

# Índice

<b>Prefácio</b>	<b>i</b>
<b>Preface</b>	<b>iii</b>
<b>1 Breve introdução ao HAREM</b>	<b>1</b>
1.1 O modelo da avaliação conjunta . . . . .	2
1.2 Entidades mencionadas . . . . .	3
1.3 A terminologia que emergiu do HAREM . . . . .	4
1.4 Um pouco de história . . . . .	4
1.4.1 A inspiração . . . . .	5
1.4.2 Avaliação de REM em português antes do HAREM . . . . .	6
1.4.3 A preparação do Primeiro HAREM . . . . .	7
1.4.4 O primeiro evento do Primeiro HAREM . . . . .	8
1.4.5 O Mini-HAREM: medição do progresso e validação estatística . . . . .	10
1.5 Uma breve descrição da participação no Primeiro HAREM . . . . .	12
1.6 Mais informação sobre o HAREM: um pequeno guia . . . . .	13
1.6.1 Ensaio pré-HAREM . . . . .	13
1.6.2 Metodologia . . . . .	13
1.6.3 A colecção dourada . . . . .	14
1.6.4 Quantificação: Métricas, medidas, pontuações e regras de cálculo . . . . .	14
1.6.5 A arquitectura e os programas da plataforma de avaliação . . . . .	14
1.6.6 Validação estatística . . . . .	14
1.6.7 Resultados do HAREM . . . . .	15
1.6.8 Discussão e primeiro balanço . . . . .	15
1.7 O presente livro . . . . .	15

<b>I</b>	<b>17</b>
<b>2 Estudo preliminar para a avaliação de REM em português</b>	<b>19</b>
2.1 Descrição da Proposta . . . . .	21
2.2 Descrição dos textos . . . . .	23
2.3 Resultados . . . . .	26
2.3.1 Identificação de entidades . . . . .	28
2.3.2 Classificação de entidades . . . . .	30
2.3.3 Quadros comparativos entre pares de anotadores . . . . .	32
2.4 Comentários finais . . . . .	32
<b>3 MUC vs HAREM: a contrastive perspective</b>	<b>35</b>
3.1 An Overview of MUC . . . . .	36
3.2 Named Entity Recognition . . . . .	37
3.3 HAREM . . . . .	38
3.4 Evaluation . . . . .	40
3.5 Final Remarks . . . . .	40
<b>4 O modelo semântico usado no Primeiro HAREM</b>	<b>43</b>
4.1 O que é semântica? . . . . .	44
4.1.1 A importância da vagueza para a semântica . . . . .	45
4.2 O que é o REM? . . . . .	46
4.2.1 Metonímia . . . . .	46
4.2.2 REM como aplicação prática . . . . .	49
4.2.3 REM como classificação semântica tradicional . . . . .	50
4.3 O ACE como uma alternativa ao MUC: outras escolhas . . . . .	51
4.4 A abordagem do HAREM como processamento da linguagem natural em geral . . . . .	53
4.5 Alguma discussão em torno do modelo de REM do Primeiro HAREM . . . . .	55
4.6 Outros trabalhos . . . . .	55
4.7 Comentários finais . . . . .	56
<b>5 Validação estatística dos resultados do Primeiro HAREM</b>	<b>59</b>
5.1 Validação estatística para REM . . . . .	61
5.2 Teste de aleatorização parcial . . . . .	62
5.2.1 Metodologia . . . . .	63

