Filozofski fakultet
Sveučilišta u Zagrebu

CLEOPATRA

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

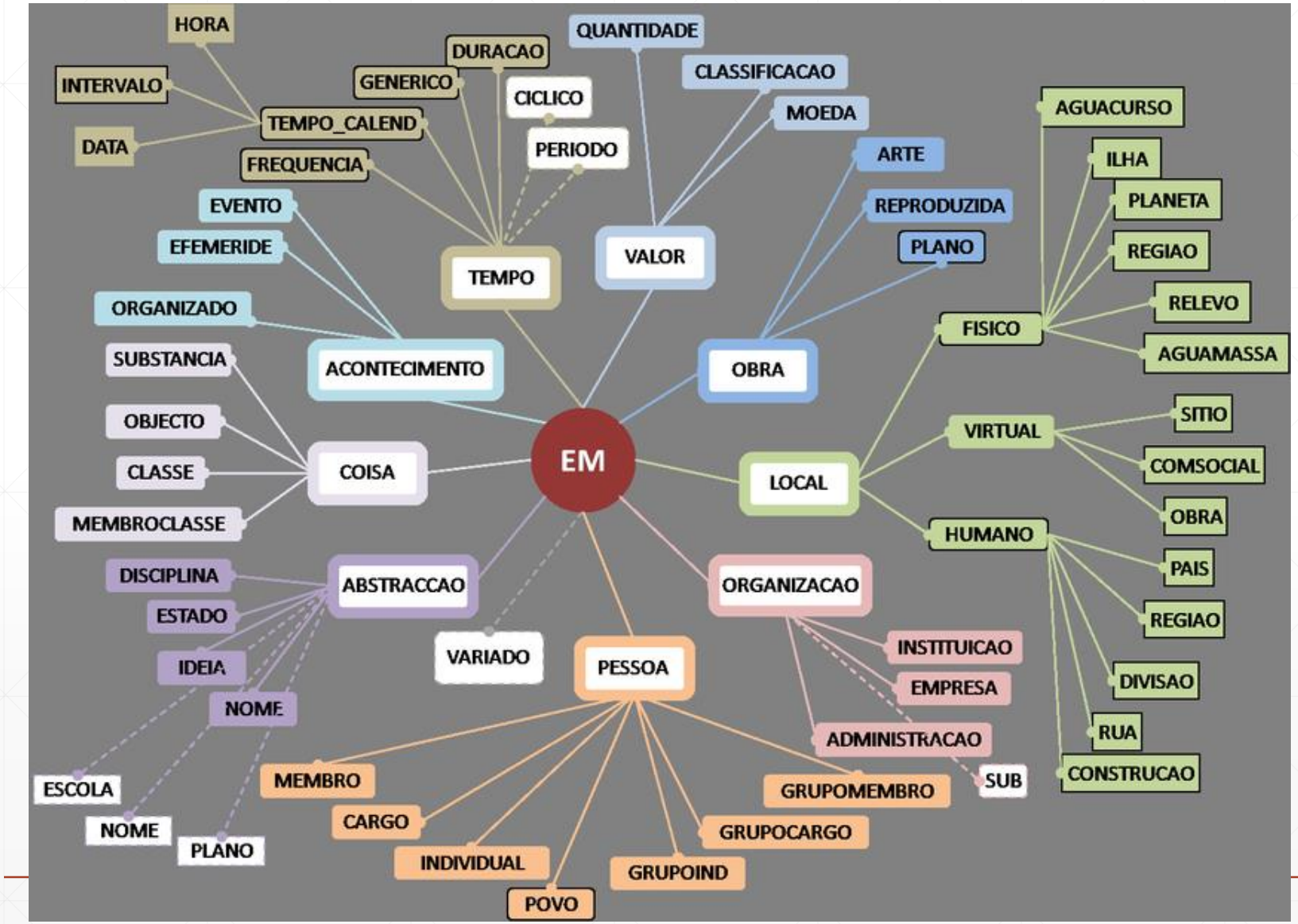
- Understand the synergy in training CRF models between information coming from parsed corpus and local grammars

- 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



Second Harem structure

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:

Type	Subtype	Number of occurrences in HAREM GS
TIME_CALENDAR	DATE	873
	HOUR	37
	INTERVAL	63
DURATION	-	56
FREQUENCY	-	71
GENERIC	-	89

