Second HAREM: New Challenges and Old Wisdom

Diana Santos[1], Cláudia Freitas[2], Hugo Gonçalo Oliveira[2], and Paula Carvalho[3]

[1]SINTEF ICT, Norway
[2]CISUC, DEI-FCTUC, Portugal
[3]DI-FCUL, Portugal
Diana.Santos@sintef.no, maclaudia.freitas@gmail.com, hroliv@dei.uc.pt, pqfcarvalho@gmail.com

Abstract. Discussion of the Second HAREM: changes to the guidelines, introduction of new tracks, improvement of evaluation measures and description of the new evaluation resources.

1 Introduction

In this paper we present the second evaluation contest of named entity recognition in Portuguese, the Second HAREM[1] which started September 2007 and whose submission period took place 14-28 April 2008. We are mainly concerned with presenting the options while designing this new event, describing the process followed and the differences compared to the First HAREM.

After a successful first event, the First HAREM, in whose workshop most participants stated their definite interest in a new edition, and which culminated with the production of a book [1], call for participation in a second edition was issued in September 2007. A record number of 22 different prospective participants or interested parties enrolled. In early 2008, 16 systems registered, although only 10 would eventually participate.

Discussion took place until November 2007, and two new tracks or tasks were proposed: (i) temporal recognition and normalization, according to extensive guidelines proposed by a group of participants, Hagège et al. [2]; and (ii) relation detection between named entities (dubbed ReRelEM), including, but not limited to, coreference identification.

As for training material, a fully annotated example collection – six fully annotated texts, ca. 1,500 words – were made available to the participants at the HAREM site in January 2008, while the golden collections from First HAREM were converted to the new format: ca. 140,000 words, corresponding to 9,000 NEs from 257 different documents. Later, a fully annotated subsection of previous material – ca. 3,500 words and 279 NEs – was also created by the authors of the time guidelines.

[1] http://www.linguateca.pt/HAREM/

A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, pp. 212–215, 2008.

2 "Old", Persistent, Features of HAREM

The most important features of the First HAREM remained: (i) its semantic model (that describes the use of a NE in context, and not its dictionary meaning) [1, chapter 4] and (ii) the flexibility of its evaluation setup (in particular the existence of selective scenarios) [3].

To introduce the semantic model to newcomers, let us take the example of continents: These are, basically, lexically covered by the five words *Oceânia*, *Ásia*, *América*, *Europa* and *África*. In addition to requiring that systems decide whether these continents are actually being mentioned and thus the words do not refer to a ship (COISA), a deity (PESSOA) or a book (OBRA), as is usual in NER contests, HAREM asks systems to decide also whether the particular mention in context refers to an entity of the physical domain (LOCAL FISICO), an entity belonging to human geography (LOCAL HUMANO), the people inhabiting that continent (PESSOA POVO), supra-national political organizations (ORGANIZACAO ADMINISTRACAO), or even just the abstract concept (ABSTRACCAO), whatever that may be. This shows that HAREM tasks are considerably more difficult, and fine-grained, than the classical NER task as represented for example by MUC [4].

As to the selective scenario property, it allows HAREM to encompass different systems with different goals and different applications in mind, enabling comparsion of every system also relative to its preferred view (its selective scenario).

3 Improvements

The problems of artificially separating identification from classification, noted and discussed in [1, chapter 7], were solved by assuming a more consistent (although more complex) identification strategy, as well as removing from the NE task the identification of objects which only accidentally include names (as is the case of *pastéis de Belém* and *bolas de Berlim*).

A number of finer distinctions in the geographic domain were also added, mirroring, in a way, the finer grained classification of time expressions.

This will allow for an empirical investigation of the possibility or need of a more detailed subcategorization in NER.

We also made explicit the difference between not knowing (a negative statement) and knowing that **none** of the explicit choices was right (a positive one). Now we have OUTRO for the latter case, while not knowing is conveyed by no value at all, and the two cases are differently scored.

There were also several changes on the technical side. The format of collections and submissions was changed to XML, to achieve easier processing and validation of the material. A validator was deployed and made public for systems to test their output previous to submission.

Instead of allowing submissions doing either identification only or classification plus identification, both tasks were merged in one syntax only; and a more consistent use of OUTRO label across CATEG, TIPO and SUBTIPO was implemented.

Substantial changes to the evaluation machinery, while preserving backward compatibility, became necessary. In fact, one the most important contributions

of Fisrt HAREM was to define a set of measures and metrics for NER, at the same time making available a set of open source programs to compute them.

Those measures, however, were based on a fixed depth of categories and types: each category had a number of types, while now we have a four level hierarchy, with everything optional. We have therefore extended and made more robust the evaluation measure, in order to account, in the same fell swoop, for everything covered by the previous measures (except for types-only).

