NLP for textual analysis of judicial proceedings

Lucie Gianola, Julien Longhi

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.

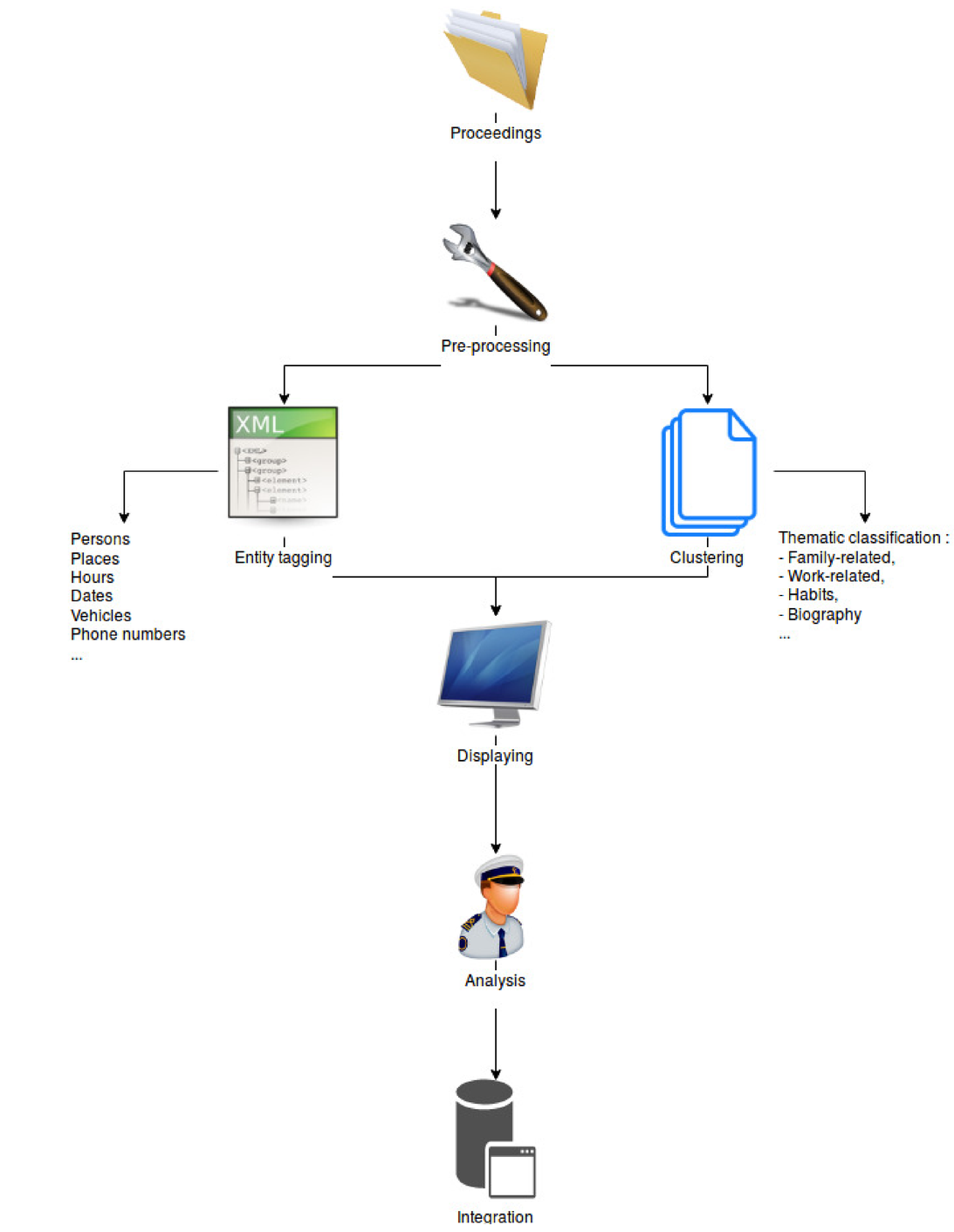
⇒ Reading and extracting entities is tedious and time consuming.

⇒ 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.

