Capítulo 6

Segundo HAREM: Balanço e perspectivas de futuro

Diana Santos, Cláudia Freitas, Hugo Gonçalo Oliveira, Paula Carvalho e Cristina Mota

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 6, p. 131–146.

Neste capítulo fazemos o balanço do Segundo HAREM, tentando documentar tanto os pontos fortes desta avaliação como as questões que, na nossa perspectiva, não foram completa ou adequadamente resolvidas. Este exercício de reflexão é feito numa perspectiva de tentar projectar o que aprendemos para o futuro, no caso de vir a existir a oportunidade de organizar uma terceira edição desta avaliação conjunta.

Sempre que tal for pertinente, referiremos o balanço do Primeiro HAREM, documentado em Santos e Cardoso (2007b), para dar uma dimensão histórica em relação ao que foi conseguido e ao que ainda ficou por fazer.

Visto que as três pistas mereceram no livro capítulos distintos e trouxeram questões e problemas diversos, resolvemos apresentar um balanço por pista, ao contrário do que fizemos no encontro do Segundo HAREM (Santos et al., 2008e). Contudo, as questões que se refiram ao HAREM como um todo serão discutidas na primeira oportunidade.

Começamos por apresentar as questões que nos parecem dever ser melhoradas ou que não correram como esperado, relatando, em seguida, as que tiveram um desfecho a (nosso) contento. Terminamos com algumas perguntas e sugestões para o futuro, após uma caracterização crítica da participação neste Segundo HAREM.

6.1 HAREM clássico: balanço geral

Em relação ao HAREM clássico, não nos podemos esquivar à seguinte autocrítica: assumimos que todo o trabalho árduo de desbravamento do texto, com a criação das directivas e estabelecimento de categorias, já havia sido feito, e que portanto o tempo que levaria o processo de anotação da colecção dourada de uma segunda edição seria muito menor.

Contudo, novos textos (e uma nova equipe) tendem, sempre, a levantar novas dúvidas de anotação, e portanto levam a refinamentos e alterações nas directivas, e, consequentemente, a novas possibilidades de anotação. Com isso, lamentamos não ter havido tempo de produzir um documento único com as directivas do Segundo HAREM, onde seria possível encontrar tanto as características e categorias adoptadas do Primeiro HAREM (e, portanto, descritas no seu âmbito) como as introduzidas no Segundo HAREM, descritas no sítio do Segundo HAREM.

Outra questão que tem sido recorrente em todas as avaliações conjuntas que a Linguateca já organizou prende-se com a dificuldade de arranjar um esquema de classificação válido e consensual para o tipo e género de textos utilizados. Com efeito, ainda não foi desta que ficámos plenamente satisfeitos com o resultado obtido (e divulgado na LÂMPADA, o pacote de recursos do Segundo HAREM).

De qualquer forma, identificámos como especialmente problemáticas as seguintes questões, que discutimos separadamente em seguida:

- Peso da identificação em relação à classificação
- Delimitação das entidades mencionadas
- Dois modelos filosóficos opostos incluídos no HAREM

Iremos também aflorar a questão da utilização do XML, apresentando depois os aspectos positivos, tais como:

- Progresso na definição da tarefa

- Recursos mais ricos e mais bem documentados
- Véus mais bem aproveitados
- Relação explícita com outras tarefas
- Desenvolvimento de ferramentas para facilitar a avaliação conjunta

6.1.1 Identificação vs. classificação

Uma das questões que tentámos resolver e melhorar em relação ao Primeiro HAREM foi evitar a separação absoluta da identificação em relação à classificação, tornando a medida única e entrando em conta com esses dois aspectos como duas faces da mesma moeda (seguindo a proposta de Santos e Cardoso (2007b)). De um ponto de vista conceptual, também quisemos promover a "identificação" como a "classificação o mais vaga possível", ou seja, uma supercategoria das dez categorias usadas no Segundo HAREM, como explicado detalhadamente no capítulo anterior).

Contudo, e malgrado essa mudança, parece-nos que a identificação ainda teve um peso demasiado grande em relação à classificação, fazendo com que sistemas que se cingiram a identificar EM fossem superiores aos que também tentaram classificá-las. Por exemplo, o vencedor em termos de precisão na classificação para o cenário total, o SeRELeP, só fez identificação, o que é, no mínimo, pouco natural.

Além disso, logo que os participantes seleccionem um subconjunto de categorias (ou seja, concorram num cenário selectivo) estão implicitamente a classificar. Isso é de sobremaneira flagrante nos casos em que concorrem apenas numa única categoria, mas também se aplica quando competem num conjunto de categorias (e não em todas).

Isto leva-nos a concluir que, a não remover o premiar da identificação simples, deveríamos garantir que esse prémio fosse ínfimo comparado com a classificação. E, como trabalho futuro, aqui fica deixado o repto de estudar várias combinações possíveis de pesos para ver qual a combinação mais adequada.

6.1.2 Delimitação das entidades mencionadas

Para simultaneamente reduzir a importância das diferentes estratégias de identificação e garantir que as EM na colecção dourada estivessem bem delimitadas – ao contrário da CD do Primeiro HAREM, em que estavam a ser reconhecidas inadequadamente como EM sequências como, por exemplo, *de Planck*, como notado em Santos e Cardoso (2007b) – usámos a seguinte estratégia:

1. Procurámos todos os casos da CD em que havia palavras em minúsculas que achámos que deviam fazer parte da entidade (que formalmente corresponde a uma entidade complexa ou multipalavra);
2. Produzimos uma lista exaustiva dessas palavras (a lista completa encontra-se no apêndice A, secção A.6) e tornámo-la pública, de modo a que todos os participantes pudessem tê-las em conta no desenvolvimento dos seus sistemas;
3. Declarámos que quaisquer outros casos não deviam ser marcados, no sentido de restringir ao máximo a identificação de minúsculas, no âmbito da tarefa específica de REM, tal como é definida no modelo do HAREM.

