Optimization of Portuguese Named Entity Recognition and Classification
Combination of Local Grammars and Conditional Random Fields Trained with Parsed Corpora

Diego Alves
Božo Bekavac
Marko Tadić

NooJ 2020 International Conference
05/06/2020
Agenda

- Objective
- Data
- Methodology
- Results
- Conclusion and Perspectives
- References
Objective

- Associate NooJ syntactic local grammars with Conditional Random Fields (CRF) probabilistic method to improve Name-Entity Recognition and Classification (NERC) for Portuguese

- Understand the synergy in training CRF models between information coming from parsed corpus and local grammars

- Focus on "TIME" category and its subcategories
Data

- Portuguese Second Harem:
  - Dataset provided for the second edition of an evaluation campaign for Portuguese, addressing named entity recognition (NER)

- Golden set (GS):
  - 129 documents manually annotated according to HAREM guidelines for NER
  - Domains: news, didactic, opinion, blog, questions, interview, legal, literary, promotional, private manuscript
  - 4053 sentences
  - 89634 tokens
Second Harem NER Hierarchy

3-level hierarchy:
- First level: 10 categories
- Second level: 36 types
- Third level: 21 subtypes

Second Harem GS:
- 7846 entities
Second Harem: „TIME” category

- 1189 occurrences in HAREM Golden Set divided in:

<table>
<thead>
<tr>
<th>Type</th>
<th>Subtype</th>
<th>Number of occurrences in HAREM GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME_CALENDAR</td>
<td>DATE</td>
<td>873</td>
</tr>
<tr>
<td></td>
<td>HOUR</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>INTERVAL</td>
<td>63</td>
</tr>
<tr>
<td>DURATION</td>
<td>-</td>
<td>56</td>
</tr>
<tr>
<td>FREQUENCY</td>
<td>-</td>
<td>71</td>
</tr>
<tr>
<td>GENERIC</td>
<td>-</td>
<td>89</td>
</tr>
</tbody>
</table>

- Our study: BIO format
A revolta histórica produz normalmente uma nova forma de pensamento quanto à forma de organização da sociedade. Assim foi com a Reforma Protestante.

No seguimento do colapso de instituições monásticas e do escolasticismo nos finais da Idade Média.
Second Harem – Train and Test

- From Second Harem GS:
  - Random selection of sentences to compose train and test sets (70/30):

- Train:
  - 2842 sentences
  - 63032 tokens
  - 808 entities

- Test:
  - 1211 sentences
  - 26602 tokens
  - 352 entities
Methodology

- **STEP 1:** CRF trained with Parsed Harem Corpus.
  - Harem Corpus automatically annotated with UDPipe tool
    - UDpipe-bosque model
  - CRF:
    - Python library: sklearn_crfsuite
    - Basic Features for all tests (token, token -1 and token +1) - Baseline:
      - Lower case
      - Upper case
      - Token starts with upper case
    - Other Features studied:
      - Part Of Speech tag
      - Morphosyntactic tag
      - Dependency parsing tag

- **STEP 2:** NooJ local grammars.
  - Identification of "TIME" category (BIO)

- **STEP 3:** CRF trained with Parsed and NooJ Features
Step 1 – CRF - Results

- CRF trained with train corpus with basic features and linguistic data:
  - „Time“ category, types and subtypes
  - BIO

<table>
<thead>
<tr>
<th>Added Feature</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.809</td>
<td>0.636</td>
<td>0.700</td>
</tr>
<tr>
<td>Part of Speech info</td>
<td>0.838</td>
<td>0.671</td>
<td>0.735</td>
</tr>
<tr>
<td>Morphosyntactic info</td>
<td>0.805</td>
<td>0.664</td>
<td>0.723</td>
</tr>
<tr>
<td>Parsing info</td>
<td>0.823</td>
<td>0.634</td>
<td>0.708</td>
</tr>
<tr>
<td>Part of Speech + Morphosyntactic info</td>
<td>0.807</td>
<td>0.670</td>
<td>0.727</td>
</tr>
<tr>
<td>Part of Speech + Dependency Parsing info</td>
<td>0.830</td>
<td>0.658</td>
<td>0.727</td>
</tr>
<tr>
<td>Part of Speech + Morphosyntactic + Dependency Parsing info</td>
<td>0.823</td>
<td>0.655</td>
<td>0.721</td>
</tr>
</tbody>
</table>

