Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers

Sophie Rosset\(^\alpha\)  Cyril Grouin\(^\alpha\)  Karën Fort\(^\beta,\gamma\)
Olivier Galibert\(^\delta\)  Juliette Kahn\(^\delta\)  Pierre Zweigenbaum\(^\alpha\)

\(\alpha\) LIMSI–CNRS, France  \(\beta\) INIST–CNRS, France  \(\gamma\) LIPN, France  \(\delta\) LNE, France

LAW 2012: 6th Linguistic Annotation Workshop
제주도, 대한민국  *(Jeju-do, Korea)* – July 12th, 2012
Introduction

Context

- **Quaero Project:**
  - Extracting information from news:
    - Proposal of a definition for *extended* and *structured* named entities; guidelines → (Rosset et al. 2011);
    - Annotation of two press corpora (1.5 million of words each one) used in two evaluation campaigns.

- **Corpus annotation:**
  - 2011: Broadcast News (BN) corpus, radio and television shows (Grouin et al. 2011; Galibert et al. 2011);
  - 2012: Old Press (OP) corpus, French newspapers from December 1890 (Galibert et al. 2012).

- Aims of this work: to compare annotations in both corpora.
Introduction

Named Entities
Text element classifiable on a semantic level:

- MUC-6: *person*, *location*, *organization*
- Numerical types: *date*, *time*, *money*
- Existing proposals:
  - finer-grained classes (*person* → *politician*, *location* → *city*);
  - new class: *product*, hierarchy w/ 200 types (Sekine 2004);
  - to fit historical data: *ships*, *regiments*, *railroads* (American Civil War).

Original objective
Answer to basic questions: *Who? What? Where? When?*
Extended Named Entities

Our definition

- New types (*products*, *functions*),
- New coverage (expressions w/o proper nouns allowed),
- Structuring of the entities:
  - **Hierarchy**: types/subtypes taxonomy;
    - Type *person*:
      - Subtype *individual*: pers.ind
      - Subtype *collective*: pers.coll
    - Special subtypes:
      - *.oth* (other subtype than those proposed)
      - *.unk* (I don’t know which subtype to use).
  - **Compositionality**: entity composed of
    - types/subtypes (out of 31),
    - components (out of 30).
Extended Named Entities

Compositionality
Each entity type includes at least one component:

pers.ind

name.first  name.last
   |       |
François  Hollande
Extended Named Entities

Compositionality

Another entity can act as a component:

Mr Fiat, General Superior of the Lazarists
Extended Named Entities

Metonymy and Antonomasia

An entity type can be used to refer to another type:

Mr Berthelot was succeeding him at rue de Grenelle
“rue de Grenelle” = “Ministry of Education”
## Some numbers

<table>
<thead>
<tr>
<th></th>
<th>BN corpus</th>
<th>OP corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Test</td>
</tr>
<tr>
<td># show/pages</td>
<td>188</td>
<td>18</td>
</tr>
<tr>
<td># words</td>
<td>1,291,225</td>
<td>108,010</td>
</tr>
<tr>
<td># entity types</td>
<td>113,885</td>
<td>5,523</td>
</tr>
<tr>
<td>mentions</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td># entities w/</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>correction</td>
<td></td>
<td></td>
</tr>
<tr>
<td># components</td>
<td>146,405</td>
<td>8,902</td>
</tr>
<tr>
<td>mentions</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td># components</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/ correction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Adaptation of the annotation

From Broadcast news to Old press

- OCRed Old Press corpus characteristics:
  - some remaining incorrectly recognized characters;
  - fixed-size columns from the original formatted text:
    → some remaining line breaks and hyphenations.

- Annotation adaptation to the Old Press corpus:
  - **Attribute “correction”**
    → annotators corrected incorrectly recognized entities:
    `<loc.adm.town correction="d’Alger"> d’Algor </loc.adm.town>`
  - **Component “noisy-entities”**
    → one or several entities combined due to a segmentation error (involves an entity boundary):
    `<noisy-entities correction=”M. Montmerqué, ingénieur”>`
    M. Montmerqué, ingénieur
    `</noisy-entities>`
Annotation evaluation

