HAREM: An Advanced NER Evaluation Contest for Portuguese

Diana Santos*, Nuno Seco†, Nuno Cardoso‡ and Rui Vilela⊕

Linguateca: nodes of *Oslo, †Coimbra, ‡XLDB in Lisbon, and ⊕Braga
diana.santos@sintef.no, nseco@dei.uc.pt, ncardoso@xldb.di.fc.ul.pt, ruivilela@di.uminho.pt

Abstract

In this paper we provide an overview of the first evaluation contest for named entity recognition in Portuguese, HAREM, which features several original traits and provided the first state of the art for the field in Portuguese, as well as a public-domain evaluation architecture.

1. Introduction

Although there is a wide awareness of NER evaluation after the MUC conferences (Hirschman, 1998), there were several points that could be improved in the evaluation setup of (Grishman and Sundheim, 1996). After all, the NE task was inspired as a follow-up of an information extraction task about a particular kind of events, and there is more to NER than that, if we take the position, as (Romacker and Hahn, 2000) do, that NER is a subtask of semantic interpretation.

HAREM[1] was organized by Linguateca and had 10 participants from 6 different countries, who submitted 15 runs (plus 6 non-official ones, i.e. sent after the deadline). Roughly, the participants had 48 hours to tag a large and varied collection (the HAREM 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, described in section 3) had been manually handcoded by the organization, according to detailed guidelines discussed with the participants, who also provided some initial annotation effort.

HAREM features several innovative traits, which we describe in more detail below:

- Separation between properly identifying (or detecting, or flagging) and classifying NEs;
- Introduction of a morphological task, given that the "same" NE may represent different things in a different gender or number;
- Taking into account vagueness and indeterminacy, both while building the golden resource and when evaluating the systems;
- Allowing the choice of a subset of semantic categories for evaluation, i.e. a kind of partial ontology morphism;
- Definition of several measures to reflect subtle distinctions, such as partial overlap, overgeneration, and distinguishing assignment of minor type vs. major type;
- Providing meta-information associated with the texts, to allow investigation on genre and Portuguese variety.

[1]HAREM stands for "HAREM - Avaliação de sistemas de Reconhecimento de Entidades Mencionadas".

As a result of the first HAREM evaluation contest (available from `http://www.linguateca.pt/HAREM/`), we can present:

- a fully documented modular architecture, whose source code is available under GPL;
- a valuable resource for the deployment of further systems in this area, the golden collection (GC), also useful to conduct research on the problem and on the evaluation methodology, as in Morfolimpíadas (Santos and Barreiro, 2004; Santos et al., 2003);
- a first state of the art for Portuguese.

As a less objective – but not less rewarding – result of our work with HAREM, we can report on the emergence of an active community dealing with NER for Portuguese.

This paper starts by describing the evaluation architecture, then the resources and finally sketches possible future developments in Portuguese NER.

2. The Evaluation Setup

Evaluation in HAREM was divided into three tasks, each capturing different aspects of the NER problem, namely (i) identification; (ii) semantic classification; and (iii) morphological classification. Participating systems were, in addition, allowed to choose a subset of the categories in which they wanted to be evaluated. By allowing a *selective* participation, systems which were fine tuned for certain purposes could be evaluated accordingly, making their scores increasingly relevant to their designers. As such, for each submitted run, we analyzed two different scenarios:

total Taking into account all categories defined in HAREM.

selective Taking into account a subset of the categories, namely those which the system is tuned to recognize.

Another important aspect of our evaluation framework is that we distinguish between two types of evaluation:

absolute Taking into account all entities in the GC and all the NEs identified by the system.

relative Taking into account only the NEs that were regarded as correct and partially correct in the identification task.