Esta última declaração, embora tornasse a tarefa igual para todos, provocou muita confusão e descontentamento por parte dos participantes, sobretudo porque não podíamos explicar o verdadeiro motivo da escolha de tais elementos (que, reiterando, era a garantia de uma delimitação perfeita na CD), em detrimento de outros lexical e semanticamente próximos ou equivalentes.

Na verdade, do ponto de vista da avaliação, seria irrelevante o modo como os sistemas tratassem todos os outros casos não contemplados na CD. Assim, é fácil perceber, agora, que o terceiro passo não só era desnecessário como até seria prejudicial caso quiséssemos estender a colecção dourada (o que não veio a acontecer), pois isso implicaria uma actualização da dita lista.

Contudo, a nossa acção também não é totalmente indefensável: de facto, uma análise mais apurada indica que temos aqui uma tensão irredutível entre a) especificar a 100% uma tarefa para todos os participantes e b) produzir directivas linguisticamente apropriadas (ou mesmo simplesmente consensuais).

De facto, não existe uma descrição formal (no sentido de rigorosa, explícita e completa) do que é uma EM em português[1], e como organização do HAREM (tanto no Primeiro como neste Segundo), nós propusemos uma descrição (quase) meramente gráfica: *Uma EM deve conter pelo menos uma letra em maiúsculas, e/ou algarismos*, ver página 214 do capítulo 16 de Santos e Cardoso (2007a))

A outra alternativa, nomeadamente a de aceitar qualquer que fosse a delimitação que os (autores dos) sistemas achassem correcta, teria a desvantagem de deixar a tarefa mal definida, e implicar que, nessas condições, um sistema seria melhor ou pior pontuado dependendo da maior ou menor proximidade que tivesse relativamente à posição dos anotadores da CD.

6.1.3 Modelos de avaliação conjunta incongruentes entre si

Talvez a maior fraqueza deste HAREM tenha sido a coexistência, no seu seio, de tarefas concebidas em termos de filosofias diferentes de avaliação conjunta, dentro de uma mesma tarefa, o HAREM clássico (no que respeita à categoria `TEMPO`).

Com efeito, no modelo do Primeiro HAREM e que pretendemos continuar neste Segundo (ver Santos (2007d); Santos et al. (2008d)), é o contexto que decide a análise a atribuir a uma dada expressão e a análise é a da pessoa (ou do grupo) que anota a colecção dourada (sem quaisquer limitações ou simplificações), que tenta reproduzir fielmente a compreensão humana dos textos em questão, resultando assim num tecto, ou melhor, no alvo daquilo que idealmente se pretende obter, mas que pode ser quase impossível de automatizar.

No modelo de avaliação conjunta proposto para a categoria `TEMPO`, por outro lado, a tarefa apresenta-se como "capaz de ser executável em seis meses de desenvolvimento" (ver página 36 do capítulo 2), ou seja, não é para representar já toda a informação que um ser humano consiga identificar[2]. No capítulo respectivo, os autores mencionam, quer como guias iniciais quer como escolhas posteriores, critérios de "minimizar interacções complexas", deixando claramente questões mais espinhosas para mais tarde. Contudo, no

[1] Estamos a referir-nos à delimitação do texto que aponta/menciona a entidade mencionada, não à entidade "real" em si (ver figura 4.1 de Santos (2007d)).

[2] Os autores até falam em progressões em pequenos passos, ou seja, a sua abordagem foi definir primeiro tarefas que lhes pareciam mais fáceis.

nosso entender, ao simplificar a tarefa em alguns casos (como a introdução da preposição na expressão temporal independentemente do seu sentido) transformam-na em algo que não é consistentemente semântico, mas uma mistura de certa forma arbitrária entre vários tipos de considerações (sintácticas, semânticas, ...).

Não nos ficaria bem estar aqui a argumentar outra vez a favor de um modelo contra o outro (visto que o modelo do HAREM já foi defendido em Santos (2007d)). O que nos interessa sublinhar é aquilo que nos parece uma situação infeliz desta "incongruência" de modelos: como resultado desta situação, os recursos de avaliação – em particular, a colecção dourada do HAREM clássico, mas também a do ReRelEM – exibem categorias anotadas segundo filosofias diferentes. Uma que se pretende "derradeira" em termos de interpretação humana; outra que representa um primeiro passo numa sequência definida pelos autores da proposta, e, na nossa opinião, de forma relativamente arbitrária: basta pensar que o que pode ser simples para um sistema pode ser complicado de alcançar por outro, ou que essa sequência impõe restrições à forma como os sistemas são construídos.

Seria assim útil que uma nova anotação "derradeira" do TEMPO fosse levada a cabo (de acordo com a filosofia do HAREM), assim como seria também interessante proceder a uma anotação mais "fácil" e preliminar das outras categorias, aliás proposta também no encontro do Segundo HAREM (assim como no do Primeiro), em que os países fossem sempre considerados LOCAL, etc.

Em conclusão, esta é apenas uma observação sobre modelos de anotação incongruentes misturados nos mesmos recursos, sem tentar sequer escolher um deles.

Embora entrando no terreno da especulação, é possível que tenha sido aliás isso que levou a que as iniciativas para o inglês de anotação temporal (veja-se por exemplo Wilson et al. (2001); Pustejovsky et al. (2003)) fossem separadas do ACE, ao contrário do que aconteceu para o português no HAREM. O tempo o dirá se a interligação das duas comunidades (em vez da sua separação) trará vantagens para a nossa língua, ou se ambas as tarefas acabarão finalmente por divergir.

6.1.4 Novo formato XML

Uma das questões apontadas por vários participantes no Primeiro HAREM e que nos comprometemos a melhorar neste Segundo estava relacionada com o uso de XML, veja-se Martins e Silva (2007) e Almeida (2007). Mas o maravilhoso mundo da padronização parece melhor ao longe... De perto, descobrimos que há várias versões dos padrões, incompatíveis entre si, isto mesmo ao nível da visualização na rede.

