NLP for textual analysis of judicial proceedings
Lucie Gianola, Julien Longhi

To cite this version:


HAL Id: halshs-01562856
https://halshs.archives-ouvertes.fr/halshs-01562856
Submitted on 17 Jul 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
NLP for textual analysis of judicial proceedings
Lucie Gianola & Julien Longhi
AGORA, Université de Cergy-Pontoise
lucie.gianola@u-cergy.fr, julien.longhi@u-cergy.fr

Issue

Department of Criminal Analysis of Pôle Judiciaire de la Gendarmerie Nationale (PJGN) supports investigation teams by realizing manual extractions of entities from proceedings (vehicles, persons, places, dates, phone numbers) to build relational and chronological diagrams through data analysis software.

⇒ Reading and extracting entities is tedious and time consuming.
⇒ Corpus includes witness and suspect interviews, pictures (crime scene, autopsy, clues, etc.), detailed phone invoices, requisitions from hospitals, toll stations...
⇒ For decades-old cold cases, documents are scanned and optical character recognition is operated, with various degrees of accuracy.
⇒ We focused on extracting information from interviews, since it represents a large source of messy, unstructured data.

Workflow

Matching temporal expressions & phone numbers

Temporal expressions (hours, dates and intervals) allow analysts to build timelines reconstructing events.

⇒ Issue : vague temporal expressions given by interviewees
⇒ Key : detection of time expression phrases : en fin d'après-midi (late afternoon), le 14 ou le 15 juillet (the 14th or 15th of July), vers le mois de mars (by March).

We built Unitex grammars [5] recovering temporal expressions. Unitex is an open source platform which allows among other things to build automaton recognizing parts of texts, as seen on figure 1. Phone numbers are matched with a regular expression.

Results

Theoretical perspectives

⇒ Detecting officer’s reformulation : neutrality, misquoting, and attention span issues,
⇒ Discriminating informations on linguistic criterias (biographical information/informations directly related to the matter investigated) [1],
⇒ Issue a good pratices guide for interviews : encouraging interviewees to be as precise as possible especially about temporal expressions.

References


Acknowledgements

This PhD thesis is funded by département du Val d’Oise and Comue Paris Seine.