Second HAREM

Advancing the State of the Art of Named Entity Recognition in Portuguese

Cláudia Freitas*, Cristina Mota, Diana Santos**,
Hugo Oliveira* and Paula Carvalho***

Linguateca, FCCN
* at Univ. of Coimbra – CISUC / DEI
**at SINTEF ICT,
***at Univ. of Lisbon = Faculty of Sciences, Lasige

LREC 2010 Conference
Valletta, Malta, May, 2010

Linguateca (www.linguateca.pt)

is a distributed network for fostering the computational processing of the Portuguese language

- **Organization of evaluation contests for Portuguese** (Morfolimpíadas, HAREM and CLEF [GeoCLEF, QA@CLEF, adhoc CLEF, GikiP, LogCLEF, GikiCLEF])
- Creation of free resources that enable sophisticated processing of Portuguese
- Monitoring and cataloguing the area

Acknowledgement

- Linguateca and HAREM were funded by the Portuguese government and the European Union with contract number 339/1.3/C/NAC, UMIC and FCCN

HAREM

- Evaluation of named entity recognition in Portuguese texts

Second HAREM

- 10 participants; 27 official runs
- New tracks:
  - recognition and normalization of temporal entities (Hagège et al., 2008)
  - detection of relations between named entities (Freitas et al., 2008, 2009)

Main features (Santos, 2007b)

I. Semantic model

→ NE classified in context

*A morte é reportada no **Diário de Notícias** do dia*
('The death is announced in Diário de Notícias of that day')
→ **LOCAL VIRTUAL COMSOC / place**

*A diferença entre o ´Jornal de Notícias´ e o ´**Diário de Notícias**'*
('The difference between Jornal de Notícias and Diário de Notícias')
→ **COISA CLASSE / thing**

*O seu pai era funcionário público do Ministério da Justiça e crítico musical do ´**Diário de Notícias**´*
('His father was an employee of the Ministry of Justice and a music reviewer for Diário de Notícias')
→ **ORGANIZACAO EMPRESA/ org**

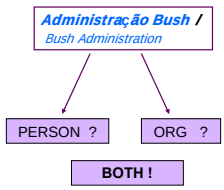
*… foi fotografado pelo **Diário de Notícias** (DN) a fumar uma cigarrilha...*
('had a picture taken by Diário de Notícias smoking a cigarette')
→ **PESSOA GRUPOMEMBRO / person**

Main features

II. Vagueness

→ NE may belong simultaneously to more than one category or type

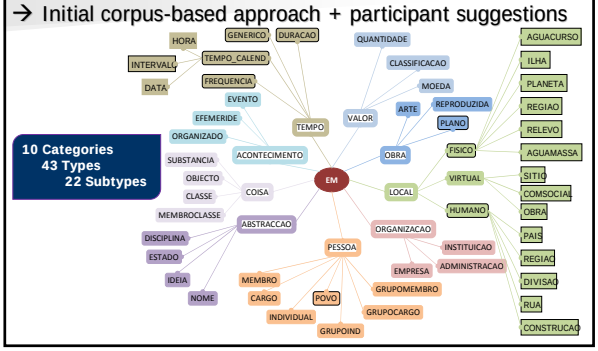
*A **Administração Bush** identifica-se com a Justiça Divina*
('**Bush Administration** takes the role of Divine Providence')



Main features

III. Categories

→ Initial corpus-based approach + participant suggestions



Main Features
IV. Embedded NEs

→ ALT mechanism

*Quantos atletas participaram nos **Jogos Olímpicos de Barcelona?** | How many athletes participated in **Barcelona Olympic Games?***

Barcelona Olympic Games → **EVENT**

Barcelona | Olympic Games

↓ **PLACE** ↓ **EVENT**

<ALT><Jogos Olímpicos de Barcelona | <Jogos Olímpicos> de <Barcelona> </ALT>

Main features
V. Evaluation setup

→ Flexibility

Participants' selective scenarios

Identification
Classification

Participant systems	SCEN	PES	ORG	LOC	OBR	ACO	ABS	COI	TEM	VAL
Cage2	Sel2	CAT	CAT	F + H			Only CATEGORY		CAT	
DobrEM	Pes			Only PLACEs (human and natural)						
PorTexTO	Temp									
Priberam	Tot									
R3M	Sel3									
REMBRANDT	Tot						Only CATEGORY and TYPE			
REMMA	Sel4			C/T					C/T	
SEI-Geo	Sel5			F + H						
SeRELeP	Tot						Normalization of temporal expressions			
XIP/L2F/ XEROX	Sel6								NORM	

New track: ReRelEM

Anaphora resolution
Mitkov, 2000; Collovoni et al., 2007; de Souza et al. 2008

Co-reference
Anaphoric chains in texts

+

Relation detection
Agichtein and Gravano, 2000; Zhao and Grishman, 2005; Culotta and Sorensen, 2004

