Autores

Antonio Toral Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.

Andrés Montoyo Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.

Bruno Martins Universidade de Lisboa, Faculdade de Ciências, LaSIGE, Portugal, *agora* Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Christian Nunes Aranha Cortex Intelligence, Brasil.

Cristina Mota Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal, *agora* Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal / L2F, INESC-ID, Portugal / New York University, EUA.

Diana Santos Linguateca, SINTEF ICT, Noruega.

Eckhard Bick VISL, Institute of Language and Communication, University of Southern Denmark, Dinamarca.

José João Dias de Almeida Departamento de Informática, Universidade do Minho, Portugal.

Luís Sarmento Linguateca, CLUP, Faculdade de Letras da Universidade do Porto, Portugal, *agora* Faculdade de Engenharia da Universidade do Porto, Portugal.

Marcirio Chaves Universidade de Lisboa, Faculdade de Ciências, LaSIGE, Portugal.

Marília Antunes Universidade de Lisboa, Faculdade de Ciências, Portugal.

names are viewed as ***mentions*** *of the underlying entities.* Não podíamos ter confirmação mais evidente para a nossa escolha de nome em português, nem demonstração mais óbvia de que o HAREM e o ACE identificaram o mesmo problema no MUC. Contudo, abordaram-no de uma forma diametralmente oposta.

O problema do MUC, que refinamos aqui, é que partia de uma definição arbitrária com base nos dois campos ligados pela semântica (a língua/forma, e o conteúdo/denotação), delimitada por um subconjunto deste último: a tarefa do MUC tinha como alvo nomes próprios (forma) com significado de organização, local, etc. (denotação), como se aprecia nas palavras de Chinchor e Marsh (1998): «*the expressions to be annotated are "unique identifiers" of entities (organizations, persons, locations), times [...] and quantities [...] The task requires that the system recognize what a string represents, not just its superficial appearance*».

O ACE escolheu o **lado do conteúdo** e pediu para — independentemente da forma — os sistemas marcarem tudo o que fosse organização, local, pessoa, etc., sem restrições de forma (podiam ser realizados linguisticamente como substantivo, pronome, nome próprio, sintagma nominal, etc.).

O HAREM, ao contrário, **escolheu o lado da forma**: partiu de tudo o que é nome próprio em português (ver capítulo 16) e pediu para os sistemas identificarem e classificarem – sem restrições de sentido numa primeira fase, mas, depois de um estudo empírico inicial – com base na classificação proposta pela organização. (Note-se, no entanto, que aceitamos uma categoria OUTRO, ou seja, não garantimos que todas e quaisquer ocorrências de nomes próprios no texto podem ser enquadrados no produto cartesiano das categorias do HAREM.)

A parecença entre as duas extensões ao MUC (ambas reconhecem o MUC como inspiração) é também visível no aumento da variedade em tipo de textos: em vez de alargar em género como fizemos no HAREM, contudo, o ACE alargou em qualidade de texto ou meio de obtenção desse texto. Além de notícias impressas, usou textos obtidos a partir de reconhecimento óptico, e de reconhecimento automático de fala. Também alargou o assunto (em vez de um único domínio, passou a ter notícias sobre vários domínios ou assuntos). Interessante que, no caso do HAREM, usámos a extensão em termos de variante e sobretudo de estilo/género textual, alargando em termos de meio ou de qualidade apenas quando tal derivava de um género textual diferente: em particular, para cobrirmos a Web, tivemos de incluir textos de pouca qualidade, e para incluir entrevistas, tivemos de recorrer à transcrição da linguagem oral.

Outra semelhança entre o ACE e o HAREM foi o aumento significativo da complexidade na anotação humana, que, de acordo com (Maynard et al., 2003a), atingiu apenas 82,2% de consenso no ACE.

Outra diferença em relação ao MUC partilhada pelo HAREM e pelo ACE é a utilização neste último duma métrica baseada em custo (Maynard et al., 2002), que, embora mais geral do que a do HAREM, tem pontos de semelhança com a medida da classificação se-

	Attributes											
	Hand-coded tagger			Internal			External			Internal & External		
Class	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*
B	60.0	68.8	0.641	**82.4**	85.8	0.841	75.9	81.0	0.784	82.1	**87.8**	**0.849**
I	64.5	73.3	0.686	80.1	76.8	0.784	73.8	70.3	0.720	**80.9**	**77.8**	**0.793**
O	97.2	95.5	0.964	98.7	98.5	0.986	98.1	97.7	0.979	**98.8**	98.4	0.986
Overall	73.9	79.2	0.763	87.0	87.0	0.870	82.6	83.0	0.827	87.2	88.0	0.876

Table 10.7: Experimental results for NED in Portuguese.

