Second HAREM:
Advancing the State of the Art of Named Entity Recognition in Portuguese

Cláudia Freitas*, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira*, Paula Carvalho****
Linguateca, FCCN, **at SINTEF ICT, ***at Univ. of Lisbon - Faculty of Sciences, Lasige, *at Univ. of Coimbra - CISUC/DEI
E-mail: claudiafreitas@puc-rio.br, cmota@ist.utl.pt, Diana.Santos@sintef.no, hroliv@dei.uc.pt, pcc@di.fc.ul.pt

Abstract

In this paper, we present Second HAREM, the second edition of an evaluation campaign for Portuguese, addressing named entity recognition (NER). This second edition also included two new tracks: the recognition and normalization of temporal entities (proposed by a group of participants, and hence not covered on this paper) and ReRelEM, the detection of semantic relations between named entities. We summarize the setup of Second HAREM by showing the preserved distinctive features and discussing the changes compared to the first edition. Furthermore, we present the main results achieved and describe the available resources and tools developed under this evaluation, namely, (i) the golden collections, i.e. a set of documents whose named entities and semantic relations between those entities were manually annotated, (ii) the Second HAREM collection (which contains the unannotated version of the golden collection), as well as the participating systems results on it, (iii) the scoring tools, and (iv) SAHARA, a Web application that allows interactive evaluation. We end the paper by offering some remarks about what was learned.

1. Introduction[1]

This paper presents Second HAREM, the second joint evaluation campaign of named entity recognition (NER) in Portuguese, which has been presented in detail, including the description of the participant systems, in a devoted book in Portuguese (Mota and Santos, 2008).

We summarize and discuss the main results achieved in this evaluation, after presenting the available resources created in its scope.

HAREM is organized by Linguateca[2], a project devoted to the fostering of the computational processing of Portuguese. First HAREM, its first edition, was initiated in September 2004. It comprised two evaluation events, and officially ended at the First HAREM Workshop in Porto, 15 July 2006 (Santos and Cardoso, 2007).

Second HAREM took place between September 2007 and September 2008, and the evaluation contest itself occurred in a temporal window from 14 to 28 April 2008. Participants had at most 48 hours to submit a maximum of four runs. A total of 27 official runs were received from 10 participating systems.

As usual in evaluation contests, participants were consulted and a consensus was reached concerning several issues:

(i) HAREM would not support embedded (or nested) NER; (ii) the text type or genre of the documents used in the HAREM collection would not be made available beforehand; (iii) the (time) performance of the different systems should be provided by participants; (iv) the organization should decide which categories, types and subtypes would be taken into account.

In this second edition, two new tracks were included: ReRelEM, which evaluated the detection of relations between named entities, including, but not limited to, co-reference resolution (Freitas et al., 2008, 2009); and the recognition and normalization of temporal entities (Hagège et al., 2008). Given that the latter was proposed and defined by a group of participants, it will not be further described here.

This paper is organized as follows: section 2 describes the main features of HAREM; section 3 provides information on the golden collections, as well as on the tools deployed; section 4 presents the evaluation measures employed in the Second HAREM, and section 5 briefly discusses the participants' performance. Finally, section 6 offers some concluding remarks.

2. Main features of HAREM

2.1 Features preserved from First HAREM

Second HAREM preserved what we considered to be the three most distinctive features of the first evaluation contest, namely:

(i) the semantic model: we asked systems to provide the semantic classification based on the use of the NE in context, going beyond its dictionary meaning;

(ii) vagueness: we addressed the fact that NE may have more than one category or type, based on the evidence that vagueness is an essential property in natural language, and it should be preserved;

(iii) the flexibility of the evaluation setup: in particular, offering selective scenarios and different evaluation modes (Santos et al., 2006).

We proceed to better motivate these three points in turn.

