CuriosiText : application web d'aide au peuplement d'ontologies métiers comme ressources lexicales basée sur Word2Vec

Meryl Bothua[1], Delphine Lagarde[1], Laurent Pierre[1]
(1) EDF Lab, 7 Boulevard Gaspard Monge, 91120, Palaiseau
meryl.bothua@edf.fr, delphine.lagarde@edf.fr, laurent.pierre@edf.fr

RESUME

Suite à la mise en place d'une chaîne traitement destinée à extraire automatiquement des actions de maintenance réalisées sur des composants dans des comptes rendus, nous avons cherché à constituer des ressources lexicales à partir de textes souvent mal normalisés sur le plan linguistique. Nous avons ainsi développé une application web, CuriosiText, qui permet de lancer un traitement Word2Vec et de peupler semi automatiquement une ontologie métier avec les termes similaires correctement détectés. Des relations métiers spécifiques peuvent également être ajoutées.

ABSTRACT

CuriosiText: a web application based on Word2Vec helping with the population of ontologies (serving as lexical resources).
After having developed a process dedicated to the extraction of maintenance actions on components from reports, we sought to constitute lexical resources. These reports usually include many lexical irregularities and lack language standardisation. We developed a web application, CuriosiText, based on Word2Vec method that helps to populate an ontology with the similar terms thus detected.

MOTS-CLES : similarité entre mots, plongement de mots, visualisation de données, extraction d'informations, peuplement d'ontologie.

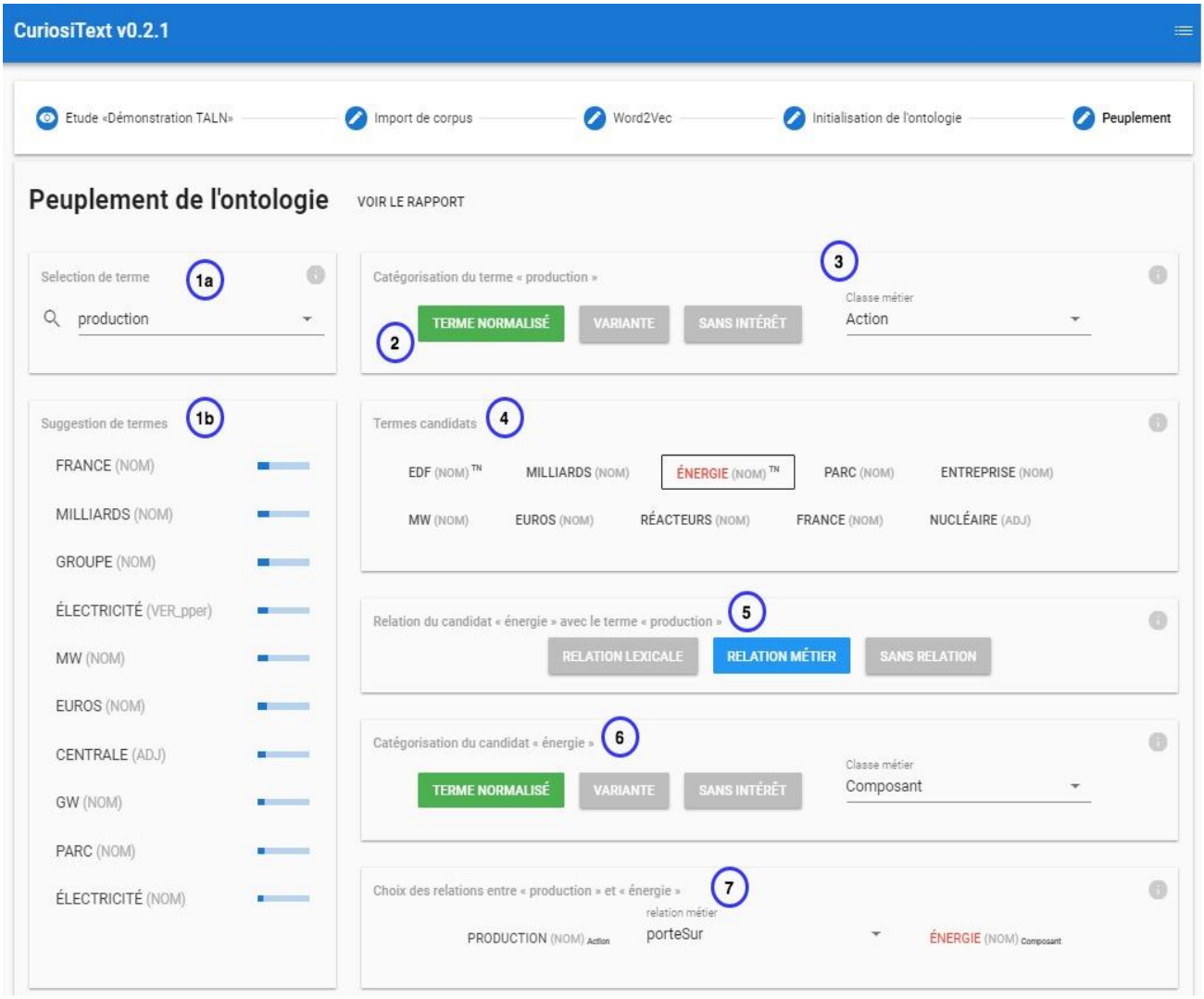
KEYWORDS: word similarity, word embedding, data visualization, information extraction, ontology population.

