Second HAREM: new challenges and old wisdom

http://www.linguateca.pt/HAREM

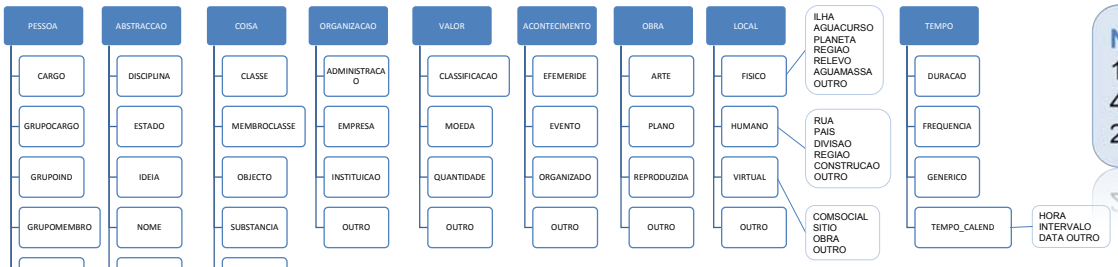
Diana Santos Diana.Santos@sintef.no **Cláudia Freitas** maclaudia.freitas@gmail.com
Hugo Gonçalo Oliveirahroliv@dei.uc.pt **Paula Carvalho** pqfcarvalho@gmail.com **

First HAREM (Santos & Cardoso, 2007)

Second HAREM

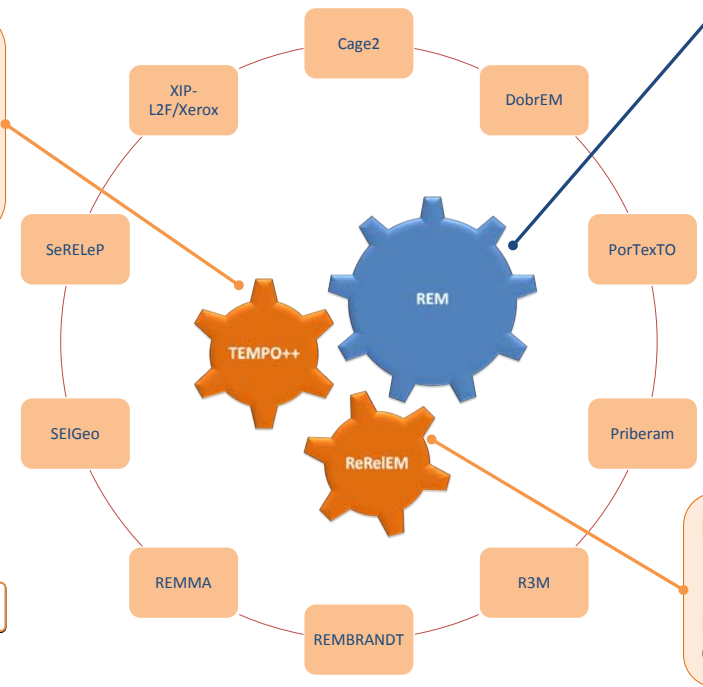
Participants
22 prospective participants
16 registered systems
10 actual systems

Call for participation — September 2007
Proposal of 3 tasks — November 2007
Release of training material — January 2008
Submission period — April 2008
Workshop — September 2008



Named entity recognition
10 categories (CATEG)
42 types (TIPO)
20 sub-types (SUBTIPO)

Time normalization
Reference
Direction
Time gap
Absolute dates
(Hagège et al., 2008)



HAREM Collection
- Golden collection
- CHAVE documents
- GeoCLEF documents
- Training material

Golden Collection
- Wikis
- Blogs
- Wikipedia entries
- Questions
- Standard web pages
- Newspaper
- Oral transcriptions
- Literary Text

Human annotation
2 annotators
Etiquet(H)AREM
Automatic comparison
Comp(H)AREM
Discussion and revision of differences
Full sequential revision
Revision per category

Relation detection
Identity
Includes / Is-included
Located-in / Place-of
Other

Improvements

- ✓No separation between identification and classification
- ✓Finer geographic distinctions
- ✓Finer grained classification of time expressions (Hagège et al., 2008)
- ✓Not knowing <EM> ≠ Knowing it is not <EM CATEG="OUTRO">
- ✓Text collections in XML
- ✓Open source scoring programs
- ✓Both GC and submissions may have classification (|) and delimitation vagueness (ALT)
- ✓Partial identification doesn't score unless covered by ALT
- ✓New evaluation measure

"Old", persistent features
- ✓ Semantic model
- ✓ Selective scenarios

$$\sum_{n=1}^{N}(1-1/num_{cat})*cat_{certa}*\alpha+(1-1/num_{tipos})*tipo_{certo}*\beta+(1-1/num_{sub})*sub_{certo}*\gamma-$$
$$\sum_{n=0}^{M}(1/num_{cat})*cat_{certa}*\delta+(1/num_{tipos})*tipo_{certo}*\varepsilon+(1/num_{sub})*sub_{certo}*\phi$$

Hagège, Caroline, Jorge Baptista & Nuno Mamede. "Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o HAREM II". 2008.

Santos, Diana & Nuno Cardoso (eds.). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. 2007.

** Other HAREM organization members: **David Cruz** dcruz@lasige.di.fc.ul.pt **Luís M. Cabral** luis.m.cabral@sintef.no **Cristina Mota** cmota@ist.utl.pt