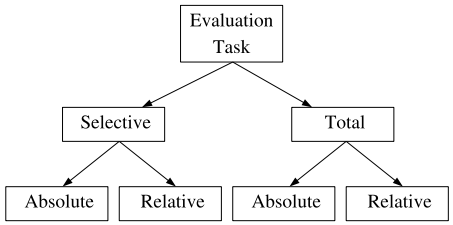

Figure 1: Evaluation configurations

In other words, absolute evaluation considers missing and spurious NEs, whereas relative evaluation does not. Figure 1 visually sums up the evaluation combinations. For comparison purposes, let us state that MUC only dealt with a total absolute combination.

2.1. Entity Alignment

Before delving into the details of evaluating each task, we must first explain how we achieve accurate alignment of the entities tagged by the system (the *target* NEs henceforth) with the entities manually tagged (the *source* NEs henceforth) in the GC. Alignment of target and source NEs is a crucial task and heavily influences the caliber of the results, as is convincingly argued by (Kehler et al., 2001).

The alignment procedure was particularly difficult in our case, since many of the participating systems modified the texts that they were only supposed to tag. In other contests, these systems would probably have been disqualified and their submission ignored, but in HAREM we tried to salvage these submissions by implementing a robust entity aligner that could cope with these issues.

In order for the reader to get a better grasp of the problem at hand, consider the two documents of figure 2, depicting the same text in the GC and as classified by a system.

Several differences become immediately apparent. In the submitted text, contractions were "expanded" (DA, DO and PELA were replaced by DE A, DE O and POR A, respectively). Spaces were introduced between non-alphanumeric and alphanumeric characters; one particular numerical token (1937) was split. Some systems introduced preceding backslash (\) before some characters (e.g. 01\/01\/1997), others ignored (and consistently tagged as normal text) meta-information such as genre and variety, provided by the organization.

These and many other modifications to the original text – we gathered a list of about 80 different types of intrusive modifications – made the alignment process quite hard. A naive approach trying to map words (tokens) in equal positions of both texts would be deemed to fail. We had, therefore, to design software that could deal with this in a robust fashion. The main idea underlying our approach was **to concentrate on the strings classified as source NEs, and find them in the systems' outputs**. This means we can ignore all words/tokens in the texts except for those that are part of source NEs.[2]

We have in addition to take into consideration the ordering of the tokens: we need to guarantee that the first occurrence of a source token aligns with the first occurrence of the same token in the target text. This requirement is achieved by numbering each occurrence of a particular token; see figure 3 for illustration. Considering the text given in figure 3, all tokens relevant to the alignment process are:

```
ST = {<1>DCC</1>, <2>DCC</2>, <3>DCC</3>,
<1>Departamento</1>, <1>Cultura</1>, <1>Científica</1>,
<1>Centro</1>, <1>Acadêmico</1>, <1>Pedro</1>,
<1>Nunes</1>, <1>CAPB</1>, <1>UNIFESP</1>, <1>EPM</1>,
<1>1</1>, <1>9</1>, <1>3</1>, <1>7</1>}.
```

Note that the year 1937 was tokenized into four tokens in order to deal with *faulty* submissions that split numerical entities (dates, numbers, ...) during the tagging process.

The same numbering scheme (considering the same tokens) can then be applied to the target text. By doing so we uniquely identify each token, facilitating the alignment process. Alignment is now a question of mapping identical tokens from each text to each other. The elements of the set ST represent the tokens that we have to identify in the target NEs.[3]

By aligning tokens we may trivially derive the NE alignments simply by considering that **the alignment of a source token with a target token entails (some) alignment of the NEs (source and target) to which the tokens belong.** Applying the above rule and numbering scheme allows us to infer, for the texts in figure 2, the alignments depicted in figure 4.

NEs alignments are obviously not always perfect one-to-one mappings. Rather, we have to deal with the following five different cases: one to one, many to one, one to many, none to one (the system has incorrectly identified a (spurious) NE), and one to none (one source NE has been missed by the system).

2.2. Identification

The identification task aimed at evaluating systems with respect to their ability to correctly limit the bounds of a named entity, **irrespective of its semantic or morphological classification**. We consider five different scores for every pairing of source and target NEs:

Correct if all tokens correctly match (except for the above mentioned stopwords)

Partially correct by Excess if at least one token matches with a token belonging to a source entity, but the target entity has an equal or greater number of tokens than the source entity.

