Named Entity Recognition for Distant Reading in ELTeC

Francesca Frontini¹, Carmen Brando², Joanna Byszuk³, Ioana Galleron⁴, Diana Santos⁵, Ranka Stanković⁶

CLARIN annual conference
5–7 October 2020
Overview of the talk

Introduction

Manual annotation
  Annotation guidelines

Software infrastructure
  BRAT
  NER & Beyond

Comparing manual and automatic annotation
  Testing Automatic NER
  Results
  Analysis

Links to CLARIN

Links
“Distant Reading for European Literary History (COST Action CA16204) is a project aiming to create a vibrant and diverse network of researchers jointly developing the resources and methods necessary to change the way European literary history is written. Grounded in the Distant Reading paradigm (i.e. using computational methods of analysis for large collections of literary texts), the Action will create a shared theoretical and practical framework to enable innovative, sophisticated, data-driven, computational methods of literary text analysis across at least 10 European languages.”

https://www.distant-reading.net
Motivation

- Most NER systems and studies are
  - not specifically concerned with literary texts
  - not dealing with 19th century texts
  - not multilingual enough
- We wanted to explore the issues in COST texts
  - what kind of entities
  - what kind of problems
- Create an infrastructure to evaluate current NER systems for literary texts in the languages of COST
Exercise

- Languages: Czech, German, English, French, Hungarian, Norwegian, Portuguese, Serbian and Slovene
- "Lightweight semantic annotation" defined: marking persons, demonyms, professions and other roles, works, places, facilities and organizations
- 5 random passages of 400 white space-delimited tokens taken from 20 novels from each language, manually annotated using brat
### Table: Current manually NE-annotated corpus

<table>
<thead>
<tr>
<th></th>
<th>DEMO</th>
<th>EVENT</th>
<th>LOC</th>
<th>ORG</th>
<th>OTHER</th>
<th>PERS</th>
<th>ROLE</th>
<th>WORK</th>
</tr>
</thead>
<tbody>
<tr>
<td>cze</td>
<td>163</td>
<td>5</td>
<td>275</td>
<td>0</td>
<td>0</td>
<td>1150</td>
<td>454</td>
<td>0</td>
</tr>
<tr>
<td>deu</td>
<td>66</td>
<td>2</td>
<td>323</td>
<td>12</td>
<td>0</td>
<td>973</td>
<td>458</td>
<td>4</td>
</tr>
<tr>
<td>eng</td>
<td>56</td>
<td>7</td>
<td>198</td>
<td>37</td>
<td>0</td>
<td>1184</td>
<td>203</td>
<td>25</td>
</tr>
<tr>
<td>fra</td>
<td>77</td>
<td>3</td>
<td>262</td>
<td>22</td>
<td>128</td>
<td>900</td>
<td>244</td>
<td>18</td>
</tr>
<tr>
<td>hun</td>
<td>29</td>
<td>7</td>
<td>152</td>
<td>20</td>
<td>0</td>
<td>1091</td>
<td>367</td>
<td>7</td>
</tr>
<tr>
<td>nor</td>
<td>4</td>
<td>8</td>
<td>83</td>
<td>25</td>
<td>3</td>
<td>990</td>
<td>201</td>
<td>10</td>
</tr>
<tr>
<td>por1</td>
<td>17</td>
<td>9</td>
<td>351</td>
<td>19</td>
<td>0</td>
<td>940</td>
<td>490</td>
<td>54</td>
</tr>
<tr>
<td>por2</td>
<td>34</td>
<td>1</td>
<td>256</td>
<td>30</td>
<td>7</td>
<td>1059</td>
<td>347</td>
<td>7</td>
</tr>
<tr>
<td>slv</td>
<td>133</td>
<td>54</td>
<td>336</td>
<td>37</td>
<td>0</td>
<td>1230</td>
<td>620</td>
<td>2</td>
</tr>
<tr>
<td>srp</td>
<td>121</td>
<td>18</td>
<td>185</td>
<td>11</td>
<td>0</td>
<td>985</td>
<td>301</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>700</td>
<td>114</td>
<td>2425</td>
<td>213</td>
<td>138</td>
<td>10514</td>
<td>3685</td>
<td>131</td>
</tr>
</tbody>
</table>

---

1. Portuguese the exercise was done twice: with canonical, modernized orthography works (por1), and with non-canonical, old-orthography works (por2)
Annotation guidelines

Literary text show a broad variety NEs with respect to more standardised non-fictional texts.

