Du TALN au LOD : Extraction d'entités, liage, et visualisation

Cédric Lopez, Matthieu Osmuk, Dana Popovici, Farhad Nooralahzadeh, Domoina Rabarijaona, Fabien Gandon, Elena Cabrio, Frédérique Segond

Du TALN au LOD : Extraction d'entités, liage, et visualisation

Cédric Lopez[1], Matthieu Osmuk[1], Dana Popovici[1], Farhad Nooralahzadeh[2], Domoina Rabarijaona[1], Fabien Gandon[2], Elena Cabrio[2], Frédérique Segond[1]

[1] VISEO TECHNOLOGIES, R&D, Grenoble, France
firstname.name@viseo.com

[2] INRIA, Wimmics, Sophia Antipolis, Nice, France
firstname.name@inria.fr

Résumé : Dans un contexte de veille stratégique, nous avons développé un prototype prenant la forme d'un plugin de navigateur ayant pour principale ambition d'enrichir les connaissances des utilisateurs naviguant sur le Web. Au fur et à mesure de la navigation sur le Web, le système peuple la base de connaissance et tisse des liens avec le Web des données ouvertes que l'utilisateur peut parcourir. Ce prototype s'appuie et démontre en pratique des techniques d'extraction d'entités d'intérêts et de leurs relations dans une page Web couplées à une représentation des connaissances extraites au format du web sémantique et liées avec des données du Linked Open Data. Finalement le plugin propose une visualisation en temps réel de l'ensemble de ces données liées en regard des pages consultées.

Mots-clés : Extraction d'entités nommmées, Données ouvertes, Ontologies, Visualisation.

Un des objectifs du laboratoire commun SMILK (Social Media Intelligence and Linked Knowledge, LabCom ANR) concerne l'étude du couplage du Traitement Automatique du Langage Naturel (TALN) au Linked Open Data (LOD). Pour atteindre cet objectif, nos recherchent portent sur : 1) extraction d'entités d'intérêts et de leurs relations dans un contenu textuel non structuré, 2) représentation des connaissances extraites, 3) liage des données extraites avec les données du LOD, 4) Visualisation et exploration des données liées.

Pour valider nos recherches et démontrer les possibilités nous avons développé un prototype qui prend la forme d'un plugin de navigateur ayant pour ambition d'enrichir les connaissances des utilisateurs naviguant sur le Web. Dans un contexte de veille stratégique, notre cas d'application se concentre sur le secteur de la cosmétique, bien représenté chez Viseo par des clients d'envergure tels que L'Oréal, L'Occitane, ou LVMH (Moët Hennessy Louis Vuitton).

Le prototype SMILK analyse les pages Web en vue d'identifier des entités pertinentes dans le domaine de la cosmétique. Celles-ci correspondent à des classes de notre ontologie ProVoc (*PROduct VOCabulary*) qui a pour vocation la publication de données liées portant sur des produits sur le Web (Lopez *et al.*, 2016), précisément : les groupes (ex : "'L'Oréal"'), leurs divisions (ex : "'L'Oréal produits grands public"'), les marques (ex : L'Oréal Paris), les gammes de produits (ex : "'Elsève"') et les noms et caractéristiques de produits (ex : "'Color Vive 200"').

La reconnaissance automatique des entités d'intérêts et de leurs relations est effectuée par Renco, notre système à base de règles linguistiques (Lopez *et al.*, 2014). Les règles développées sont de type lexico-syntaxique, fondées sur les principes de (McDonald, 1996) et (Hearst, 1992), tenant en compte le contexte droit et gauche de l'entité pour la désambiguïser. Un résultat de cette étape de reconnaissance d'entités est illustré à la Figure 1.

Les entités repérées sont ensuite liées au LOD (précisément DBpedia) par notre système de liage des données fondé sur le framework Dexter (Ceccarelli *et al.*, 2013), pour lequel nous

FIGURE 1 – Un résultat de l'extraction des entités d'intérêts sur une page Web de Vogue.fr.

