

Conteúdo

Prefácio	i
Lista de autores	iii
Glossário	v
Enquadramento e historial do Segundo HAREM	1
Historial detalhado	5
I O HAREM pela organização	9
1 Segundo HAREM: Modelo geral, novidades e avaliação	11
1.1 Filosofia do HAREM	12
1.2 Esquema de anotação no Segundo HAREM	15
1.2.1 Sintaxe das anotações	15
1.2.2 Classificação das EM	16
1.3 Melhorias no Segundo HAREM	17
1.3.1 Delimitação e classificação das EM	18
1.3.2 Representação sistemática das análises alternativas	18
1.4 Recursos	19
1.4.1 Constituição das colecções do Segundo HAREM	20
1.4.2 Processo de anotação da CD	20
1.5 Resultados da avaliação	24
1.5.1 Sistemas participantes	25
1.5.2 Resultados	26

2	Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e o futuro	33
2.1	Introdução	34
2.1.1	Generalidades	34
2.1.2	Motivação da proposta	35
2.1.3	Questões operacionais da proposta	36
2.2	Proposta para o Segundo HAREM	37
2.2.1	Delimitação das ET	37
2.2.2	Delimitação das ET complexas	38
2.2.3	TEMPO_CALEND	40
2.2.3.1	Data	40
2.2.3.2	Hora	42
2.2.3.3	Intervalo	42
2.2.3.4	Duração	43
2.2.3.5	Frequência	44
2.2.3.6	ET genéricas	44
2.3	Normalização	45
2.3.1	Normalização de datas referenciais	45
2.3.2	Normalização da DURACAO	46
2.4	A experiência do Segundo HAREM	47
2.5	Próximos passos e perspectivas futuras	47
2.5.1	TEMPO_CALEND	49
2.5.2	Novo subtipo=DATA	49
2.5.2.1	subtipo=INTERVALO	50
2.5.2.2	Novo subtipo=COMPLEXO	51
2.5.3	DURACAO	51
2.5.3.1	tipo=DURACAO subtipo=INTERVALO	51
2.5.4	FREQUENCIA	51
2.5.5	Outras sugestões	53
2.5.5.1	Not_Norm	53
2.5.5.2	Indefinição (ou vagueza)	53
2.6	Conclusões	53
3	É tempo de avaliar o TEMPO	55

3.1	Anotação da colecção dourada	56
3.1.1	Opções relativas aos atributos do HAREM clássico de TEMPO	57
3.1.1.1	Delimitação da entidade quando a expressão temporal verifica os critérios 1 e 2-6	57
3.1.1.2	Delimitação da entidade quando a expressão temporal é constituída por DATA e HORA	58
3.1.1.3	Classificação como GENERICO	59
3.1.1.4	Classificação como DURACAO	60
3.1.1.5	Classificação de expressões iniciadas por <i>há</i>	61
3.1.1.6	Ausência de anotação relativa a TEMPO	62
3.1.2	Opções relativas aos atributos do TEMPO estendido	63
3.1.2.1	Tensão entre dois tipos de DATA	63
3.1.2.2	Expressões com valor de data sem nenhum dos campos ANO-MES-DIA especificado	64
3.1.2.3	Preenchimento de VAL_DELTA e VAL_NORM na ausência total de informação	64
3.2	O TEMPO em números no Segundo HAREM	65
3.3	Avaliação	68
3.3.1	Sistemas participantes	70
3.3.2	Resultados	70
3.4	Sugestões para o futuro da avaliação do TEMPO	73
3.4.1	Medida de avaliação	73
3.4.2	Estudos empíricos ilumináveis pela LÂMPADA	74
3.4.3	Opiniões diferentes sobre o REM temporal, ou melhor, sobre o RET	74
4	Relações semânticas do ReReEM: além das entidades no Segundo HAREM	77
4.1	Relações do ReReEM: o que anotar	79
4.1.1	Identidade	79
4.1.2	Relação de inclusão	80
4.1.3	Relação de localização, ou de ocorrência em	81
4.1.4	Relação outra e outras relações	82
4.2	Relações do ReReEM: como anotar	84
4.2.1	Relações múltiplas entre EM	85
4.2.2	ReReEM e análises alternativas (ALT)	85

