A complex evaluation architecture for HAREM

Nuno Seco[1], Diana Santos[2], Nuno Cardoso[3], and Rui Vilela[4]

Linguateca nodes of Coimbra[1], Oslo[2], Lisbon[3] and Braga[4]
nseco@dei.uc.pt, diana.santos@sintef.no, ncardoso@xldb.di.fc.ul.pt,
ruivilela@di.uminho.pt

Abstract. In this paper we briefly describe the evaluation architecture and the measures employed in HAREM, the first evaluation contest for named entity recognition in Portuguese. All programs are publically available for experimentation.

1 Introduction

Named Entity Recognition (NER) is nowadays regarded as a fundamental building block in the larger endeavour of understanding natural language by the NLP comunity. The recognition of NER as a separate task started with the MUC conferences [1] (more precisely MUC-6) and has ever after been considered in other contests of the same genre such as (e.g. ACE).

HAREM[2] features several original traits and provided the first state of the art for the field in Portuguese, joining 10 different NER systems for Portuguese. We took into consideration most of the points mentioned in [3] that should be considered in future contests, such as (1) *domain independence* of the systems being tested, (2) *portability*, meaning that systems could be fine-tuned (or retargeted) to specific class of event and (3) encouraging work on deeper semantic understanding. Our goal with HAREM was to take the first step towards building an evaluation framework that could facilitate all these aspects.

Participants had to tag a large and varied collection, with 1202 documents (over 466,000 words) from 8 different genres and several varieties of Portuguese, of which a smaller part (the HAREM Golden Collection) had been manually hand-coded by the organizers, according to detailed guidelines discussed with the participants. We conceptually separated the NER process in two phases (even if most systems do not implement it this way): one first identifies an NE and then attributes some meaning to it in accordance with the surrounding context. In HAREM we took the classification process a step further by including a *morphological* task where NEs where assigned their respective gender and number in context. Another feature worth noting is that semantic classification was again divided in two conceptual steps (categories and types), in order to more precisely pin down the intended meaning of the NE (see [4] for details).

2 Evaluation Facets and Options

Evaluation was carefully studied in order to provide the participants with the most relevant information possible. We hold the view that ceiling effects are as

relevant in evaluation contests as are baselines. Therefore, we were extremely careful in maintaining vagueness during manual annotation of the golden collection either by allowing several categories to mark a single entity or by employing the ALT tag which permitted alternative delimitations of NEs (see [4] for details). Another interesting note is that we also allowed participants to choose the set of categories and types that they wanted to be evaluated in (the selective scenario).

All these options had significant consequences on the complexity of the resulting architecture, as can be seen in Figure 1. In a nutshell, we had to implement

- alignment of the system output with the golden collection: including finding the best alignment among all alternative choices of ALT;
- restriction of comparison to different sets of categories (a kind of ontology mapping);
- several different evaluation measures to reflect all these subtle distinctions.

The outcome is a modular architecture for NER evaluation, providing a valuable resource for further studies in the field.

Evaluation was divided into three tasks each capturing different aspects of the NER problem, namely (i) Identification; (ii) Semantic classification; and (iii) Morphological classification. These and other aspects of evaluation and the correponding metrics are clearly explained in [2]. It will suffice to say here that NEs in terms of **identification** can either be considered *Correct*, *Partially correct by Excess*, *Partially correct by Shortage*, *Spurious* or *Missing*. We note that partially correct NEs correspond to NEs that systems identified and that overlapped in terms of tokens with NEs in the GC. As will be mentioned in section 3 partial alignments can be evaluated in several ways.

Semantic evaluation in our framework was captured through the use of four different measures. Since in the GC we marked entities with a category and a type, obviously two of those measures individually took into account each of these axes, in which the possible values are *Correct*, *Spurious* or *Missing*. The other two measures combined category and type in an overall score. In terms of **morphological** classification, NEs were scored as either *Correct*, *Partially correct*, *Wrong*, *Missing*, *Spurious* or *Overspecified*.

3 Software

Automation of the evaluation task was achieved through the use of specialized evaluation software. Since our evaluation scheme differs from the MUC scenario and consequently from their software [5] we had to implement our own evaluator. The software is implemented in a pipelined architecture all programs read from a file (the initial or an already partially processed submission), perform some processing and output a new file, which could then be studied and searched for unexpected situations, initially unforeseen. The HAREM evaluation architecture is presented in figure 1.