De facto, tal "solução" acabou por criar mais problemas do que resolveu: não só foi preciso reformatar as antigas CD (que disponibilizámos para treino), como levou a que parte significativa dos programas tivesse de ser reescrita.

É no entanto preciso reconhecer que podemos ter sido demasiado cautelosos na migração para XML, tendo dado lugar a um híbrido ainda mais difícil de caracterizar e processar. Em particular, parece-nos agora que a nossa notação dos ALT deveria ter sido generalizada à vagueza da classificação, no sentido de que A|B (sintaxe do Primeiro HAREM), passada agora para CATEG="A|B" no Segundo HAREM, deveria sim ter sido transformada em algo como uma das seguintes alternativas:

1. <ALT ID="x">
   <ALTN><EM CATEG="OBRA">...</EM></ALTN>

```
<ALTN><EM CATEG="LOCAL">...</EM></ALTN>
</ALT>
```

2.
```
<ALT>
<ALTN><EM ID="x" CATEG="OBRA"></ALTN>
<ALTN><EM ID="y" CATEG="LOCAL"></ALTN>
</ALT>
```

No entanto, a primeira alternativa não permitia atribuir diferentes identificações (ID) a alternativas com diferentes delimitações, enquanto a segunda parecia impor diferentes identificações ao que nós reputamos uma **única** EM.

Outra questão que terá de ser melhor equacionada é o uso de espaços para permitir mais de um valor no mesmo atributo (usado no ReRelEM), e que não é boa prática.

Confessamos, assim, a necessidade de mais uma ronda para definir, em XML, SGML ou ainda outro formalismo, uma representação que seja simultaneamente adequada e fácil de processar, abarcando a marcação da vagueza tanto na classificação como na identificação, assim como a problemática do encaixe, ou seja, da recursividade na definição das EM.

6.1.5 Progresso na definição da tarefa e nos desafios

Contudo, de uma forma geral, pensamos que é indiscutível que o Segundo HAREM representou um claro progresso em relação ao Primeiro, sob várias perspectivas.

Em primeiro lugar, algumas das limitações detectadas foram colmatadas, levando a que a medida de avaliação fosse melhor motivada (ver Santos et al. (2008d) e sobretudo o capítulo 5), aliás em paralelo com a especificação da tarefa: `CATEG` e `TIPO` passaram a ser opcionais e EM uma marcação a nível mais elevado, como já mencionado.

Outra questão que facilita a referência posterior e a discussão de exemplos concretos é o facto de cada EM ter um identificador único.

Também nos orgulhamos, do ponto de vista linguístico, de termos avançado significativamente na descrição das EM em português, em particular na especificação das combinações de `ALT` sistemáticas, apresentadas no apêndice D. Estabelecemos assim uma primeira lista, com base em corpos, de combinações entre EM de diversos tipos que nos pareceram produtivas em português.

Isto relaciona-se com o facto de termos alargado a interpretação dos `ALT`, que passaram a identificar consistentemente todas as EM possíveis e não apenas a maior. Além disso, a avaliação dos `ALT` deixou de ser feita por critérios quantitativos cegos em termos de fracção do número de palavras coincidentes, como acontecia no Primeiro HAREM (Santos et al., 2007), para passar a sê-lo em termos do conteúdo previamente anotado.

Finalmente, o termos facultado material de treino para as três tarefas também permitiu uma definição mais clara do que se esperava dos sistemas.

Contudo, não queremos de forma alguma dar a ideia de termos coberto todos os pormenores ou questões levantadas pela análise do material. Em particular, há duas áreas em que estamos conscientes de que é preciso mais trabalho, nomeadamente o que respeita a coordenação de EM, e a classificação de entidades que se refiram a meios de comunicação social.

Tabela 6.1: Contabilização da informação adicional no Segundo HAREM, nas três pistas

Número de subtipos marcados	2480 [7769]
Número de ALT	255
Número de EM de TEMPO só em minúsculas	426
Número de EM do TEMPO com atributos do tempo estendido	192
Número de atributos do TEMPO estendido	372
Número de relações semânticas (entre facetas)	614

6.1.6 Recursos mais ricos, mais bem revistos e documentados

Talvez o resultado com maior impacto deste Segundo HAREM sejam os recursos linguísticos criados, que passaram por um crivo apertado quer da equipe da organização quer, em alguns casos, dos próprios participantes.

Em primeiro lugar, houve uma grande preocupação de fundamentação da tarefa e de melhoria e correcção dos problemas identificados nas CD do Primeiro HAREM.

Depois, podemos afirmar que as CD foram muito bem revistas (sob vários ângulos e por várias pessoas), e que as opções tomadas (após a definição das tarefas) foram bem documentadas. De facto, houve muita revisão e consideração das divergências, linguísticas e de interpretação, que os textos suscitaram, tendo-se procedido aliás à compilação de diversa informação para estudos futuros, relativa a dúvidas e discordâncias, e que foi na sua maioria tornada pública na LÂMPADA. A maior parte dos casos problemáticos, aliás mais uma vez seguindo as recomendações do Primeiro HAREM (Santos e Cardoso, 2007b), foi marcada (em 68 casos) como OMITIDO, de forma a não basearmos a comparação dos sistemas em casos desviantes.

Outra melhoria relativamente evidente, mas que nos parece útil não passar em branco neste balanço, é o facto de termos marcado os recursos com bastante mais informação do que no Primeiro HAREM.