	395
5.2.2	Aplicação ao HAREM . . . . . 64
5.3	Experiências com o tamanho da colecção . . . . . 67
5.3.1	Seleccção dos blocos . . . . . 68
5.3.2	Resultados da experiência . . . . . 68
5.4	Resultados . . . . . 69
5.4.1	Conclusões . . . . . 76
<b>6</b>	<b>O HAREM e a avaliação de sistemas para o reconhecimento de entidades geográficas em textos em língua portuguesa</b> . . . . . <b>79</b>
6.1	Conceitos e trabalhos relacionados . . . . . 80
6.2	Proposta para futuras edições do HAREM . . . . . 81
6.2.1	Classificação semântica refinada para as EM de categoria LOCAL . . . . . 82
6.2.2	Geração de anotações para ontologias geográficas padrão . . . . . 82
6.2.3	Possibilidade de considerar sub-anotações e anotações alternativas . . . . . 83
6.2.4	Desempenho computacional . . . . . 85
6.3	Conclusões . . . . . 86
<b>7</b>	<b>Balanco do Primeiro HAREM e futuro</b> . . . . . <b>87</b>
7.1	Uma retrospectiva das opções tomadas . . . . . 88
7.1.1	Uma dependência infeliz entre a classificação e a identificação . . . . . 88
7.1.2	Avaliação da identificação baseada em categorias de classificação . . . . . 89
7.1.3	Cenários relativos vistos por outra perspectiva . . . . . 90
7.1.4	Inconsistência nas medidas usadas . . . . . 90
7.1.5	Tratamento dos problemas incluídos em texto real . . . . . 91
7.2	Receitas para uma nova avaliação conjunta fundamentada . . . . . 91
7.3	Alguns futuros possíveis . . . . . 93
<b>II</b>	<b>95</b>
<b>8</b>	<b>O sistema CaGE no HAREM - reconhecimento de entidades geográficas em textos em língua portuguesa</b> . . . . . <b>97</b>
8.1	Conceitos e trabalhos relacionados . . . . . 99
8.2	Os recursos lexicais usados pelo sistema CaGE . . . . . 100
8.3	Reconhecimento e desambiguação de referências geográficas . . . . . 105

8.3.1	Operações de pré-processamento . . . . .	105
8.3.2	Identificação de referências geográficas . . . . .	106
8.3.3	Desambiguação de referências geográficas . . . . .	107
8.3.4	Geração de anotações para a ontologia . . . . .	108
8.4	Experiências de avaliação no Mini-HAREM . . . . .	109
8.5	Conclusões . . . . .	111
<b>9</b>	<b>O Cortex e a sua participação no HAREM</b>	<b>113</b>
9.1	Filosofia . . . . .	114
9.2	Classificação de entidades mencionadas no Cortex . . . . .	115
9.3	A participação do Cortex no HAREM . . . . .	118
9.4	A participação do Cortex no Mini-HAREM . . . . .	119
9.5	Cortex 3.0 . . . . .	122
9.6	Conclusões . . . . .	122
<b>10</b>	<b>MALINCHE: A NER system for Portuguese that reuses knowledge from Spanish</b>	<b>123</b>
10.1	The MALINCHE System . . . . .	124
10.1.1	Named Entity Delimitation . . . . .	125
10.1.2	The features . . . . .	127
10.1.3	Named Entity Classification . . . . .	128
10.1.4	The machine learning algorithm . . . . .	129
10.2	Named Entity Recognition in Portuguese . . . . .	131
10.2.1	Results on NED . . . . .	132
10.2.2	Results on NEC in Portuguese . . . . .	132
10.3	Final remarks . . . . .	135
<b>11</b>	<b>Tackling HAREM's Portuguese Named Entity Recognition task with Spanish resources</b>	<b>137</b>
11.1	System Description . . . . .	138
11.1.1	Feature sets . . . . .	140
11.2	Experiments and discussion . . . . .	142
11.3	Conclusions . . . . .	144