Creation of a mini reference corpus

- **Selection** of a sub-corpus from the training corpus
- **Annotation** by 2 teams of 2 annotators (A₁, A₂, B₁, B₂)
- **Adjudication:**
  1. within each team: A₁+A₂ / B₁+B₂
  2. from the previous ones: A+B
  3. with the annotated sub-corpus: AB+sub-corpus
→ **mini-reference corpus.**

Inter-annotator Agreement

<table>
<thead>
<tr>
<th>Which markables? See (Grouin et al. 2011)</th>
<th>BN</th>
<th>OP</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure <em>(highest possible bound)</em></td>
<td>0.845</td>
<td>0.799</td>
</tr>
<tr>
<td>All annotated entities as markables <em>(κ, lowest possible bound)</em></td>
<td>0.713</td>
<td>0.647</td>
</tr>
</tbody>
</table>
Comparisons

Broadcast News vs. Old Press annotation campaigns

- **Source material** (more problems in OP corpus):
  - OCR errors that do not appear:
    - “touché” (touched) instead of “Fouché” (last name)
  - combined entities: “M. Montmerqué, ingénieur”

- **Language** (OP corpus is more difficult):
  - Specific languages: religious language, abbreviations;
  - Cultural context: geographical divisions from 1890.
    - Tonkin: country (loc.adm.nat) or region (loc.adm.reg)?
  - Annotation difficulties: boundary delimitation more difficult:

```
org.adm           loc.adm.nat
    kind     qualifier     name                 name
Comité consultatif d’hygiène publique de France
```

*Consultative committee for public hygiene of France*
Comparisons

Broadcast News vs. Old Press corpora

- **Statistical test** (Welch Two Sample t-test) to compare distribution of types across the corpora:

Figure: 19 entity types with $p < 0.001$, ranked by decreasing order of significance (top: BN corpus; bottom: OP corpus)
Comparisons

Broadcast News vs. Old Press corpora

- **Statistical test** (Welch Two Sample t-test) to compare distribution of components across the corpora:

![Diagram of comparison](image)

**Figure:** 17 components with $p < 0.001$, ranked by decreasing order of significance (top: BN corpus; bottom: OP corpus)
## Comparisons

Structure differences across corpora

<table>
<thead>
<tr>
<th>PATTERN</th>
<th>BN</th>
<th>OP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type <code>&lt;pers.*&gt;</code> (person)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- composed of <code>&lt;name.first/&gt;</code> and <code>&lt;name.last/&gt;</code></td>
<td>52%</td>
<td>6%</td>
</tr>
<tr>
<td>- includes a <code>&lt;name.first/&gt;</code></td>
<td>69%</td>
<td>19%</td>
</tr>
<tr>
<td>- composed of <code>&lt;title/&gt;</code> and <code>&lt;name.last/&gt;</code></td>
<td>2%</td>
<td>34%</td>
</tr>
<tr>
<td>- includes a <code>&lt;title/&gt;</code></td>
<td>8%</td>
<td>44%</td>
</tr>
<tr>
<td>Type <code>&lt;org.*&gt;</code> (organization)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- <code>&lt;org.adm&gt;</code> <code>&lt;kind/&gt;</code></td>
<td>6%</td>
<td>29%</td>
</tr>
</tbody>
</table>
Comparisons

Broadcast News vs. Old Press corpora

- **Automatic classification** based upon the distribution of types and components (73 tag ratios) across the corpora:

<table>
<thead>
<tr>
<th>Method</th>
<th>FP</th>
<th>FN</th>
<th>FP+FN</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Rule</td>
<td>22</td>
<td>12</td>
<td>34</td>
<td>0.919</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>0.983</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0.993</td>
</tr>
<tr>
<td>SVM</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**Table:** Classification based on tag ratio (Weka toolbox)

- The ratios are discriminant enough to determine the corpus a document belongs to.
Conclusion and perspectives

- Same annotation scheme used in two corpora:
  - similar overall sizes (# tokens, # types and components)
  - but different annotation times.
- Comparisons made possible due to the structured definition;
- Human annotation process more difficult in OP;
- Future work:
  - further studies of comparison,
  - detecting relations between information,
  - new corpora annotation (w/ parallel FRE/ENG corpora).
- The corpora will soon be made available for free to the scientific community through ELDA catalogue.

Acknowledgements

This work was partly realized as part of:

- Quaero project funded by OSEO, French State agency for innovation
- ETAPE project funded by ANR, French National Agency for Research