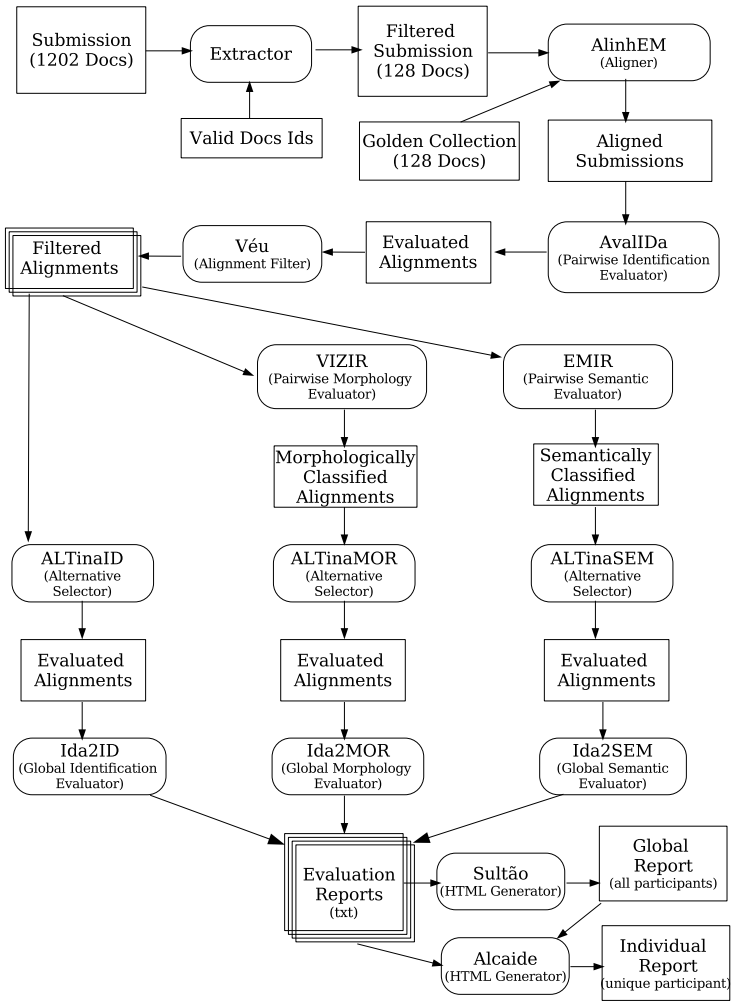

Fig. 1. Overall HAREM Evaluation Architecture

The first step is to extract, from the HAREM collection, the subpart for which there is a key (the golden collection). After alignment we individually evaluate each target NE the according to the rules described. In this step we simply attach information to every alignment attributing a score to each aligned NE.

Since we allowed participants to choose which NE categories they wanted to be evaluated in (the selective scenario), the alignment filter removes all the categories that should be discarded. Note that this step is not as trivial as may first seem, as we must account for situations where an NEs can be considered spurious or missing depending on the type of filter applied. It is worthwhile noting that the filter can also be configured to ignore partial alignments (MUC style) or, in

the case of several NEs aligning with one NE, just considering one NE as partially correct instead of all of them (we are grateful to Beth Sundheim (p.c.) for this suggestion). By default, our filter considers all partial alignments for evaluation purposes (for the exact metrics see http://www.linguateca.pt/HAREM/).

At this point in the processing we have several files; one corresponding to the total scenario and the others to several selective scenarios (note that we also use the filter to select subparts of the task – by textual genre, by language variety, by single semantic category, etc). Each file produced will then follow three different evaluation paths. Since the alignments produced have only been individually evaluated according to the identification criteria, they have to be submitted to the pairwise morphology evaluator and the pairwise semantic evaluator as well.

Finally, after the best alternatives have been chosen, overall scores for precision, recall, over-generation and under-generation are computed for each task, and individual and global (comparative) HTML and PDF reports produced.

4 Concluding remarks

Compared to MUC and other evaluation contests for NER, the architecture devised and deployed in HAREM represents progress, because it adds a number of degrees of freedom to experiment with. Some studies concerning Portuguese and different text genres have already been carried out.

More significant perhaps, is the fact that the whole architecture (and the programs implementing it) is publicly available. Together with the GC, any researcher can develop and test NER systems for Portuguese. Ample opportunity for reuse will, anyway, occur in the next HAREM contests.

5 Acknowledgements

This work was supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia, co-financed by POSI.

References

1. Hirschman, L.: The evolution of evaluation: lessons from the message understanding conference. Computer Speech and Language **12** (1998) 281–305
2. Santos, D., Seco, N., Cardoso, N., Vilela, R.: Harem: An advanced NER evaluation contest for portuguese. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, Genova, Italy (2006)
3. Grishman, R., Sundheim, B.: Message understanding conference-6: a brief history. In: Proceedings of the 16th conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1996) 466–471
4. Santos, D., Cardoso, N.: A golden resource for named entity recognition for portuguese. In: This Volume, Itatiaia, Rio de Janeiro, Brasil, Springer (2006)
5. Douthat, A.: The message understanding conference scoring software users manual. In: Proceedings of the 7th Message Understanding Conference (MUC-7). (1998)