



Chapitre d'actes

2012

Published version

Open Access

This is the published version of the publication, made available in accordance with the publisher's policy.

---

## Extraction d'information basée sur une ontologie pour les besoins du Web Sémantique

---

Nebhi, Kamel

### How to cite

NEBHI, Kamel. Extraction d'information basée sur une ontologie pour les besoins du Web Sémantique.  
In: Actes des 23es Journées Francophones d'Ingénierie des Connaissances- IC 2012. Paris (France).  
[s.l.] : [s.n.], 2012.

This publication URL: <https://archive-ouverte.unige.ch/unige:24082>

# Extraction d'information basée sur une ontologie pour le Web Sémantique

Kamel Nebhi

LABORATOIRE D'ANALYSE ET DE TECHNOLOGIE DU LANGAGE (LATL)  
Université de Genève, Suisse  
kamel.nebhi@unige.ch

**Résumé** : Dans cet article, nous présentons un système d'extraction d'information pour le français basé sur une ontologie et fondé sur une approche à base de règles. Ce papier détaille le fonctionnement de notre application et présente les premiers résultats obtenus à partir d'un corpus journalistique.

**Mots-clés** : Extraction d'information, Ontologies, Web Sémantique, Linked Data

## 1 L'EIBO

La réalisation du Web Sémantique (WS) à grande échelle implique l'annotation généralisée de documents Web à l'aide de bases de connaissances ontologiques. Depuis plusieurs années, de nombreux travaux de recherche démontrent clairement que l'Extraction d'Information (EI) est essentielle à l'automatisation de ce processus (Brewster & Wilks, 2009).

L'« Extraction d'Information Basée sur les Ontologies » (EIBO) est un sous-domaine de l'EI. Cette approche a été conçue il y a quelques années et a donné lieu à de nombreuses publications (Wimalasuriya & Dou, 2010). La plupart de ces travaux sont consacrés à des systèmes pour la langue anglaise. Nous remarquons également qu'ils n'ont pas été intégrés dans le schéma général du WS. Pour résoudre ce problème, nous proposons un système d'EIBO pour le français permettant d'établir une relation entre les entités nommées d'un texte, l'ontologie DBpedia et la banque de données du même nom.

## 2 Description du Système et Expérience

### 2.1 Architecture

Notre système d'EIBO a été créé à partir de la plateforme GATE <sup>1</sup>.

L'application est composée de 3 éléments :

- le pré-traitement linguistique
- les dictionnaires
- les grammaires

Le pré-traitement linguistique est constitué des ressources suivantes : tokeniser et délimiteur de phrases. Pour l'étiquetage grammatical, nous utilisons l'outil TreeTagger pour le français. Les dictionnaires sont composés d'environ 42000 entrées extraites de Wikipédia et DBpedia. Parmi les ressources fournies par GATE, il est possible d'utiliser un automate à états finis basé sur le langage JAPE pour « Java Annotation Patterns Engine ». JAPE permet d'écrire des règles basées sur l'étiquetage grammatical, les dictionnaires ou une chaîne de caractères particulière.

L'exemple de la Figure 1 permet de catégoriser la séquence *Liverpool* comme une ville en se basant sur les dictionnaires. Le système détermine ensuite la classe, l'instance et les *Linked Data* de cette entité nommée :

```
class : http://dbpedia.org/ontology/City
inst : http://dbpedia.org/ontology/#Liverpool
linked-data : http://dbpedia.org/data/Liverpool
```

```
Rule: CityLookup
({ Lookup.minorType == city }):cityLabel
-->
:cityLabel {
  String city = stringFor(doc, cityLabelAnnots);
  String baseUrl = "http://dbpedia.org/"
  newFeatures.put("class", baseUrl + "ontology/City");
  newFeatures.put("inst", baseUrl + "ontology/#" + city);
  newFeatures.put("linked-data", baseUrl + "data/" + city);
}
```

FIGURE 1 – Exemple d'une règle écrite avec JAPE

---

1. GATE est une plateforme d'ingénierie linguistique développée à l'Université de Sheffield et permettant de construire ou de réutiliser des outils et ressources linguistiques.

## 2.2 Expérience

L'évaluation des systèmes d'EIBO nécessite d'introduire des mesures de similarités aux calculs traditionnels pour déterminer les distances ontologiques entre les termes. Pour l'expérience<sup>2</sup>, nous avons utilisé la Balanced Distance Metric (BDM) (Maynard *et al.*, 2008). Un corpus d'évaluation de 40 dépêches du journal *Le Monde* a été annoté manuellement afin de créer une référence. Nous l'avons ensuite traité automatiquement et comparé les résultats avec la référence. La F-mesure augmentée du système est de 94%.

|              | $P_{Augm}$  | $R_{Augm}$  | $F_{Augm}$  |
|--------------|-------------|-------------|-------------|
| Lieu         | 0.92        | 0.96        | 0.94        |
| Organisation | 0.95        | 0.96        | 0.95        |
| Personne     | 0.96        | 0.93        | 0.94        |
| <b>Total</b> | <b>0.94</b> | <b>0.95</b> | <b>0.94</b> |

FIGURE 2 – Evaluation sur le corpus *Le Monde* de 2012

## 3 Conclusion

Nous avons présenté un système d'EIBO pour le français intégré dans le schéma général du WS et dont les performances sont très honorables. Dans nos futurs travaux, nous tenterons d'améliorer le système en y intégrant un outil permettant la désambiguïsation des entités nommées.

## Références

- BREWSTER C. & WILKS Y. (2009). *Natural Language Processing as a Foundation of the Semantic Web*. Delft, Netherlands : now Publishers Inc.
- MAYNARD D., PETERS W. & LI Y. (2008). Evaluating evaluation metrics for ontology-based applications : Infinite reflection. In *LREC*.
- WIMALASURIYA D. C. & DOU D. (2010). Ontology-based information extraction : An introduction and a survey of current approaches. *J. Information Science*.

---

2. Nous repérons 3 types d'entités nommées (Lieu, Organisation, Personne).