2.1.1 The semantic model

As expounded in Santos (2007b), let us take the following case:

(1) *A morte é reportada no Diário de Notícias do dia* ('The death is annouced in Diário de Notícias')

(2) *A diferença entre o ´Jornal de Notícias´ e o ´Diário de Notícias'* ('The difference between Jornal de Notícias and Diário de Notícias')

(3) *O seu pai era funcionário público do Ministério da Justiça e crítico musical do ´Diário de Notícias´* ('His

[1] The present list of authors is in alphabetical order, all have contributed equally to HAREM and this paper.
[2] http://www.linguateca.pt/

father was an employee of the Ministry of the Justice and a music reviewer for Diário de Notícias')

(4) … *foi fotografado pelo Diário de Notícias (DN) a fumar uma cigarrilha*... ('had a picture taken by Diário de Notícias smoking a cigarette')

As shown by examples 1-4, respectively, reference to a name such as *Diário de Notícias* or *Jornal de Notícias* can be understood as a place (LOCAL VIRTUAL COMSOC), as an object (COISA CLASSE), as a (private) organization (ORGANIZACAO EMPRESA) or as a person or group of people standing for their role as interviewers or recipients of information (PESSOA GRUPOMEMBRO).

So, instead of classifying the instances of that named entity as newspaper, or mass media (its dictionary meaning), HAREM required their meaning in context.

This shows that the HAREM task is considerably more difficult, and fine-grained, than the classical NER task, as performed for example in MUC (Grishman and Sundheim, 1996). For a detailed comparison with MUC and the NE CoNLL shared task, see Santos (2007a).

Another argument to go beyond pre-established dictionary meanings is the strong contextual dependence of natural language expressions. Indeed, while there are cases where it is not difficult to agree about the semantic value of an entity out of context (like the news agency just discussed), in many cases the situation is not clear-cut, as shown by examples 5-6 below. What is the "real" meaning of *Big Bang* out of context: a theory (abstraction) or an explosion (event)?

(5) *É duvidoso que o modelo do Big Bang tivesse sido recebido com tanto interesse*... ('It is hard to believe that the Big Bang model would have been received…')

(6) *O que causou a explosão do Big Bang?* ('What caused the Big Bang explosion?) [3]

2.1.2 Vagueness

In HAREM, NE can receive more than one tag, whenever the context where it occurs does not allow deciding for only one of them. We thus opt for preserving the vagueness present in the natural language formulation, since we believe that its arbitrary resolution or simplification implies a real loss of information. For example, in example 7:

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

the entity *Administração Bush* can be interpreted as both a group of people (PESSOA GRUPOMEMBRO) and an organization (ORGANIZACAO ADMINISTRACAO). In fact, this is even warranted by cases where anaphoric relations later select different parts/facets of a vague entity, as example 8 shows:

(8) *Com a proclamação da Carta, temos a obrigação e a oportunidade de dar aos quase 500 milhões de cidadãos a ideia de uma Europa [LOCAL|PESSOA] unida. (...) Dentro e fora da Europa [LOCAL], "temos o dever de sempre defender a dignidade e os direitos humanos", concluiu.* ('With this Declaration, we have the obligation and opportunity to give to almost 500 million European citizens the idea of a united Europe. (…) Inside and outside Europe "we must defend dignity and human rights", he concluded').

In (8), the first mention of *Europa* (Europe) means both the place (LOCAL) and the European citizens (PESSOA). The second mention of *Europa*, however, refers only to its geographical (LOCAL) facet.

2.1.3 Flexibility of the evaluation setup

In HAREM, participants could opt to compete in selective scenarios. In other words, they could select the set of categories, types and subtypes in which to be evaluated. This way, HAREM was able to encompass many different systems with different goals and different applications in mind, and in addition to compare those systems for the general HAREM task, we were also able to compare every system relative to its preferred view.

Finally, we emphasize that the HAREM categories (to which we refer loosely as the "HAREM ontology") were defined via a corpus-based approach, that is, instead of starting from a set of predefined categories, these were chosen after human analysis of text (Santos, 2007b).

Due to the high participation and the little request for changes, most categories and types from First HAREM remained unchanged, as shown in Figure 1.

