Named Entity Recognition for Distant Reading in Several Languages

Ranka Stanković¹, Diana Santos², Francesca Frontini³, Tomaž Erjavec⁴, Carmen Brando⁵

DH_Budapest
25–27 September 2019
Overview of the talk

Introduction

Some analyses
  Annotation guidelines
  Interesting issues
  Problems detected
  Methodological choices
  Inter-annotator agreement
  Professions mentioned
  Representativity

Software infrastructure
  BRAT
  NER & Beyond

Further work
“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, 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: Preliminary data on main classes of NE (as of May 21st, 2019, 20:44)

<table>
<thead>
<tr>
<th></th>
<th>eng</th>
<th>fra</th>
<th>por</th>
<th>hun</th>
<th>nor</th>
<th>slv</th>
<th>srp</th>
<th>cze</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>53093</td>
<td>52031</td>
<td>53958</td>
<td>58825</td>
<td>52900</td>
<td>53348</td>
<td>67232</td>
<td>52650</td>
</tr>
<tr>
<td>NE</td>
<td>1744</td>
<td>1669</td>
<td>1883</td>
<td>1805</td>
<td>1344</td>
<td>2412</td>
<td>1962</td>
<td>3424</td>
</tr>
<tr>
<td>DEMO</td>
<td>56</td>
<td>77</td>
<td>17</td>
<td>29</td>
<td>4</td>
<td>133</td>
<td>121</td>
<td>163</td>
</tr>
<tr>
<td>EVENT</td>
<td>7</td>
<td>3</td>
<td>9</td>
<td>7</td>
<td>8</td>
<td>54</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>FAC</td>
<td>75</td>
<td>70</td>
<td>103</td>
<td>65</td>
<td>41</td>
<td>128</td>
<td>23</td>
<td>108</td>
</tr>
<tr>
<td>ORG</td>
<td>37</td>
<td>22</td>
<td>19</td>
<td>20</td>
<td>25</td>
<td>37</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>PERS</td>
<td>1186</td>
<td>900</td>
<td>940</td>
<td>1091</td>
<td>990</td>
<td>1230</td>
<td>995</td>
<td>1150</td>
</tr>
<tr>
<td>PLACE</td>
<td>126</td>
<td>192</td>
<td>248</td>
<td>87</td>
<td>42</td>
<td>208</td>
<td>163</td>
<td>167</td>
</tr>
<tr>
<td>ROLE</td>
<td>203</td>
<td>244</td>
<td>490</td>
<td>367</td>
<td>201</td>
<td>620</td>
<td>301</td>
<td>454</td>
</tr>
<tr>
<td>WORK</td>
<td>25</td>
<td>18</td>
<td>54</td>
<td>7</td>
<td>10</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Quantitative data - all classes

NER Annotations per Language

- Tag:
  - DEMO
  - EVENT
  - FAC
  - MISC
  - ORG
  - OTHER
  - PERS
  - PLACE
  - QUANTITY
  - ROLE
  - ROLEMISC
  - TEMPORAL
  - WORK

Language:
- cze
- eng
- fra
- hun
- nor
- por
- slv
- srp

Annotations:
- cze: 3200
- eng: 1800
- fra: 1200
- hun: 2400
- nor: 1400
- por: 1000
- slv: 2000
- srp: 1200
Quantitative data - PERS class

Number of Annotations per Language for PERS

Languages

PERS

PERS_GENDER-Collective

PERS_GENDER-Female

PERS_GENDER-Male

PERS_REALITY-Fiction

PERS_REALITY-Real

Languages

srp

slv

por

nor

hun

den

fra

eng

Cze

No of Annotations

5

10

cze

eng

fra

hun

nor

por

slv

srp

n

0

250

500

750

1000

1250

0

250

500

750

1000

1250
Quantitative data - ROLE class

Number of Annotations per Language for ROLE

<table>
<thead>
<tr>
<th>Languages</th>
<th>ROLE</th>
<th>ROLE-Military</th>
<th>ROLE-Nobility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>srp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slv</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>por</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fra</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eng</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cze</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Languages

<table>
<thead>
<tr>
<th>ROLE-Office</th>
<th>ROLE-Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>srp</td>
<td></td>
</tr>
<tr>
<td>slv</td>
<td></td>
</tr>
<tr>
<td>por</td>
<td></td>
</tr>
<tr>
<td>nor</td>
<td></td>
</tr>
<tr>
<td>hun</td>
<td></td>
</tr>
<tr>
<td>fra</td>
<td></td>
</tr>
<tr>
<td>eng</td>
<td></td>
</tr>
<tr>
<td>cze</td>
<td></td>
</tr>
</tbody>
</table>