Estamos a referir-nos não só ao facto de as categorias LOCAL e TEMPO terem subtipos que foi necessário preencher, como ao facto de que, devido às novas directivas do TEMPO terem um critério muito mais abrangente (não exigirem algarismos ou nomes próprios), ter havido um número muito maior de EM a classificar. Na tabela 6.1 apresentamos uma contabilização da informação adicional presente no conjunto das três CD como um todo.[3]

6.1.7 Cenários selectivos melhor aproveitados

Outra melhoria a que já fizemos menção no capítulo anterior é termos levado até às últimas consequências a questão dos cenários selectivos neste Segundo HAREM, ou seja, foi finalmente implementada a visão dos cenários selectivos (realizados pelo Véus) como constituindo ontologias distintas.

Assim, além de ser possível, como no Primeiro HAREM, comparar cada sistema segundo as suas próprias condições, os chamados cenários selectivos de avaliação possibilitaram a comparação dos vários sistemas entre si, ao fornecer a possibilidade de ver **todos** os sistemas segundo todos os ângulos individuais (representados pelos cenários selectivos de participação de cada sistema).

[3] No caso dos subtipos, o número dentro de parêntesis rectos indica o número de subtipos se se tiver em conta a vagueza.

A Figura 6.1 ilustra o caso de um sistema que participa num determinado cenário (o seu cenário selectivo de participação, à direita) a ser avaliado noutro cenário de avaliação (proposto por outro sistema, por exemplo, ou considerado de interesse por outra razão – indicado pelas caixas de contorno sólido na ontologia que se encontra do lado esquerdo). De acordo com esse cenário de avaliação, o sistema seria apenas avaliado relativamente aos elementos da ontologia que têm igualmente contorno sólido.

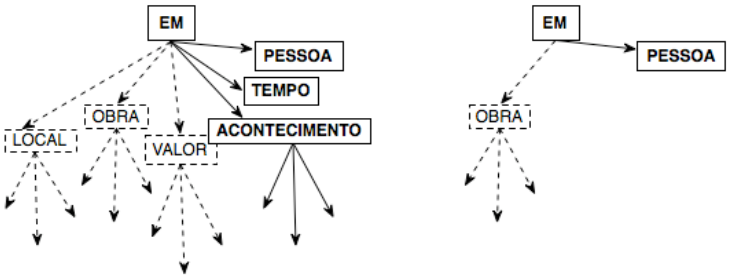

Figura 6.1: Cenários selectivos de participação e de avaliação vistos como alinhamento de ontologias

6.1.8 Potencialidades de investigação do valor do REM noutras áreas

Além de um recurso valioso para REM, a constituição da nova colecção do HAREM e dos resultados dos sistemas – disponibilizados na LÂMPADA – permite fazer investigação com base num recurso comum em várias áreas afins, nomeadamente:

- recolha de informação geográfica (RIG)
- resposta automática a perguntas (RAP)

Como brevemente referido no capítulo 1, a maior parte dos textos que constituem a colecção foram retirados da colecção CHAVE (Santos e Rocha, 2005), melhor dizendo do monte de documentos atribuído aos tópicos do GeoCLEF 2007 (Mandl et al., 2008).[4] Criámos assim uma colecção em que, para cada tópico, os sistemas anotaram com EM (e talvez também com relações) todos os documentos relevantes, assim como alguns dos irrelevantes que pertenciam ao monte. Tal colecção permite medir, de forma rigorosa, a (ir)relevância do REM para RIG, se se contabilizar as ocorrências relevantes para a tarefa em jogo.

Uma questão semelhante põe-se em relação à influência ou necessidade de identificar EM nas perguntas, no âmbito da resposta automática a perguntas. Com efeito, é costume assumir que o REM aplicado ao texto das perguntas é uma vantagem óbvia em RAP, muitas vezes trocando a compreensão da pergunta pela própria resposta, já conhecida do sistema.

[4] *Monte* em RI é um conjunto de documentos que foram considerados relevantes para responder a um determinado tópico por um grupo de sistemas participantes numa avaliação conjunta, conjunto esse que é verificado por juízes humanos, que atribuem o julgamento relevante/não relevante a cada documento do monte. Veja-se Rocha e Santos (2007a) e Gonzalez et al. (2007).

Quisemos neste HAREM investigar a possibilidade de anotar as perguntas sem contexto (assumindo ignorância total em relação à resposta), não só para ver até onde conseguíamos ir, como para fornecer exemplos mais diversificados de casos onde os sistemas REM são chamados a dar uma classificação, como é o caso do QA@CLEF (Giampiccolo et al., 2008). Os resultados da anotação humana estão na colecção dourada e podem ser investigados por todos os interessados nesta problemática. (Trabalho relacionado no âmbito da RAP é por exemplo Roberts e Hickl (2008), que definem uma hierarquia complexa de tipos de respostas.)

6.1.9 Ferramentas para auxiliar o HAREM

Pensamos também ser necessário salientar que o trabalho de organização do Segundo HAREM deu origem a uma maior e mais sofisticada panóplia de ferramentas, sistemas e serviços que foram disponibilizados aos participantes e ao público em geral.

Em primeiro lugar, foi posto à disposição de todos os participantes (e do público em geral) um validador (como serviço na rede) que permitia o teste atempado da sintaxe das saídas dos sistemas e das suas consequentes participações no HAREM.[5]

Em segundo lugar, todo o processo de envio de participações para o Segundo HAREM também foi monitorizado e apoiado por um sistema automático.[6]

Em terceiro lugar, após essa ideia ter surgido durante o Encontro do Segundo HAREM, foi desenvolvido um sistema na rede que permite fazer o teste e avaliação posterior (de acordo com as CD do Segundo HAREM) de novas participações (não oficiais).[7]

Finalmente, e como já referido, foi tornada pública uma ferramenta para ajudar à edição e criação de colecções douradas, o Etiquet(H)AREM, que foi, além disso, usada no âmbito da criação das próprias CD.[8]

6.2 Pista do TEMPO: algumas observações

Apesar de no capítulo 2 ter sido feito um balanço desta pista pelos proponentes, julgamos ser necessário chamar a atenção para outros aspectos ligados à vertente organizativa que nos coube em mãos.

