Capítulo 1

Breve introdução ao HAREM

Diana Santos e Nuno Cardoso

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 1, p. 1–16, 2007

Este capítulo apresenta o HAREM, tentando constituir algo interessante para leitores sem conhecimento prévio da área, passando por pessoas interessadas e conhecedoras do paradigma de avaliação conjunta, até aos próprios participantes no HAREM. Apresentamos a motivação para a realização do HAREM e consequente publicação deste volume, ao nível da necessidade de avaliação na área do processamento computacional da língua portuguesa em geral, e as razões que motivaram a escolha da área específica do reconhecimento das entidades mencionadas.

Prosseguimos com uma breve descrição sobre o evento que inspirou o HAREM, o MUC, assim como toda a história da organização do HAREM.

Depois de esclarecermos a terminologia e fixarmos as designações **HAREM**, **Primeiro HAREM** e **Mini-HAREM**, descrevemos o Primeiro HAREM em detalhe.

Essa descrição abarca, com o respectivo calendário:

- O trabalho preparatório;
- A criação dos recursos de avaliação;
- A organização da primeira avaliação;
- A organização do Mini-HAREM.

Produzimos depois um pequeno guia sobre onde encontrar mais documentação sobre o HAREM, fazendo uma espécie de inventário das publicações associadas, e terminamos o capítulo com uma pequena apresentação do presente livro, que marca a última contribuição do Primeiro HAREM.

1.1 O modelo da avaliação conjunta

Há poucos anos atrás, o processamento do português estava numa fase pré-científica, em que os (poucos) trabalhos publicados relatavam no máximo a sua própria auto-avaliação. Isso impedia, na prática, a reprodução dos resultados, inibindo o progresso na área e impedindo a formação de uma verdadeira comunidade científica que pudesse comparar abordagens e métodos aplicados a uma tarefa comum.

Essa situação foi identificada como um dos principais entraves ao progresso do processamento computacional da nossa língua em Santos (1999), e tem vindo a ser progressivamente modificada através da actuação da Linguateca nesse campo (Santos, 2007a).

A Linguateca possui três eixos de actuação: a informação, os recursos e a avaliação.[1] Nesta última vertente, promovemos desde o início o modelo da avaliação conjunta, tendo

[1] Para uma panorâmica da Linguateca através dos tempos veja-se entre outros Santos (2000, 2002); Santos et al. (2004); Santos e Costa (2005); Santos (2006c), assim como a lista de publicações constantemente actualizada no sítio da Linguateca.

organizado as Morfolimpíadas em 2002-2003 (Santos et al., 2003; Costa et al., 2007) e participando anualmente na organização do CLEF para o português desde 2004 (Rocha e Santos, 2007). Em 2005 iniciámos a organização do HAREM, a que se refere o presente volume e capítulo.

Ao possibilitar a comparação de diferentes abordagens de uma forma justa e imparcial, estas avaliações conjuntas fomentam o desenvolvimento de melhores sistemas e contribuem para a melhoria do desempenho destes. Além disso, permitem definir em conjunto uma área e avaliar e comparar tecnologias diferentes, além de fixarem e tornarem público um conjunto de recursos para avaliar e treinar sistemas no futuro. Para uma defesa alongada deste paradigma, veja-se Santos (2007b).

1.2 Entidades mencionadas

"Entidades mencionadas" (EM) foi a nossa tradução (ou melhor, adaptação) do conceito usado em inglês, *named entities*, e que literalmente poderá ser traduzido para "entidades com nome próprio".

A tarefa que nos propusemos avaliar era a de reconhecer essas entidades, atribuindo-lhes uma classificação (dentre um leque de categorias previamente definido e aprovado por todos) que representaria o significado daquela ocorrência específica da entidade no texto em questão.

Nós vemos o reconhecimento de entidades mencionadas (REM) como um primeiro passo na análise semântica de um texto. Separámos esse reconhecimento em duas subtarefas separadas: a **identificação** (de que uma dada sequência de palavras constitui uma EM) e a **classificação** (a que categoria semântica essa EM pertence, naquele contexto).

A razão para abordarmos esta tarefa foi a nossa convicção de que o REM é parte integrante da maioria dos sistemas inteligentes que processam e interpretam a língua, tais como sistemas de extracção de informação, de resposta automática a perguntas, de tradução automática, ou de sumarização de textos. Visto que a qualidade do REM nestes sistemas influencia decisivamente o seu resultado final, estamos convencidos de que a organização de avaliações específicas sobre REM pode beneficiar fortemente o progresso nestas tarefas.

A tarefa de REM necessita de uma clarificação das bases semânticas e pragmáticas do processamento de linguagem natural que não são necessariamente consensuais ou explícitas, pelo que a delimitação precisa do conceito de entidade mencionada e da sua operacionalização prática veio fazer correr muita tinta. O capítulo 4 deste livro é dedicado precisamente a este assunto, que não será portanto abordado aqui.

