Entity and relation extraction from web content

André Pires

Supervision:
Sérgio Nunes
José Devezas

What are entities?

Mark, CEO of Facebook,
went to Porto, in 2010.

[Mark]$_{Person}$, CEO of
[Facebook]$_{Organisation}$,
went to [Porto]$_{Location}$, in [2010]$_{Time}$.

ANT project

- Entity-oriented search engine
- Index UP entities (students, staff, departments, etc.)
- Focus on UP news (in SIGARRA), for this project

http://ant.fe.up.pt

Current entity extraction method

- Made with selectors (XPath and CSS)
- Page structure dependant
- Huge work effort and very time consuming
- **Doesn’t work** in SIGARRA’s news

How to improve the extraction?

- Evaluate state-of-the-art methods
- Optimise for the University of Porto domain
  - Using a Portuguese corpus to train the model

Main approaches

- Hand-coded techniques
  - Rule-based
  - Dictionary-based
- Machine learning
  - HMM
  - MEMM
  - CRF
  - SVM
- Ontology-based (OBIE)

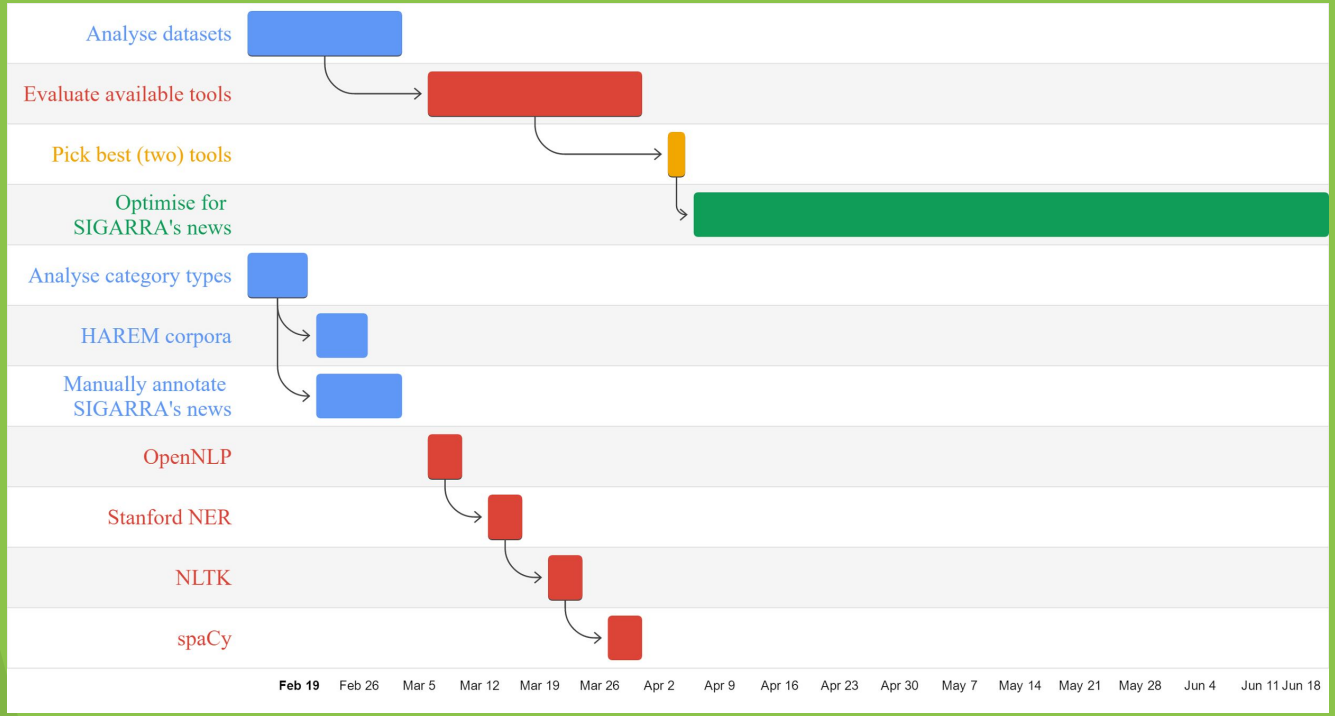
Main evaluation conferences

Conference	Years	Number of entities types	Languages
MUC	1996, 1998	7	English
CoNLL	2002, 2003	4	Spanish, Dutch, English, German
ACE	2003	7	English, Arabic, Chinese
HAREM	2005, 2008	10	Portuguese

What to use?

- Datasets
  - HAREM
  - SIGARRA's news
- Tools
  - OpenNLP
  - Stanford NER
  - NLTK
  - spaCy
- Evaluation
  - CoNLL

Work plan



- Analyse datasets
- Analyse tools
- Pick tools
- Optimise tools

Work plan

- Analyse PT available datasets
  - Analyse entity categories
  - Manually annotate SIGARRA's news
  - Analyse HAREM collection
- Analyse and evaluate available tools, documenting the process
  - OpenNLP
  - Stanford NER
  - NLTK
  - spaCy
- Pick (two) tools with best performance
- Tune them for SIGARRA's news

What is the outcome?