Já foi identificado como um aspecto problemático neste Segundo HAREM (cf. capítulo 3) o facto de os proponentes da pista do TEMPO não terem levado a cabo a própria criação dos recursos e dos programas de avaliação, o que trouxe à organização do HAREM não só bastante trabalho adicional, mas sobretudo bastante insegurança sobre a própria anotação realizada.

Não vamos portanto referir outra vez esta questão, a não ser para indicar que tal situação deverá ser evitada de futuro. De facto, por muita boa vontade que um determinado grupo, neste caso a Linguateca, possa ter em prestar um bom serviço à comunidade, parece-nos pouco apropriado (digamos até mesmo desmotivante) criar recursos de avaliação e fazer escolhas que vão contra as nossas próprias opiniões, como aconteceu com esta pista.

Consideramos, pois, que os proponentes de pistas independentes da pista geral (em termos de filosofia e objectivos a atingir) deverão: ou deixar a organização ter a última

[5] Este validador foi desenvolvido por David Cruz e Luís Miguel Cabral.
[6] O sistema foi desenvolvido por Luís Miguel Cabral.
[7] Este sistema foi desenvolvido por Nuno Cardoso.
[8] Essa ferramenta foi desenvolvida por Hugo Gonçalo Oliveira.

palavra na decisão dos critérios, ou serem eles a tomar em ombros toda a sua organização (nomeadamente no que se refere à criação e anotação de recursos linguísticos e criação e implementação dos próprios programas de avaliação).

Dito isto, gostávamos de mais uma vez salientar que tal não é uma crítica aos três proponentes do grupo do TEMPO – que desde o início, aliás, nos disseram claramente que seriam participantes – e aos quais estamos gratos por esta colaboração. Esta conclusão é simplesmente fruto de uma experiência que não poderíamos ter, nem eles, antes de a levar a cabo.

Pensamos ser contudo inegável que a pista do TEMPO enriqueceu o HAREM e que levou a um progresso notável na descrição das expressões temporais e sua normalização em português.

6.3 ReRelEM: primeiro balanço

Se apreciarmos agora a pista do ReRelEM, temos de concordar a posteriori com a Renata Vieira de que enveredámos por uma tarefa demasiado ambiciosa no âmbito de um piloto.

Com efeito, carregámos com as complexidades inerentes ao HAREM, sem tentar proceder a qualquer tipo de simplificação: o `ALT`, a vagueza, os cenários selectivos do HAREM, etc. e definimos sobre essa tarefa, já de si complexa, uma outra completamente nova.

Além disso, tivemos de comparar participantes muito diversos. Diríamos mesmo completamente distintos (ver capítulo 4). Enquanto o REMBRANDT seguiu à risca o que esperávamos, os outros dois sistemas participantes divergiram substancialmente, e de forma completamente inesperada para a organização. Por um lado, o SeRELeP não enviou a classificação (apenas tentou identificar as relações); por outro, o SEI-Geo escolheu apenas uma relação (`inclui`), competindo portanto num cenário selectivo do ReRelEM, e além disso sem identidade.[9]

Com esta variedade de participação, cedo nos demos conta de que a forma de avaliação que tínhamos inicialmente proposto era demasiado ingénua e simples, e que muitas outras questões tinham de ser equacionadas. Revemos assim aqui as principais questões discutidas entre os membros da organização[10], algumas das quais ainda sem resposta.

6.3.1 A expansão das participações

Ao aplicarmos as regras associadas a cada relação à participação de um dado sistema, podemos estar a desdobrar os seus erros em muito mais relações erradas. Será que vale a pena fazer a avaliação também sem expansão, assumindo que apenas o que é explicitamente marcado pelo sistema deve ser pontuado, ou esperando que os sistemas tenham, eles próprios, o seu mecanismo de expansão? Após acesa discussão, acabámos por considerar que, de acordo com os pressupostos do HAREM, em que o que interessa é a semântica, teríamos de expandir (levar às últimas consequências em termos de compreensão do texto) tanto as corridas dos sistemas como a CD.

[9] De facto, ao conceber o ReRelEM como uma extensão de detecção de co-referência, ou seja, marcação de identidade entre EM, não tínhamos considerado sequer a possibilidade de haver sistemas que não tratassem primeiro da identidade.

[10] Convém ressalvar que esta discussão se deu internamente e não conjuntamente com os participantes, como seria desejável, uma vez que decorreu durante o processo de anotação da CD e de desenvolvimento dos programas de avaliação.

6.3.2 Relação com a vagueza

Embora inicialmente tivéssemos falado em relações entre EM, demo-nos conta de que diferentes facetas de uma EM vaga poderiam entrar em relações distintas com outra EM, e que a única forma de ter tudo convenientemente anotado era, no caso das EM vagas, descer ao nível da faceta durante a anotação da CD do ReRelEM.

Mencionamos novamente esta problemática, já discutida no capítulo 4, porque deu origem à repetição total do trabalho de anotação da CD do ReRelEM, contrastando com o tratamento dos `ALT`, referido nas próximas linhas.

Além disso, pôs de certa forma em causa precisamente a nossa escolha inicial de ter um único ID por EM e não por faceta, o que é algo que terá de ser mais bem pensado em futuras edições, além de ter exigido mais um conjunto de ferramentas para lidar com esta situação.

6.3.3 O que fazer aos ALT?

Embora tenhamos conseguido um processo satisfatório, ainda que trabalhoso, de lidar com a vagueza da classificação no ReRelEM, o mesmo não aconteceu em relação aos `ALT`. Ou seja, não respondemos ainda de forma conclusiva à pergunta: como é que a formulação de alternativas de identificação (que muitas vezes também redundam em mudanças de classificação) interage com a especificação de relações?