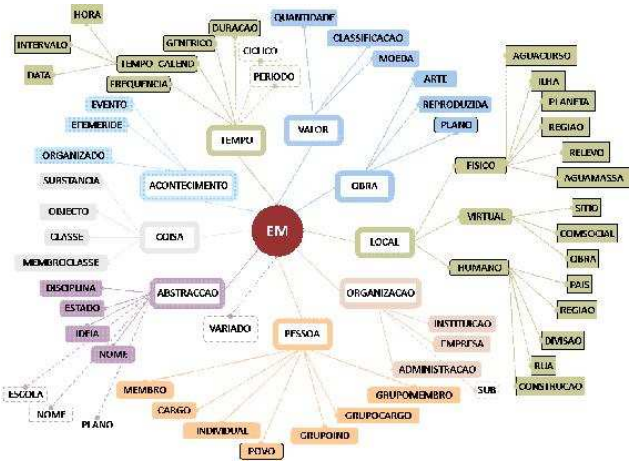

Figure 1: Categories, types and subtypes in Second HAREM

2.2 Features introduced in Second HAREM

More than merely repeat the previous format, we tried to advance the state of the art and foster systems' advances with Second HAREM. We have thus improved some features and proposed new challenges, to which we turn now.

One important improvement in Second HAREM concerned the systematic annotation of embedded NE that take part of larger entities, through the ALT mechanism. In the example below

[3] Of course the argument for this semantic model can also apply to any natural language, but we stick to Portuguese because it was for this language that it was originally conceived and discussed.

(9) *Quantos atletas participaram nos Jogos Olímpicos de Barcelona?* (*How many athletes participated in Barcelona Olympic Games?*)
we consider that two alternative analyses are motivated: (i) the whole entity *Barcelona Olympic Games*, an event, and (ii) the embedded entities Barcelona (LOCAL - place) and *Olympic Games* (ACONTECIMENTO - event). So, instead of deciding arbitrarily for the widest possible NE, we classified both as possible correct analyses in the golden collection, and required – or better, encouraged – systems to do the same (providing ALT in their output). Example 10 shows the exact output desired:

(10) <ALT><Jogos Olímpicos de Barcelona | <Jogos Olímpicos> de <Barcelona></ALT>

Since this was a new feature, two different evaluation modes (strict and relaxed) were offered to deal with this.

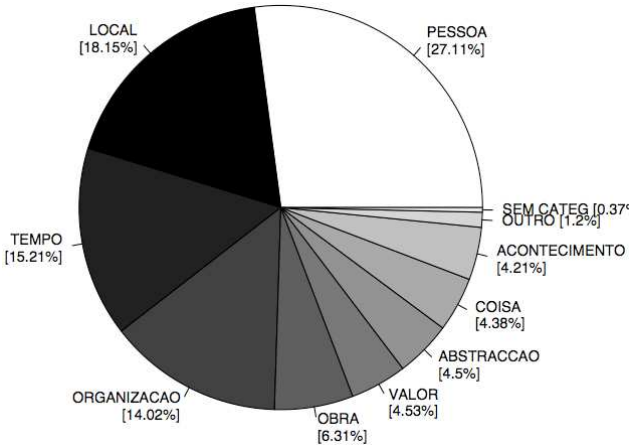

Figure 2: Category distribution in Second HAREM's golden collection

As already mentioned, we also provided two pilot tasks under the scope of Second HAREM, creating separate golden collections for each. The temporal task will not be discussed here, since we were not the proposal's authors, and it overlapped with the main track's golden collection – by adding a set of new attributes to the temporal NEs (corresponding to the TEMPO category).

ReRelEM, however, will be presented here also since it was crucially related to all categories (but TEMPO). ReRelEM was concerned with the automatic detection of relations between named entities in a document.

Since we were not aware of any empirical study (for Portuguese or any other language) that actually described which were the most relevant or frequent relations, we made an exploratory study in order to find the most frequent and less controversial relations in texts. We identified four basic relation types: `identidade` (identity), `incluido/inclui` (inclusion), `ocorre-em/sede-de` (location), and `outra` (other) (which was later on explicitly detailed into twenty two different relations).

As explained in Freitas et al. (2009), we found out that human annotation of the `outra` ('other') relation was more reliable and intelligible for human beings if it was specified which specific relation. We have also had to use the several different categories of vague NE to clearly specify the relation, as discussed in connection with examples (7) and (8).