- People are often referred to by profession or their origin, as well as by family relations only ("Maman")...
- Fictional characters may be present and animals and objects may have proper names.
- Additional types of entities, which are particularly interesting for the purposes of literary and cultural analysis, require annotation, such as titles of works of art, books, publications; literary movements, ...

Classification is sometimes difficult and there is a risk of proliferation of categories and sub-categories. E.g. "Vive la Réforme! à bas Giuzot!" (here "Réforme" refers to a proposal of reform of electoral law which the politician Giuzot opposed).
Manual annotation infrastructure: BRAT

Despenharem-se outra vez no abismo da ignorância! Meu Deus, meu Deus, compadece-te d'elles!</p>

Não desesperes assim!»</p>

Verás se alguém torna a fallar n'isto!</p>

Decorre um anno depois que Manuel Félix e sua mulher se tinham separado do padre. Este continuava quasi entretendo, em Santarém; os seus passeios limitavam-se a andar encostado a um pau, á roda do cercado ou quintal povoado de bananeiras, araticús, mangueiras e, laranjeiras, que havia na casa do irmão. As horas do calor atavam-se as redes nas arvores; o padre deitava-se n'uma d'ellas, lamentando sempre o seu estado e a perda dos seus índios das cachoeiras ; e Romualdo sentava-se n'outra, consolando-o com palavras aitectuosas, ou lendo-lhe, n'algum livro favorito, cousas de religião e de moral.</p>

De padre que o substituisse nunca ninguém lhe fallou mais. Os espíritos andavam preoccupados, no Pará e na corte, com as noticias da revolução de 1820, em Portugal. Quem podia lembrar-se, vendo abalados pela base os alicerces das velhas sociedades europeas, de que havia um rio no Brazil chamado Tapajós, onde uns pobres índios necessitavam de quem os ajudasse a completar a obra da sua redempção ?! A sabedoria humana trabalhava em grande n'esse momento; não podia descer a miudezas; ocupava-se do geral, e não do particular ; pensava em reformar nações, e não em civilizar aldeias. O vento que andava no mundo sacudia as cabeças, e fazia cair d'ellas opiniões que espantavam os próprios que as enunciavam. Era 1820 o precursor das novas ideias, que annos depois transformavam a colónia em pais independente e a metrópole em terra de homens livres. Era e governos, diante da gravidade das circumstancias, tratavam da própria salvação, 6 não da alheia. O padre Félix comprehendeu isto ; deu a sua causa por perdida, e, sem se resignar nem esquecer d'ella, deixou de esperar o remédio, que sempre suspeitou que lhe não mandariam.
NER & Beyond

Annotation tagset mapping and transliteration

Description of a procedure for harmonisation of two annotation tag sets:

1. User uploads a set of source annotated documents with any tagset with one or more document with gold (target) annotation tagset
2. System reads both tagsets and offers mappings (for each tag, some option should be selected)
3. System generates source documents in a new annotation scheme
4. User downloads the results

Upload a .zip file with following structure:

```
<table>
<thead>
<tr>
<th>f.zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>gold/</td>
</tr>
<tr>
<td>md1.ann</td>
</tr>
<tr>
<td>mdm2.ann</td>
</tr>
<tr>
<td>to_map/</td>
</tr>
<tr>
<td>x1.ann</td>
</tr>
<tr>
<td>x2.xml</td>
</tr>
<tr>
<td>x3.txt</td>
</tr>
</tbody>
</table>
```

Select an archive:

```
Choose file No file chosen
```

Upload

---

Convert BRAT to CoNLL.02 format (.py source)

Upload a .zip file with following structure:

```
<table>
<thead>
<tr>
<th>f.zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1.ann</td>
</tr>
<tr>
<td>f1.txt</td>
</tr>
<tr>
<td>f2.ann</td>
</tr>
<tr>
<td>f2.txt</td>
</tr>
</tbody>
</table>
```

Select an archive:

```
Choose file No file chosen
```

Select tokenizer:

```
Unspecified
```

Upload

---

Convert XML with NE tags to CONLL.02 format

Upload a .zip file with following structure:

```
<table>
<thead>
<tr>
<th>files.zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1.xml</td>
</tr>
<tr>
<td>f2.xml</td>
</tr>
</tbody>
</table>
```