A única coisa que nos pareceu sempre clara é que não fazia sentido a declaração de relações entre alternativas de um mesmo `ALT`. Mas a interacção entre um texto marcado com `ALT` e a formulação de relações entre essas EM e o resto do texto não foi ainda considerada seriamente de um ponto de vista linguístico, e constitui naturalmente trabalho de reflexão futuro.

De facto, neste primeiro ReRelEM limitámo-nos a aceitar como certas todas as relações, independentemente de estarem "repetidas" dentro de diferentes alternativas ou não. Ou seja, se *Universidade de Lisboa* aparecesse como EM duas vezes em alternativas diferentes dentro de um mesmo `ALT`, e se em ambas as vezes estivesse relacionada com outra EM (por exemplo *Universidade*, algumas frases mais tarde) essa relação seria contada como certa, ou como errada, duas vezes.

6.3.4 O que fazer a participações inconsistentes?

Outra questão que se nos pôs foi como lidar com a marcação de relações que, depois de expandidas, levassem a uma contradição.

Neste primeiro ReRelEM, para o cálculo dos resultados ignorámos completamente esse aspecto (simplesmente fazendo a expansão enquanto for possível), mas estamos plenamente conscientes de que ainda falta especificar o que fazer, e como pontuar, nesses casos.

De momento apenas conseguimos postular que as EM em questão deviam ser consideradas negativamente para atribuir a classificação à tarefa do ReRelEM, mas, naturalmente, é necessário precisar a forma como isso deve ser feito, e desenvolver mais uma ferramenta auxiliar de detecção de inconsistências entre relações marcadas[11].

[11] Uma funcionalidade embrionária já se encontra no `Expandidor`, que detecta alguns tipos de inconsistências se invocado com a opção `-ver_inconsistencias`.

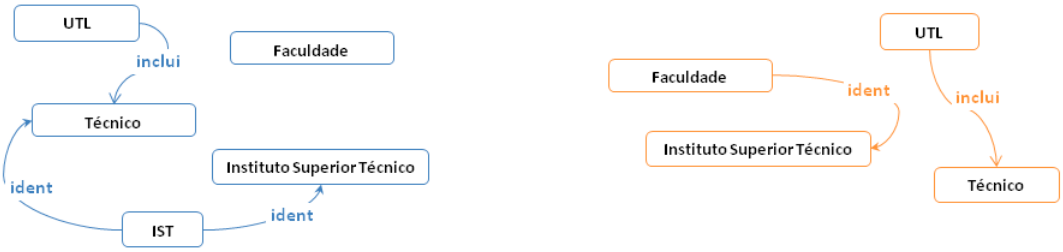

Figura 6.2: Exemplo de relações na CD e numa participação

6.3.5 Que sentido faz a comparação?

Embora tal seja relativamente evidente, pôs-se-nos com mais acuidade a pergunta fundamental de como comparar o incomparável: se um sistema só procura reconhecer a identidade e outro só reconhece a localização, o que têm os dois em comum? Ou melhor, o que têm as duas relações em comum para poderem ser comparadas?

O problema aqui é ainda mais agudo do que no HAREM em geral, porque as próprias relações podem implicar graus de complexidade muito diferentes.

6.3.6 A identidade é diferente?

Finalmente, uma das considerações que nos tomou mais tempo, e à qual acabámos por não dar seguimento, foi a intuição de que a relação de identidade era diferente e devia ser separadamente analisada, antes de pontuar as outras relações. Com base nessa ideia, aplicámos uma medida de avaliação do agrupamento[12] obtido a partir das relações de identidade, para medir o desempenho de um sistema quanto ao reconhecimento da identidade (por outras palavras, para medir os grupos obtidos a partir das relações de identidade propostas pelo sistema comparando-os com os grupos obtidos a partir da colecção dourada).

O problema é que, depois desse primeiro agrupamento (em que substituíamos as EM pelo grupo a que pertenciam), não conseguimos fazer sentido das relações (que não a identidade) propostas pelos sistemas, e compará-las com as relações na CD.

Veja-se a figura 6.2, com um exemplo de relações marcadas na CD e numa participação fictícia.

A figura 6.3 apresenta o resultado do agrupamento para o caso da figura 6.2. Se substituirmos as EM por um representante do agrupamento, como representar, e pontuar, por exemplo, a relação entre *UTL* e o *Técnico* proposta pelo sistema? De facto, como ilustra a figura 6.4, o agrupamento contituído por *Técnico* na participação faz parte do agrupamento que está envolvido na relação de inclusão na CD, o que poderia ser suficiente para considerar a relação como correcta. No entanto, em vez de um agrupamento estar incluído no outro, poderíamos ter uma sobreposição parcial entre agrupamentos. Continuando com a nossa participação fictícia, imagine-se que o sistema tinha estabelecido a relação *UTL* `inclui` *Instituto Superio Técnico* em vez de *UTL* `inclui` *Técnico*. Será que nesse caso a relação também deve ser considerada correcta?

[12] Agrupamento é a nossa tradução para o termo inglês *clustering*.