Finally, the annotation of relations between entities also led to the development of a set of specifically dedicated tools whose applicability may transcend ReRelEM or HAREM.

3. Second HAREM resources

As usual in the evaluation contests and other activities created in the scope of Linguateca, everything is free for the community (not only for the participants), and we take special care in making our resources public and reusable. So, we have created two kinds of resources: annotated material, and tools, some of which also provided as services on the Web, which we will describe here.

3.1 The golden collection for the main track

The golden collection (GC) of Second HAREM heavily included new text genres such as blogs, wikis, and encyclopedia (Wikipedia) text, as well as questions used for QA evaluation, in addition to the more traditional kinds of newspaper text and usual Web pages. Comparing with First HAREM, oral transcriptions and literary text were far more scarcely used. Figure 2 provides a quantitative distribution of the 7,847 NEs contained in the GC by NE category.

Each document of the GC (and of the larger HAREM collection of which the GC is a subset, see below) is unambiguously identified by its document identification value, which is followed by the following set of features: (i) language variety (Brazil or Portugal); (ii) text genre (see Figure 3); and (iii) source. The GC also contains comments signaled by the COMENT attribute, provided for further study, including cases of disagreement among annotators, and mistakes detected during annotation.

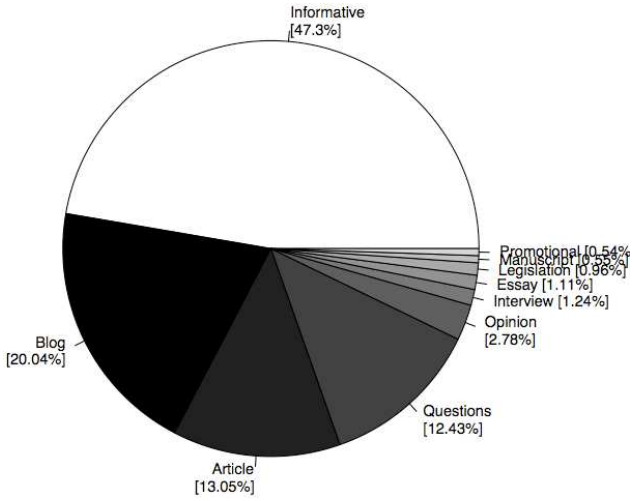

Figure 3: Genre of Second HAREM's golden collection

For example, the comment "2/3" indicates that the NE classification (category, type or subtype) was not assigned consensually by the annotation team, but was the result of

a majority vote. For the record, there were 223 cases of disagreement after prolonged discussion, of which 101 could not even be annotated by majority voting (and hence were marked to be ignored during scoring). The annotation process and conflict resolution has already been detailed in (Santos et al., 2008, Carvalho et al., 2008), so we redirect the reader to these works.

3.2 The Second HAREM collection

The Second HAREM collection includes 1,040 documents, and was composed by adding to the GC documents all training material provided beforehand (in order to investigate later whether significant performance differences would be detectable). Then, all remaining text came from the CHAVE collection, which contains Portuguese and Brazilian newspapers from 1994 and 1995 (Santos and Rocha, 2005). However, the choice of the CHAVE documents was not random, rather, the texts were chosen from GeoCLEF's pool (Mandl et al., 2008), in the following way: for each of the 25 topics corresponding to the 2007 edition, all relevant documents were included, as well as ten irrelevant ones -- our goal was to create in this way a unique resource to study the influence of NER for geographical information retrieval.

3.3 The golden collection for ReRelEM

For the actual contest, and given the lack of time to create a larger resource, ReRelEM's GC was a subset of HAREM's GC, including 12 documents with 4,417 words and 573 NEs. It describes 6,790 relations, which were manually annotated (1436 identity; 1612 inclusion; 1232 placement; 2510 other). Further details can be found in (Freitas et al., 2008, 2009).

