Stage de master 2 / Graduate internship
Distant supervision for event extraction from a newswire corpus

Keywords: natural language processing, text mining, machine learning, distant supervision.

Ideal starting date: March/April 2017

Duration: 4-6 months

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

Location: LIMSI, Orsay, Univ. Paris-Saclay[1]

1 Context

1.1 ANR Project ASRAEL

Information and communication society led to the production of huge volumes of content. This content is still generally non-structured (text, images, videos) and the promises of a "Web of Knowledge" are still long ahead. This situation evolves with the development of Open Data portals or resources such as DBPedia, that have made easier the access to information stored in databases (economic or demographic statistics, world knowledge contained in Wikipedia infoboxes, etc). However, most of the knowledge is still produced by textual data. Among the information concerned by the difficulty of accessing textual data, those related to events are of great interest, notably in the context of the emergence of data journalism. Data journalism has been fed until now by publicly available, statistical data, but it has paradoxically made only little use of the very journalistic materials that are events. The project ASRAEL aims at bridging this gap.

Our proposal comes within the scope of the general scientific framework of information extraction (IE). We aim at extracting events from a large set of

[1] `https://www.limsi.fr/en/access`

textual documents, without prior knowledge about them, and at populating and publishing a knowledge base of events. This knowledge base will be the support of a dedicated event search engine.

1.2 Event extraction

We define event in a traditional information extraction way. An event is a structured representation of something that happens, with a nucleus, a spatio-temporal context and some arguments. The "event type" gathers comparable instances of events, as "earthquake", "election" or "car race". Arguments are attribute/value pairs that characterize an event type (for an earthquake, its location, date, magnitude, casualties...). A template is the set of arguments that can describe an event type (earthquake template, election template). The generic representation of an event is based on the rule of the "5 Ws" (What, Who, Where, When, Why) that prevails in the "Anglo-Saxon" way of writing articles. This rule stipulates that a good description of an event must make these five elements explicit.

In automatic information extraction, the information about "Who", "Where" and "When" are extracted by a traditional and quite generic named entity recognition approach. On the other hand, the "What" is very domain-specific. For this reason, traditional IE systems lean on templates predefined by experts and identify events in texts with either rule-based systems or statistical models. 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.

2 Description

The global aim of the ASRAEL project is to build a fully-unsupervised event extraction system. However, the goal of this proposed internship can be seen as an intermediate goal, seeking at reducing the amount of necessary supervision in event extraction.

Agence France Presse (AFP) is one of the partners of the project. They provide us with their newswire article corpus from 2004 to present, as well as textual chronologies of events and a few structured datasets containing the attributes of events of the same kinds (for example, a list of plane crashes, together with their date, location, plane type, casualties, cause, etc.).

The intern will work on a distantly supervised system aiming at consolidating and updating such datasets. The different steps of such a system will be the following:

1. Use structured instances of events as described in the existing datasets as

seed for a bootstrapping approach;

2. Find textual descriptions of these events in the newswire corpus;
3. Build a classifier from these descriptions;
4. Run the classifier on the entire corpus to find new instances or news descriptions of existing instances;
5. Build an update procedure for the analysis of new articles.

Two main differences exist between the proposed approach and existing distant supervision approaches [1, 3]:

- The eventive nature of the relations, making them temporally constrained and not always true (also explored in [2]);
- The fact to some attributes may not been named entities (e.g. the cause of a crash).

3 Application

We are particularly interested in candidates with a solid background in computer science and strong programming skills, having a good knowledge of machine learning and/or natural language processing.

As most of the data are in French, knowledge of French basics is a plus.

Applications should include:

- Cover letter outlining interest in the position
- Names of two referees
- Curriculum Vitae (CV)

The intern will be given a "bonus" (was 546,01 € in 2016) + half a "Navigo" (or "Imagine R") pass.

Contact for questions and applications:

Xavier.Tannier[at]limsi.fr

References

[1] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, pages 1003–1011, Suntec, Singapore, 2009. Association for Computational Linguistics.

[2] Kevin Reschke, Martin Jankowiak, Mihai Surdeanu, Christopher Manning, and Daniel Jurafsky. Event Extraction Using Distant Supervision. In *Proceedings of the 9th International Language Resources and Evaluation (LREC'2014)*, Reykjavik, Iceland, May 2014.

[3] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, September 2015. Association for Computational Linguistics, Morristown, NJ, USA.