Un système expert fondé sur une analyse sémantique pour l'identification de menaces d'ordre biologique

Cédric Lopez[1], Aleksandra Ponomareva[1], Cécile Robin[2], André Bittar[2], Paolo Curtoni[2], Xabier Larrucea[3], Frédérique Segond[1], Marie-Hélène Metzger[4]

(1) Viseo Technologies, 4 avenue Doyen Louis Weil, Grenoble (France)
(2) Holmes Semantic Solutions, 12-14, rue Claude Genin, Grenoble (France)
(3) Tecnalia, Parque tecnologico de Bizkaia, Edif. 202, Zamudio (Espagne)
(4) Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne (France)

cedric.lopez@viseo.com, aleksandra.ponomareva@viseo.com, robin@ho2s.com, bittar@ho2s.com, curtoni@ho2s.com, xabier.larrucea@tecnalia.com, frederique.segond@viseo.com, marie-helene.metzger@chu-lyon.fr

Résumé. Le projet européen TIER (Integrated strategy for CBRN – Chemical, Biological, Radiological and Nuclear – Threat Identification and Emergency Response) vise à intégrer une stratégie complète et intégrée pour la réponse d'urgence dans un contexte de dangers biologiques, chimiques, radiologiques, nucléaires, ou liés aux explosifs, basée sur l'identification des menaces et d'évaluation des risques. Dans cet article, nous nous focalisons sur les risques biologiques. Nous présentons notre système expert fondé sur une analyse sémantique, permettant l'extraction de données structurées à partir de données non structurées dans le but de raisonner.

Abstract.

An Expert System Based on a Semantic Analysis for Identifying Biological Threats
The European project TIER (Integrated strategy for CBRN – Chemical, Biological, Radiological and Nuclear - Threat Identification and Emergency Response) aims at developing a comprehensive and integrated strategy for emergency response in case of chemical, biological, radiological and nuclear danger, as well as explosives use, based on threat identification and risk assessment. In this article, we focus on the biological risks. We introduce our business rules management system based on a semantic analysis, that enables the extraction of structured data from unstructured data with the aim to make reasoning.

Mots-clés : TIER, SGRM, Système de Gestion de Règles Métier, analyse sémantique.
Keywords: TIER, BRMS, Business Rules Management System, semantic analysis.

1 Introduction

L'un des objectifs du projet européen TIER (*Integrated strategy for CBRN Threat Identification and Emergency Response*) est d'identifier les menaces et les risques chimiques, biologiques, radiologiques et nucléaires. Nous nous focalisons dans un premier temps sur les risques biologiques à des fins de détection de menaces liées au bioterrorisme. Le défi consiste à structurer les informations recueillies depuis différentes sources de données non structurées du Web concernant l'apparition ou l'évolution des pathologies infectieuses relevant des catégories A à C définies par les *Centers for Disease Control and Prevention* (CDC) et de leurs informations relatives (symptômes, nombre de cas détectés, cas mortels, etc.).

2 Approche

L'originalité de notre approche consiste à utiliser un BRMS (*Business Rules Management System*) pour la gestion de règles linguistiques dites « de transition ». Celles-ci s'appuient sur une analyse syntaxique et sémantique pour générer des données structurées en vue de leur intégration dans la base de connaissance. Un BRMS est un outil composé d'un moteur de règles et de l'environnement nécessaire permettant de les manipuler. On peut ainsi définir une base de connaissances et raisonner sur des faits à partir d'un ensemble de règles métier, ici nos règles de transition. L'outil *open source* Drools (Browne, 2009) a été retenu dans le cadre de ce projet. Notre approche est constituée de 5 étapes :