We have later on extended the manual annotation of semantic relations to the remaining documents of the HAREM's GC and made it available to the public from http://www.linguateca.pt/HAREM/coleccoes/CDSegundoHAREMReRelEM.xml. This exercise allowed us not only to validate the previous relations, but also to offer a robust resource to the NLP community that deals with Portuguese processing.

As expected, the annotation of new texts provided a refinement of the original relations, and we achieved a final set of 24 relation types, shown in Table 1.

Of the 7,847 NEs annotated in the GC, 3,776 are related to some other NE, and are responsible for 4,803 relations manually annotated. Their distribution in terms of categories is shown in table 2.

In ReRelEM's GC each NE has a unique ID, so that a relation is indicated by additional attributes: COREL (containing the ID of the related entity) and TIPOREL (displaying the name of the relation), both added to the NE that corresponds to one of the arguments of the relation. A NE can be associated with one or more NEs through several semantic relations. When the relation holds between vague NEs, the annotation is somewhat different, since we make explicit which facet of the vague NE takes part in the relation.

Relations are also made available in a RDF-like triple format automatically computed by the tools we describe below.

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

Table 1. ReRelEM relation types in HAREM's GC. In bold are the ones that the systems had to explicitly name. The others were under OUTRA.

Relations per category	#
ABSTRACCAO	255
ACONTECIMENTO	168
COISA	175
LOCAL	960
OBRA	274
ORGANIZACAO	783
OUTRO	25
PESSOA	1286
TEMPO	192
VALOR	19

Table 2. ReRelEM relations, before expansion, by simple categories in HAREM's GC

3.4 Tools for Second HAREM

Although conceptually the differences between the First and the Second HAREM are insignificant, the addition of the ReRelEM and temporal tracks together with the new ALT format, and a refined evaluation measure, resulted in significant new programming, which is documented in detail in (Gonçalo Oliveira et al., 2008).
Also, in connection with a more distributed annotation procedure, some tools to help linguists to annotate and compare annotations were also developed, see for example Etiquet(H)AREM (Carvalho and Gonçalo Oliveira, 2008). Finally, the relation visualization and processing also required specific programming.
All of this is available in the LÂMPADA package, http://www.linguateca.pt/HAREM/PacoteRecursosSegundoHAREM.zip, together with the participant system's outputs.

3.5 The SAHARA service

We were also aware – from our experience of organizing previous evaluation contests – that many of the participants would not use the tools because their installation might bring problems or because they had not the time to even try it out.
So, this time, we also provided a service that allows researchers to use the whole setup and just concentrate on the development of their systems, SAHARA (Gonçalo Oliveira and Cardoso, 2009), available from http://www.linguateca.pt/SAHARA/.
The user can input new runs and select a lot of different options for scoring against the golden collection(s), in several scenarios, even choosing his individual sets of categories or types, and check his relative performance against the official runs.

4. Evaluation measures

The changes mentioned in the previous sections prompted a set of improvements and updates to the evaluation machinery as well.

4.1 Measure for the main track

In fact, one of the most relevant contributions of the First HAREM was to define a set of measures and metrics for NER (Santos et al., 2007), together with making available a set of open source programs that computed them (Seco et al., 2006).
Those measures, however, were based on a fixed depth of categories and types: each category had a number of types, while in Second HAREM we provided a four level hierarchy, with everything optional.
We have therefore enlarged and made the evaluation measure more robust, in order to account, in the same fell swoop, for everything covered by the previous measures (except for types-only, which we now consider irrelevant).
The new (single) measure for the Second HAREM is thus an extension of the combined measure of First HAREM, accounting for the existence of subtypes and for the optionality of all values, as well as dealing more adequately with vague NEs (i.e., NEs with N categories):

$$\text{HAREM score} = 1 + \text{sum}_N((1-W_{cat}) * cat_{certa} * \alpha + (1-W_{tipos}) * tipo_{certa} * \beta + (1-W_{sub}) * sub_{certa} * \gamma) - \text{sum}_M(W_{cat} * cat_{esp} * \alpha + W_{tipos} * tipo_{esp} * \beta + W_{sub} * sub_{esp} * \gamma)$$