- Our study: BIO format

```xml
<DOC DOCID="H2-dftre765">
  <P>Fatores Demográficos e Econômicos Subjacentes</P>
  <P>
    A revolta histórica produz normalmente uma nova forma de pensamento quanto à forma de organização da sociedade. Assim foi com a
    <EM ID="H2-dftre765-1" CATEG="ABSTRACCAO|ACONTECIMENTO" TIPO="IDEIA|EFEMERIDE">Reforma Protestante</EM>
    . No seguimento do colapso de instituições monásticas e do escolasticismo
    nos finais da <EM ID="H2-dftre765-102" CATEG="OUTRO" COMENT="DUVIDA_DIRECTIVASTEMPO">Idade Média</EM>
    na
    <EM ID="H2-dftre765-37" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="DIVISAO">Europa</EM>
```

Second Harem xml file

```
Anos    B-DURACAO
e   I-DURACAO
anos    I-DURACAO
de  O
destruição  O
compulsiva  O
de  O
interlocutores  O
,   O
alguns  O
moderados   O
,   O
e   O
de  O
opressão    O
,   O
fizeram O
do  O
povo    O
palestiniano    O
uma O
nação   O
de  O
pedintes    O
.   O
```

Second Harem BIO file

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

```
1   Assim   assim   ADV _   _   2   advmod   _   _
2   foi foi ADV _   _   0   root    _   _
3   com com ADP _   _   5   case    _   _
4   a   o   DET _   Definite=Def|Gender=Fem|Number=Sing|PronType=Art    5   det _   _
5   Reforma reforma PROPN   _   Gender=Fem|Number=Sing  2   obl _   _
6   Protestante Protestante PROPN   _   Number=Sing 5   flat:name   _   _
7   .   .   PUNCT   _   _   2   punct   _   _
```

Parsed Second Harem example

Step 1 – CRF - Results

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

Added Feature	Precision	Recall	F1-measure
Baseline	0.809	0.636	0.700
Part of Speech info	**0.838**	**0.671**	**0.735**
Morphosyntactic info	0.805	0.664	0.723
Parsing info	0.823	0.634	0.708
Part of Speech + Morphosyntactic info	0.807	0.670	0.727
Part of Speech + Dependency Parsing info	0.830	0.658	0.727
Part of Speech + Morphosyntactic + Dependency Parsing info	0.823	0.655	0.721

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

	Precision	Recall	F1-measure
Train set	0.870	0.661	0.751
Test set	0.847	0.660	0.741

Evaluation of NooJ annotations

Step 3 – CRF + NooJ - Results

- NooJ tags combined with POS information:

Added Feature	Precision	Recall	F1-measure
Baseline	0.809	0.636	0.700
Part of Speech info	0.838	0.671	0.735
NooJ tags	**0.867**	**0.725**	**0.757**
Part of Speech + NooJ tags	0.797	0.675	0.700

Step 3 – CRF + NooJ – Detailed Results

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

Tag	Precision	Recall	F1-measure
B-DURACAO	0.333	0.059	0.100
I-DURACAO	0.333	0.089	0.140
B-FREQUENCIA	1.000	0.545	0.706
I-FREQUENCIA	1.000	0.516	0.681
B-GENERICO	0.500	0.138	0.216
I-GENERICO	0.800	0.082	0.148
B-TEMPO_CALEND-DATA	0.841	0.821	0.831
I-TEMPO_CALEND-DATA	0.828	0.857	0.842
B-TEMPO_CALEND-HORA	1.000	0.167	0.286
I-TEMPO_CALEND-HORA	0.750	0.194	0.308
B-TEMPO_CALEND-INTERVALO	0.778	0.438	0.560
I-TEMPO_CALEND-INTERVALO	0.778	0.404	0.532

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

- Freitas, C., Mota, C., Santos, D., Oliveira, H.G., Carvalho, P., Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese, LREC, 2010.
- Mota, C., Carvalho, P., Barreiro, A., Port4NooJ v3.0: Integrated Linguistic Resources for Portuguese NLP, LREC, 2016.
- Pirovani, J., Oliveira, E., Portuguese Named Entity Recognition using Conditional Random Fields and Local Grammars, LREC, 2018.
- Sklearn-crfsuite website: https://sklearn-crfsuite.readthedocs.io/en/latest/index.html (Last visited on: 29/05/2020).
- Sutton, C., McCallum, A., An Introduction to Conditional Random Fields, Foundations and Trends in Machine Learning, vol. 4, n.4, p.267-373, 2012.

The work presented has received funding from the European Union's Horizon 2020 research and innovation programme under theMarie Sk lodowska-Curie grant agreement no. 812997 and under the name CLEOPATRA (Cross-lingual Event-centric Open Analytics Research Academy).

Questions?