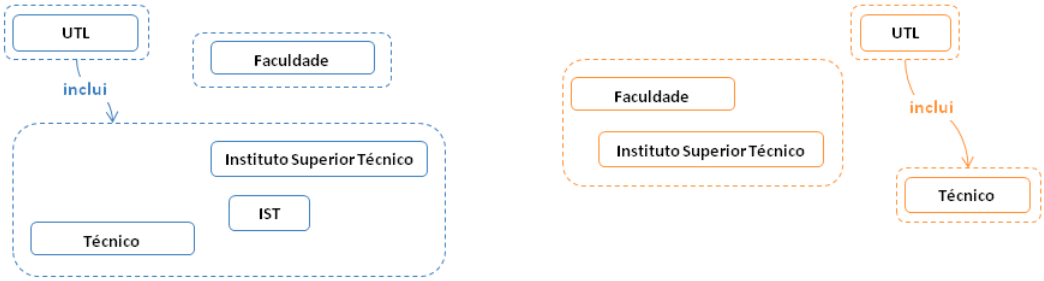

Figura 6.3: Exemplo de agrupamento na CD e numa participação

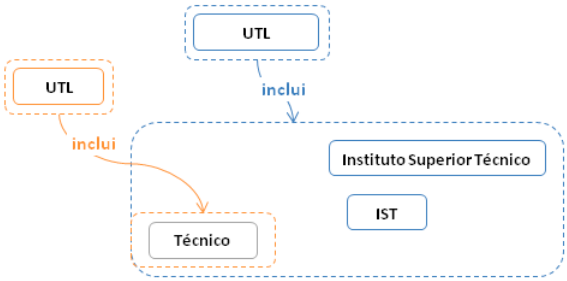

Figura 6.4: Tentativa de avaliação de relações (que não a identidade) com base em agrupamentos

Não tendo conseguido, de facto, arranjar uma solução satisfatória para o emparelhamento dos grupos e para a consequente avaliação das relações (não identidade) entre eles, acabámos por desistir de tratar de forma diferente a identidade.

Contudo, se a única relação do ReRelEM fosse esta (encontrando-nos portanto em presença de uma avaliação de co-referência simples), as medidas do agrupamento ainda nos pareceriam uma boa alternativa à opção tomada, e colocamo-las aqui na figura 6.5 para delas dar conta aos leitores.

6.3.7 Progresso na área da semântica computacional

Embora tendo tropeçado em várias dificuldades imprevistas, o que se reflectiu no atraso da publicação dos resultados (pode dizer-se que durante alguns meses fomos aumentando a sofisticação do tratamento das corridas quase de semana para semana), não podemos deixar de nos orgulhar por termos proposto, e avaliado, uma tarefa mais complicada do que qualquer outra por nós conhecida em termos de avaliação conjunta de qualquer língua.

Estamos convencidos de que a tarefa exploratória do ReRelEM que levou à explicitação dos tipos de relações entre EM é, do ponto de vista linguístico e computacional, inovadora (veja-se o capítulo 4), assim como as decisões de avaliação, embora preliminares.

Também produzimos material que é interessante estudar em profundidade, e até cruzar informação entre as várias pistas ou tarefas, embora o tamanho do recurso dourado para o ReRelEM seja inquestionavelmente pequeno.

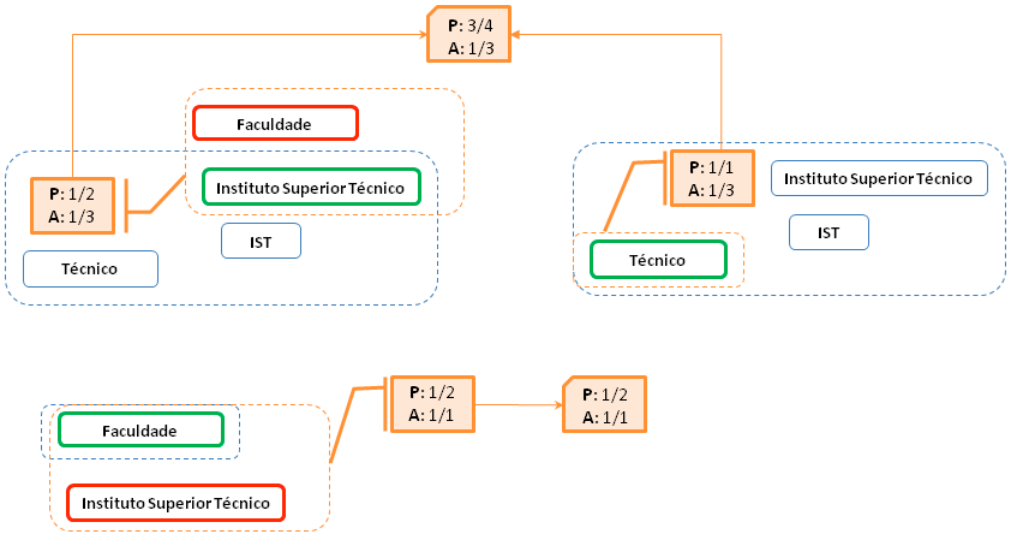

Figura 6.5: Classificação da medida de agrupamento (avaliando só identidade)

6.4 O HAREM tem futuro?

Nesta secção, um pouco distinta das anteriores, porque olhamos para o futuro e não para o passado, tentamos equacionar as vantagens de um terceiro HAREM e as recomendações que podemos deixar para os futuros organizadores, como foi feito em Santos e Cardoso (2007b).

A primeira pergunta a formular é: existe uma comunidade de REM em português que pretende continuar? Ou de cada vez que uma nova avaliação conjunta ocorre aparecem maioritariamente novos grupos, porque os anteriores já progrediram para outras tarefas?[13] Dado que não houve quase nenhuma sobreposição entre o Primeiro e o Segundo HAREM em termos da massa dos participantes, a pergunta parece-nos pertinente. De certa forma, o facto de termos tido vários novos participantes (e até bastante interesse internacional, que infelizmente depois não se concretizou em termos de participação), é cancelado pelo facto de que muitos antigos participantes não responderam à chamada.

Contudo, e dada a conclusão (em ambos os Encontros) de que a maioria dos grupos estaria interessada em continuar, seria necessário organizar uma terceira edição de avaliação para tirar a prova.

Muito brevemente, façamos contudo uma panorâmica da comunidade que respondeu presente neste HAREM, para indagar, pelo menos parcialmente, quais os objectivos científicos de cada participante:[14]

- a comunidade de RI(G) parece ser a mais estável, com três participantes (Cage, SEI-Geo e REMBRANDT, embora só o primeiro tenha participado no Primeiro HAREM)

[13] De acordo com a formulação do Marcirio Chaves, porque consideram que a área do REM já se encontra resolvida para o português.