Select an archive:

```
Choose file No file chosen
```

Upload
Web portal NER & Beyond: tools

At University of Belgrade and Jerteh, http://nerbeyond.jerteh.rs/

▶ Conversion tools among different annotation formats
  ▶ BRAT to CoNLL02
  ▶ XML with NE tags to CONLL02
  ▶ XML with NE tags to BRAT
  ▶ BRAT (ann & txt) to XML
  ▶ CONLL02 to BRAT

▶ Automatic Named entity recognition (and annotation)
  ▶ spaCy (eng, fra, ita, nld, ger, por, srp, spa, multi )
  ▶ Stanford (eng, ger, srp)
  ▶ CVNER (ser)

▶ Computation of statistics based on BRAT .ann format
▶ NER evaluation with the Gemini tool
  ▶ Precision and recall tables
  ▶ Visual comparison in HTML
▶ Annotation tagset harmonization (mapping and transliteration)
Example of comparison of gold (manual, blue) and automatic (SpaCy, pink) annotation

And at this moment entered the room the young nobleman whom we have before mentioned, accompanied by an individual who was approaching perhaps the termination of his fifth lustre but whose general air rather betokened even a less experienced time of life. **Tall**, with a well-proportioned figure and a graceful carriage, his countenance touched with a sensibility that at once engages the affections. **Charles Egremont** was not only admired by that sex, whose approval generally secures men enemies among their fellows, but was at the same time the favourite of his own.

“Ah, Egremont! come and sit here,” exclaimed more than one banqueter.

“I saw you waltzing with the little **Bertie**, old fellow,” said Lord **Fitzheron**, “and therefore did not stay to speak to you, as I thought we should meet here. I am to call for you, mind.”

“How shall we all feel this time to-morrow?” said **Egremont**, smiling.

“The happiest fellow at this moment must be **Cockie Graves**, ” said Lord **Milford**, “He can have no suspense. I have been looking over his book, and I defy him, whatever happens, not to lose.”

“Poor **Cockie**,” said **Mr Berners**: “he has asked me to dine with him at the Clarendon on Saturday.”

“**Cockie** is a very good,” **Cockie**,” said Lord **Milford**, “and any gentleman sportsman present wishes to give seven to two, I will take him to any arm.”

“My book is made up,” said **Egremont**: “and I stand or fall by its merits.”
Testing Automatic NER

- Case study on four collections only, in English, French, Portuguese and Serbian
- For each collection, we tested two tools: one common for all (spaCy) and another one language specific
- BRAT outputs were compared to annotations produced by these tools
- Evaluation of string detection was strict (segments must match exactly)
## Testing Automatic NER - results