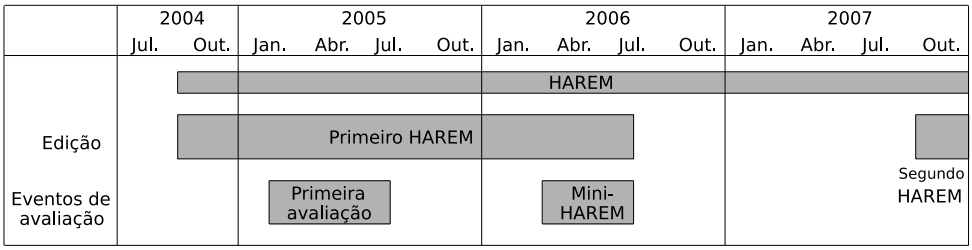

Figura 1.1: Diagrama temporal das edições e eventos de avaliação do HAREM.

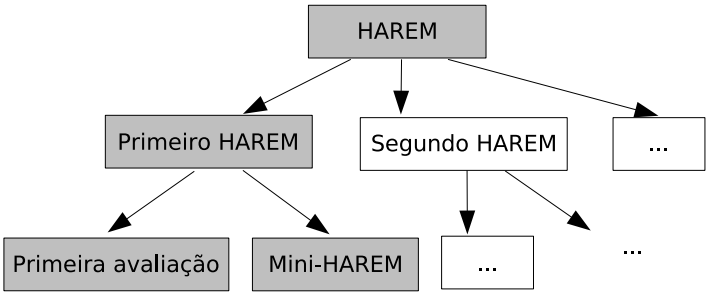

Figura 1.2: Terminologia usada no HAREM. Os eventos cobertos no presente livro estão marcados a cinzento.

1.3 A terminologia que emergiu do HAREM

Antes de prosseguirmos com uma análise histórica do desenvolvimento do HAREM, é essencial fixar a terminologia que vai ser usada neste livro e que foi surgindo muito pouco sistematicamente com as variadas fases da história do HAREM.

Assim sendo, a Figura 1.1 fornece um diagrama temporal das etapas do HAREM, enquanto que a Figura 1.2 indica graficamente as inclusões dos variados termos no contexto do HAREM.

1.4 Um pouco de história

Não fomos certamente os primeiros a achar que a detecção dos significados (ou categorias ontológicas) de nomes próprios seria uma sub-tarefa passível de avaliação separada. Cabe aqui contudo indicar como surgiu a inspiração, e até admitir que, no processo seguido, nem todas as outras fontes teoricamente possíveis de ser inspiradoras (porque já passadas) foram consultadas.

1.4.1 A inspiração

Foi o MUC (Message Understanding Conference), uma avaliação conjunta na área de extracção de informação (EI) existente desde 1987 (Hirschman, 1998), que propôs pela primeira vez, na sua sexta edição, que a tarefa de REM fosse medida de uma forma independente, após ter sido considerada durante vários anos como uma parte da tarefa mais geral de extrair informação de um texto (Grishman e Sundheim, 1996).

Embora os resultados da tarefa de REM, tal como definida pelo MUC, se tivessem situado a níveis muito altos de desempenho (mais de metade dos participantes obtiveram medidas F superiores a 90%), o que foi considerado um resultado comparável ao dos seres humanos, nem todos os investigadores aceitaram que isso indicava que a tarefa de REM já estava resolvida (veja-se por exemplo Palmer e Day (1997); Mikheev et al. (1999)). Por um lado, havia a questão da língua: "resolvido" para o inglês não significa resolvido para todas as línguas. Por outro lado, era preciso avaliar que métodos ou recursos eram necessários para essa tarefa.

Assim, após o MUC, vários outros eventos de avaliação focando o REM se seguiram, como o MET (Merchant et al., 1996), a tarefa partilhada do CoNLL (Sang, 2002; Sang e Meulder, 2003) ou o ACE (Doddington et al., 2004).

Enquanto o MET adoptou directamente a tarefa do MUC aplicando-a a japonês, espanhol e chinês, a tarefa partilhada do CoNLL procurou fomentar a investigação em sistemas de REM independentes da língua, usando textos em flamengo, espanhol, inglês e alemão mas reduzindo significativamente a grelha de classificação, que passou a conter apenas quatro categorias semânticas: `LOC` (local), `ORG` (organização), `PER` (pessoa) e `MISC` (diversos), simplificando portanto ainda mais a tarefa.