Partially correct by Shortage if at least one token matches with a token belonging to a source entity, but the target entity has a smaller number of tokens than the source entity.

Spurious if the target entity has no counterpart in the GC.

Missing if the source entity has no corresponding target entity, i.e., the particular GC entity has not been identified by the system.

[2] This statement is not entirely true, as we will need to consider the tokens that belong to target NEs in order to determine spurious NEs, but this is a relatively straightforward step.

[3] Because most changes occurred in contractions and other grammatical words, we created a stopword list of tokens which we ignore in the alignment process.

```
HISTÓRICO Esta seção traz de volta um pouco da longa
história do <ORGANIZACAO TIPO="SUB"
MORF="M,S">DCC</ORGANIZACAO>. O <ORGANIZACAO TIPO="SUB"
MORF="M,S">DCC</ORGANIZACAO> - <ORGANIZACAO TIPO="SUB"
MORF="M,S">Departamento de Cultura Científica do Centro
Acadêmico Pedro Nunes </ORGANIZACAO> (<ORGANIZACAO
TIPO="SUB" MORF="M,S">DCC/CAPB</ORGANIZACAO>), órgão
responsável pela representação e encaminhamento
científico dos alunos da
<ORGANIZACAO TIPO="INSTITUICAO" MORF="F,S">UNIFESP/EPM
</ORGANIZACAO>, fundado em <TEMPO TIPO="DATA">1937
</TEMPO>, atua junto aos alunos promovendo vários cursos
extracurri-culares, palestras, conferências e discussões
de interesse à área médica.
```

→

```
HISTÓRICO Esta seção traz de volta um pouco de a longa
história de o <ORGANIZACAO TIPO="SUB"
MORF="M,S">DCC</ORGANIZACAO> . O <ORGANIZACAO TIPO="SUB"
MORF="M,S">DCC - Departamento de Cultura</ORGANIZACAO>
<ORGANIZACAO TIPO="SUB" MORF="M,S">Científica de o
Centro Acadêmico Pedro Nunes </ORGANIZACAO>
( DCC / CAPB) , órgão responsável
por a representação e
encaminhamento científico de os alunos de a
UNIFESP / EPM , fundado em
19<TEMPO TIPO="DATA">37</TEMPO>,
atua junto aos alunos promovendo vários cursos
extracurri-culares,palestras, conferências e discussões
de interesse à área médica .
```

Figure 2: The same document in the golden collection (to the left) and tagged by a system (to the right).

```
HISTÓRICO Esta seção traz de volta um pouco da longa história do <ORGANIZACAO TIPO="SUB" MORF="M,S"><1>DCC</1>
</ORGANIZACAO>. O <ORGANIZACAO TIPO="SUB" MORF="M,S"><2>DCC</2></ORGANIZACAO>-<ORGANIZACAO TIPO="SUB"  MORF="M,S">
<1>Departamento</1> de <1>Cultura</1> <1>Científica</1> do <1>Centro</1> <1>Acadêmico</1> <1>Pedro</1> <1>Nunes</1>
 </ORGANIZACAO> (<ORGANIZACAO TIPO="SUB" MORF="M,S"><3>DCC</3>/<1>CAPB</1></ORGANIZACAO>), órgão responsável pela
```

Figure 3: The previous document with all tokens inside NE's duly numbered.

In the case of partial identifications we associate a weight to these entities that reflects their contribution to the overall score. The weight is calculated according to equation:

$$0.5 * (nc/nd) \quad (1)$$

where nc and nd correspond to the number of common tokens and the number of distinct tokens, respectively. This metric essentially measures the degree of similarity between two NEs, while the factor of 0.5 guarantees that the sum of two or more partially correct entities is always less than 1. Going back to our example in figure 4 we can see that the source NE *Departamento de Cultura Científica do Centro Acadêmico Pedro Nunes* is aligned with two target NEs, *Departamento de Cultura* and *Científica do Centro Acadêmico Pedro Nunes*. Accordingly, our metric yields the score 0.17 for the first NE and 0.33 for the second. Note that, if we had not employed the 0.5 factor, the sum of the two would yield 1, which is the score assigned to exact (precise) identification – obviously inadequate in this case.