M is the number of spurious classifications in the participant's run and N is the number of classification in the GC, both according to the selective scenario. The final score for each system is obtained by summing over all NEs (the suffix *certa* is 1 when it is right, 0 when wrong, the suffix *esp* takes 1 when spurious, 0 when not), and comparing with the maximum possible score given the system's output (precision) or the golden collection material (recall). The weights (W_{cat},W_{tipos},W_{sub}) are simply the inverse of the number of different categories, types, etc. More weight is given to a choice among a higher number of alternatives, and different weights have been experimented with to produce better discrimination among systems. By setting all weights to 0, the formula measures simple identification.
Also, by providing a consistent catchall category/type/subtype OUTRO in the HAREM grid, we were able to express the difference between ignorance (no value provided) and explicit disagreement (using OUTRO) and evaluate them differently.

4.2 Measures for ReRelEM

In ReRelEM, our first concern was to make a clear separation between the evaluation of relations and the evaluation of NE detection. Therefore, relations established between incorrect or misclassified NEs were not considered and the first step carried out by the evaluation chain was thus removing them both from the GC and the runs. Furthermore, in order to make the annotation task easier to the systems and, especially, to the GC annotators, each document was not required to have all possible relations explicitly annotated, but only a set from where all the implicit relations could be inferred. This was achieved by applying a set of symmetry and transitivity rules to the original set of relations, both in the GC and in all runs. After this step, all implicit relations were made explicit, right before computing the system's score.
These rules, as well as the evaluation process of ReRelEM, are detailed in Freitas et al. (2009).
Relations annotated by the system were then compared with the ones in the GC, and each triple <NE relation NE> was scored as correct, missing or incorrect. Only those triples which linked the correct NE and whose relation was well classified were considered correct.
Then, one point is assigned to each correct relation and none to incorrect or missing relations, which allowed us to compute precision, recall and F-measure.

5. Participation and results

There were ten participants in the main track of Second HAREM, of which three also participated in the ReRelEM pilot task, producing 27 runs altogether (as

previously mentioned, each participant could submit at most four runs).

A curious fact, but nonetheless a natural consequence of allowing selective scenarios, is that only two systems (Priberam and REMBRANDT[4]) recognized the complete set of categories, types and subtypes; all other systems opted for different subsets of the classification tree. See Table 1.3 in Carvalho et al. (2008). A similar variation happened in RelRelEM regarding the types of relations recognized.

Of the ten systems, only one (R3M) adopted a machine learning approach (specifically, co-training); the others relied on hand-coded rules in combination with dictionaries, gazetteers, and ontologies. Two of them (REMBRANDT and REMMA) made use of the Portuguese Wikipedia, in different ways. This shows that the community dedicated to NER in Portuguese hasn't embraced machine learning techniques, contrary to the situation for English. This was also observed in the First HAREM, where out of nine systems only two (NERUA and MALINCHE[5]), that were originally developed for Spanish, were trained based on previously annotated corpora.

