Stage de master 2 / Graduate internship
Automatic Classification of Claims from Political Debates and Declarations

Keywords: natural language processing, text mining, machine learning, computational journalism, fact-checking.

Ideal starting date: March/April 2017

Duration: 4-6 months

Advisor: Xavier Tannier (LIMSI-CNRS)

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

Context (ANR Project ContentCheck)

Fact checking is the task of assessing the factual accuracy of claims, generally made by public figures such as politicians, entrepreneurs, etc. Fact-checking is part and parcel of journalists' everyday work, either while working independently on an article, or as part of vetting done in the newsroom before publication, to prevent the publication of innacurate information. Modern fact-checking is faced with a triple revolution in terms of scale, complexity, and visibility: many more claims are made and disseminated through Web and social media, they represent a complex reality and their investigation requires using multiple heterogeneous data source. Our project[2] brings together academic labs with expertise in data management, natural language processing, automated reasoning and data mining, and a fact-checking team of journalists from a major French Web media.

In recent years, journalists and the computer scientists started talking to each other in order to identify which technologies could help journalists' everyday work. This space of exchanges is known as *Computational Journalism* [1]. At this level, it encompasses very diverse uses and tools such as learning how to correctly or better use a database, producing simple spreadsheet-based visuali-

[1] `https://www.limsi.fr/en/access`
[2] `https://team.inria.fr/cedar/contentcheck/`

sations that are however personalized or better adapted to Web format, optical character recognition for scanned texts in order to conduct computer-based keyword search or using statistics to analyse public data from an interesting point of view, thus highlighting interesting trends. These latter uses are referred to by the term *Data Journalism* [2].

Description

The process of fact checking requires many challenging steps; one of them is to separate factual claims from opinions, beliefs, hyperboles, questions, etc. and to discern which are "check-worthy", *i.e.* deserve to be considered and checked by the journalists [3].

The intern will work on this particular task: (s)he will build a tool extracting automatically the check-worthy claims and classifying them between different predefined classes (such as "doubtful number", "doubtful fact", "opinion", "contextualization needed", etc.), in order to make the watch easier for the journalist.

Examples of claims to classify could be:

- "40 % de la taxe ont été détourné pour rémunérer le capital d'une société italienne privée" (number to check)
- "25 % du chiffre d'affaires d'Amazon se fait le dimanche." (number to check) "On peut continuer à ne vouloir laisser travailler que les multinationales anglo-saxonnes qui paient peu d'impôts dans notre pays le dimanche mais ça n'est pas la bonne solution." (opinion, need for contextualization)
- "J'ai été le premier avec Wolfgang Schäuble à signer une lettre pour que nous soyions capables de mettre en place cette coopération renforcée à onze." (fact to check)
- "Je veux abroger le droit du sol" (possible contradiction with a former claim by the same person)
- "Je ne peux pas accepter que les États-Unis soient devenus du point de vue de l'énergie indépendants grâce au gaz de schiste et que la France ne puisse pas profiter de cette nouvelle énergie" (opinion, need for contextualization)

Dataset

A labeled dataset in French will be provided by our partners from the newspaper Le Monde. It will contains political claims coming from different sources:

- Newspaper articles
- Debates
- Speeches
- Twitter and other social networks
- etc.

Approach

We will model this problem as a classication task and follow a supervised learning approach to tackle it.

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] Sarah Cohen, James T. Hamilton, and Fred Turner. Computational Journalism. *Communications of the ACM*, 54(11):66–71, 2011.

[2] Jonathan Gray, Lucy Chambers, and Liliana Bounegru. *The Data Journalism Handbook*. O'Reilly, 2012.

[3] Naeemul Hassan, Chengkai Li, and Mark Tremayne. Detecting Checkworthy Factual Claims in Presidential Debates. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015)*, Melbourne, Australia, October 2015.