O ACE, pelo contrário, propôs a pista de *EDT - Entity Detection and Tracking*, em que o objectivo é fazer o reconhecimento de entidades, quer sejam quer não mencionadas através de um nome próprio, o que alarga consideravelmente a dificuldade da tarefa. O REM passa pois no ACE a compreender todo o reconhecimento semântico de entidades, sejam elas descritas por nomes comuns, próprios, pronomes, ou sintagmas nominais de tamanho considerável. Além disso, há um alargamento significativo das categorias usadas, como são exemplos as categorias armas, veículos ou instalações (em inglês, *facilities*), assim como a definição de uma "supercategoria" para locais+organizações, chamada "entidade geopolítica".

Deve ser referido que a inspiração directa e mais importante para o HAREM foi o MUC, e o nosso interesse de delimitarmos o problema em português e para o português, fez-nos duvidar ou não levar suficientemente a sério as iniciativas multilingues. Quanto ao ACE, foi tarde demais que soubemos das actividades deste, o que teve como consequência não nos termos inspirado nele para a organização do HAREM.

Por outro lado, convém lembrar que, em 2003 e 2004, altura em que surgiram várias

iniciativas de problematização e alargamento do REM, tais como o encontro de Guthrie et al. (2004), a Linguateca já estava em pleno no meio da organização do HAREM (ou do ensaio pré-HAREM), que será descrito em seguida.

1.4.2 Avaliação de REM em português antes do HAREM

O HAREM começou a ser planeado em Junho de 2003, por ocasião do Encontro AvalON.[2] Além de constituir o encontro final das Morfolimpíadas (Santos et al., 2003; Costa et al., 2007), nesse encontro foram discutidas e preparadas várias outras iniciativas, tendo sido lançadas as bases para um plano organizado de avaliações conjuntas em português, coadjuvado por uma comunidade científica interessada em participar em futuros iniciativas de avaliação semelhantes. Assim, foram convidadas várias pessoas a apresentar propostas concretas, uma das quais, da responsabilidade da Cristina Mota, era o culminar de um ensaio que visava medir ou auscultar o problema do REM em português.

Com efeito, esta investigadora tinha organizado nos meses antecedentes um ensaio, mais tarde documentado em Mota et al. (2007) e agora mais profusamente no capítulo 2 do presente livro, cujo objectivo era medir precisamente a dificuldade da tarefa de REM, abordando várias questões que ainda não tinham sido consideradas (ou, pelo menos, documentadas) em eventos anteriores.

O ensaio mostrou que:

- Muitos investigadores marcaram manualmente os textos usando uma hierarquia de classes semânticas bem mais vasta do que as hierarquias estipuladas por exemplo pelo MUC, o que mostra que a sua concepção de REM era diferente da reflectida pelos eventos de avaliação em REM da altura.

- A discordância entre anotadores era significativa, não só na interpretação do que é uma EM, mas também na identificação e na classificação das EM. Uma possível ilação a retirar foi a necessidade de incorporar o conceito de vagueza, quer na identificação quer na classificação, de forma a poder entrar em conta com as divergências, num ambiente de avaliação onde se mede e pontua o desempenho dos sistemas.

A apresentação das conclusões desse ensaio desencadeou uma discussão muito produtiva e participada sobre várias questões no encontro AvalON, tendo vários grupos sugerido que se começasse pelo REM geográfico. Contudo, pareceu-nos demasiado redutor cingir a futura tarefa de REM apenas à categoria dos locais em português, até porque um dos aspectos interessantes da avaliação seria medir a "confundibilidade" de nomes de locais com outras entidades.

[2] O Encontro AvalON, `http://www.linguateca.pt/avalon2003/`, foi um encontro sobre avaliação conjunta organizado pela Linguateca, que decorreu como um encontro satélite da 6ª edição do PROPOR em Faro (Mamede et al., 2003).

Este estudo serviu de inspiração para a organização do HAREM, que acabou por não incluir como organizadora a própria iniciadora do processo por razões relacionadas com a dedicação exclusiva desta nesse período à sua tese de doutoramento, e pelo facto de, além disso, pretender participar no HAREM, como veio a acontecer (veja-se o capítulo 15).

Embora tenhamos divergido em muitas questões da proposta original da Cristina Mota, é indubitavelmente a este ensaio que o HAREM mais deve a sua génese.

1.4.3 A preparação do Primeiro HAREM

O Primeiro HAREM teve o seu início oficial em Setembro de 2004, com um anúncio e chamada à participação através de mensagens nas listas e por mensagens directas aos já conhecidos possíveis interessados, saídos do ensaio inicial e da lista sobre avaliação mantida pela Linguateca.

Os autores do presente capítulo expuseram nessa altura a intenção da Linguateca de desenvolver uma metodologia nova para avaliar o REM, usando uma colecção de textos de diferentes géneros textuais e de várias variantes (a colecção do HAREM – CH), como base para criar uma colecção dourada (CD), ou seja, uma colecção devidamente anotada por seres humanos e que constituiria a bitola de comparação utilizada no HAREM.