It could be argued that NEs should be atomic, and so it makes little sense to consider two target NEs as partially correct each instead of the large NE as missing. This may be right in some cases; however, it is undeniable that some NEs are intrinsically compositional, so that allowing for partial credit seems reasonable. Consider again the NE *Departamento de Cultura Científica do Centro Acadêmico Pedro Nunes* (a department (DCC) inside a larger organization (CAPN) named after the scientist Pedro Nunes): if a system identifies the two "institutions" separately, it seems to deserve some credit. Conversely, if it identifies *Pedro Nunes* as a PERSON and fails to find any organization, although it still gets some credit in the identification task[4], it is considered wrong in the other tasks.

2.3. Semantic classification

Semantic classification is the assignment of a semantic category and type to every identified NE from a predefined set of categories and types, diplayed in Table 2. Systems could use the OUTRO ("other") type if they did not want to perform subcategorization for some categories, or use the selective option to choose a subset of types.

[4]Our evaluation package has the possibility of both strategies (HAREM and MUC-style) to allow for experimentation.

Evaluation of semantic classification, which takes place after evaluating the identification task, considers the semantic tags assigned to the target NEs. As illustrated above, the same NE can be considered correct in the identification task, but missing or spurious in terms of semantic content. Consider also the alignments of figure 5. In terms of identification, the first alignment is considered correct, but considering the semantic content of the second (given by the tags employed), the ORGANIZACAO NE is missing in the system's output while the LOCAL target NE is spurious.

We provide four different metrics for measuring a systems' performance in the semantic classification task, which can be conceived as complementary evaluation viewpoints:

Only categories Only categories are evaluated, types are ignored.

Only types Given the subset of NEs with correct categories, this measure evaluates the types submitted.

Combined Categories and types are evaluated simultaneously. This metric tries to model the intuition that the more types there are, the more information a correct choice brings, and therefore a higher reward is deserved. The formula is:

$$\begin{cases} 0, & \text{if category is incorrect;} \\ 1, & \text{if category is correct and type is incorrect;} \\ 2 - \frac{1}{n}, & \text{if both category and type are correct} \end{cases}$$

where n represents the number of types that the given category could have.

Flat Considers the pair (category, type) as the tag, and only if the same pair appears in the target NE is it considered correct. This is stricter than simple category evaluation, since if the category is correct but the type is not, then the NE is considered incorrectly classified.

2.4. Morphological Classification

Morphological classification of NEs as defined by HAREM consists of filling the MORF attribute of identified NEs with values for the gender and number attributes, respectively F,M,? and S,P,?, with "?"standing for underspecified.

```
<ORGANIZACAO TIPO="SUB" MORF="M,S"><1>DCC</1></ORGANIZACAO> ---> [<ORGANIZACAO TIPO="SUB" MORF="M,S"><1>DCC</1>
</ORGANIZACAO>]
<ORGANIZACAO TIPO="SUB" MORF="M,S"><2>DCC</2></ORGANIZACAO> ---> [<ORGANIZACAO TIPO="SUB" MORF="M,S"><2>DCC</2>
</ORGANIZACAO>]
<ORGANIZACAO TIPO="SUB" MORF="M,S"><1>Departamento</1> de <1>Cultura</1> <1>Científica</1> do <1>Centro</1>
<1>Acadêmico</1> <1>Pedro</1> <1>Nunes</1></ORGANIZACAO> ---> [<ORGANIZACAO TIPO="SUB" MORF="M,S"><1>Departamento</1>
de <1>Cultura</1></ORGANIZACAO>, <ORGANIZACAO TIPO="SUB" MORF="M,S"><1>Científica</1> do <1>Centro</1>
<1>Acadêmico</1> <1>Pedro</1> <1>Nunes</1></ORGANIZACAO>]
<ORGANIZACAO TIPO="SUB" MORF="M,S"><3>DCC</3>/<1>CAPB</1> </ORGANIZACAO> ---> [null]
<ORGANIZACAO TIPO="INSTITUICAO" MORF="F,S"><1>UNIFESP</1>/<1>EPM</1> </ORGANIZACAO> ---> [null]
<TEMPO TIPO="DATA"><1>1</1><1>9</1><1>3</1><1>7</1></TEMPO> ---> [<TEMPO TIPO="DATA"><1>3</1><1>7</1></TEMPO>]
<ESPURIO><1>cursos extracurriculares</1></ESPURIO> ---> [<ORGANIZACAO TIPO="SUB"><1>cursos extracurriculares</1>
</ORGANIZACAO>]
```

