Chatbot for OCDE statistic databases

(Ioana Manolescu and Xavier Tannier)

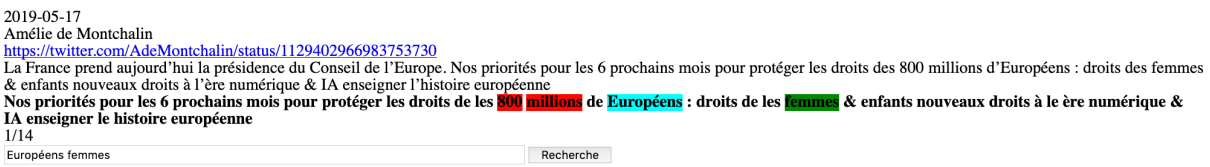
Statistic data sources are compiled across the world by local, national, or international organizations. They provide valuable information on topics such as economy, education, health etc. Statistic data sources are used to elaborate public policies, as well as to assess policy effectiveness through its measurable results, in particular in media analyses.

The Organisation for Economic Co-operation and Development (**OECD**) is an international organisation whose goal is to shape policies that foster prosperity, equality, opportunity and well-being for all. Together with governments, policy makers and citizens, OECD works on establishing international norms and finding evidence-based solutions to a range of social, economic and environmental challenges.

Together with the UN, the European Central Bank, the International Monetary fund and other organizations, OECD has put forward on a standard, called **SDMX**, for describing and accessing statistic databases. This allows developing Web applications based on such statistics.

OCDE is currently interested in developing a 'Guided search' or 'Research assistant' **chatbot**, capable to engage a conversation with the bot in order to find relevant data (typically, 'what is the population of Utrecht?'…). The bot would ask questions to specify the user's need ('do you mean Utrecht the city or Utrecht the province?' 'Which is the year of reference you are looking for?', etc.) and take the person to the wished result (and possible switch to human assistance if unsuccessful). The final result of the conversation can be one figure, one trend, one graph or table, one file. Ideally, caveats and disclaimers attached to the response would also be promoted to the user.
Such development could contribute to bring new services to facilitate access to reference, quality data, and empower citizens, NGOs and media in cross-checking facts more easily.

In the past few years, Inria and U. Paris Sorbonne have carried research on a similar project: (i) given a keyword query, e.g., "Unemployment France 2015", find the closest datasets published by INSEE for that query; if possible, find within a dataset the precise value that the user is looking for [1]; (ii) identify, in a text statement, all the statistic claims which refer to the INSEE dataset [2]. A tool has been developed (cf. screen shot below) that puts these algorithms to work on a corpus of tweets selected by Les Décodeurs, with whom we collaborated within the ContentCheck project.


2019-05-17
Amélie de Montchalin
https://twitter.com/AdeMontchalin/status/1129402966983753730
La France prend aujourd'hui la présidence du Conseil de l'Europe. Nos priorités pour les 6 prochains mois pour protéger les droits des 800 millions d'Européens : droits des femmes & enfants nouveaux droits à l'ère numérique & IA enseigner l'histoire européenne
Nos priorités pour les 6 prochains mois pour protéger les droits de les 800 millions de Européens : droits de les femmes & enfants nouveaux droits à le ère numérique & IA enseigner le histoire européenne
1/14
Européens femmes
Recherche

The purpose of the internship is to investigate the construction of a prototype chatbot (or at a minimum, a keyword search engine) for SDMX statistics, and to validate its interest as a proof of concept with OECD (contact: Eric Anvar, Head of Smart Data Practices and Solutions). This will leverage natural language processing, information extraction, and data management techniques and algorithms. The Python code base of the previously developed work [1, 2] can be used as a starting point.

Supervisors: Ioana Manolescu (Inria and Ecole Polytechnique, ioana.manolescu@inria.fr, Turing building, office 1029) and Xavier Tannier (Paris Sorbonne Université, xavier.tannier@sorbonne-universite.fr)

References

1. "Search for Truth in a Database of Statistics", Tien-Duc Cao, Ioana Manolescu, Xavier Tannier, WebDB workshop, co-located with ACM SIGMOD, 2018
2. "Extracting statistical mentions from textual claims to provide trusted content" by Tien-Duc Cao, Ioana Manolescu and Xavier Tannier, in the NLDB conference 2019