Automated Geoparsing of Paris Street Names in 19th Century Novels

L. Moncla, M. Gaio, T. Joliveau, and Y-F. Le Lay
Introduction

Background

Methodology

Conclusion
Introduction

Objectives

Objectives of the project

- Retrieve, map and analyze the occurrences of place names in fictional novels

Corpus of novels

- 31 French novels
- Published between 1800 and 1914
- Action occurs wholly or partly in Paris
OUTLINE

1 Introduction

2 Background
   Digital humanities
   Named Entity Recognition

3 Methodology

4 Conclusion
Background

Digital humanities

Spatial analysis of literary texts

- Generation of data sets
- Spatial representation of social or spatial relationships
- Visualisation and interpretation of texts (historical, novels, ...) 

Use of NLP for text mining

- Geoparsing texts
- Named Entity Recognition (NER) and Toponym resolution
Background

Named Entity Recognition

NER approaches

- Data-driven: machine learning
- Knowledge-based: heuristic and handcrafted rules

Named Entity (NE)

- Pure and descriptive proper names
- Absolute and relative spatial named entities
Background

Named Entity Recognition

Extended Named Entity (ENE)

- Entity built with a proper name and may be composed of one or more concepts
- Several levels of overlapping

1. a. *la rue* **La Fayette**
   ‘La Fayette Street’

2. b. *le quartier de la* **Goutte-d’Or**
   ‘Goutte-d’Or neighbourhood’
Related work

Extended Named Entity

**Figure** – Feature structure representing an ENE of level 1

**Figure** – Feature structure representing an ENE of level 1
Automated Geoparsing of Street Names
L. Moncla

OUTLINE

1 Introduction

2 Background

3 Methodology
   Overview
   Results
   Combine NLP and textometric analysis

4 Conclusion
Methodology
Overview

Main objective

Automatically retrieve street names from novels published in the 19th century and whose action occurs in Paris.

Three steps

- Extract spatial named entities
- Locate these entities using historical sources and gazetteers
- Create maps adapted to the specificities of literary spaces
Methodology
Extracting street names via CQL requests

TXM platform

- http://sf.net/projects/txm
- Implements lexicometric methods for content analyses of text corpora
- CQL request find occurrences of specific entities and TXM produces a concordancer

CQL requests used in TXM

```
[lemma="rue"%cd][word!="\.|\,|\;|\!|\?|\s|\-|une|\-|ou\ainsi|et|aurait|\-l"%c]? [word!="\.|\,|\;|\!|\?|\s|\-|une|\-|ou\ainsi|et|aurait|\-l"%c] ? [word="\p{Lu}."& word!=\Ça|Ah|O|Venez|Et|M|L."
```

```
[frlemma="rue"%cd] [word!="\p{P}+" ] ? [word!="\p{P}+" ] ? [word!="\p{P}+" ] ? [word="\p{Lu}."" ]
```
Methodology
Named entity recognition and classification

PERDIDO NER processing chain

- [http://erig.univ-pau.fr/PERDIDO/](http://erig.univ-pau.fr/PERDIDO/)
- A geographically oriented NER system
- Retrieve, tag and extract extended named entities
- Produce annotated XML files and a concordancer

Cascaded finite-state transducers

- Transducers are developed and processed with the Unitex platform
Methodology
Named entity recognition and classification

XML/TEI output produced by the PPC

```xml
<placeName>
  <geogName type="R" subtype="ST">
    <geogFeat>
      <w lemma="rue" type="N">rue</w>
    </geogFeat>
    <w lemma="de" type="PREP">de</w>
    <name>
      <w lemma="Rivoli" type="NPr">Rivoli</w>
    </name>
  </geogName>
</placeName>
```

- PERDIDO annotates also geo-semantic information (spatial relations, motion verbs, ...)

Automated Geoparsing of Street Names
L. Moncla

GeoHumanities’17 – 13/22
Results
Extraction of street names

Evaluation scores

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F$_1$-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TXM</td>
<td>98.3</td>
<td>98.5</td>
<td>98.4</td>
</tr>
<tr>
<td>PERDIDO</td>
<td>99.7</td>
<td>99.0</td>
<td>99.3</td>
</tr>
</tbody>
</table>

Classified errors

<table>
<thead>
<tr>
<th></th>
<th>TXM</th>
<th>PERDIDO</th>
</tr>
</thead>
<tbody>
<tr>
<td># of results</td>
<td>2607</td>
<td>2583</td>
</tr>
<tr>
<td>false positive</td>
<td>44</td>
<td>7</td>
</tr>
<tr>
<td>false negative</td>
<td>39</td>
<td>26</td>
</tr>
<tr>
<td>malformed</td>
<td>127</td>
<td>4</td>
</tr>
</tbody>
</table>
## Results

*Extraction of street names*

### Most frequent geographical feature types

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Occurrences</th>
<th>Feature type</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>rue</td>
<td>2583</td>
<td>quartier</td>
<td>106</td>
</tr>
<tr>
<td>boulevard</td>
<td>257</td>
<td>porte</td>
<td>105</td>
</tr>
<tr>
<td>maison</td>
<td>200</td>
<td>place</td>
<td>81</td>
</tr>
<tr>
<td>faubourg</td>
<td>149</td>
<td>bois</td>
<td>75</td>
</tr>
<tr>
<td>hôtel</td>
<td>134</td>
<td>avenue</td>
<td>68</td>
</tr>
<tr>
<td>pont</td>
<td>123</td>
<td>barrière</td>
<td>62</td>
</tr>
<tr>
<td>quai</td>
<td>122</td>
<td>route</td>
<td>58</td>
</tr>
</tbody>
</table>

- Use of a lexicon or a geographic ontology
- 112 feature types are found in the corpus
Combining NLP and textometric analysis

Interoperability between NER and textometric tools

- Building a fully automatic and more generic process
- The XML/TEI files produced by PERDIDO are compatible with TXM
- TXM provides innovative analytical corpus tools
Textometric analysis
Preliminary results

Occurrences of the four most frequent street names
Textometric analysis

Preliminary results

Distribution of street names normalized by number of words
Textometric analysis

Preliminary results

Map showing the number of occurrences of street names
Conclusion

Retrieving street names from novels

- The evaluation validate the choice of the NER method for the automatic process
- Results show the interest of combining NLP approaches and textometric analysis tools

Further work

- Finish the other steps (creating maps once the gazetteer will be finished)
- Use more geo-semantic annotations provided by PERDIDO
  - semantic content associated with the spatial named entities
- Visualization of displacements of a character and the representation of the temporal dynamics of places
Thank you for your attention

CONTACT

Ludovic Moncla (IRENav)
ludovic.moncla@ecole-navale.fr

Mauro Gaio (LIUPPA)
mauro.gaio@univ-pau.fr

Thierry Joliveau (EVS)
thierry.joliveau@univ-st-etienne.fr

Yves-François Le Lay (EVS)
yves-francois.le-lay@ens-lyon.fr