Figure 4: Alignments: Golden collection to the left; system's output to the right

```
<ORGANIZACAO TIPO="SUB" MORF="M,S">DCC</ORGANIZACAO> ---> [<LOCAL TIPO="VIRTUAL" MORF="M,S">DCC</LOCAL>]
<ORGANIZACAO TIPO="SUB" MORF="M,S">DCC</ORGANIZACAO> ---> [<ORGANIZACAO TIPO="EMPRESA" MORF="M,S">DCC</ORGANIZACAO>]
```

Figure 5: Semantically correct and incorrect alignments

Note that not all NEs are supposed to require morphological classification: it simply does not make sense to classify URLs, postal addresses, movie names, etc. according to number and gender. We have thus made a choice in the GC as to which NEs should have morphological information, and we disregard in HAREM whatever MORF values the systems may have come up with in those cases.

For the evaluation of morphological classification, we compute two distinct scoring fields and a combined one (Gender, Number, Combined). For each of the previous fields, a given alignment can be considered:

Correct If the two values are equal and the NE was identified correctly;

Partially correct If the two values are equal, the NE was identified as partially correct, and the two NEs share (agree in) their beginning;

Wrong If a wrong specific value was provided;

Missing Several different cases are characterized as missing: (i) if the NE was missing in the identification task; (ii) if the system did not assign the attribute `MORF`, even though it properly identified the NE; (iii) if the system classified as underspecified a case which had a specific value in the GC; or (iv) if a partially correctly identified NE does not share its first token with the one it aligns with in the GC.

Spurious If a given NE does not exist at all in the GC but has been morphologically classified by the system;

Overspecified If the system produced a specific value when in the golden collection the NE was considered morphologically underspecified.

An example of morphological assessment is presented in figure 6. In the first case, morphological classification is considered correct. In the next case, the system has produced a gender different from the one in the golden collection, therefore it scores gender as wrong, but number as correct. The final case, in which a system is not able to assign gender or number to an NE, which was classified for both properties in the GC, is dealt with by considering both attributes missing.

3. Building the Golden Collection

HAREM's Golden Collection (henceforth GC) is a collection of texts from several origins (table 1) and genres (figure 7), in which NEs have been identified, semantically classified and morphologically tagged in context, and revised independently, according to a large set of directives approved (and discussed) by all participants.

Origin	Text extracts	Words	NEs
Portugal	63	33 618	2 550
Brazil	60	42 073	2 274
Asia	3	2 864	233
Africa	3	1 253	75

Table 1: Language variety distribution

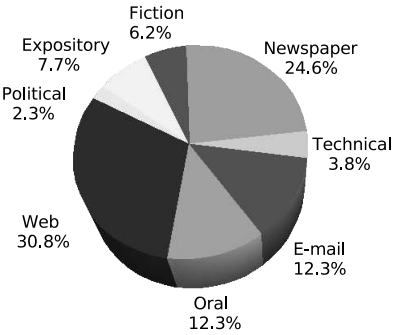

Figure 7: Genre distribution in the GC

Some important design decisions for this evaluation resource were:

- Use a preliminary corpus-based assessment to decide on the final categories relevant for Portuguese (instead of forcing a top down approach)
- Select types (subdivision of the main categories) which had some linguistic marking in Portuguese: for example, the distinction between unique works and not reproducible ones, or between an atomic event or a large one;
- Do not recursively annotate NEs;[5]

[5]No participating system actually accepted embedded NEs, so there was no point in making the whole evaluation setup unnecessarily complex.

```
<PESSOA TIPO="INDIVIDUAL" MORF="M,S">Marcelo Calixto
</PESSOA> ---> [<EM MORF="M,S">Marcelo Calixto</EM>]
:[Correcto]
<LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Pedro Leopoldo
</LOCAL> ---> [<EM MORF="M,S">Pedro Leopoldo</EM>]:
[Correcto]
<LOCAL TIPO="GEOGRAFICO" MORF="F,S">Costa Africana
</LOCAL> ---> [<EM MORF="?,?">Costa Africana</EM>]:
[Correcto]
```

→

```
<EM MORF="M,S">Marcelo Calixto</EM>  ---> [<EM MORF=
"M,S">Marcelo Calixto</EM>]:[(Género: Correcto 1)
(Número: Correcto 1) (Combinada: Correcto 1)]
<EM MORF="M,S">Pedro Leopoldo</EM>  ---> [<EM MORF=
"F,S">Pedro Leopoldo</EM>]:[(Género: Incorrecto 0)
(Número: Correcto 1) (Combinada: Incorrecto 0)]
<EM MORF="F,S">Costa Africana</EM>  ---> [<EM MORF=
"?,?">Costa Africana</EM>]:[(Género: Em Falta 0)
(Número: Em Falta 0) (Combinada: Em Falta 0)]
```

Figure 6: Examples of morphological evaluation after alignment

- Classify NEs in context, in a more refined way than usual (see below for discussion on this point).

Category	Type	English gloss	Nr.
PESSOA	INDIVIDUAL	individual person	856
21.5%	CARGO	title	79
BP: 58.75%	MEMBRO	members	10
BR: 72.72%	GRUPOIND	group of people	10
	GRUPOCARGO	group of titles	19
	GRUPOMEMBRO	group of members	137
ORGANIZACAO	ADMINISTRACAO	administration	224
18.0%	INSTITUICAO	institution	462
BP: 51.01%	EMPRESA	company	230
BR: 62.72%	SUB	sub-organization	61
TEMPO	DATA	date	335
8.5%	HORA	time	39
BP: 77.68%	PERIODO	period	62
BR: 69.79%	CICLICO	cyclic	5
LOCAL	CORREIO	address	17
25.0%	ADMINISTRATIVO	administrative	906
BP: 68.03%	GEOGRAFICO	geographic	86
BR: 73.91%	VIRTUAL	virtual	126
	ALARGADO	extended	161
OBRA	PRODUTO	product	74
4.0%	REPRODUZIDA	reproducible work	89
BP: 20.58%	ARTE	unique work	10
BR: 18.85%	PUBLICACAO	publication	51
ACONTECIMENTO	EFEMERIDE	unique	23
2.5%	ORGANIZADO	large event	62
BP: 50.76%	EVENTO	atomic event	45
BR: 46.61%			
ABSTRACCAO	DISCIPLINA	subject	228
8.5%	MARCA	brandname	36
BP: 45.43%	ESTADO	condition	34
BR: 38.04%	ESCOLA	school	14
	IDEIA	ideal	45
	PLANO	plan	40
	OBRA	complete works	4
	NOME	name	76
COISA	OBJECTO	object	39
1.6%	SUBSTANCIA	substance	9
BP: 25.38%	CLASSE	class	37
BR: 40.74%			
VALOR	CLASSIFICACAO	classification	62
9.5%	QUANTIDADE	amount	370
BP: 84.82%	MOEDA	money	53
BR: 79.69%			
VARIADO	OUTRO	other	42
0.9%			

Table 2: HAREM categories and types, their distribution in the GC, and best precision and recall, BP and BR. Whenever a NE is considered vague between A_1 and A_2, we incremented both counters.