As categorias semânticas seriam criadas por todos os participantes a partir da análise cuidada dos textos, e as directivas seriam continuamente aperfeiçoadas à medida que se progredia na tarefa de anotação da colecção dourada.

Nessa altura estabeleceu-se um grupo inicial de interessados, que se declararam participantes ou apenas observadores (por exemplo, interessados no problema mas que não tinham intenções ou condições de desenvolver um sistema REM para participar). Tivemos dez observadores, quatro dos quais participaram no exercício de anotação manual inicial (Débora Oliveira, Elisabete Ranchhod, John Cullen e Jorge Baptista), pelo qual manifestamos aqui a nossa gratidão.

Após coligir uma colecção de textos para a CD, o primeiro passo foi a divisão da CD em vários pedaços. A 26 de Outubro de 2004 foi entregue aos participantes (ou observadores) um pedaço diferente para o anotarem manualmente no prazo de duas semanas, seguindo uma proposta inicial de regras de etiquetagem e um conjunto inicial de categorias semânticas, meramente indicativas. Os participantes nessa anotação cooperativa foram mesmo instados a alargar ou mesmo "desobedecer" às directivas, e partilhar os seus argumentos com o resto da comunidade.

Com esta actividade, tentámos atingir vários objectivos:

- Em primeiro lugar, os participantes e observadores familiarizaram-se de imediato com as dificuldades da tarefa, nomeadamente a vagueza[3] da identificação e da classificação semântica, e a escolha das categorias e tipos semânticos a usar na hierarquia

[3] Sobre a questão da ubiquidade da vagueza em linguagem natural, ver Santos (1997).

final, que abranja adequadamente as EM reconhecidas. Desta forma, as discussões conjuntas em torno da metodologia do HAREM deixaram o reino do abstracto e foram muito mais produtivas e orientadas para os reais requisitos da tarefa em questão.

- A participação activa dos participantes e observadores nas etapas da organização da primeira avaliação do Primeiro HAREM tentou garantir que este correspondesse às necessidades da comunidade, e que os seus objectivos fossem ouvidos e levados em conta na metodologia em desenvolvimento. Ou seja, tentámos chegar a uma metodologia que traduzisse o que a comunidade entendia por REM em português, e que estaria implementada nos seus sistemas, evitando o erro de estipular uma tarefa desfasada da realidade que se pretende avaliar. Se tal foi ou não cabalmente conseguido, poderá ser julgado pelos capítulos de discussão no presente volume.

Durante o processo de anotação dos pedaços, várias dúvidas e casos "difíceis" (ou, simplesmente, casos que causaram discordâncias) foram debatidos, servindo de base para elaborar a primeira revisão às directivas, cuja discussão, pelos participantes, observadores e público em geral, teve como prazo final o dia 5 de Novembro de 2004. Os pedaços anotados foram entregues até ao dia 19 de Novembro de 2004.

Estes pedaços voltaram a ser reunidos numa verdadeira CD anotada, que foi exaustivamente revista por quatro anotadores da Linguateca: os autores do presente capítulo, Anabela Barreiro e Susana Afonso. Contudo, é preciso confessar que, no processo de revisão, as directivas não deixaram de ser aperfeiçoadas, quando assim achámos oportuno. A 16 de Dezembro de 2004, foi distribuído aos participantes um pedaço da CH etiquetado conforme as directivas em vigor, para poderem adaptar os seus sistemas e familiarizarem-se com o formato a empregar no HAREM. Até 10 de Janeiro de 2005, a organização dedicou-se aos aspectos associados com a medição dos sistemas, nomeadamente as directivas de avaliação e a definição da arquitectura de avaliação. Contudo, a CD continuou a ser revista aturadamente, com alterações pontuais às directivas oportunamente divulgadas. Entre 10 de Janeiro e 14 de Fevereiro de 2005 não foram realizadas mais alterações, para que se pudesse dar tempo aos participantes para adaptar os seus sistemas às directivas oficiais do HAREM.

1.4.4 O primeiro evento do Primeiro HAREM

O primeiro evento de avaliação teve início no dia 14 de Fevereiro de 2005. Os dez participantes (descritos na Tabela 1.1), oriundos de seis países diferentes (Brasil, Dinamarca, Espanha, França, México e Portugal), receberam a CH sem anotações, que tinham de devolver, marcada automaticamente passadas 48 horas. Foram-nos enviadas 18 saídas dentro do prazo e 3 saídas fora do prazo (não-oficiais, portanto).