Fact extraction
World knowledge

Investigate which relations could be found in texts

Devise a pilot task to compare systems that recognize those relations

=

ReRelEM

Reconhecimento de Relações entre Entidades Mencionadas

Relation detection between named entities

Relation inventory

- Identity (ident)

✓ foi fundada em 1131 por **D. Telo** (**São Teotónio**)
It was founded in 1132 by **D. Telo** (**São Teotónio**)

✘ Os adeptos do **Porto** invadiram a cidade do **Porto** em júbilo
The (FC) **Porto** fans invaded the (city of) **Porto**, very happy

- Inclusion (inclui (*includes*) / incluido (*included*))

Hamilton, colega de Alonso na **McLaren**
Lewis Hamilton, Alonso's team-mate in **McLaren**

- Placement (ocorre-em (*occurs-in*) / sede-de (*place_of*))

GP Brasil – Não faltou emoção em Interlagos no **Circuito José Carlos Pace** desde a primeira volta…
GP Brasil – There was no lack of excitement in Interlagos at the **José Carlos Pace Circuit**

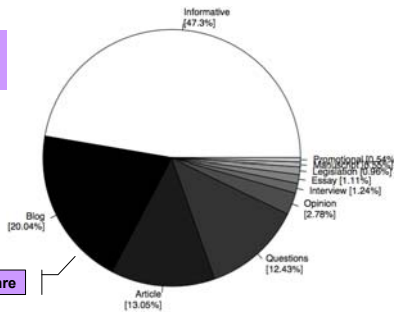
Relation inventory

- Other (`outra`)

Relation / gloss	#
vinculo-inst / affiliation	936
obra-de / work-of	300
participante-em / participant-in	202
ter-participacao-de / has-participant	202
relacao-familiar / family-tie	90
residencia-de / home-of	75
natural-de / born-in	47
relacao-profissional / professional-tie	46
povo-de / people-of	30
representante-de / representative-of	19
residente-de / living-in	15
personagem-de / character-of	12
periodo-vida / life-period	11
propriedade-de / owned-by	10
proprietario-de / owner-of	10
representado-por / represented-by	7
praticado-em / practised-in	7
outra-rel / other	6
nome-de-ident / name-of	4
outra-edicao / other-edition	2

Second HAREM Collection

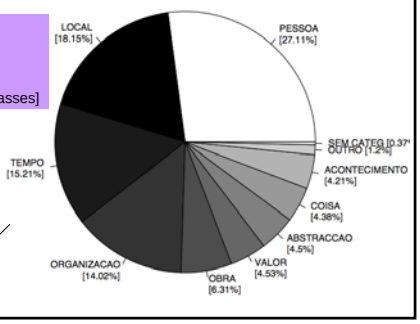
DOCS: 1,040
Paragraphs: 15,737
Words: 670,610



Distribution by text genre

Second HAREM Golden Collection

DOCS: 129
Paragraphs: 2,274
Words: 147,991
NEs: 7,847
Vague NEs: 633 [52 classes]



NE distribution

ReRelEM Golden Collection – full version

DOCS: 129
Paragraphs: 2,274
Words: 147,991
NE: 7,847
Relations: 4,803

ReRelEM relation types

Relations that the systems had to explicitly name

Relations under OUTRA/OTHER

Relation type	#
autor_de/obra_de (authorship)	142
causador_de (agent)	22
consequencia_de (result_of)	1
data_de /datado_de (date of)	105
data_morte (death date)	10
data_nascimento (birth date)	5
ident (identity)	**2229**
inclui/incluido (inclusion)	**854**
local_nascimento_de/natural_de (birth place)	142
localizado_em/localizacao_de (place of)	24
nome_de/nomeado_por (name-of)	56
ocorre_em/sede_de / (location)	**358**
outra_edicao (other edition)	3
outrarel (other relation)	93
participante_em/ter_participacao_de (participation-in)	153
periodo_vida (lifetime)	5
personagem_de (character of)	14
praticado_em/pratica_se/praticante_de/praticado_por (practicing)	99
produtor_de/produzido_por (manufacturing)	50
proprietario_de/propriedade_de (ownership)	39
relacao_familiar (kinship relation)	88
relacao_profissional (professional relation)	17
residente_de/residencia_de (place of residence)	19
vinculo_inst (affiliation)	275
TOTAL	4803

ReRelEM Golden Collection – full version

ReRelEM relations per category

Relations per category	#
ABSTRACCAO/ abstraction	255
ACONTECIMENTO/ event	168
COISA / thing	175
LOCAL / place	960
OBRA / title	274
ORGANIZACAO / org	783
OUTRO / other	25
PESSOA / person	1286
TEMPO / time	192
VALOR / value	19

