PhD Subject: Unsupervised Template Acquisition for Open Event Extraction

Keywords: *Natural Language Processing, Text Mining, Knowledge Extraction, Event Extraction, Unsupervised Learning*

Ideal starting date: Jan. 2016

Duration: 3 years (fully funded)

Advisors: Xavier Tannier (LIMSI-CNRS), Olivier Ferret (CEA LIST)

Location: LIMSI-CNRS, Orsay (Paris-Saclay), France (Doctoral School**:** STIC - Paris Saclay)

Context (ANR Project ASRAEL)

The PhD candidate will join the ANR Project ASRAEL, starting January 2016. Partners of this project are LIMSI-CNRS (leader), CEA LIST, AFP and EURECOM.

This project takes place in the context of data journalism, which has been fed until now by publicly available, statistical data but has paradoxically made only little use of the very journalistic materials that are events. The objective of the ASRAEL project is to bridge this gap. It comes within the scope of the general scientific framework of Information Extraction (IE) and aims at extracting events from a large set of textual documents, without prior knowledge about them, for populating and publishing a knowledge base of events. This knowledge base will be the support of a dedicated event search engine.

Traditional IE systems lean on templates predefined by experts and identify events in texts with either rule-based systems or statistical models trained from annotated corpora. However, in the general domain, where the huge number of possible events makes the manual definition of these templates impossible, information retrieval ("bag of words") methods take over, but do not provide a structured answer. The ASRAEL project aims to overcome these limits by tackling the following challenges:

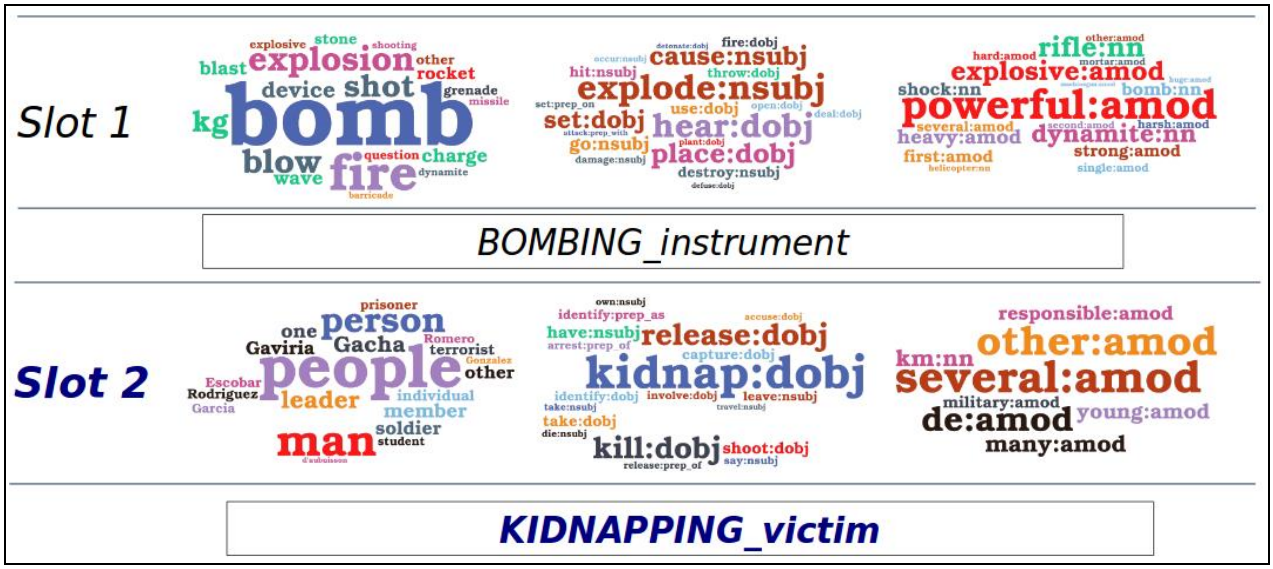
- Discover automatically event templates from very large text corpora, and populate a knowledge base dedicated to events. This implies a mixture of supervised and non-supervised approaches, which is necessary as soon as one consider such a generic problem.
- Use this knowledge base in order to build an event aggregator and a semantic search engine. With this engine, a user (either journalist or end-user) will be able to query for an event type (e.g. earthquake) and provide filters on attribute values (location = Turkey, magnitude > 8, etc.). The knowledge base will also be published following the linked data principles for other to re-use.