- Provide context to the search engine
- Improve search results
- Reduce work effort and time consumed
- Provide a scalable tool

- New way of extracting entities and relations from web content
- Provide a better search engine to the community

André Pires
ei12058@fe.up.pt

CoNLL evaluation

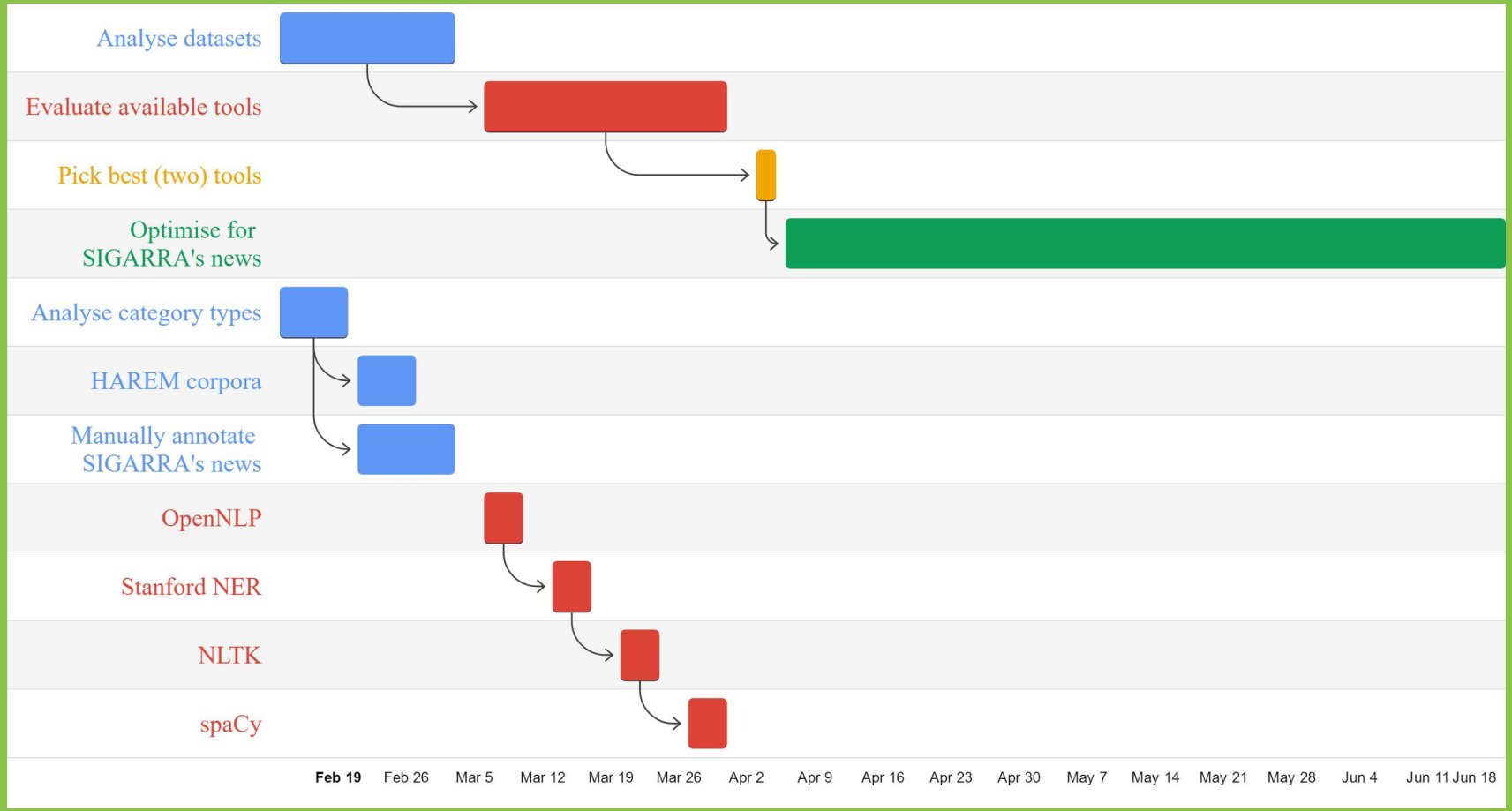
$$precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}$$

$$f - measure = \frac{2 * precision * recall}{precision * recall}$$

TP - *True Positive*, only for **exact matches**

Work plan dates

Task Name	Start Date	End Date	Dur...	Prede...
Analyse datasets	**13-02-2017**	**03-03-2017**	**15d**	
Analyse category types	13-02-2017	20-02-2017	6d	
HAREM corpora	21-02-2017	27-02-2017	5d	2
Manually annotate SIGARRA's news	21-02-2017	03-03-2017	9d	2
Analyse available tools	**06-03-2017**	**31-03-2017**	**20d**	**1**
OpenNLP	06-03-2017	10-03-2017	5d	
Stanford NER	13-03-2017	17-03-2017	5d	6
NLTK	20-03-2017	24-03-2017	5d	7
spaCy	27-03-2017	31-03-2017	5d	8
Pick best two tools	**03-04-2017**	**05-04-2017**	**3d**	**5**
Optimise for use case	06-04-2017	19-06-2017	53d	10


Analyse datasets
Evaluate available tools
Pick best (two) tools
Optimise for SIGARRA's news
Analyse category types
HAREM corpora
Manually annotate SIGARRA's news
OpenNLP
Stanford NER
NLTK
spaCy
Feb 19
Feb 26
Mar 5
Mar 12
Mar 19
Mar 26
Apr 2
Apr 9
Apr 16
Apr 23
Apr 30
May 7
May 14
May 21
May 28
Jun 4
Jun 11
Jun 18

Work plan