Sistema	Participante	Instituição
CaGE	Mário J. Silva, Bruno Martins e Marcirio Chaves	Grupo XLDB, Universidade de Lisboa
Cortex	Violeta Quental	PUC-Rio/CLIC
ELLE	Isabel Marcelino	Pólo da Linguateca no LabEL
Malinche	Thamar Solorio	INAOE
NERUA	Óscar Ferrández, Zornitsa Kozareva, Antonio Toral, Rafael Muñoz e Andrés Montoyo	Universidade de Alicante
PALAVRAS-NER	Eckhard Bick	University of Southern Denmark
RENA	Edgar Alves e José João Dias de Almeida	Universidade do Minho
RSN-NILC	Graça Nunes, Ricardo Hasegawa e Ronaldo Martins	NILC
SIEMÊS	Ana Sofia Pinto, Luís Sarmento e Luís Miguel Cabral	Pólo do Porto da Linguateca
Stencil/NooJ	Cristina Mota e Max Silberztein	IST e LASELDI, Université de Franche-Comté

Tabela 1.1: Participantes na primeira avaliação do Primeiro HAREM

Passados mais dois dias, a colecção dourada (CD) (ou seja, o subconjunto anotado da colecção HAREM, CH) foi divulgada aos participantes, para eles próprios, se assim o desejassem, analisar as soluções e eventualmente alertar para possíveis erros.

Era tempo para desenvolver a plataforma de avaliação (capítulo 19 e Seco et al. (2006)), na qual, além dos autores do presente capítulo, participaram Nuno Seco e Rui Vilela.

O HAREM inspirou-se nas métricas de avaliação do MUC para a avaliação comparativa das saídas dos sistemas (Douthat, 1998). Contudo, foram introduzidos diversos melhoramentos para lidar com várias questões não contempladas no MUC, tais como a vagueza, a separação entre a avaliação da identificação e a da classificação semântica (categorias e tipos), o conceito de correcção parcial, e a avaliação separada por cenários distintos. Além disso, foram também aproveitados alguns conceitos da experiência anterior das Morfolimpíadas, tal como a distinção entre medidas absolutas e relativas (Santos et al., 2003; Costa et al., 2007). As métricas de avaliação, bem como as medidas, regras e as pontuações usadas no cálculo do desempenho dos sistemas, foram publicadas a 29 de Setembro de 2005. A última redacção desse texto (mas sem mudanças em relação à substância) encontra-se no capítulo 18 deste livro.

A 22 de Abril de 2005, foi apresentada aos participantes uma primeira arquitectura da plataforma de avaliação, permitindo a avaliação por cenários, e implementando na totalidade as directivas de avaliação entretanto colocadas públicas. Também nesta fase, os participantes podiam acompanhar o trabalho desenvolvido e opinar sobre as regras de avaliação e a pertinência das medidas, já com a ajuda dos exemplos concretos disponibili-

zados com a documentação dos programas.

A 20 de Maio de 2005 foram enviados aos participantes os primeiros resultados do HAREM, respeitantes à tarefa de identificação. Os resultados globais, devidamente anonimizados, foram tornados públicos a 9 de Junho de 2005. Uma semana depois, eram divulgados os resultados relativos à classificação morfológica.

É preciso mais uma vez salientar que as directivas de avaliação foram continuamente revistas (e tornadas mais pormenorizadas), pois, à medida que se desenvolviam os programas de avaliação, algumas situações particulares iam sendo detectados e resolvidos.

A grande demora na publicação dos resultados ficou no entanto também a dever-se ao facto de quase todas as saídas submetidas ao HAREM não respeitarem as regras de etiquetagem, o que levou à necessidade de normalizar manualmente as saídas enviadas, e interagir com os participantes no sentido de resolver estes problemas.

Assim sendo, só a 6 de Setembro de 2005 (sensivelmente sete meses após os participantes terem enviado o resultado dos seus sistemas) é que foi possível divulgar os resultados finais da tarefa de classificação semântica, juntamente com uma revisão ligeira dos valores para a tarefa de identificação, que não apresentou alterações significativas na ordenação dos participantes. Seguiram-se os resultados da tarefa da classificação morfológica, publicados em 29 de Setembro de 2005. Finalmente, o processo foi dado por concluído com o envio dos resultados individuais, para todas as tarefas, aos participantes, a 28 de Outubro de 2005.

1.4.5 O Mini-HAREM: medição do progresso e validação estatística

Considerando que os resultados do HAREM já não representavam fielmente o estado dos sistemas concorrentes, e que o atraso na publicação destes tinha resultado em alguma desmotivação da comunidade, resolvemos repetir, ainda dentro do Primeiro HAREM, a comparação entre os sistemas que estivessem dispostos a enviar novas saídas. Uma vez que a arquitectura de avaliação se encontrava concluída e os programas prontos, livremente disponíveis e amplamente testados com os mesmos sistemas que iriam participar, não se previam atrasos substanciais na publicação dos resultados da nova avaliação conjunta.