4.2.3	ReReLEM e a vagueza do HAREM	86
4.2.4	Simetria, inversão e transitividade	86
4.3	A coleção dourada do ReReLEM	88
4.4	Avaliação	91
4.4.1	Processo de avaliação	92
4.4.2	Sistemas participantes	92
4.4.3	Resultados	93
4.5	Considerações finais	95
5	Avaliação à medida no Segundo HAREM	97
5.1	Avaliação do HAREM clássico	98
5.1.1	Pontuações	98
5.1.2	Uma única medida	98
5.1.3	Cenários selectivos	100
5.1.4	Avaliação de ALT	102
5.2	Avaliação da pista do TEMPO	102
5.3	Avaliação do ReReLEM	104
5.3.1	Pontuações e medidas	104
5.3.2	Expansão de relações	104
5.3.3	Seleccção de alinhamentos	105
5.4	Métricas	105
5.5	Vista geral da arquitectura	105
5.5.1	Formato das colecções	106
5.5.2	Os módulos	106
5.6	Módulos de avaliação do HAREM clássico	108
5.6.1	Alinhador	108
5.6.1.1	Formato da saída	108
5.6.1.2	Etiquetas ALT	109
5.6.1.3	Etiquetas OMITIDO	109
5.6.2	Avaliador de alinhamentos	109
5.6.3	Véus	110
5.6.3.1	Representação dos cenários selectivos	110
5.6.3.2	Formato da saída	110
5.6.3.3	Exemplo de aplicação de filtros pelo Véus	111

5.6.4	Organizador de ALT	111
5.6.5	Listador de espúrios	112
5.6.6	Avaliador da classificação	112
5.6.6.1	Formato da saída	112
5.6.7	Seleccionador de ALT	115
5.6.8	Resumidor das classificações	115
5.6.9	Gerador de resultados	115
5.6.10	Gerador de relatórios individuais	115
5.7	Módulo de avaliação da pista do TEMPO	116
5.8	Módulos de avaliação do ReReLEM	118
5.8.1	Conversão de notação	118
5.8.2	Expandidor de relações	121
5.8.2.1	Expansão	121
5.8.2.2	Compatibilidade de facetas	123
5.8.3	Seleccionador de alinhamentos	123
5.8.4	Normalizador de identificadores (ID)	123
5.8.5	Tradutor de alinhamentos para triplas	123
5.8.6	Véus para o ReReLEM	126
5.8.7	Avaliador de relações	126
5.8.7.1	Resumidor das classificações do ReReLEM	126
5.8.8	Gerador de resultados do ReReLEM	126
5.8.9	Visualizador de relações	126
5.9	Observações finais	128
6	Segundo HAREM: Balanço e perspectivas de futuro	131
6.1	HAREM clássico: balanço geral	132
6.1.1	Identificação vs. classificação	133
6.1.2	Delimitação das entidades mencionadas	133
6.1.3	Modelos de avaliação conjunta incongruentes entre si	134
6.1.4	Novo formato XML	135
6.1.5	Progresso na definição da tarefa e nos desafios	136
6.1.6	Recursos mais ricos, mais bem revistos e documentados	137
6.1.7	Cenários selectivos melhor aproveitados	137
6.1.8	Potencialidades de investigação do valor do REM noutras áreas	138