<b>12 Functional aspects on Portuguese NER</b>	<b>145</b>
12.1 Recognizing MWE name chains . . . . .	146
12.2 Semantic typing of name tokens: Lexematic versus functional NE categories . . .	149
12.2.1 Micromapping: Name type rules based on name parts and patterns . . . .	151
12.2.2 Macromapping: Name type rules based on syntactic propagation . . . . .	151
12.3 Evaluation . . . . .	152
12.4 Conclusion: Comparison with other systems . . . . .	154
<b>13 RENA - reconhecedor de entidades</b>	<b>157</b>
13.1 Descrição do RENA . . . . .	159
13.1.1 Estrutura interna do RENA . . . . .	159
13.1.2 Ficheiros de configuração . . . . .	161
13.2 Participação no HAREM . . . . .	166
13.3 Subsídio para a discussão sobre futuras edições . . . . .	167
13.3.1 Uso de documentos seguindo XML . . . . .	167
13.3.2 Uso claro e expansível de metadados nas colecções . . . . .	168
13.3.3 Questões ligadas à estrutura classificativa usada . . . . .	168
13.3.4 Sugestão para futuras edições . . . . .	172
13.4 Conclusões e trabalho futuro . . . . .	172
<b>14 O SIEMÊS e a sua participação no HAREM e no Mini-HAREM</b>	<b>173</b>
14.1 A participação no HAREM . . . . .	175
14.2 A segunda versão do SIEMÊS . . . . .	177
14.2.1 Bloco de regras “simples” . . . . .	179
14.2.2 Bloco de pesquisa directa no REPENTINO . . . . .	179
14.2.3 Bloco de emparelhamento de prefixo sobre o REPENTINO . . . . .	179
14.2.4 Bloco de semelhança sobre o REPENTINO . . . . .	180
14.2.5 Bloco posterior de recurso . . . . .	182
14.3 A participação no Mini-HAREM . . . . .	182
14.3.1 A decomposição da avaliação . . . . .	183
14.3.2 Resultados globais . . . . .	185
14.3.3 Os melhores componentes por categoria . . . . .	186
14.3.4 Alguns comentários . . . . .	187
14.4 Conclusões . . . . .	188

<b>15 Em busca da máxima precisão sem almanaques: O Stencil/NooJ no HAREM</b>	<b>191</b>
15.1 O que é o NooJ? . . . . .	194
15.1.1 Características dos recursos . . . . .	195
15.1.2 Processamento linguístico de textos . . . . .	196
15.2 O que é o Stencil? . . . . .	196
15.2.1 Organização dos recursos e forma de aplicação . . . . .	197
15.2.2 Utilização de regras precisas . . . . .	198
15.2.3 Utilização de regras combinatórias . . . . .	200
15.2.4 Consulta simples dos dicionários de nomes próprios extraídos . . . . .	201
15.3 Participação no HAREM . . . . .	202
15.3.1 HAREM vs. Mini-HAREM . . . . .	203
15.3.2 Resultados . . . . .	204
15.3.3 Problemas e dificuldades . . . . .	207
15.4 Comentários finais . . . . .	208
<b>III</b>	<b>209</b>
<b>16 Directivas para a identificação e classificação semântica na colecção dourada do HAREM</b>	<b>211</b>
16.1 Regras gerais de etiquetagem . . . . .	212
16.1.1 Recursividade das etiquetas . . . . .	213
16.1.2 Vagueza na classificação semântica . . . . .	213
16.1.3 Vagueza na identificação . . . . .	213
16.1.4 Critérios de identificação de uma EM . . . . .	214
16.1.5 Relação entre a classificação e a identificação . . . . .	215
16.1.6 Escolha da EM máxima . . . . .	216
16.2 Categoria PESSOA . . . . .	216
16.2.1 Tipo INDIVIDUAL . . . . .	216
16.2.2 Tipo GRUPOIND . . . . .	217
16.2.3 Tipo CARGO . . . . .	218
16.2.4 Tipo GRUPOCARGO . . . . .	218
16.2.5 Tipo MEMBRO . . . . .	219
16.2.6 Tipo GRUPOMEMBRO . . . . .	219