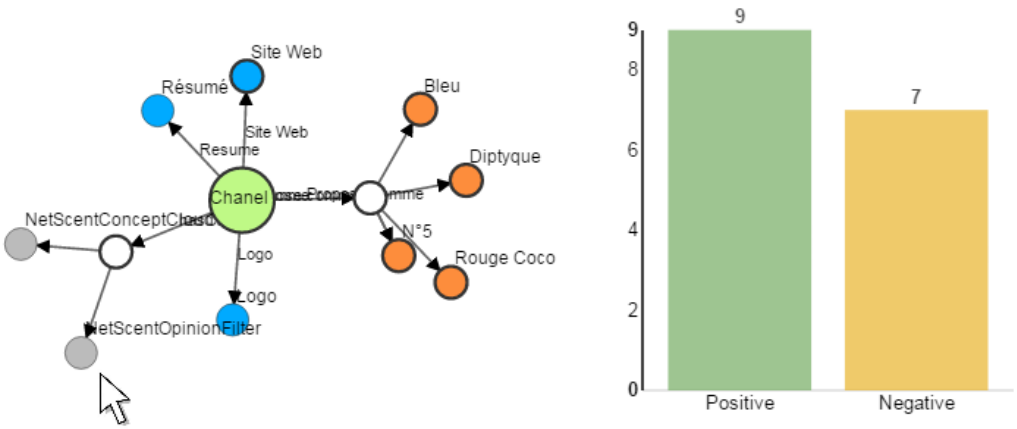

FIGURE 2 – Liage des entités extraites avec les données de bases de connaissances privés et publiques. En bleu, données provenant de DBpedia ; en gris : données provenant de NetSent ; en orange : données de notre base de connaissance privée

avons retenu notre approche de propagation sémantique (Nooralahzadeh *et al.*, 2016). Le système peuple ainsi semi-automatiquement notre base de connaissance au format RDF au fur et à mesure de la consultation des pages Web. Une validation manuelle peut être effectuée par l'administrateur de la base de connaissance afin d'assurer que les données qui y sont intégrées sont correctes.

Un exemple de graphe RDF, généré à la volée par le plugin, est montré en Figure 2. Au cours de la navigation dans le graphe, plusieurs liens apparaissent : vers notre base de connaissance privée, vers DBpedia, ou vers NetSent, une base de connaissance fournissant des résultats d'analyse d'opinion réalisée sur des forums de cosmétiques et développée en collaboration avec Holmes Semantic Solutions [1].

Par ailleurs, notre base de connaissance peut être explorée en utilisant notre outil SMILK Viewer qui repose sur le serveur RDF Jena Fuseki. L'accès au graphe de connaissances s'opère en choisissant une entité particulière dans la liste des entités catégorisées par type (groupes, divisions, marques, etc.). L'utilisateur peut ensuite facilement y naviguer et découvrir les connaissances recueillies par le plugin au fil des pages Web parcourues, ainsi que des données DBpedia et Netscent s'y rapportant. Par exemple, en sélectionnant la marque *Chanel*, apparaissent différents produits de ladite marque dont *N° 5* de type *eau de parfum*. En cliquant sur *N° 5*, le graphe permet de prendre connaissance de 3 de ses ambassadeurs dont *Audrey Tautou* pour laquelle le système nous informe qu'elle est une actrice française née le 9 août 1976 à Beaumont.

Dans cette démonstration nous naviguerons sur des pages Web évoquant des produits pour montrer comment le plugin reconnait les entités qui y sont mentionnées et collecte des données supplémentaires. Nous verrons ainsi les traitements effectués sur leur texte et parcourrons le graphe construit et visualisé en regard de pages démontrant l'apport de ces sources extérieures à l'augmentation de la compréhension des pages visitées.

Remerciements. Ce travail est réalisé dans le cadre du Laboratoire Commun SMILK financé par l'ANR (ANR-13-LAB2-0001).

Références

CECCARELLI D., LUCCHESE C., ORLANDO S., PEREGO R. & TRANI S. (2013). Dexter : an open source framework for entity linking. In *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*, p. 17–20 : ACM.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, p. 539–545 : Association for Computational Linguistics.

LOPEZ C., NOORALAHZADEH F., CABRIO E., SEGOND F. & GANDON F. (2016). Provoc : une ontologie pour décrire des produits sur le web. In *Actes d'IC'16*, p. to appear.

LOPEZ C., SEGOND F., HONDERMARCK O., CURTONI P. & DINI L. (2014). Generating a resource for products and brandnames recognition. application to the cosmetic domain. In *Actes d'LREC'14*, p. 2559–2564.

MCDONALD D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *Corpus processing for lexical acquisition*, p. 21–39.

NOORALAHZADEH F., LOPEZ C., CABRIO E., GANDON F. & SEGOND F. (2016). Adapting semantic spreading activation to entity discovery in text. In *Actes d'NLDB'16*, p. to appear.

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