A este novo evento de avaliação chamou-se o Mini-HAREM, e a participação no dito foi restrita apenas aos participantes do primeiro evento. O Mini-HAREM empregou a mesma metodologia do HAREM – com excepção de algumas pequenas alterações nas categorias. Muito brevemente,

- o tipo `PRODUTO` da categoria `OBRA` foi suprimido;
- o tipo `MEMBROCLASSE` foi adicionado à categoria `COISA`;
- os URL e os endereços de correio electrónico deixaram de ser considerados EM.

Os participantes foram evidentemente informados com antecedência destas ligeiras mudanças, mas não de qual colecção de textos os seus sistemas iriam classificar. De facto, foi distribuída aos participantes a mesma CH; a diferença residia no uso de uma nova CD. A constituição desta segunda CD usada no Mini-HAREM, a que chamamos CD 2006, é semelhante à da primeira CD, chamada CD 2005, e os seus documentos são disjuntos.

O Mini-HAREM teve os seguintes objectivos (mais detalhados em Cardoso (2006a)):

- A obtenção de mais dados sobre cada sistema participante: ao rever/anotar manualmente mais uma parcela da CH, conseguimos o dobro do material no qual podemos basear a avaliação, ao concatenar as duas CD.
- A obtenção de material para a validação estatística dos resultados dos sistemas participantes (ver capítulo 5): com dois eventos usando a mesma colecção, pode-se medir os sistemas sobre duas colecções douradas e sobre o conjunto destas (ao todo, três recursos de avaliação).
- A medição da evolução dos sistemas ao longo do tempo (desde a altura do primeiro evento até ao Mini-HAREM medeou um ano).
- Uma melhor caracterização do estado da arte em REM para o português.

Para evitar que problemas inesperados na formatação dos resultados dos sistemas atrasassem novamente esta comparação, para o Mini-HAREM foi também desenvolvido um verificador de sintaxe das saídas (ver secção 19.2.1),que permitia que os participantes verificassem se a marcação produzida pelos seus sistemas estava conforme as regras do HAREM e os requisitos dos programas de avaliação do mesmo, antes de enviarem as saídas oficialmente para o HAREM.

Com os programas de avaliação e de geração de relatórios já desenvolvidos, o Mini-HAREM decorreu com maior rapidez. A chamada à participação foi realizada no início de 2006, e o Mini-HAREM foi marcado para o dia 3 de Abril de 2006. Infelizmente, nem todos os participantes no Primeiro HAREM se mostraram interessados, e alguns sistemas tinham mudado de mãos ou sido completamente reestruturados.

O Mini-HAREM contou assim apenas com cinco participantes (descritos na Tabela 1.2), metade dos participantes originais, mas que enviaram 20 saídas, todas oficiais. Os participantes tiveram igualmente um prazo de 48 horas para devolver a colecção do HAREM devidamente etiquetada, um prazo que terminou no dia 5 de Abril de 2006, ao meio-dia, hora de Lisboa.

Não obstante ter sido facultado o validador e termos informado os participantes dos problemas no caso do evento anterior, foi necessário mesmo assim rever manualmente as saídas e corrigir a sua sintaxe para que pudessem ser processadas.

Assim, dois meses depois, a 9 de Junho de 2006, foram divulgados os resultados globais do Mini-HAREM, e os relatórios individuais enviados aos participantes. A comparação dos

Sistema	Participante	Instituição
CaGE	Mário J. Silva, Bruno Martins e Marcirio Chaves	Grupo XLDB, Universidade de Lisboa
Cortex	Violeta Quental e Christian Nunes	PUC-Rio
SIEMÊS2	Luís Sarmento	FEUP/Pólo do Porto da Linguateca
SMELL	Elisabete Ranchhod e Samuel Eleutério	LabEL
Stencil-NooJ	Cristina Mota e Max Silberztein	L2F/INESC e LASELDI, Université de Franche-Comté

Tabela 1.2: Participantes na segunda avaliação do Primeiro HAREM, o Mini-HAREM

dois resultados foi apresentada no Encontro do HAREM no Porto, a 15 de Julho de 2006 (Cardoso, 2006b), além de ser pormenorizadamente discutida em Cardoso (2006a).

1.5 Uma breve descrição da participação no Primeiro HAREM

A participação no Primeiro HAREM foi muito variada, englobando desde sistemas desenvolvidos de raiz para participar no HAREM, como o SIEMÊS (ver capítulo 14) e o ELLE (Marcelino, 2005), até sistemas que participaram "de raspão" para verificar ou estudar questões relativamente marginais, tais como o reconhecimento de entidades geográficas apenas, como o CaGE (capítulo 8), ou a simples identificação de entidades mencionadas através de métodos de aprendizagem automática, como o MALINCHE (capítulo 10).