16.3	Categoria ORGANIZACAO . . . . .	220
16.3.1	Tipo ADMINISTRACAO . . . . .	220
16.3.2	Tipo EMPRESA . . . . .	221
16.3.3	Tipo INSTITUICAO . . . . .	221
16.3.4	Tipo SUB . . . . .	221
16.4	Categoria TEMPO . . . . .	223
16.4.1	Tipo DATA . . . . .	223
16.4.2	Tipo HORA . . . . .	224
16.4.3	Tipo PERIODO . . . . .	224
16.4.4	Tipo CICLICO . . . . .	225
16.5	Categoria ACONTECIMENTO . . . . .	225
16.5.1	Tipo EFEMERIDE . . . . .	226
16.5.2	Tipo ORGANIZADO . . . . .	226
16.5.3	Tipo EVENTO . . . . .	226
16.6	Categoria COISA . . . . .	227
16.6.1	Tipo OBJECTO . . . . .	227
16.6.2	Tipo SUBSTANCIA . . . . .	227
16.6.3	Tipo CLASSE . . . . .	227
16.6.4	Tipo MEMBROCLASSE . . . . .	228
16.7	Categoria LOCAL . . . . .	228
16.7.1	Tipo CORREIO . . . . .	229
16.7.2	Tipo ADMINISTRATIVO . . . . .	229
16.7.3	Tipo GEOGRAFICO . . . . .	230
16.7.4	Tipo VIRTUAL . . . . .	230
16.7.5	Tipo ALARGADO . . . . .	231
16.8	Categoria OBRA . . . . .	232
16.8.1	Tipo REPRODUZIDA . . . . .	232
16.8.2	Tipo ARTE . . . . .	232
16.8.3	Tipo PUBLICACAO . . . . .	233
16.9	Categoria ABSTRACAO . . . . .	233
16.9.1	Tipo DISCIPLINA . . . . .	234
16.9.2	Tipo ESTADO . . . . .	234
16.9.3	Tipo ESCOLA . . . . .	234

16.9.4	Tipo MARCA . . . . .	234
16.9.5	Tipo PLANO . . . . .	235
16.9.6	Tipo IDEIA . . . . .	235
16.9.7	Tipo NOME . . . . .	236
16.9.8	Tipo OBRA . . . . .	236
16.10	Categoria VALOR . . . . .	236
16.10.1	Tipo CLASSIFICACAO . . . . .	236
16.10.2	Tipo MOEDA . . . . .	237
16.10.3	Tipo QUANTIDADE . . . . .	238
16.11	Categoria VARIADO . . . . .	238
<b>17</b>	<b>Directivas para a identificação e classificação morfológica na colecção dou- rada do HAREM</b>	<b>239</b>
17.1	Regras gerais da tarefa de classificação morfológica . . . . .	240
17.1.1	Género (morfológico) . . . . .	241
17.1.2	Número . . . . .	241
17.1.3	Exemplos de não atribuição de MORF na categoria LOCAL . . . . .	241
17.1.4	Exemplos de não atribuição de MORF na categoria TEMPO . . . . .	241
17.2	Regras de atribuição de classificação morfológica . . . . .	242
17.2.1	Exemplos na categoria LOCAL . . . . .	242
17.2.2	Exemplos na categoria ORGANIZACAO . . . . .	243
17.2.3	Exemplos na categoria PESSOA . . . . .	243
17.2.4	Exemplos na categoria ACONTECIMENTO . . . . .	244
17.2.5	Exemplos na categoria ABSTRACCAO . . . . .	244
<b>18</b>	<b>Avaliação no HAREM: métodos e medidas</b>	<b>245</b>
18.1	Terminologia . . . . .	246
18.1.1	Pontuações . . . . .	246
18.1.2	Medidas . . . . .	246
18.1.3	Métricas . . . . .	246
18.1.4	Cenários de avaliação . . . . .	247
18.2	Tarefa de identificação . . . . .	248
18.2.1	Pontuações . . . . .	249
18.2.2	Métricas . . . . .	249