6.1.9	Ferramentas para auxiliar o HAREM	139
6.2	Pista do TEMPO: algumas observações	139
6.3	ReReLEM: primeiro balanço	140
6.3.1	A expansão das participações	140
6.3.2	Relação com a vagueza	141
6.3.3	O que fazer aos ALT?	141
6.3.4	O que fazer a participações inconsistentes?	141
6.3.5	Que sentido faz a comparação?	142
6.3.6	A identidade é diferente?	142
6.3.7	Progresso na área da semântica computacional	143
6.4	O HAREM tem futuro?	144
II O HAREM pelos participantes		147
7	O sistema CaGE no Segundo HAREM	149
7.1	Descrição do sistema	151
7.1.1	Os dicionários e o almanaque usados pelo sistema CaGE	151
7.1.2	Funcionamento geral do sistema	152
7.1.3	Aplicações práticas do sistema CaGE	155
7.2	Experiências no HAREM e análise dos resultados	156
7.3	Conclusões	158
8	PorTexTO: sistema de anotação/extracção de expressões temporais	159
8.1	Descrição do sistema	161
8.1.1	Módulo Anotador	161
8.1.2	Módulo Processador de co-ocorrências	163
8.2	Participação no Segundo HAREM	165
8.3	Resultados da participação no Segundo HAREM	166
8.4	Conclusões e trabalho futuro	168
9	Adaptação do sistema de reconhecimento de entidades mencionadas da Pri-beram ao HAREM	171
9.1	Descrição do sistema	172
9.1.1	Adaptação do sistema ao Segundo HAREM	174

9.2	Análise dos resultados da participação no Segundo HAREM	175
9.2.1	Resultados do HAREM clássico	175
9.2.2	Resultados da pista do TEMPO	177
9.3	Conclusões e trabalho futuro	178
10	R3M, uma participação minimalista no Segundo HAREM	181
10.1	Descrição do sistema R3M	182
10.1.1	Identificação	184
10.1.1.1	Detecção de candidatos a EM	184
10.1.1.2	Detecção do contexto da EM	185
10.1.2	Extracção de características	187
10.1.3	Classificação	187
10.1.4	Co-treino	188
10.1.5	Propagação	189
10.2	Resultados	190
10.3	Comentários finais	192
11	REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em	
	Relações e ANálise Detalhada do Texto	195
11.1	Inspiração para o REMBRANDT	196
11.2	Anatomia do REMBRANDT	197
11.3	SASKIA	199
11.3.1	Pré-processamento da Wikipédia	199
11.3.2	Estratégia de classificação	200
11.4	Regras gramaticais	202
11.4.1	Propriedades das cláusulas	203
11.4.2	Aplicação das regras	204
11.4.3	Tribunal de EM	205
11.5	Detecção de relações entre EM	206
11.6	Resultados no Segundo HAREM	207
11.6.1	Corridas	208
11.6.2	Resultados na tarefa de REM	208
11.6.3	Resultados na tarefa de DRE	209
11.7	Conclusões e trabalho futuro	210

12 REMMA - Reconhecimento de Entidades Mencionadas do MedAlert	213
12.1 A Wikipédia como fonte de conhecimento para REM	215
12.1.1 Estrutura básica	216
12.1.2 Redirecção	216
12.1.3 Páginas de desambiguação	216
12.1.4 Categorias	217
12.1.5 Ligações internas	217
12.2 O sistema REMMA	218
12.2.1 A plataforma base - UIMA	218
12.2.2 A arquitectura	218
12.2.2.1 Classificação com base em regras e almanaques	220
12.2.2.2 Classificação com recurso à Wikipédia	221
12.2.2.3 Anotadores VALOR, TEMPO e Minúsculas	223
12.3 Resultados no Segundo HAREM	224
12.3.1 Usar a Wikipédia tem potencial para melhor desempenho?	224
12.3.2 Para a Wikipédia todas as categorias nascem iguais?	224
12.3.3 Esta abordagem é competitiva?	225
12.3.4 Comparação com o REMBRANDT	226
12.4 Discussão	227
12.5 Conclusão e trabalho futuro	228
13 Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM	231
13.1 Trabalhos relacionados	232
13.2 O SEI-Geo	233
13.2.1 Geo-ontologias utilizadas pelo SEI-Geo	235
13.2.2 Algoritmos de identificação e classificação de locais	236
13.2.3 Reconhecimento de relações semânticas entre EM – ReReLEM	239
13.3 Descrição das corridas	239
13.4 Análise dos resultados	240
13.5 Discussão	243
13.6 Conclusões	244