<table>
<thead>
<tr>
<th></th>
<th>Cat</th>
<th>Correct</th>
<th>Missing</th>
<th>Spurious</th>
<th>Precision</th>
<th>Recall</th>
<th>Excess</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEM-fra</td>
<td>LOC</td>
<td>73</td>
<td>112</td>
<td>84</td>
<td>0.465</td>
<td>0.395</td>
<td>0.535</td>
</tr>
<tr>
<td></td>
<td>PERS</td>
<td>82</td>
<td>512</td>
<td>115</td>
<td>0.416</td>
<td>0.138</td>
<td>0.584</td>
</tr>
<tr>
<td>SPACY-fra</td>
<td>LOC</td>
<td>103</td>
<td>78</td>
<td>468</td>
<td>0.180</td>
<td>0.569</td>
<td>0.820</td>
</tr>
<tr>
<td></td>
<td>PERS</td>
<td>329</td>
<td>194</td>
<td>297</td>
<td>0.526</td>
<td>0.629</td>
<td>0.474</td>
</tr>
<tr>
<td>PALAVRAS-por1</td>
<td>LOC</td>
<td>223</td>
<td>63</td>
<td>44</td>
<td>0.835</td>
<td>0.780</td>
<td>0.165</td>
</tr>
<tr>
<td></td>
<td>PERS</td>
<td>816</td>
<td>90</td>
<td>86</td>
<td>0.905</td>
<td>0.901</td>
<td>0.095</td>
</tr>
<tr>
<td>SPACY-por1</td>
<td>LOC</td>
<td>225</td>
<td>84</td>
<td>440</td>
<td>0.338</td>
<td>0.728</td>
<td>0.662</td>
</tr>
<tr>
<td></td>
<td>PERS</td>
<td>465</td>
<td>256</td>
<td>374</td>
<td>0.554</td>
<td>0.645</td>
<td>0.446</td>
</tr>
<tr>
<td>PALAVRAS-por2</td>
<td>LOC</td>
<td>151</td>
<td>67</td>
<td>91</td>
<td>0.624</td>
<td>0.693</td>
<td>0.376</td>
</tr>
<tr>
<td></td>
<td>PERS</td>
<td>857</td>
<td>133</td>
<td>285</td>
<td>0.750</td>
<td>0.866</td>
<td>0.250</td>
</tr>
<tr>
<td>SPACY-por2</td>
<td>LOC</td>
<td>157</td>
<td>57</td>
<td>396</td>
<td>0.284</td>
<td>0.734</td>
<td>0.716</td>
</tr>
<tr>
<td></td>
<td>PERS</td>
<td>569</td>
<td>236</td>
<td>393</td>
<td>0.591</td>
<td>0.707</td>
<td>0.409</td>
</tr>
<tr>
<td>Stanford-eng</td>
<td>LOC</td>
<td>98</td>
<td>100</td>
<td>126</td>
<td>0.438</td>
<td>0.495</td>
<td>0.563</td>
</tr>
<tr>
<td></td>
<td>PERS</td>
<td>649</td>
<td>535</td>
<td>399</td>
<td>0.619</td>
<td>0.548</td>
<td>0.381</td>
</tr>
<tr>
<td>SPACY-eng</td>
<td>LOC</td>
<td>98</td>
<td>100</td>
<td>170</td>
<td>0.366</td>
<td>0.495</td>
<td>0.634</td>
</tr>
<tr>
<td></td>
<td>PERS</td>
<td>536</td>
<td>648</td>
<td>240</td>
<td>0.691</td>
<td>0.453</td>
<td>0.309</td>
</tr>
<tr>
<td>SrpNER-srp</td>
<td>LOC</td>
<td>107</td>
<td>78</td>
<td>19</td>
<td>0.849</td>
<td>0.578</td>
<td>0.151</td>
</tr>
<tr>
<td></td>
<td>PERS</td>
<td>718</td>
<td>267</td>
<td>158</td>
<td>0.820</td>
<td>0.729</td>
<td>0.180</td>
</tr>
<tr>
<td>SPACY-srp</td>
<td>LOC</td>
<td>57</td>
<td>128</td>
<td>104</td>
<td>0.354</td>
<td>0.308</td>
<td>0.646</td>
</tr>
<tr>
<td></td>
<td>PERS</td>
<td>553</td>
<td>432</td>
<td>315</td>
<td>0.637</td>
<td>0.561</td>
<td>0.363</td>
</tr>
</tbody>
</table>

**Table:** Results of the strict evaluation, per language and category.
A strict evaluation of detection is often penalising for PERS, because of honorifics which we chose to include in our annotation.

It is further complicated by the fact that the XML annotated input was processed as such by tools which often expect plain text.

In most cases, LOC seems to be less problematic for the pre-trained models.
The NE corpus, as part of the larger annotated ELTeC corpus, will have to adhere to the FAIR requirements, for long term preservation, visibility, accessibility. A deposit in a repository (possibly Textgrid) is envisaged at the end of the project.

Accessibility could be further enhanced by making the corpus part of the CLARIN federated content search.

Finally, CLARIN collaboration is welcome for the creation of better adapted NE tools which can overcome the problems that we have identified for state of the art tools.
Useful links

▶ The COST Action web page: https://www.distant-reading.net
▶ The ELTeC corpus:
  https://distantreading.github.io/ELTeC/index.html
▶ The current NE annotated sub-corpus
  http://brat.jerteh.rs/#/eltec-simplified/
Acknowledgements

Presented research is supported by COST Action 16204 – Distant Reading for European Literary History

* * *

We thank Lou Burnard for creating the samples, Tajda Liplin, Zala Vidic, Karolina Zgaga, Michael Preminger, Tonje Vold, Emma Takács, Silvie Cinková, Klára Macháčková, and Cvetana Krstev for annotating the texts, and Branislava Šandrih for help in creating the sites at the University of Belgrade, and all our colleagues at the COST action Distant Reading for European Literary History.
Thank you!