	401
18.2.3 Exemplo detalhado de atribuição de pontuação . . . . .	250
18.2.4 Identificações alternativas . . . . .	251
18.3 Tarefa de classificação semântica . . . . .	257
18.3.1 Medidas . . . . .	257
18.3.2 Pontuações . . . . .	257
18.3.3 Métricas . . . . .	260
18.3.4 Exemplo detalhado de atribuição de pontuação . . . . .	265
18.4 Tarefa de classificação morfológica . . . . .	271
18.4.1 Medidas . . . . .	271
18.4.2 Pontuações . . . . .	271
18.4.3 Métricas . . . . .	273
18.5 Apresentação dos resultados . . . . .	277
18.5.1 Resultados globais . . . . .	277
18.5.2 Resultados individuais . . . . .	279
<b>19 A arquitectura dos programas de avaliação do HAREM</b>	<b>283</b>
19.1 Sinopse da arquitectura . . . . .	284
19.2 Descrição pormenorizada de cada módulo . . . . .	286
19.2.1 Validador . . . . .	286
19.2.2 Extractor . . . . .	288
19.2.3 AlinhEM . . . . .	288
19.2.4 AvalIDa . . . . .	294
19.2.5 Véus . . . . .	295
19.2.6 ALTinaID . . . . .	296
19.2.7 Ida2ID . . . . .	296
19.2.8 Emir . . . . .	299
19.2.9 AltinaSEM . . . . .	301
19.2.10 Ida2SEM . . . . .	301
19.2.11 Vizir . . . . .	303
19.2.12 AltinaMOR . . . . .	304
19.2.13 Ida2MOR . . . . .	304
19.2.14 Sultão . . . . .	305
19.2.15 Alcaide . . . . .	305
19.3 Comentários finais . . . . .	306

<b>20 Disponibilizando a CD do HAREM pelo AC/DC</b>	<b>307</b>
20.1 O projecto AC/DC	308
20.1.1 A criação de um corpus novo no AC/DC	309
20.1.2 IMS-CWB, o sistema subjacente	309
20.2 Disponibilizando a CD do HAREM como corpus	310
20.2.1 Opções gerais de codificação	311
20.2.2 O atributo EM	311
20.2.3 Atributos relativos às categorias e tipos das EM	313
20.2.4 O atributo prem para compatibilizar contagens por palavras e por EM	314
20.2.5 Atributos relativos ao texto	315
20.2.6 Atributos relativos à classificação morfológica	316
20.2.7 Atributos relativos à anotação sintáctica do AC/DC	316
20.3 Vagueza	317
20.3.1 Vagueza na classificação (categorias ou tipos com )	317
20.3.2 Vagueza na identificação: as etiquetas <ALT>	318
20.4 Dados quantitativos	319
20.5 Observações finais	325
<b>A Resultados do Primeiro HAREM</b>	<b>329</b>
<b>B Lista de entidades classificadas no ensaio pré-HAREM</b>	<b>337</b>
<b>C Tabelas de valores <math>p</math></b>	<b>349</b>
<b>D Documentação técnica da plataforma de avaliação</b>	<b>355</b>
D.1 Instalação e configuração	355
D.2 Utilização	356
D.2.1 Extractor	356
D.2.2 AlinhEM	357
D.2.3 AvalIDa	357
D.2.4 Véus	358
D.2.5 AltinaID	359
D.2.6 Ida2ID	359
D.2.7 Emir	360
D.2.8 AltinaSEM	360

	403
D.2.9 Ida2SEM . . . . .	360
D.2.10 Vizir . . . . .	361
D.2.11 AltinaMOR . . . . .	361
D.2.12 Ida2MOR . . . . .	361
D.2.13 Sultão . . . . .	361
D.2.14 Alcaide . . . . .	363
D.3 Ficheiro de configuração do HAREM, harem.conf . . . . .	364
<b>E Exemplos da invocação dos programas de avaliação</b>	<b>365</b>
E.1 Exemplos do Emir . . . . .	365
E.2 Exemplos do Vizir . . . . .	367
<b>Referências</b>	<b>369</b>
<b>Índice</b>	<b>393</b>