14 Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas	247
14.1 Trabalhos relacionados e motivação	248
14.2 SeRELeP: Sistema de reconhecimento de RElações em textos de Língua Portuguesa	249
14.2.1 Visão geral	249
14.2.2 Reconhecimento de relações entre entidades mencionadas	254
14.2.3 Resultados	256
14.3 Considerações finais	258
15 Reconhecimento de entidades mencionadas com o XIP: Uma colaboração entre a Xerox e o L2F do INESC-ID Lisboa	261
15.1 XIP: Uma ferramenta para o processamento lexical, sintáctico e semântico	262
15.1.1 Ilustração	262
15.1.2 Desenvolvimento do módulo de REM	264
15.1.2.1 Integração do REM no processamento geral do português	264
15.1.2.2 Tratamento incremental da informação linguística	264
15.2 Léxico e pré-processamento	264
15.2.1 O que é uma entrada lexical no XIP?	264
15.2.2 Dois tipos de léxicos	265
15.2.2.1 Léxico pré-existente	265
15.2.2.2 Léxico definido no XIP	266
15.2.3 Adaptação do pré-processamento	266
15.3 Gramáticas locais para o REM	267
15.3.1 Expressão de gramáticas locais em XIP	267
15.3.2 Delimitação de EM complexas	268
15.3.3 Utilização de contexto imediato	268
15.4 Últimas fases de processamento das EM	269
15.4.1 Particionamento	269
15.4.2 Dependências	269
15.4.3 Generalizando o contexto para classificar EM	270
15.4.4 Propagação	272
15.5 Resultados e perspectivas	274

Apêndices	277
A Segundo HAREM: Directivas de anotação	277
A.1 Motivação para as presentes directivas	278
A.2 Questões de delimitação	278
A.2.1 Desaparecimento de entidades complexas	279
A.2.2 Tratamento mais convencional de expressões com várias palavras	279
A.2.3 Introdução de intervalos de valores como EM	280
A.3 Mudanças por categoria	280
A.3.1 VALOR	280
A.3.2 VARIADO	280
A.3.3 PESSOA	280
A.3.4 ORGANIZACAO	280
A.3.5 LOCAL	281
A.3.6 ACONTECIMENTO	282
A.3.7 OBRA	282
A.3.8 ABSTRACCAO	283
A.3.9 COISA	283
A.4 Elenco de categorias do Segundo HAREM	284
A.5 Segundo HAREM: sintaxe	284
A.6 Lista de minúsculas	285
B Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o Segundo HAREM	289
B.1 Preâmbulo	290
B.2 Motivação da proposta	290
B.3 Proposta	291
B.3.1 Categoria TEMPO	291
B.3.1.1 Definição da entidade de tipo TEMPO	291
B.3.2 TIPO = “TEMPO_CALEND”	296
B.3.2.1 SUBTIPO = “DATA”	297
B.3.2.2 Expressões de datas relativas: dois tipos de referências considerados	298
B.3.2.3 Atributo TEMPO_REF	299

B.3.2.4	Atributos SENTIDO e VAL_DELTA	300
B.3.2.5	SUBTIPO = "HORA"	301
B.3.2.6	SUBTIPO = "INTERVALO"	301
B.3.3	TIPO = "DURACAO"	302
B.3.4	TIPO = "FREQUENCIA"	302
B.3.5	TIPO = "GENERICICO"	302
B.3.6	Atributo VAL_NORM	303
B.3.6.1	Atributo VAL_NORM para expressões de subtipo DATA absoluta	303
B.3.6.2	Atributo VAL_NORM para expressões de tipo HORA	304
B.3.6.3	Atributo VAL_NORM para expressões de tipo DURACAO	304
B.4	Resumo das principais modificações	305
B.5	Alguns exemplos de anotação	305
B.6	Adenda	307
C	ReReLEM - Reconhecimento de Relações entre Entidades Mencionadas. Segundo HAREM: proposta de nova pista	309
C.1	Directivas para anotação das relações entre EM	310
C.1.1	Regras gerais de integração da pista no HAREM	311
C.1.2	Relações múltiplas de uma dada EM	311
C.1.3	Equivalência entre relações	311
C.1.4	Opcionalidade de marcação de TIPOREL no caso de identidade	312
C.2	Tipos de relações a marcar	312
C.2.1	Relação de identidade	312
C.2.2	Relação de inclusão	313
C.2.3	Relação de localização, ou de ocorrência em	314
C.2.4	Outras relações	315
C.2.5	Relações entre EM vagas	316
C.2.6	Quadro-resumo das categorias por tipo de relações a marcar	317
D	CrITÉrios de ALT no Segundo HAREM	319
E	Exemplário do Segundo HAREM	323
E.1	PESSOA	324
E.1.1	INDIVIDUAL	324
E.1.2	CARGO	324