1. **Constitution d'un jeu de données textuelles** (à partir du Web : Direction générale de la Santé, Institut de veille sanitaire, …) et **définition du modèle de données.** Les médecins impliqués dans TIER ont identifié les libellés des germes correspondant aux catégories A à C du CDC et ont défini les entités d'intérêt pour caractériser l'événement : date de début et date de fin de l'événement, germe impliqué dans l'événement, nombre de personnes infectées, nombre de personnes décédées, lieu (pays ou région) de l'événement, *etc*. Dans la suite, nous extrayons automatiquement des données correspondant à ce modèle.
2. **Reconnaissance des entités d'intérêts.** Celle-ci est fondée sur une approche hybride (symbolique et statistique). Concernant la partie symbolique, nous avons conçu une grammaire composée de lexiques et de règles basées sur l'analyse morphosyntaxique. Par exemple, la présence du lemme « cas » implique souvent la mention d'une quantité (nombre de cas mortels d'un évènement par exemple). La partie statistique se concentre sur la localisation des évènements biologiques par le biais d'un classifieur de type CRF (*Conditional Random Field statistical modelling method*) (Lefferty et al., 2001).
3. **Analyse sémantique.** La détection des relations sémantiques entre les prédicats et leurs arguments s'opère sur un graphe de dépendances syntaxiques (fourni par HOLMES[1]) et se fait par l'application d'un ensemble de grammaires de transformation de graphe. Les grammaires exploitent les informations linguistiques présentes sur les nœuds (tokens) du graphe pour convertir les dépendances syntaxiques en relations sémantiques telles que AGENT, CAUSE, LOCALIZATION, MANNER, MODALITY, NEGATION, *etc.* et des relations temporelles, telles que AFTER, BEFORE, DURING, *etc.*
4. **Développement des règles.** Fondées sur les résultats de la reconnaissance d'entités d'intérêts et sur l'analyse sémantique, les règles permettent de générer des données structurées candidates pour peupler la base de connaissance. Nos règles « de transition », développées avec Java/MVEL dans Drools, ont pour objectif de transformer le résultat de l'analyse linguistique en des objets correspondants au modèle précédemment défini, permettant ainsi de peupler notre base de connaissance. Nos règles s'appuient à la fois sur les entités d'intérêt détectées, et sur des éléments linguistiques d'ordre syntaxique et sémantique. Par exemple, pour la phrase « En conséquence, on considère désormais que 18 cas de fièvre hémorragique à virus Ebola et 6 décès ont été notifiées », on pourra appliquer la règle suivante : « si l'annotation sémantique CASE_NB (nombre de cas) a une relation de préposition avec le nom du germe, alors le nom du germe et le nombre de cas sont intégrés dans le même fait candidat.
5. **Sélection des candidats pertinents et peuplement de la base de connaissance.** À chaque règle est associé un score de saillance : plus la règle est précise (i.e. contraignante), plus le score est élevé. Une forte précision des faits extraits est assurée en imposant la présence de certaines relations sémantiques. Au contraire, des contraintes faibles permettent de gagner en rappel au détriment de la précision.
   La sélection des faits est réalisée via le calcul d'un score qui permet de classer les faits par ordre de pertinence. Ce score dépend de la saillance des règles appliquées. Le seuil S a pour objectif de distinguer les faits pertinents des faits non pertinents.

L'évaluation a été effectuée sur 166 faits extraits automatiquement dans 50 textes jusque-là non exploités. Nous avons annoté manuellement chaque fait selon deux classes : « pertinent » et « non pertinent ». Un fait est considéré « pertinent » lorsque les informations sont cohérentes avec l'information véhiculée dans le texte (dans les autres cas, le fait est jugé « non pertinent »). Nous avons utilisé les mesures de micro-moyenne de précision, de rappel et de F-score en variant le seuil *S* du score de saillance. Les résultats indiquent que le meilleur F-score (0,73) est atteint pour S=0,65, avec une précision (0,74) et un rappel (0,73) équivalents.

Remerciements

Avec le soutien financier du programme Prévenir et combattre la criminalité (ISEC) Commission européenne – DG Affaires Intérieures.

Références

BROWNE P. (2009) *JBoss Drools Business Rules*, Packt Publishing, pp. 304, ISBN 1847196063, 2009.

LAFFERTY, J., MCCALLUM, A., PEREIRA, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001* (ICML 2001), Williamstown, MA, USA.

[1] http://www.ho2s.com/fr/