1 Contexte

Dans le contexte de transition numérique d'EDF et dans sa volonté d'exploiter au mieux ses données, il est aujourd'hui indispensable d'explorer des méthodes pour traiter la masse d'informations textuelles contenues dans les documents techniques, les données relatives à la relation client, ou encore les archives documentaires. Des chaînes de traitements pour de l'extraction de connaissances sont mises en place pour traiter ces contenus textuels. L'automatisation du processus offre un gain de temps conséquent et met en exergue des éléments auparavant noyés dans la masse de données. Afin de rendre pertinent ce type de chaînes de traitement, il est nécessaire de développer en amont différentes ressources lexicales. Afin de faciliter la création de telles ressources, nous avons développé une application web, CuriosiText. Celle-ci intègre la méthode Word2Vec pour extraire des synonymes, des abréviations, des mots mal orthographiés ou encore des phénomènes de multilinguisme. Nous rendons ensuite possible le peuplement d'ontologies que nous utilisons comme ressources lexicales alimentant les chaînes d'extraction de connaissance.

2 L'application CuriosiText

CuriosiText est une application web qui permet de charger des données, de lancer un traitement Word2Vec et de peupler manuellement à partir des termes similaires détectés une ontologie modélisée en amont. Cette ontologie peut être exportée pour être intégrée en tant que ressource lexicale au sein d'une chaîne d'extraction de connaissance. Nous proposerons dans cette démonstration de présenter l'ensemble des fonctionnalités de cet outil pour le peuplement de l'ontologie.



1a L'utilisateur peut requêter un terme (ici « **production** »).

1b Il peut aussi choisir un terme parmi les suggestions (termes les plus fréquents de Word2Vec).

2 Il ajoute une classe lexicale (ici **Terme Normalisé**, le terme étant bien écrit).

3 Il ajoute une classe métier issue de son ontologie métier chargée en amont (ici **Action**).

4 Les candidats de Word2Vec sont retournés et l'utilisateur choisi un terme (ici **« énergie »**).

5 Il spécifie la relation qui lie le terme requêté et le terme candidat (ici **Relation Métier**).

6 Il définit alors la classe lexicale du candidat choisi ainsi que sa classe métier (ici **Terme Normalisé** et **Composant)**. La classe métier est automatiquement héritée du terme requêté si l'on a choisi à l'étape 5 de créer une Relation Lexicale.

7 Il précise enfin la relation métier qui lie les termes (ici **porteSur**). Si l'on a choisi à l'étape 5 de créer une Relation Lexicale, il est possible d'ajouter un lien de synonymie, de variation du terme (cas des abréviations et des fautes d'orthographe) ou de bruit.

Références

MCKEE G. T., MALVERN D. & RICHARDS B. J. (2000). MEASURING VOCABULARY DIVERSITY USING DEDICATED SOFTWARE.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). EFFICIENT ESTIMATION OF WORD REPRESENTATIONS IN VECTOR SPACE. CORR.

MOTIK B., NENOV Y., PIRO R., HORROCKS I. & OLTEANU D. (2014). PARALLEL MATERIALISATION OF DATALOG PROGRAMS IN CENTRALISED, MAIN-MEMORY RDF SYSTEMS. IN C. E. BRODLEY & P. STONE, EDS., PROC. OF THE 28TH AAAI CONF. ON ARTIFICIAL INTELLIGENCE (AAAI 2014), P. 129–137, QUÉBEC CITY, QUÉBEC, CANADA : AAAI PRESS.

SCHMID H. (1994). PROBABILISTIC PART-OF-SPEECH TAGGING USING DECISION TREES. *Insertion/Caractères spéciaux*, onglet *Caractères spéciaux*).