E.1.3	GRUPOCARGO	325
E.1.4	GRUPOMEMBRO	325
E.1.5	MEMBRO	326
E.1.6	GRUPOIND	326
E.1.7	POVO	326
E.2	ABSTRACCAO	327
E.2.1	DISCIPLINA	327
E.2.2	ESTADO	327
E.2.3	IDEIA	327
E.2.4	NOME	328
E.3	ACONTECIMENTO	328
E.3.1	EFEMERIDE	328
E.3.2	ORGANIZADO	328
E.3.3	EVENTO	329
E.4	COISA	329
E.4.1	CLASSE	329
E.4.2	MEMBROCLASSE	329
E.4.3	OBJECTO	330
E.4.4	SUBSTANCIA	330
E.5	LOCAL	330
E.5.1	HUMANO	330
E.5.1.1	PAIS	330
E.5.1.2	DIVISAO	330
E.5.1.3	REGIAO	331
E.5.1.4	CONSTRUCAO	331
E.5.1.5	RUA	331
E.5.1.6	OUTRO	332
E.5.2	FISICO	332
E.5.2.1	AGUAMASSA	332
E.5.2.2	AGUACURSO	332
E.5.2.3	RELEVO	333
E.5.2.4	PLANETA	333
E.5.2.5	REGIAO	333

E.6	VIRTUAL	334
	E.6.0.6 COMSOCIAL	334
	E.6.0.7 SITIO	334
	E.6.0.8 OBRA	334
E.7	OBRA	334
	E.7.1 ARTE	334
	E.7.2 PLANO	335
	E.7.3 REPRODUZIDA	335
E.8	ORGANIZACAO	335
	E.8.1 ADMINISTRACAO	335
	E.8.2 EMPRESA	336
	E.8.3 INSTITUICAO	336
E.9	VALOR	336
	E.9.1 CLASSIFICACAO	336
	E.9.2 MOEDA	336
	E.9.3 QUANTIDADE	337
E.10	Exemplos de vagueza	337
F	Manual do Etiquet(H)AREM	339
	F.1 Requisitos básicos na utilização do programa	340
	F.2 Lista de notações a utilizar	340
	F.3 Manuseamento do programa propriamente dito	340
	F.4 Menus do Etiquet(H)arem	342
G	SAHARA - Serviço de Avaliação HAREM Automático	347
	G.1 Primeiro passo: validação	348
	G.2 Configuração da avaliação	348
	G.2.1 Selecção das pistas	348
	G.2.2 Escolha dos cenários	350
	G.2.3 Selecção do modo de avaliação	350
	G.2.4 Escolha da colecção dourada	351
	G.3 Apresentação dos resultados	351
H	Apresentação detalhada das colecções do Segundo HAREM	355

I	Resumo de resultados do Segundo HAREM	379
I.1	Resultados do HAREM clássico	380
I.1.1	Avaliação estrita de ALT	380
I.1.2	Avaliação relaxada de ALT	392
I.2	Resultados da pista do TEMPO	394
I.2.1	HAREM clássico na CD do TEMPO	394
I.2.2	TEMPO completo	395
I.2.3	TEMPO sem normalização	396
I.2.4	TEMPO só normalização	397
I.3	Resultados do ReReLEM	398
I.3.1	HAREM clássico na CD do ReReLEM	398
I.3.2	ReReLEM no cenário total	400
I.3.3	ReReLEM no cenário 5	402
	Bibliografia	405
	Conteúdo	423