Table 2 describes the population of the GC in terms of the HAREM categories, providing also the best results per category.

There are in principle two ways to NE-annotate a corpus: (i) to consider the main category a given NE represents, and use it; or (ii) to try to make a finer classification on how they are used in text. Two examples should make this distinction clear: the case of countries, and of newspapers. Form (i), which we believe to be mainstream, would NE-classify any reference to a country as COUNTRY (or LOCATION), and any reference to a newspaper as NEWSPAPER (or ORGANIZATION). For us, this is a much simpler task than to decide in which role a given proper name is mentioned in a given context, which was the path taken in HAREM.

It is well known that different kinds of proper names have a number of distinct roles related to their semantics, that is: a country is generally associated with a geographical location, a political organization, **and** an abstract property; while a reference to a newspaper typically reflects the many different roles of: place of publication, workplace, product, organization, or even a reporter representing the newspaper. Generalizing further, any concrete thing can also be employed to refer to a place in space; any event can also be used in place of a date, or to refer to its organization committee, or even metonymically to its spokesman.

In HAREM we wanted to compare systems in a meaningful task, not a middle/intermediate task that would require further interpretation and processing to be useful. We are convinced that names of countries as geographical locations are clearly separated from names of countries as political entities, and that most applications would rather be able to select which of these two. Also, a person interested in finding out which companies are lately firing their employees would not like to see any other reference to a newspaper than the ones in which it is referred to as a company.

We were in any case aware that this decision would make considerably harder the task of the systems; conversely, it led to the creation of a incomparably richer and more complex resource. To create it also resulted in much work, and not every decision – although documented in detail – is consensual. In order to give an idea of what kinds of decisions had to be taken (detailed in (Santos and Cardoso, 2006)), we list here some:

- we only allow a fixed small set of uncapitalized NE beginnings: titles and address forms (such as in *major Otelo* and *senhor Alves*) and disease hyperonyms (such as in *doença de Alzheimer*, and *sindroma de Down*);
- if a title is followed by a name of a person (who has this title), consider only one NE, as in *Presidente do Parlamento Europeu Stefano Prodi*;
- acronyms in parentheses after a given name are considered as a new NE (inside the parentheses). This allows us to distinguish these from cases of a particular NE also having parentheses or acronyms in its name.

Also, we hold the view that ceiling effects are as relevant in evaluation contests as are baselines. Accordingly, during manual annotation of the golden collection we were extremely careful in maintaining vagueness (OR categories) in the annotation, both in semantic classification (where we marked NEs as $< A_1|A_2|A_n >$), in identification, where

we used an <ALT> tag to mark alternative identification solutions, and in morphology, where the "?" category is treated as different from no morphological classification provided.

Size	GC	HAREM Collection
Words	92 761	520 752
Text extracts	129	1 202
Named entities	5 132	ca. 40 000
Vague NEs in classification	131	ca. 1000
Vague NEs in identification	65	ca. 500

Table 3: Statistics of the collections

We therefore provide a quantitative description of the cases where indeterminacy was our best annotation choice in table 3, discriminating the number of NEs with <ALT>, and the number of NEs semantically classified with one or more alternatives in the GC.

4. Concluding remarks

We believe that methodological questions and architecture description are more interesting for an international audience than reporting the actual results obtained by the participants, each of which received a 100-pages report on the individual and comparative performance of their system, but the (anonymized) results are publically available on the HAREM website for anyone interested.

However, over and above the results of the systems, as organizers, we wanted to investigate some questions with our setup, most notably the influence of genre in the NER task, and produce some ranking of the categories: are there some more difficult than others? Figures 8 and 9 graphically illustrate some initial results (Seco et al., 2006).