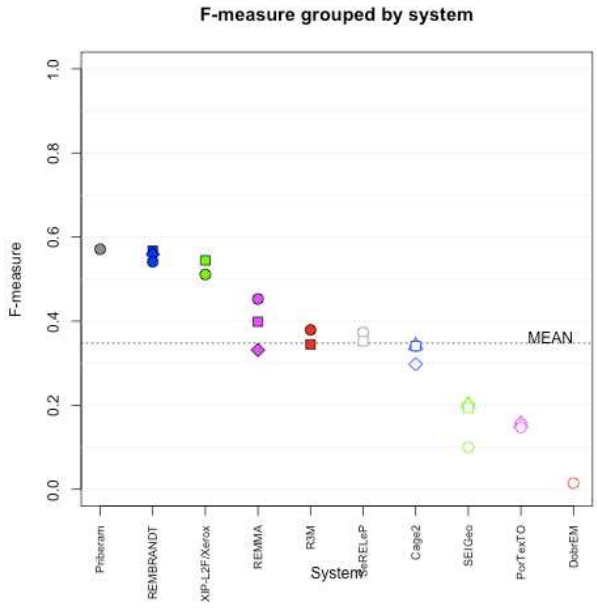

Figure 4: F-measure of the participating systems in the main HAREM track

Figure 4 displays the systems' results according to the F-measure, the harmonic mean of precision and recall, for the NER task. As can been seen, the best performing system is a commercial product (from Priberam), which in any case has a very close performance to REMBRANDT's best run: the former uses a multilingual ontology combined with lexical-semantic contextual rules, whereas the second exploits Wikipedia as knowledge source, combined with grammatical rules that describe internal and external evidence about the named entities. The comparison of the remaining systems is not as straightforward because all participated in different selective scenarios. In fact, the evaluation by selective scenarios only provides a completely fair evaluation in the case where the evaluation scenario is contained in the participation scenarios; otherwise, systems that correspond exactly to the evaluation scenario may have a slight advantage.

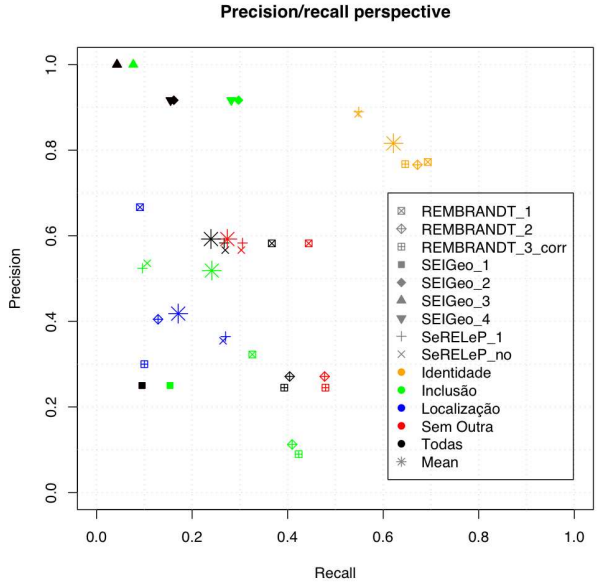

Figure 5: F-measure of the participating systems in ReRelEM

In Figure 5 we can see the precision plotted as a function of the recall for the three systems that participated in the ReRelEM task (*Todas*), as well as the precision and recall for the different relations (*identidade*, *inclusão* and *locate-in*) and all relations without Outra (*Sem Outra*). Again, we stress that those systems chose to recognize different types of relations, so it is hard to conclude about their relative merits.

6. Concluding remarks

In this paper we presented the main features of Second HAREM. Although we could not produce an uncontroversial and conclusive state of the art for Portuguese NER – in fact, in the two editions of HAREM there was very little overlap among participants, and two of the common participants had even rewritten their systems from scratch – we were at least able to provide an hopefully interesting and important resource for empirical studies and for training of future systems.

While we believe the importance of this for the Portuguese-language processing community is beyond doubt, we hope that, by sharing these data with the international community as well, we may both influence

[4] For each system see the corresponding chapter in Mota and Santos (2008).

[5] Again for each system see the corresponding chapter in Santos and Cardoso (2007).

other languages' processing and receive feedback from similar or related initiatives for other languages.

One interesting subject that such multilingual comparison may rise is the possibility to discover relevant differences in attention (and therefore frequency of mention) of different categories. For example, Germanic languages may give more precise descriptions of places and mention more place NEs while Romance languages may have more abstractions named.

Also and is well known from e.g. translation theory, different languages differ in cohesive devices, so the relations they tend to make explicit or leave implicit will plausibly differ. It is our contention that only by comparing different resources created from scratch for different languages such tendencies can reliably be uncovered.

7. Acknowledgements

Linguateca has throughout the years been jointly funded by the Portuguese Government, the European Union (FEDER and FSE), under contract ref. POSC/339/1.3/C/NAC, MCTES, UMIC and FCCN.

We thank Nuno Cardoso for his development of SAHARA, and all people who participated or helped organize HAREM through the two editions.

Cristina Mota is grateful for the appreciation and encouragement of her work provided by the NYU Proteus group.