Evaluation HAREM

HAREM score = 1 + sumN((1-Wcat) * catcerta* α + (1- Wtipos) * tipocerta*β + (1-Wsub) * subcerta*γ) – sumM(Wcat* catesp*α + Wtipos* tipoesp*β + Wsub* subesp*γ)

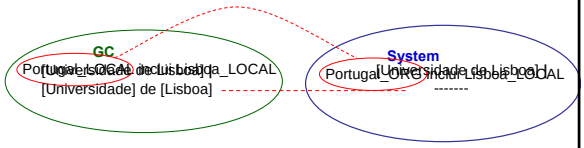
N = number of classification in the GC
M = number of spurious classifications in the participant's run
Wcat = 1/number of categories in the scenario; Wtipo=1/number of types...
α, **β**, **γ** = weights for categories (1), types (0.5) and subtypes (0.25)
(cat, tipo, sub)certa = 1, when it is right; = 0 when wrong
(cat, tipo, sub)esp= 1, when spurious ; = 0 when not

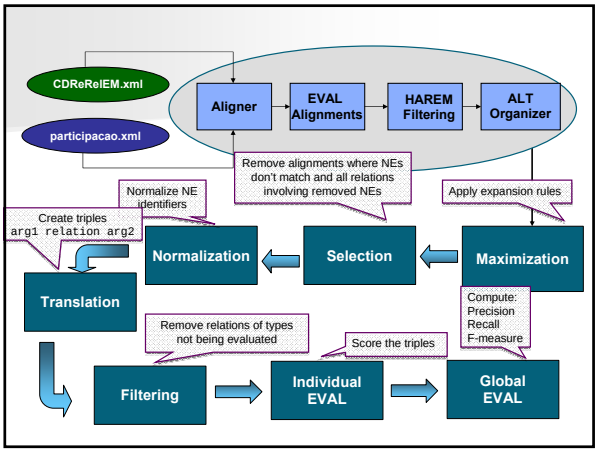
Evaluation ReRelEM

- Evaluate JUST the relations (not the NE)

Relations with mismatched arguments were ignored

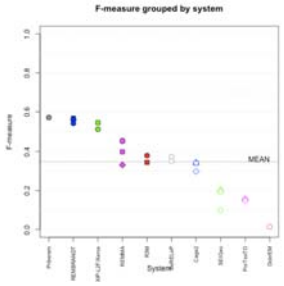
Alternative segmentations were ignored





Participation and results HAREM

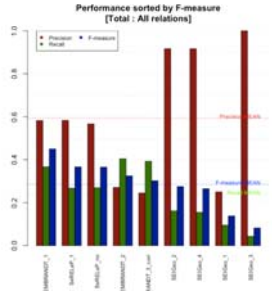
- Only two systems (Priberam and REMBRANDT) tried to recognize the complete set of categories;
- Only one system (R3M) adopted a machine learning approach; the others relied on hand-coded rules + dictionaries, gazetteers, and ontologies;
- Two of them (REMBRANDT and REMMA) made use of the Portuguese Wikipedia, in different ways



Participation and results ReRelEM

System	NE task	Relations
Rembrandt	all	all
SeRelEP	only identification	all but outra
SeiGeo	only LOCAL detection	inclusion

- Answer complex questions based on Wikipedia (PhD work in progress)
- Develop a hot news portal based on NEs
- Evaluate a system for ontology creation (PhD work)



Second HAREM Resources

Second HAREM Collection and its metadata
+
Second HAREM Golden Collection (GC) including ReReLEM
+
Extended TEMPO Golden Collection
+
ReRelEM triples
+
Evaluation programs
+
System runs
+
Documentation
=

LÂMPADA – Second HAREM Resource Package
http://www.linguateca.pt/HAREM/PacoteRecursosSegundoHAREM.zip

SAHARA and AC/DC: further access to HAREM and ReRelEM resources

- Sahara web service (Gonçalo Oliveira & Cardoso, 2009), http://www.linguateca.pt/SAHARA/
  - Submit new runs and...
    - select different options for scoring against the GC(s);
    - use several scenarios;
    - check the relative performance against the official runs.
- AC/DC, interaction with the parsed GC (Rocha & Santos, 2007) http://www.linguateca.pt/ACDC/

Discussion

- Undeniable relevance for Portuguese processing community, but of possible interest to a wider audience
- Multilingual comparison
  - →Are there relevant differences regarding categories?
  - →Do cohesive devices differ between languages?
  - →Differences between explicit / implicit relations
- Relationship with QA
  - →Questions for QA@CLEF as one text genre
- Relationship with GIR
  - →Use of GeoCLEF pool documents in the Second HAREM collection, that allow detailed assess of the importance of NER for this application

Comments and reuse welcome!

- Studies of NER and RD difficulty for Portuguese, by text genre
- Studies of other subjects that may involve NE
- Training material
- Further linguistic analysis
- Conversion to other formats/theories