Architecture and Systems for Monitoring Hospital Acquired Infections inside Hospital Information Workflows. Proceedings of the Second Workshop on Biomedical Natural Language, Bulgaria, 43-48.

Soualmia LF, Griffon N., Grosjean J., Darmoni S. (2011) Improving Information Retrieval by Meta-Modelling Medical Terminologies. 13th conference on Artificial Intelligence in Medicine (AIME) : Springer, Heidelberg; 215-219.

Zhu F., Patumcharoenpol P., Zhang C., Yang Y., Chan J., Meechai A., Vongsangnak W., Shen B. (2013) Biomedical text mining and its application in cancer research. J Biomed Inform; 46 : 200-211