[14] Não estamos naturalmente a dizer – o que seria trivialmente falso, dado o HAREM – que estas comunidades não se intersectam! Pelo contrário, a maior parte dos sistemas pode dizer-se que pertence a mais do que uma.

- a comunidade de resposta automática a perguntas (RAP) teve grande protagonismo e dos melhores resultados (Priberam e XIP-L2F/XEROX)
- a comunidade do processamento do tempo juntou-se ao HAREM (XIP-L2F/XEROX e PorTexto)
- a comunidade de extracção de informação (EI) também se mostrou interessada, com quatro participantes (REMMA, R3M, DobREM e SEI-Geo)
- a comunidade da co-referência também participou com bons resultados, embora só com um participante (SeRELeP)

Primaram pela ausência, contudo, alguns dos principais actores na semântica computacional do português, e em particular ambos os vencedores da primeira edição, o PALAVRAS-NER e o Cortex. Embora ambos tenham invocado falta de tempo, ou melhor, outras prioridades, é significativo que não achassem que valia a pena tornar a participar. De facto, apenas um sistema repetiu a participação, o CaGe, embora dois participantes antigos aparecessem com novos sistemas (SEI-Geo e R3M), o que é naturalmente positivo.

Outra questão que nos preocupa é a cada vez menor participação de grupos brasileiros nas actividades da Linguateca. Mais uma vez só tivemos um grupo do Brasil[15], apesar de termos sempre o maior cuidado em organizar as avaliações de forma a que as duas variantes estivessem igualmente bem representadas. Isto pode dever-se ao calendário limitado com que fomos forçados a organizar este Segundo HAREM, mas é algo que devemos considerar com mais cuidado em futuras avaliações.

Além disso, é preciso também reconhecer que a participação nas discussões de preparação ficou muito aquém das nossas expectativas. A maior parte dos participantes não quis fazer jus à caracterização "conjunta", que devia ser parte integrante de uma avaliação conjunta, preferindo aceitar as regras (de qualquer das tarefas) sem debate. Muito provavelmente seria necessária uma reunião presencial (como foi o caso nas Morfolimpíadas (Costa et al., 2007)) para pôr todas as pessoas à volta de uma mesa a discutir casos concretos.

Obviamente que outra pergunta associada a uma eventual continuação do HAREM é a de identificar diferentes alternativas de continuidade. Visto que todos os recursos são finitos, porque não organizar (ou participar, conforme o ponto de vista) uma avaliação noutra área? Por exemplo integrando como parte ou constituinte o próprio REM, mas não o fazendo o objecto principal... Ou seja, porque não fazer o processo inverso do MUC, que começou com extracção de informação em geral e especializou para tarefas mais delimitadas? O HAREM poderia ter começado com essas tarefas mais delimitadas e desenvolver no sentido de extrair mais informação, em forma de gabaritos (as "templates" do MUC).

Ainda outra forma de evoluir/mudar o enquadramento do HAREM seria acoplá-lo ou associá-lo a uma avaliação internacional que contivesse mais línguas. Nesse caso a comunidade óbvia seria o CLEF (Rocha e Santos, 2007a; Braschler e Peters, 2004).

Vamos contudo nas linhas que se seguem assumir que irá existir um Terceiro HAREM – ainda e só para o português – e cujo foco seja ainda a classificação de EM e de relações entre elas, para podermos fornecer algumas recomendações para a futura edição, com base na nossa experiência:

[15] No Primeiro HAREM, tivemos apenas um grupo brasileiro, o CorTex, que foi aliás o vencedor do Mini-HAREM, embora inicialmente mais grupos tenham indicado interesse, aliás como na presente edição.

- Parece-nos aconselhável manter as tarefas do HAREM clássico apenas com modificações pontuais (marcando-as claramente nas directivas anteriores, mas refazendo-as e publicitando-as com tempo para haver uma discussão até presencial das mesmas) para permitir uma comparação de progresso do Segundo para o Terceiro HAREM.
- Uma reunião presencial de discussão, ou mesmo várias, parece obviamente importante para obter um consenso inicial, assim como esclarecer muitas coisas que podem não ser óbvias a participantes pela primeira vez.
- Sugerimos que os participantes sejam envolvidos na escolha dos textos que pertencem à colecção do Terceiro HAREM (embora a escolha dos textos da CD tenha de ser secreta e feita pela organização), para permitir que essa colecção responda aos interesses de investigação da comunidade.
- Interessaria obter uma "garantia" de participação, ou um prémio de participação, que diminuísse o grau de desistência dos participantes inscritos. Uma hipótese de "garantia" poderia ser o enviarem-nos uma versão do seu sistema que seria usado se não conseguissem participar à última hora.

Seja como for, não nos parece necessário nem apropriado começar desde já a organizar uma nova iniciativa neste campo.

De facto, estamos conscientes de que os próprios resultados postos à disposição de toda a comunidade permitem, ou mesmo exigem, estudos aprofundados, que vão desde validação estatística a comparação entre as duas edições do HAREM, antes de ser apropriada a organização de uma nova edição.

Esperamos por isso que muitos investigadores possam beneficiar do trabalho já feito e identificar questões interessantes em relação ao processamento semântico da nossa língua, além do mero treino e desenvolvimento de melhores sistemas para esta tarefa específica. Tanto do lado mais linguístico da descrição da língua, como do lado mais computacional do desenvolvimento de ferramentas para explorar os recursos complexos criados pela anotação humana, como do lado da metodologia de avaliação e da reflexão sobre as conclusões estatitisticamente válidas sobre o desempenho dos sistemas, muito ainda há para fazer.

Agradecimentos

Agradecemos a Jorge Baptista, Marcirio Chaves e Mírian Bruckschen os comentários a versões anteriores deste capítulo, que nos ajudaram a melhorá-lo significativamente.