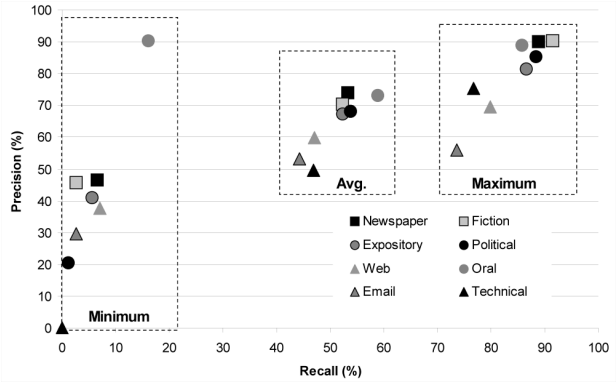

Figure 8: Precision vs Recall for text genres

Currently (one year after the event) we are organizing the first HAREM sequel, with a new golden collection, to study both the systems' evolution and the statistical reliability of the compiled material (Cardoso, 2006).

Reflections around the building of the golden resource suggest that even more information should have been manually coded, such as annotation of the deviant cases in terms of spelling or punctuation (too many capitals, too little capitals); as well as a thorough explicit encoding of anaphoric or reduced NEs.

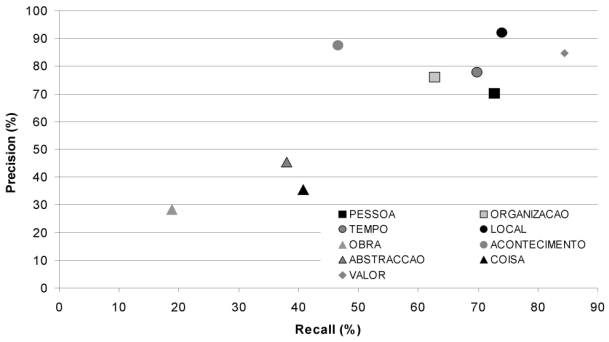

Figure 9: Precision vs Recall for Categories (Best)

5. Acknowledgements

The authors wish to thank Fundação para a Ciência e Tecnologia for the grant POSI/PLP/43931/2001, co-financed by POSI. We are also grateful to all participants of HAREM, in particular to Cristina Mota for her initial effort in the field and the many stimulating discussions.

6. References

Nuno Cardoso. 2006. Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Master's thesis, FEUP, Porto, Portugal. In preparation.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics*, pages 466–471, Morristown, NJ, USA. ACL.

Lynette Hirschman. 1998. The evolution of evaluation: lessons from the message understanding conference. *Computer Speech and Language*, 12(4):281–305.

Andrew Kehler, John Bear, and Douglas Appelt. 2001. The need for accurate alignment in natural language system evaluation. *Computational Linguistics*, 27(2):247–248.

Martin Romacker and Udo Hahn. 2000. An empirical assessment of semantic interpretation. In *Proceedings of NAACL*, pages 327–334, Seattle, Washington.

Diana Santos and Anabela Barreiro. 2004. On the problems of creating a consensual golden standard of inflected forms in Portuguese. In Lino et al, editor, *Proceedings of LREC'2004*, pages 483–486, Lisbon.

Diana Santos and Nuno Cardoso. 2006. A golden resource for named entity recognition in Portuguese. In *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006*, Itatiaia, Rio de Janeiro, Brasil, 13th-17th May. Springer.

Diana Santos, Luís Costa, and Paulo Rocha. 2003. Cooperatively evaluating portuguese morphology. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, and Maria das Graças Volpe Nunes, editors, *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003*, pages 259–266, Faro, Portugal, June. Springer.

Nuno Seco, Diana Santos, Nuno Cardoso, and Rui Vilela. 2006. A complex evaluation architecture for HAREM. In *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006*, Itatiaia, Rio de Janeiro, Brasil, 13th-17th May. Springer.