No meio do espectro tivemos sistemas já existentes, que faziam portanto já alguma forma de REM completo, mas sem necessariamente conceberem o problema do REM como implementado no HAREM (aliás, isso nunca aconteceu), tais como o PALAVRAS-NER (capítulo 12), o Stencil-NooJ (capítulo 15), o NERUA (capítulo 11) ou o Cortex (capítulo 9). Podemos contudo ainda subdividir os sistemas entre aqueles que tentaram de certa forma adaptar o seu funcionamento para participar no HAREM e aqueles que se ficaram por experimentar — sem adaptação -– até onde o seu sistema original conseguia ir, dada a tarefa de avaliação proposta.

Ao contrário das Morfolimpíadas, em que todos os sistemas pertenciam à categoria de sistemas já existentes e bem desenvolvidos, antes da avaliação conjunta, o HAREM parece-nos ter conseguido estimular interesse específico e novo no problema, não só devido ao facto de terem de facto surgido sistemas novos, como pelo interesse unânime em participar em novas edições, expresso por todos os participantes no Encontro do HAREM, e que esperamos poder confirmar-se na prática num futuro breve.

Mais uma vez por oposição às Morfolimpíadas, também temos de reconhecer que não conseguimos que o HAREM cobrisse outras zonas limítrofes. Ou seja, enquanto que

um radicalizador e um corrector ortográfico também participaram nas Morfolimpíadas, desta forma aumentando o âmbito desta avaliação conjunta, a nossa tentativa de alargar o HAREM ao simples reconhecimento de nomes próprios em texto falhou, visto que o NILC (o único sistema que tinha concorrido sob esta perspectiva) preferiu retirar-se por achar que esta última tarefa era demasiado distinta para fazer sentido ser englobada numa avaliação de REM.

1.6 Mais informação sobre o HAREM: um pequeno guia

Ao longo dos mais de três anos de trabalho da Linguateca na área de REM, foi sendo criada documentação variada, não só a nível das páginas na rede no sítio da Linguateca, como também sob a forma de diversos artigos e apresentações e uma tese de mestrado, todos eles sobre o HAREM.

Neste livro parece-nos mais indicado mencionar onde se encontra a informação em relação aos variados temas, em vez de a repetir, embora tenhamos tentado incluir neste volume as especificações fundamentais do HAREM, ao republicar as directivas de anotação e a descrição das medidas, respectivamente nos capítulos 16, 17 e 18.

1.6.1 Ensaio pré-HAREM

O estudo organizado pela Cristina Mota e que inspirou o HAREM foi inicialmente documentado em Mota et al. (2007), por ocasião do livro dedicado ao paradigma de avaliação conjunta (Santos, 2007a). O capítulo 2 constitui uma documentação mais pormenorizada, em que podemos seguir a experiência de anotação de textos do CETEMPúblico e do CETENFolha, que contou com a colaboração de nove investigadores e que foi fundamental para detectar muitos dos problemas que vieram a ser tratados no HAREM.

1.6.2 Metodologia

Quase todos os artigos ou apresentações relativos ao HAREM dão bastante ênfase às inovações metodológicas, quer na definição da própria tarefa, quer na forma de a avaliar. Veja-se pois Santos et al. (2006), Santos (2006a), Santos (2006b) e Seco et al. (2006) para formas diferentes de apresentar o HAREM nessa perspectiva. No capítulo 3 podemos encontrar uma comparação detalhada entre a metodologia do HAREM, e a metodologia adoptada pelo MUC, enquanto o capítulo 4 discute a questão específica do modelo semântico contrastando-o com o do MUC e o do ACE.

De qualquer forma, um prato forte de quase todos os capítulos da parte de discussão do presente volume são as questões metodológicas.

1.6.3 A colecção dourada

Uma parte importante da metodologia refere-se ao conjunto das soluções presentes na CD. Em Santos e Cardoso (2006) detalha-se a criação e as características da CD, bem como a motivação subjacente à decisão em adoptar um leque mais diversificado de categorias e de tipos, e como a vagueza se encontra codificada nas etiquetas usadas pelo HAREM.

Para conhecer a fundo as categorias e as opções utilizadas na criação das colecções douradas, é imprescindível consultar as directivas (capítulos 16 e 17 deste volume). Visto que os sistemas de REM participantes podiam escolher se participavam na classificação semântica, na classificação morfológica, ou em ambas, sendo apenas obrigatória a tarefa de identificação, dividimos as directivas em duas. Como tal, durante a avaliação, a tarefa de identificação encontrava-se descrita em ambos os documentos.

Finalmente, o capítulo 4 de Cardoso (2006a) destila as CD usadas, nomeadamente na sua composição por géneros textuais, categorias semânticas e variantes. Muito desse material foi republicado no capítulo 20 deste volume.