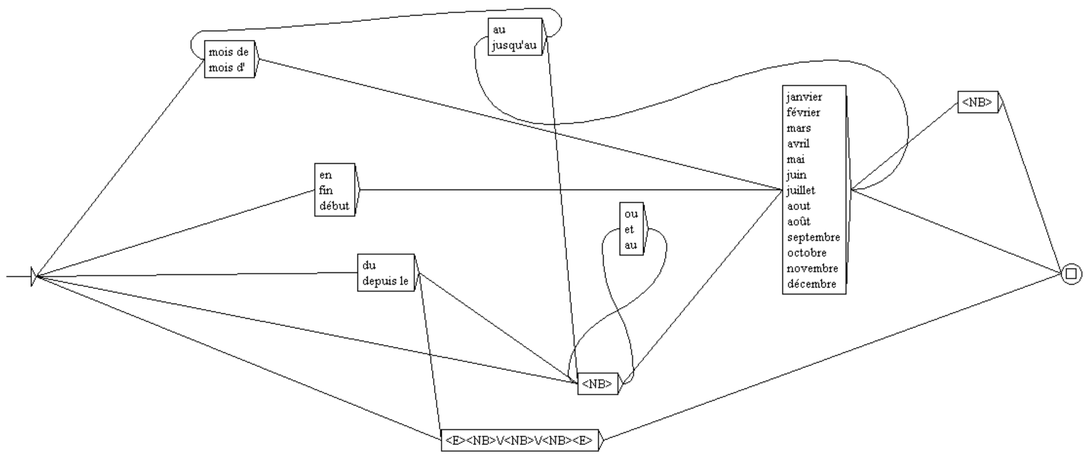


Figure 1: Unitex graph matching dates

Results

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="auditions.xsl"?>
<procedure>
<audition>
<texte>
Pièce : <filename>A1 03 - Audition M**** F****</filename>
QUESTION : Pouvez vous nous communiquer à nouveau votre emploi du temps
  complet de la journée du <DATE>21 août 2009</DATE>?
REPONSE : Le matin je me suis levé à <HORAIRE>3H15</HORAIRE>, j'ai pris un
  café, je me suis habillé, j'ai embrassé Monique et je suis parti travailler
  vers <HORAIRE>3H45</HORAIRE>, je suis arrivé à l'Intermarché de MURET vers
  <HORAIRE>4H05</HORAIRE>. J'ai travaillé jusqu'à <HORAIRE>11H15</HORAIRE>
  environ, je suis allé acheter des pinces pour les nappes et des tournevis au
  Bricorama a coté de l'Intermarché. Je suis ensuite rentré à la maison vers
  <HORAIRE>11 H45</HORAIRE>, Alice et Joe étaient là. Monique était parti
  sortir les deux filles. Monique est rentrée vers <HORAIRE>12H30</HORAIRE>,
  j'avais fait des escalopes de veaux aux champigons, que j'avais ramassé la
  veille dans la forêt de CUGNAUX vers le gros chêne, en l'attendant.
</texte>
</audition>
</procedure>
```

Treatments provide a XML and tagged file, machine-readable. Once tagged, text is displayed in a web browser through XSL and CSS style-sheet (figure 2). Entities are highlighted, and phone numbers extracted to another page, listing them in a table (figure 3).

QUESTION : Pouvez vous nous communiquer à nouveau votre emploi du temps complet de la journée du **21 août 2009**?
REPONSE : Le matin je me suis levé à **3H15**, j'ai pris un café, je me suis habillé, j'ai embrassé Monique et je suis parti travailler vers **3H45**, je suis arrivé à l'intermarché de MURET vers **4H05**. J'ai travaillé environ, je suis allé acheter des pinces pour les nappes et des tournevis au Bricorama a coté de l'Intermarché. Je suis ensuite rentré à la maison vers **11 H45**, Alice et Joe étaient là. Monique était parti sortir les deux filles. Monique est rentrée vers **12H30**, j'avais fait des escalopes de veaux aux champigons, que

Figure 2: Highlighted text displayed in a web browser

Audition	Numéros
A1 03 - Audition M	01
A1 04 - Audition N	06 06 03
A1 06 - Audition P	06. 04.

Figure 3: Phone numbers table

Drawbacks

⇒ Unitex graphs work as a rule-based system : a human person has to enter all occurrences and searched forms in the graph.

⇒ Graphs will fail to detect occurrences that contain typing errors.

⇒ Statistical approaches are less harsh, but require a lot of training data.

Operational perspectives

- Improving Unitex grammars and trying open search engines such as ElasticSearch,
- Introducing machine learning techniques : training algorithms on questions as anchors to infer topic, generalizing to larger portions of text, clustering documents according to main subject,
- High-end achievement with temporal expressions : automatically draw time lines from text extraction,
- Designing an interface allowing analysts to browse through documents efficiently.

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

[1] Eensoo E. and Valette M. Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments. In Gilles Sérasset Georges Antoniadis, Hervé Blanchon, editor, *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, pages 367–374, 2012.

[2] Tapaswi M., Bäuml M., and Stiefelhagen R. Storygraphs : Visualizing character interactions as a timeline. In *Conference on Computer Vision and Pattern Recognition*, 2014.

[3] Lafon P. Sur la variabilité des formes dans un corpus. *Mots*, (1):127–165, 1980.

[4] Marchand P. Améliorer la négociation de crise. *Le Journal du CNRS*, (286):25, 2016. interview by Laure Cailloce.

[5] Paumier S. Unitex 3.1 user manual, 2016. http://www-igm.univ-mlv.fr/ unitex.

Acknowledgements

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