No of Annotations

Languages:
- ing
- cze
- eng
- fra
- hun
- nor
- por
- slv
- srp

n:
- 2.5
- 5.0
- 7.5
- 10.0
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).
Interesting issues

Interesting issues that emerged and which our guidelines have tried to address:

- **ROLE** is used to identify professions and job titles (Archbishop of Canterbury; Sa Majesté Britannique), and **ROLEMISC** for family relations and epitaphs
- Mentions of gods and other supernatural entities, also named animals, were annotated as **PERS** (persons)
- Swear words or exclamations containing proper names were not considered as named entities ("Oh my God!")
- Fixed expressions or idioms which included proper names, like **Éva leányá** (Eva’s daughter) in Hungarian to describe “woman” were tagged as **MISC**
Problems detected

Some issues remain open such as:

▶ It is sometimes difficult to distinguish types of places, e.g. distinguish between a region and a country, or a village and a city

▶ We created a separate class for FACILITIES covering “localised artefacts”, e.g. Opera, Theater, Prison, etc.

▶ So far no distinction has been made between forms of addressing and naming the addressee, both have been marked PERS
Further problems

- Identification of fictional places and characters, which was initially thought as a requirement for literary analysis, was problematic.
- Difficulty for deciding whether determinants make part of the named entity (e.g. *l’institut Ouly* or *institut Ouly*) or possessive pronouns in family epithets (e.g. *mon oncle*), a known issue in NER and recurrent in literary texts; this sometimes leads to an inconsistent annotation.

```
<p n="FR00201650">Le lendemain, j'entrai à l'institut Ouly. </p>

Assez, <span>mon oncle</span>, au nom du ciel ! dit Josée <span>phine</span> en joignant les mains.
```
Methodological choices

- **Recursivity**: We now allow nested annotations, e.g.
  `<pers><role>General</role> Junot</pers><role>ministre de<place>Belgique</place></role>`

- **Capitalisation**: given that not all languages follow the same capitalisation rules, the use of this feature is highly problematic in annotation guidelines. In particular, family relations or professional groups and individuals are annotated independently of capitalisation. On the other hand, in some languages capitalisation may be a clue for deciding whether an adjective is part of a NE, as in "Old John".
What about inter-annotator agreement?

Measuring inter-annotator agreement (IAA) is common practice in NLP to estimate the reliability of annotations and possible when multiple independent annotators are available, several measures exist and are implemented such as F-measure, Cohen’s Kappa, ...

▶ unfortunately, most COST annotator teams (one per language) could not perform parallel annotation

▶ a small experiment was possible with two French annotators and an excerpt of the French ELTEC data set: 15 paragraphs of texts considered by some of the team members as difficult to annotate
Results on French sample

- Some results:
  - *PERS*: observed agreement: 0.69; Cohen’s kappa: -0.16
  - *PLACE*: observed agreement: 0.64; Cohen’s kappa: -0.13
  - $k=1$ is perfect agreement, $k=0$ means agreement is equal to chance, $k=-1$ means perfect disagreement
  - for the other NER categories, number of annotations is insufficient for conclusive IAA calculation

- Same segments are annotated, with only two differences in terms of NE boundaries (1. F: "Marseillaise" C: "La Marseillaise", 2. C: "représentant Charbonnel", F: "Charbonnel")

- Most differences are in the choice of different categories

- This experience helped to contribute to the NER annotation guidelines but needs to be repeated with more data and new, independent annotators
## Professions mentioned