1.6.4 Quantificação: Métricas, medidas, pontuações e regras de cálculo

Embora também apresentadas junto com a metodologia do HAREM (e portanto delineadas nos artigos e capítulos mencionados acima), a apresentação pormenorizada das medidas e métricas do HAREM é feita no capítulo 18, compreendendo as pontuações por cada alinhamento, as regras para lidar com alternativas de identificação, as várias medidas contempladas para cada tarefa, e as métricas usadas para a atribuição de um valor de desempenho às saídas dos sistemas.

1.6.5 A arquitectura e os programas da plataforma de avaliação

A arquitectura da plataforma de avaliação do HAREM foi apresentada em Seco et al. (2006), e detalhada na secção 4.3.3 de Cardoso (2006a). No capítulo 19 apresenta-se a documentação detalhada e definitiva de todos os programas que fazem parte da arquitectura proposta, cujo código fonte se encontra também disponível desde a realização do Mini-HAREM.

1.6.6 Validação estatística

A tarefa de validação estatística aos resultados do HAREM foi o assunto principal da tese (Cardoso, 2006a), onde se descreve o método estatístico utilizado, a metodologia de validação, a sua adaptação aos requisitos do HAREM, e onde se demonstra que o tamanho das colecções usadas nos eventos HAREM é suficiente para comparar adequadamente os sistemas. O capítulo 5 do presente volume resume o trabalho de validação estatística efectuado.

1.6.7 Resultados do HAREM

No capítulo 5 (página 69) e na secção 5.3 de Cardoso (2006a), faz-se uma primeira análise dos resultados globais do HAREM, fornecendo um primeiro panorama de REM em português. Uma selecção dos próprios resultados encontra-se como apêndice deste volume.

1.6.8 Discussão e primeiro balanço

O encontro presencial do HAREM constituiu um primeiro balanço da iniciativa, quer do ponto de vista da organização, quer do ponto de vista dos participantes. As contribuições (ver sítio do Encontro do HAREM) e a discussão ocorrida formaram o ponto de partida para o presente volume, que passamos a descrever brevemente.

1.7 O presente livro

Após variadas reformulações, decidimos dividir o livro em três partes:

1. a parte relacionada com o REM em português;
2. a parte de descrição conjuntural dos sistemas participantes no Primeiro HAREM;
3. a parte de documentação desta primeira avaliação conjunta.

A primeira parte é a que pode ser mais interessante de um ponto de vista teórico, porque descreve questões quer de organização quer de conteúdo de uma avaliação conjunta que são pertinentes para o futuro da área. Não é, contudo, possível nem desejável ficar a um nível de abstracção tão elevado que impeça o leitor de compreender de que tipo de sistemas e/ou problemas estamos a falar.

Para isso é fundamental consultar e compreender a documentação dos próprios sistemas e a explicação dos princípios de funcionamento subjacentes, que constitui a segunda parte do livro, e que poderá servir não só para ilustrar a grande variedade de abordagens e preocupações do leque de participantes, mas também para inspirar a criação de novos sistemas ou a reutilização de técnicas de outros sistemas.

A terceira e última parte é, em grande parte, uma mera republicação das directivas utilizadas, mas a que se juntaram dois capítulos originais: o primeiro sobre a arquitectura dos programas de avaliação, e o segundo sobre a disponibilização das colecções douradas através do projecto AC/DC (Santos e Sarmento, 2003).

Finalmente, pensamos ser necessário que fique fixado e empacotado em forma de livro a destilação do que foi o Primeiro HAREM: as directivas seguidas na anotação da CD e as medidas e métodos de cálculo empregues. Não porque achamos que devam permanecer imutáveis e usadas sempre daqui para a frente, mas porque é preciso que possam ser facilmente referidas (e eventualmente revogadas, ou melhoradas) em futuras edições do HAREM.

Agradecimentos

Embora tenhamos acabado por escrever este capítulo apenas no nosso nome, não queremos deixar de reconhecer que a organização do Primeiro HAREM foi partilhada, em maior ou menor grau, com o Nuno Seco, o Rui Vilela, a Anabela Barreiro, a Susana Afonso e o Paulo Rocha.

E que, claro, sem os participantes e/ou observadores do HAREM não teria havido HAREM.

Quanto ao texto propriamente dito, estamos muito gratos a todos os investigadores que se deram ao árduo trabalho de rever com toda a atenção a nossa primeira versão, e cujas sugestões e recomendações nos levaram a mudanças por vezes substanciais. Foram eles, por ordem alfabética, António Teixeira, Cristina Mota, Daniel Gomes, Eugénio Oliveira, Graça Nunes, Jorge Baptista, Luís Costa e Paulo Gomes. Esperamos que possam reconhecer as melhorias que eles próprios sugeriram.

Este texto, assim como o trabalho que descreve, insere-se no âmbito do trabalho da Linguateca, financiada através dos projectos POSI/PLP/43931/2001 e POSC 339/1.3/C/NAC, e co-financiada pelo POSI.