Summary of the INTERREG DECRIPT project for qualitative market research interviews

21st of July 2020

General goal of the project

Automatically process textual data for their governance, security and use in the field of prospective, by being able to automatically identify and protect personal data in natural language texts (office documents, messages, etc.). The specific skill developed is to understand the semantics of a text thanks to linguistic artificial intelligence. The semantic model that will be proposed thus goes further than a rule engine.

Within this framework three pilot applications will be developed.

Project Issue: Data Security Issues

- Ensuring the security of data has become essential for their collection and use, including for texts that may contain personal data.
- An important reference is the new GDPR regulation ("General Data Protection Regulation"), which came into force in May 2018. The GDPR requires that any company (European and foreign) must be able to prove that the personal data it holds are protected and, above all, unusable in case of theft. Examples of protected data: IBAN, telephone numbers, various identifiers, biometric data, camera recordings, etc.
- Other data protection legislation concerns companies and organisations. Switzerland is awaiting its future revised data protection law. In the USA, the California Consumer Privacy Act (CCPA) has been adopted in other states.
- The implementation of data protection regulations creates new requirements for the identification of personal data. Thus, a new job is being created: "Data Protection Officer" (DPO), which consists mainly of deleting or masking / offending the data of persons identified in documents. More generally, the DPO ensures that the company complies with the applicable regulations.
- This has created a new need for automatic tools to identify and mask/offset protected data, including in texts, in order to facilitate companies' compliance with the legislation. Tracing becomes mandatory.
- Bringing economic actors into compliance with the GDPR, beyond compliance with the regulation, may be the origin of a new "contract of trust" with customers, partners, employees ... satisfied with the protection granted to their data.

Project specificity: Linguistic modelling approach

- Linguistic work will determine how data specific to individuals and of value to users are expressed in textual documents.
- A linguistic model - meta-model - for data retrieval will be developed to be used to automate the identification of this data.

Pilot Application 1: Data Governance

- The data governance application must ensure that procedures are in compliance with the GDPR, while identifying valuable data. Data governance applications are designed to enable the digital enterprise to minimize the risk of exposure.
- A first specific goal of the data governance application is to extend current data governance to textual data, by "mapping" between the semantics of textual data and the semantics of structured enterprise data. A second specific goal of the data governance application is to enable audits on the use of personal data and to guarantee the protection of personal data.

Pilot Application 2: Identification and masking of personal data

- The objective of the application is to facilitate and automate the anonymisation (or “obfuscation”) of all types of personal data at the heart of text (numeric) zones, of various origins and forms (customer notice, document, e-mail, etc). This issue concerns all organisations (private and public) in their approach to data protection (particularly in the application of the GDPR), whether internally, for the purpose of outsourcing a processing operation, or for their outgoing flows. The main issues are that personal data can be of different natures and origins, and are complex to locate and type when they are found within a text. The application must therefore ensure the detection, localisation, typing, and then masking of personal data within a text field and according to various selectable procedures (substitution, encryption or “obfuscation”). As a result, the application produces anonymised texts, replacing personal data with values that no longer carry any information, moreover, these values do not allow the original information to be retrieved.
- Terms used in practice for this type of operation: obfuscation, anonymisation, un-identification, de-personalisation, data masking, pseudonymisation, data cloaking, data masquerading, tokenisation, data obfuscation, data privacy, data sanitisation, data scrambling, data de-identification. The reverse operation is called re-identification or de-anonymisation.

Pilot Application 3: Detection of unnamed entities

- The objective of the unnamed entity identification application is to detect, in a text, an entity: the name of a person (natural or legal), organisation, place or product/service, and to disambiguate it, even when this entity is not explicitly named. Example: "City at the end of the lake" should be able to refer to "Geneva".
- A specific goal of this unnamed entity detection application is to integrate the identification of entities not explicitly named in an information analysis process, as well as to evaluate its quality and real impact. To this end, a model and a linguistic approach will be used, which will allow for better results in terms of accuracy. In the end, the analysis of large volumes of information will be accelerated.

Project organization and planning

- INTERREG Swiss-French Project funded by the public authorities concerned (FEDER, Interreg Swiss Federal, Canton of Bern, Canton of Geneva, Canton of Wallis) and the partners of the project.
- Partners of the project :
  - Leader France : University of Franche-Comté in Besançon.
  - Leader Switzerland : University of Applied Sciences (documentary information) in Geneva.
  - Other french partner : ERDIL in Besançon.
  - Other Swiss partners : University of Applied Sciences Arc (business management) in Neuchâtel and Global Data Excellence SA in Geneva
- Project start : 1 January 2020
  Planned completion of project : 31 December 2021
- Collaboration is ensured with IBL (Institute for International Business Law of the Faculty of Law of the University of Fribourg), whose current research focuses on "Sharing Economy & Law and Law & Artificial Intelligence", with direct applications to GDPR.

Market research

The market research aims to identify the needs and restrictions in terms of data detection, governance and obfuscation among private, para-public and public user companies and organisations, as well as data analysis service providers. All organisations handling data in the broadest sense are concerned, as these data include not only names, but also addresses and identifiers of all kinds.

The needs and restrictions for which the identification is targeted by the market research concern the recognition and processing of data relating to individuals, whether such data is explicitly named, such as a telephone number, or is implicit, as data can be represented in various ways, by a synonym, a description, an abbreviation, etc. The data can also be used to identify a person or a group of people.

Market research also focuses on business models for making the best use of the economic values that are created by the data.

21st of July 2020