We have also taken measures to give a more adequate treatment to vagueness in the evaluation, accepting and expecting submissions to also produce more than one classification (|) or delimitation (ALT).

In parallel, we removed the partial identification feature of First HAREM (which was blind) by consistently annotating, with the ALT feature, all possible meaningful parts. Only those should be rewarded.

The new measure, an extension of the combined measure of First HAREM, accounts for the existence of subtypes and for the optionality of all values, as well as dealing more adequately with vague NEs (with N categories): $1 + \sum_{n=1}^{N} (1 - 1/num_{cat}) * cat_{certa} * \alpha + (1 - 1/num_{tipos}) * tipo_{certo} * \beta + (1 - 1/num_{sub}) * sub_{certo} * \gamma - \sum_{n=0}^{M} (1/num_{cat}) * cat_{espuria} * \delta + (1/num_{tipos}) * tipo_{espurio} * \epsilon + (1/num_{sub}) * sub_{espurio} * \phi$

4 The New HAREM Collection and Its Annotation

As to the constitution of the golden collection for the Second HAREM, i.e., the subset of the collection that will be used as comparison stock, this time – accompanying the flow of time – we included, and made heavy use of, new genres such as blogs, wikis, encyclopedia (Wikipedia) entries, and questions (such as used in question answering), in addition to the more traditional kinds of newspaper text and standard Web pages. Oral transcriptions and literary text, due to the difficulty of obtaining this kind of text, were far more scarcely used.

To create the full collection that includes the golden collection and is provided for the systems to analyse, we used a different strategy from the previous HAREM, in which we had tried to mirror the genre distribution of the golden collection. This time we just included in the full collection all training materials already available, while all remaining material came from the CHAVE collection [5], newspapers from 1994-1995. The texts themselves were chosen from the recall base of last GeoCLEF: for each topic, all the relevant documents and ten irrelevant documents were taken from the Portuguese pool.

Human annotation of the golden collection was performed in several stages: (i) initial independent annotation of each text by two annotators, using a specially developed tool, Etiquet(H)AREM[2] [6] (ii) automatic comparison using another tool, comp(H)AREM; (iii) discussion and revision of these differences; (iv) full sequential revision; (v) revision per category.

[2] Available from http://linguateca.dei.uc.pt/index.php?sep=recursos

When disagreement showed that several different interpretations of the same text are possible, use of vagueness was encouraged. A detailed set of guidelines, as well as specific difficulties in interpreting Portuguese text, was gathered.

5 New Challenges

One of the hallmarks of Linguateca's activities, fully reflected in HAREM, is not to repeat merely what has been done for other languages. On the contrary, we are keen on innovation and on doing first class research, which means that we try to do both something original **and** conceived with Portuguese in mind. We firmly believe that progress in NLP can be done in any language. So, when considering the possibility of offering co-reference as a new track in HAREM – and resources and time available would prohibit us to do something similar to ARE[3] for Portuguese, we would not do a simple identification of NEs referring to the same entity as in MUC or ACE [7]. Rather, we proposed the task of identifying the most frequent (and less controversial) relations among NEs.

Acknowledgments. This work was done in the scope of the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC. We are grateful to David Cruz and Luís Miguel Cabral for participating in the organization of HAREM and to Luís Costa for relevant comments.

References

1. Santos, D., Cardoso, N. (eds.): Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca (2007)
2. Hagège, C., Baptista, J., Mamede, N.: Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o HAREM II (2008)
3. Santos, D., Seco, N., Cardoso, N., Vilela, R.: HAREM: An Advanced NER Evaluation Contest for Portuguese. In: Calzolari, N., et al. (eds.) Proceedings of LREC 2006, 22-28 May 2006, pp. 1986–1991 (2006)
4. Grishman, R., Sundheim, B.: Message understanding conference-6: a brief history. In: Proceedings of the 16th conference on Computational linguistics, pp. 466–471. Association for Computational Linguistics, Morristown (1996)
5. Santos, D., Rocha, P.: The key to the first clef with portuguese: Topics, questions and answers in chave. In: Peters, C., Clough, P., Gonzalo, J., Jones, G. (eds.) Multilingual Information Access for Text, Speech and Images: Resu lts of the Fifth CLEF Evaluation Campaign, pp. 821–832. Springer, Heidelberg (2005)
6. Carvalho, P., Oliveira, H.G.: Manual de utilização do Etiquet(H)AREM (2008)
7. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The Automatic Content Extraction (ACE) Program: Tasks, Data and Evaluation. In: Lino, M.T., et al. (eds.) Proceedings of LREC 2004, Lisbon, Portugal, ELRA, 26-28 May 2004, pp. 837–840 (2004)

[3] See http://clg.wlv.ac.uk/events/ARE/