### Table: Most mentioned professions

<table>
<thead>
<tr>
<th>Lang</th>
<th>Types</th>
<th>Professions (word forms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cze</td>
<td>195</td>
<td>kněz (priest) 5, rychtáři (magistrate) 5, plavec (swimmer) 6, knížete (knight) 6, rychtář (magistrate) 7, knížecí (knight) 11, kníže (knight) 12</td>
</tr>
<tr>
<td>eng</td>
<td>91</td>
<td>Sir 6, Dr. 6, Secretary 6, Colonel 7, General 8, President 10, Professor 13, Queen 13, Lord 16</td>
</tr>
<tr>
<td>fra</td>
<td>147</td>
<td>domestique (servant) 5 colonel (colonel) 6 comte (count) 6 capitaine (captain) 9 général (general) 12</td>
</tr>
<tr>
<td>hun</td>
<td>249</td>
<td>tanár (teacher, professor) 7 tiszttartó (steward, curator, bailiff) 7 gróf (count) 8 huszár (hussar, cavalry man) 10 úr (sir, lord, mr.) 11 grófnő (countess) 11 ur (sir, lord, mr.) 15</td>
</tr>
<tr>
<td>Lang</td>
<td>Types</td>
<td>Professions (word forms)</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>nor</td>
<td>116</td>
<td>Kapellanen (chaplain) 6, Kaptein (captain) 6, Professoren (professor) 6, Unge-Konsulen (consul) 6, Styrmanden (manager) 7, Konsulen (consul) 10, Kapteinen (captain) 11</td>
</tr>
<tr>
<td>por</td>
<td>221</td>
<td>carcereiro (jailer) 9, conselheiro (counsellor) 9, pagem (page) 9, tenente (lieutenant) 9, bispo (bishop) 10, doutor (doctor) 13, padre (priest) 17, fidalgo (nobleman) 18</td>
</tr>
<tr>
<td>slv</td>
<td>386</td>
<td>kapitan (captain) 7, sodnik (judge) 7, kmetje (peasants) 7, grajščak (nobleman) 7, grof (count) 8, vitez (knight) 13, doktor (doctor) 15, profesor (professor) 16</td>
</tr>
<tr>
<td>srp</td>
<td>301</td>
<td>лекар (physician) 5, ћата (junior clerk) 5, учитељ (teacher) 5, кмет (village leader) 6, паша (pasha) 6, биров (policeman) 6, Газда (boss) 6, дуђанџија (owner of the small shop) 6, поп (priest) 7</td>
</tr>
</tbody>
</table>
How representative are these data?

Would our numbers be comparable if we had annotated a different set of texts? A small experiment was done with Portuguese:

<table>
<thead>
<tr>
<th></th>
<th>por</th>
<th>por2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>53958</td>
<td>55429</td>
</tr>
<tr>
<td>NE</td>
<td>1883</td>
<td>1649</td>
</tr>
<tr>
<td>DEMO</td>
<td>17</td>
<td>34</td>
</tr>
<tr>
<td>EVENT</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>FAC</td>
<td>103</td>
<td>78</td>
</tr>
<tr>
<td>ORG</td>
<td>19</td>
<td>30</td>
</tr>
<tr>
<td>PERS</td>
<td>940</td>
<td>990</td>
</tr>
<tr>
<td>PLACE</td>
<td>248</td>
<td>169</td>
</tr>
<tr>
<td>ROLE</td>
<td>490</td>
<td>333</td>
</tr>
<tr>
<td>TEMPORAL</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>WORK</td>
<td>54</td>
<td>7</td>
</tr>
</tbody>
</table>

Table: Comparing two sets of texts

Classics vs. non-classics?
Manual annotation infrastructure: BRAT

1. Manual annotation infrastructure: BRAT

2. NER Guidelines | NER&Beyond | Path: /eltec-gold/por2/POR0046_FraGAmo_Selvagens_sample

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

4. <p n="POR0046398">— Não desesperes assim !</p>

5. <p n="POR0046399">— Verás se alguém torna a fallar n'isto!</p>

6. <p n="POR0046400">Decorre a um anno depois que Manuel Félix e sua mulher se tinham separado do padre. Este continuava quasi entrevido,

7. em Santarém; os seus passeios limitavam-se a andar encostado a um pau, á roda do cercado ou quíntal povoado de bananeiras, araticús,

8. 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,

9. lamentando sempre o seu estado e a perda dos seus índios das cacheorbas; 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>

10. De padre que o substituisse nunca ninguém lhe fallou mais. Os espíritos andavam preoccupados, no Pará e na corte,

11. com as noticias da revolução de 1820, em Portugal. Quem podia lembrar-se, vendo abalados pela base os aliceres 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 redemção ?! 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 precursores das novas ideias, que annos depois transformavam a colônia em paiz independente e a metrópole em terra de homens livres. Rei e governos, diante da gravidade das circunstancias, tratavam da própria salvação, 6 não da alheia. O padre Félix compreendeu 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.</p>
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>.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](#)

Download Sample Archive

---

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

Upload a .zip file with following structure:

```
<table>
<thead>
<tr>
<th>.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](#)

Download Sample Archive

---

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](#)

Download Sample Archive
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 an...
Further work

- Include more languages (German, Italian, Croatian,...)
- Testing more tools for automatic annotation and models
- Resolve open issues in format harmonisation (conversions restriction)
- Compare results from different tools with our gold datasets
- Entity Linking (for places)
- A way forward would be to build on a service-based digital infrastructure such as GERBIL (General Entity Annotation Benchmark Framework - http://aksw.org/Projects/GERBIL.html)
References and Bibliography


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, Ioanna Galleron, 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!