Description

In Information Extraction, a template generally defines a specific type of events (*e.g.* earthquakes) associated with roles (called slots or event roles) hold by entities (for earthquakes, their location, date, magnitude and the damages they caused). Template filling, which is the main IE task in this context, consists in assigning event roles to individual textual mentions for *a priori* defined templates. When such templates don't already exist, an alternative is to perform *template induction*, which is the task of learning these templates with no supervision from unlabeled text. More specifically, the idea is to group together entities corresponding to the same role in an event based on the similarity of the relations that these entities hold with predicates. For example, in a corpus about terrorist attacks, entities that are objects of verbs *to kill*, *to attack* can be grouped together and characterized by a role named VICTIM.

A first approach to this task was developed by LIMSI and CEA LIST [1], continuing the trend of generative models proposed earlier [A,B]. In the context of the ANR ASRAEL Project, the PhD candidate will have to overcome the limitations of this first work by tackling the following challenges:

- Improving the quality of template induction in order to build an event knowledge base from news articles
- Using temporal and spatial textual information to improve the models
- Achieving better scalability and robustness
- Applying template induction models to French journalistic data
- Aggregating results and building a semantic search engine from the learned knowledge



Slots learned by our generative models [1]

References

Partners relevant publications

[1] Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, Romaric Besançon.
Generative Event Schema Induction with Entity Disambiguation (ACL-IJCNLP 2015, Beijing, China).

[2] Emanuela Boros, Romaric Besançon, Olivier Ferret and Brigitte Grau.
Event Role Extraction using Domain-Relevant Word Representations (EMNLP 2014, Doha, Qatar)

[3] Xavier Tannier, Véronique Moriceau.
Building Event Threads out of Multiple News Articles (EMNLP 2013, Seattle, USA).

[4] Rémy Kessler, Xavier Tannier, Caroline Hagège, Véronique Moriceau, André Bittar.
Finding Salient Dates for Building Thematic Timelines (ACL 2012, Jeju Island, Republic of Korea).

[5] Wei Wang, Romaric Besançon, Olivier Ferret and Brigitte Grau
Filtering and clustering relations for unsupervised information extraction in open domains (CIKM 2011, Glasgow, UK).

[6] Ludovic Jean-Louis, Romaric Besançon and Olivier Ferret.
Text Segmentation and Graph-based Method for Template Filling in Information Extraction (IJCNLP 2011, Chiang Mai, Thailand).

Other relevant publications (selection)

[A] N. Chambers, *Event Schema Induction with a Probabilistic Entity-Driven Model* (EMNLP 2013, Seattle, USA).

[B] K. J. Chi Cheung, H. Poon, L. Vanderwende, *Probabilistic Frame Induction* (NAACL 2014)

[C] N. Chambers and D. Jurafsky, *Template-Based Information Extraction without the Templates* (ACL-HLT 2011, Portland, USA).

[D] L. Qiu, M-Y. Kan and T-S Chua, *Modeling Context in Scenario Template Creation* (IJCNLP 2008, Hyderabad, India).

[E] M. Regneri, A. Koller and M. Pinkal, *Learning Script Knowledge with Web Experiments* (ACL 2010, Uppsala, Sweden).

[F] C. A. Bejan, *Unsupervised Discovery of Event Scenarios from Texts* (FLAIRS 2008, Coconut Grove, USA)

[G] N. Chambers and D. Jurafsky, *Unsupervised Learning of Narrative Schemas and their Participants* (ACL-IJCNLP 2009, Singapore).

[H] N. Chambers and D. Jurafsky, *Unsupervised learning of narrative event chains* (ACL 2008, Hawaii, USA).

[I] M. Freedman, L. Ramshaw, E. Boschee, R. Gabbard, G. Kratkiewicz, N. Ward and R. Weischedel.
Extreme Extraction – Machine Reading in a Week (EMNLP 2011, Edinburgh, UK).

Application

We are particularly interested in candidates with a background in one or several of the following areas:

- Natural Language Processing
- Text Mining
- Unsupervised Machine Learning, especially hierarchical bayesian models
- Information Retrieval

Applicants should have a Masters degree (or obtain it in the near future) in Computer Science, Natural Language Processing or Machine Learning. Knowledge of French is not a prerequisite, but as some of the data are in French, eagerness to learn the basics of the language is a plus.

Applications should include:

- Cover letter outlining interest in the position
- Names of two referees
- Curriculum Vitae (CV) with publications (if applicable)
- Copy of MA degree

Contacts for questions and applications:

Xavier.Tannier[at]limsi.fr
Olivier.Ferret[at]cea.fr