



## Sujet :

### **Extraction supervisée de relations entre entités nommées à une large échelle**

**Niveau** : Bac + 5 - 3<sup>ème</sup> année d'école d'ingénieur – Master

**Durée du stage** : 4 à 6 mois

### **Contexte**

Le sujet de stage proposé se situe globalement dans le domaine du Traitement Automatique des Langues (TAL) et se focalise plus précisément sur l'une de ses branches applicatives les plus actives, l'extraction d'information. Celle-ci a pour objectif de repérer automatiquement dans des textes les entités caractéristiques d'un domaine ainsi que les relations intervenant entre ces entités, ceci dans le but d'alimenter une base de connaissances ou une base de données.

Les entités considérées dans ce cadre sont plus précisément appelées entités nommées et dans le cas le plus général, correspondent à des noms de personnes, de lieux, d'organisations ou à des entités numériques telles que des dates, des montants financiers ou des mesures. Les relations entre ces entités peuvent être dans les cas les plus complexes des relations n-aires allant jusqu'à la notion d'événement. Par exemple, un événement de rachat d'une entreprise par une autre est représentable par une relation du type :

Achat\_entreprise

*société acheteuse* : ORG

*société achetée* : ORG

*montant* : MONEY

*date* : DATE

où *société acheteuse* définit le rôle d'une entité et ORG, son type.

Dans le cadre du stage, seules des relations binaires seront considérées. Le processus d'extraction d'information peut dans ce cas se résumer aux deux étapes suivantes :

- détection des entités nommées ;
- détection des relations entre les entités identifiées.

À titre d'exemple, pour le passage :

“With a father from <loc>Kenya</loc> and a mother from <loc>Kansas</loc>, <pers>President Obama</pers> was born in <loc>Hawaii</loc> on <date>August 4, 1961</date>.”

ces deux étapes donnent le résultat suivant si l'on s'intéresse aux données de naissance d'une personne :

Détection des entités nommées

Noms de lieux : Kenya, Kansas, Hawaii

Noms de personnes : President Obama

Date : August 4, 1961

Détection des relations entre entités

Lieu\_naissance : *bornIn*(President Obama, Hawaii)

Date\_naissance : *bornOn*(President Obama, August 4, 1961)

### **Objectifs du stage**

De nombreux travaux ont été réalisés sur la détection des entités nommées et comparés lors de plusieurs campagnes d'évaluation (shared task CoNLL 2002 et 2003, ACE ...). Le laboratoire LIC2M du CEA LIST possède en outre, au travers de sa plate-forme LIMA, des outils de traitement linguistique intégrant la reconnaissance d'entités nommées « générales ».

Le stage se concentrera donc sur la phase d'extraction de relations, pour laquelle le niveau de performance des systèmes actuels reste à améliorer. C'est particulièrement le cas lorsque l'objectif est de couvrir un ensemble large de types de relations.

Le stage s'effectuera dans la perspective de l'évaluation KBP (Knowledge Base Population) de la campagne TAC 2009 (Text Analysis Conference) et en reprendra les caractéristiques et les données. Plus précisément, cette évaluation vise à rassembler des informations factuelles concernant des entités relevant de trois grands types : personnes, organisations et entités géopolitiques.

Ces informations factuelles prennent la forme de relations appartenant à 42 types possibles (date et lieux de naissance, âge, religion, nombre d'employés, fondateur, etc).

Le LIC2M dispose déjà d'outils d'extraction de relations au sein des phrases, fondés sur la notion de patron linguistique. Un tel patron peut être vu comme une forme d'expression régulière intégrant des éléments de différents niveaux de généralité (mots, catégories grammaticales, « joker » ...) et permettant de valider la présence effective d'une relation entre deux entités nommées trouvées dans une phrase. Par exemple, le patron `<maladie> * traiter * par DET <traitement>` permet de valider la présence de la relation `[traitement]—(traiter)—[maladie]` dans les deux cas suivants :

`<maladie>` se traite par une `<traitement>`  
`<maladie>` est traitée efficacement par le `<traitement>`

Le LIC2M dispose également des outils permettant d'apprendre automatiquement ces patrons à partir de corpus annotés.

Le stagiaire aura tout d'abord en charge l'application de cet existant à l'échelle du grand nombre de relations considérées dans KBP. L'accent sera mis sur l'utilisation de données d'apprentissage bruitées du fait de l'impossibilité de valider manuellement de larges ensembles d'apprentissage pour un tel nombre de relations. Deux autres problématiques importantes seront ensuite abordées :

- le filtrage des relations extraites, en s'appuyant notamment sur des méthodes d'apprentissage statistique (machines à vecteurs de support (SVM)) ;
- l'extension de l'ensemble des patrons appris pour une relation par l'exploitation de données issues du Web. L'objectif est ici d'acquérir à partir d'exemples sondes de nouvelles formulations d'un type de relations ou des paraphrases de formulations déjà rencontrées.

### Bibliographie

Task Description for Knowledge-Base Population at TAC 2009, <http://apl.jhu.edu/~paulmac/kbp/090601-KBPTaskGuidelines.pdf>

Automatic Content Extraction (ACE) Evaluation, <http://www.itl.nist.gov/iad/mig/tests/ace/>

Mintz, M., Bills, S., Snow, R. & Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, August, Suntec, Singapore.

Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang and Ji-Rong Wen. 2009. StatSnowball: a Statistical Approach to Extracting Entity Relationships. 18th international World Wide Web conference (WWW 2009).

César de Pablo-Sanchez, Juan Perea, Isabel Segura-Bedmar, Paloma Martinez. 2009. The UC3M team at the Knowledge Base Population task.

Culotta, A., McCallum, A. & Betz, J. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Morristown, NJ, USA.

### Domaines de spécialité requis :

**Erreur ! Référence non valide pour un signet.**, apprentissage statistique

### Moyens informatiques mis en œuvre :

Langages : C++, langage de script de type Perl ou Python

Logiciel ou environnement : Linux



**Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)**

**Centre :** Fontenay-aux-Roses

**Pôle ou Direction :** DRT - LIST

**Dépt/Service/Labo :** DTSI/SRCI/LIC2M :

Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue

**Web :** <http://www-list.cea.fr/>

**Encadrement**

Nom de l'ingénieur responsable :	Olivier Ferret	Tél. :	+33 1 46 54 96 45
	Romarc Besançon		+33 1 46 54 80 16
E-mail :	<a href="mailto:olivier.ferret@cea.fr">olivier.ferret@cea.fr</a>	Fax :	+33 1 46 54 75 80
	<a href="mailto:romarc.besancon@cea.fr">romarc.besancon@cea.fr</a>		
Secrétariat :	DTSI/SRCI	Tél. :	+33 1 46 54 91 17
Nom du chef de laboratoire :	Olivier Mesnard	Tél. :	+33 1 46 54 73 38