8. References

Paula Carvalho and Hugo Gonçalo Oliveira. "Manual de Utilização do Etiquet(H)AREM". 29 April 2008, . http://www.linguateca.pt/aval_conjunta/HAREM/ManualUtilEtiquetHAREM.pdf

Paula Carvalho, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas and Cristina Mota. 2008. Segundo HAREM: Modelo geral, novidades e avaliação. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008, pages 11-31.

Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, and Cristina Mota. 2008. Relações Semânticas do ReRelEM: além das entidades no Segundo HAREM. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008, pages 77-96.

Cláudia Freitas, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira and Paula Carvalho. 2009. Detection of relations between named entities: report of a shared task. In *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, SEW-2009* (Boulder, Colorado, USA, June 4, 2009), pp. 129-137.

Hugo Gonçalo Oliveira, Cristina Mota,, Cláudia Freitas, Diana Santos and Paula Carvalho. 2008. Avaliação à Medida no Segundo HAREM. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008, pages 97-129.

Hugo Gonçalo Oliveira and Nuno Cardoso. 2009. SAHARA: an online service for HAREM Named Entity Recognition Evaluation. In *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)* (São Carlos, Brazil, September 8-11, 2009).

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th Conference on Computational linguistics*, pp. 466–471, Morristown, NJ, USA. Association for Computational Linguistics.

Caroline Hagège, Jorge Baptista and Nuno J. Mamede. 2008. Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e o futuro. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008, pages 33-54.

Thomas Mandl, Fredric Gey, Giorgio di Nunzio, Nicola Ferro, Mark Sanderson, Diana Santos and Christa Womser-Hacker. 2008. An evaluation resource for Geographical Information Retrieval. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008). (Marrakech, 28-30 May 2008), European Language Resources Association (ELRA), s/pp.

Cristina Mota and Diana Santos, editors. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Diana Santos, Cláudia Freitas, Hugo Gonçalo Oliveira and Paula Carvalho. 2008. Second HAREM: new challenges and old wisdom. In Antonio Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira and Paulo Quaresma, editors, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)* Vol. 5190, (Aveiro, Portugal, 8-10 September 2008), Springer Verlag, pp. 212-215.

Diana Santos. 2007a. Evaluation in natural language processing. Course at *European Summer School on Language, Logic and Information (ESSLLI 2007)*, Dublin, August 2007.

Diana Santos. 2007b. O modelo semântico usado no Primeiro HAREM. In Diana Santos and Nuno Cardoso, editors, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pages 43–57.

Diana Santos and Nuno Cardoso. 2006. A Golden Resource for Named Entity Recognition in Portuguese. In Renata Vieira, Paulo Quaresma, Maria da Graça Volpe Nunes, Nuno J. Mamede, Cláudia Oliveira and Maria Carmelita Dias, editors, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006*. Itatiaia, Brazil, May 2006, pp. 69–79, Berlin/Heidelberg. Springer.

Diana Santos and Nuno Cardoso, editors. 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca.

Diana Santos and Nuno Cardoso. 2007. Balanço do primeiro HAREM e perspectivas de trabalho futuro. In Diana Santos and Nuno Cardoso, editors, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pages 87–94.

Diana Santos, Nuno Cardoso and Nuno Seco. 2007. Avaliação no HAREM: Métodos e medidas. In Diana Santos and Nuno Cardoso, editors, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pages 245–282.

Diana Santos, Nuno Seco, Nuno Cardoso and Rui Vilela. 2006. HAREM: An Advanced NER Evaluation Contest for Portuguese. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik and Daniel Tapias editors, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)* (Genova, Italy, 22-28 May 2006), pp. 1986-1991.

Diana Santos and Paulo Rocha. 2005. The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum*, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers. Berlin/Heidelberg: Springer, Lecture Notes in Computer Science, 2005, pp. 821-832.

Nuno Seco, Diana Santos, Rui Vilela and Nuno Cardoso. 2006. A Complex Evaluation Architecture for HAREM. In Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira and Maria Carmelita Dias, editors, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006)* LNAI 3960, 13-17 de Maio de 2006, Berlin/Heidelberg : Springer Verlag, pp. 260-263.