10.2.1 Results on NED

In this section we report our results of NED in Portuguese. We describe the distribution of instances over classes for the Portuguese corpus in Table 10.6. As the goal is to explore to what extent our method can be applied to similar languages, we did not make any particular changes to our system. The method is applied in the same way as it was applied previously to Spanish, results for Spanish can be found at Solorio (2005). Experimental results on NED are presented in Table 10.7. These results are averaged using a 10-fold cross validation[3]. We can observe that the hand-coded tagger achieved surprisingly high classification measures, it reached an F measure of 0.763. We believe that these results reveal that the two languages share some characteristics, among them the orthographic features: in Portuguese it is also conventional to write proper names with the first letter in uppercase. On the other hand, note also that the behavior of the two types of features differs greatly from that observed for Spanish. The internal features have better results than the external, for Spanish we observed that external features achieved better results than the internal ones. A plausible explanation to this is that, given that the hand-coded tagger misclassified more instances in the Portuguese case, then it is harder for the SVM, trained with the output of the hand-coded tagger, to learn the task in this somehow noisier setting. Nonetheless, SVM did improve the accuracy of the hand-coded tagger, and even more relevant for us, the combination of the two types of features yielded the best results. In this setting, our method is still the best option to achieve higher precision and recall on NED in Portuguese.

10.2.2 Results on NEC in Portuguese

We have shown that our proposed solution works well for Portuguese NED, now we need to evaluate how well this solution works for NEC in Portuguese. In this case the classifi-

[2] **Editors' note.** Note that the author does not apply in the chapter the measures used for HAREM elsewhere in this book, but rather defines her own, such as accuracy per word. Also she uses a small subset of the first golden collection, not the full golden collection.

[3] Since this is a classification task where we need to assign to every word one out of three possible classes, we measure per word accuracies.

17.1.1 Género (morfológico)

Consideramos que o género de uma EM pode ter três valores:

M: EM com género masculino.

F: EM com género feminino.

?: Para os casos em que o género é indefinido.

17.1.2 Número

Consideramos que o número de uma EM pode ter três valores:

S: EM no singular.

P: EM no plural.

?: Para os casos em que o número é indefinido.

17.1.3 Exemplos de não atribuição de `MORF` na categoria `LOCAL`

Em alguns casos particulares do tipo `VIRTUAL`, o atributo `MORF` foi omitido, devido ao facto de não ser possível avaliar morfologicamente números de telefone.

```
Certo: <LOCAL TIPO="VIRTUAL">(48) 281 9595</LOCAL>
```

Os casos que possuam a etiqueta `MORF` são, pelo contrário, geralmente casos em que a entidade é de outro tipo básico, mas é empregue no contexto na acepção de `LOCAL`.

```
Certo: Como capturar da <LOCAL TIPO="VIRTUAL" MORF="F,S">Internet</LOCAL>...
Certo: uma ordem do governo local publicada na "<LOCAL TIPO="VIRTUAL"
       MORF="F,S">Gazeta de Macau</LOCAL>" ordenava...
Certo: E só depois da publicação no '<LOCAL TIPO="VIRTUAL" MORF="M,S">
       Diário da República</LOCAL>' é que tomou-se conhecimento do traçado.
```

17.1.4 Exemplos de não atribuição de `MORF` na categoria `TEMPO`

Nos tipos `PERIODO` e `DATA` há casos distintos em que são aplicados o atributo `MORF`.

As datas especificadas em termos de anos ou de dias não possuem nunca a etiqueta `MORF`.

```
Certo: Este ano de <TEMPO TIPO="PERIODO">1982</TEMPO> deve...
Certo: <TEMPO TIPO="PERIODO">1914-1918</TEMPO>...
Certo: ia ser a <TEMPO TIPO="DATA">17 de Dezembro</TEMPO> porque saiu...
Certo: Em <TEMPO|TEMPO TIPO="DATA|PERIODO">91</TEMPO>, foram angariados...
```

```
<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorense</ACONTECIMENTO> --->
[<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorense</ACONTECIMENTO>]
</ALT1>
<ALT2>
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Aves</PESSOA> --->
[<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorense</ACONTECIMENTO>]
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Campomaiorense</PESSOA> --->
[<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorense</ACONTECIMENTO>]
</ALT2>
</ALT>
```

Apresentamos agora um exemplo de alternativas para um alinhamento do tipo **nenhum para nenhum** ou do tipo **um para nenhum**, uma ou nenhuma EM na CD, para o caso em que na CD esteja `<ALT> Monárquico | <EM> Monárquico </EM> <ALT>` e a saída do sistema tenha sido `Monárquico`:

```
<ALT>
<ALT1>
</ALT1>
<ALT2>
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA> ---> [null]
</ALT2>
</ALT>
```

Finalmente, eis um exemplo de alternativas para um alinhamento do tipo **nenhum para um** ou do tipo **um para um**, uma ou nenhuma EM na CD, para o caso em que na CD esteja `<ALT> Monárquico | <EM> Monárquico </EM> <ALT>` e a saída do sistema tenha sido `<EM> Monárquico </EM>`:

```
<ALT>
<ALT1>
<ESPURIO>Monárquico</ESPURIO> --->
[<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA>]
</ALT1>
<ALT2>
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA> --->
[<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA>]
</ALT2>
</ALT>
```

Etiquetas <OMITIDO>

A etiqueta `<OMITIDO>` foi introduzida na versão 2.1 da CD de 2005, em plena avaliação do HAREM, por se ter achado necessário ignorar certos excertos de texto sem qualquer interesse do ponto de vista linguístico, sem interferir com a avaliação do HAREM. Assim, as