Results from CRF training using Linguistic Features from Parsed Harem file
Step 2 – NooJ local grammar - Results

- NooJ resources:
  - Portuguese general dictionary (PT-Dict.nod) available in NooJ website
- Local grammar:
  - 22 graphs
  - Identification of Time expressions corresponding to „Time” category in Second HAREM (BIO).
    - B-TEMPO, I-TEMPO, O tags

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train set</td>
<td>0.870</td>
<td>0.661</td>
<td>0.751</td>
</tr>
<tr>
<td>Test set</td>
<td>0.847</td>
<td>0.660</td>
<td>0.741</td>
</tr>
</tbody>
</table>

Evaluation of NooJ annotations
Step 3 – CRF + NooJ - Results

- NooJ tags combined with POS information:

<table>
<thead>
<tr>
<th>Added Feature</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.809</td>
<td>0.636</td>
<td>0.700</td>
</tr>
<tr>
<td>Part of Speech info</td>
<td>0.838</td>
<td>0.671</td>
<td>0.735</td>
</tr>
<tr>
<td>NooJ tags</td>
<td>0.867</td>
<td>0.725</td>
<td>0.757</td>
</tr>
<tr>
<td>Part of Speech + NooJ tags</td>
<td>0.797</td>
<td>0.675</td>
<td>0.700</td>
</tr>
</tbody>
</table>
Step 3 – CRF + NooJ – Detailed Results

- Results for each tag for the best CRF Model (with NooJ tags as feature):

<table>
<thead>
<tr>
<th>Tag</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-DURACAO</td>
<td>0.333</td>
<td>0.059</td>
<td>0.100</td>
</tr>
<tr>
<td>I-DURACAO</td>
<td>0.333</td>
<td>0.089</td>
<td>0.140</td>
</tr>
<tr>
<td>B-FREQUENCIA</td>
<td>1.000</td>
<td>0.545</td>
<td>0.706</td>
</tr>
<tr>
<td>I-FREQUENCIA</td>
<td>1.000</td>
<td>0.516</td>
<td>0.681</td>
</tr>
<tr>
<td>B GENERICO</td>
<td>0.500</td>
<td>0.138</td>
<td>0.216</td>
</tr>
<tr>
<td>I GENERICO</td>
<td>0.800</td>
<td>0.082</td>
<td>0.148</td>
</tr>
<tr>
<td>B-TEMPOCALEND-DATA</td>
<td>0.841</td>
<td>0.821</td>
<td>0.831</td>
</tr>
<tr>
<td>I-TEMPOCALEND-DATA</td>
<td>0.828</td>
<td>0.857</td>
<td>0.842</td>
</tr>
<tr>
<td>B-TEMPOCALEND-HORA</td>
<td>1.000</td>
<td>0.167</td>
<td>0.286</td>
</tr>
<tr>
<td>I-TEMPOCALEND-HORA</td>
<td>0.750</td>
<td>0.194</td>
<td>0.308</td>
</tr>
<tr>
<td>B-TEMPOCALEND-INTERVALO</td>
<td>0.778</td>
<td>0.438</td>
<td>0.560</td>
</tr>
<tr>
<td>I-TEMPOCALEND-INTERVALO</td>
<td>0.778</td>
<td>0.404</td>
<td>0.532</td>
</tr>
</tbody>
</table>
Conclusions and Perspectives

- Part of Speech is the most relevant linguistic information when training CRF for NER
- NooJ pre-annotation can replace POS feature with better results
- Part of Speech and NooJ pre-annotation → Negative synergy (inferior to baseline)

- Disambiguation local grammars may increase NooJ precision and therefore enhance final results
- Use NooJ to pre-annotate following „Time” detailed structure may enhance problematic cases:
  - DURATION
  - GENERIC
References


Questions?