Capítulo 7

Balanço do Primeiro HAREM e perspectivas de trabalho futuro

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 7, p. 87–94, 2007.

Neste capítulo iremos analisar em pormenor algumas opções tomadas no início do HAREM e que, vendo agora em retrospectiva, constatamos que errámos ou que não escolhemos, pelo menos, a alternativa mais apropriada.

Globalmente, fazemos um balanço francamente positivo do HAREM, não só pela participação e entusiasmo da comunidade em relação à iniciativa, mas também por ter levado a bom porto a avaliação em REM idealizada pelo estudo preliminar descrito no capítulo 2. Questões como a vagueza, a anotação em contexto, a adopção de uma categorização semântica consensual, ou a utilização de textos de diferentes proveniências e variantes, foram pela primeira vez introduzidas em avaliações conjuntas de REM. Adicionalmente, fomentámos a discussão no seio da comunidade, em torno da melhor metodologia de avaliação dos seus sistemas, o que resultou em contribuições importantes e que, acreditamos, fará do HAREM uma referência importante para avaliações conjuntas futuras na área.

O capítulo começa com uma autocrítica ao HAREM, referindo alguns tópicos sobre os quais temos actualmente uma opinião diferente em relação ao que foi feito. Esta análise pretende garantir que essas opções sejam documentadas e corrigidas em próximas avaliações conjuntas no âmbito do HAREM, fomentando uma reflexão da comunidade em seu redor. De seguida, apresentamos algum trabalho que achamos que será da maior utilidade efectuar, com base naquilo que foi feito no HAREM e no Mini-HAREM, antes de começar a organizar novas rondas de avaliação conjunta, mesmo que tal implique algum atraso na organização da segunda edição. Só nessa altura fará sentido, na nossa opinião, escolher o caminho futuro a seguir como comunidade, sobre o qual fazemos alguns comentários na terceira parte.

7.1 Uma retrospectiva das opções tomadas

7.1.1 Uma dependência infeliz entre a classificação e a identificação

Uma das opções que hoje admitimos não ter sido feliz diz respeito à separação da tarefa de REM em dois passos: identificação e classificação. Esta modularidade, apesar de ser interessante e até ter permitido que outros participantes pudessem também aproveitar o HAREM para tarefas relacionadas, como foi o caso nas Morfolimpíadas (Santos et al., 2003; Costa et al., 2007), em que além de analisadores morfológicos também participaram radicalizadores e verificadores ortográficos, transmitiu infelizmente uma aparência de independência entre os passos que na realidade não existiu, se tomarmos em conta a forma como as categorias do HAREM foram concebidas.

Ou seja, ao termos considerado que a delimitação correcta de certo tipo (semântico) era `COISA` ou `VALOR`, estamos já, ao nível da tarefa da identificação, a pressupor (nesses casos) uma classificação implícita correcta para podermos atribuir uma identificação correcta, dado que definimos directivas de identificação separadas para essa categoria (ver

capítulo 18):

```
Comprei uma flauta <COISA TIPO="CLASSE">de Bisel</COISA>
Tem um comprimento de <VALOR TIPO="QUANTIDADE">60 metros</VALOR>.
```

Embora isto não aconteça na maioria dos casos[1], ou seja, para as outras categorias a independência é real, essa dependência invalida conceptualmente a separação.

7.1.2 Avaliação da identificação baseada em categorias de classificação

Um outro ponto em que foi nebulosa a contribuição do HAREM foi a nossa escolha de apresentar relatórios de desempenho cujas medidas de identificação se encontravam discriminadas por categoria (semântica), o que produziu uma confusão generalizada entre os participantes. A necessidade ou mesmo o interesse de efectuar e apresentar esse tipo de relatórios precisa assim de ser repensada.

A ideia subjacente à geração desses relatórios era a seguinte: em paralelo com a apresentação de resultados considerando, por exemplo, apenas texto literário, ou apenas da Web, ou apenas da variante brasileira, era possível também mostrar os resultados segundo os vários conjuntos de categorias: só pessoas, só obras, só coisas, etc, e fazer as mesmas medições. Isto é equivalente a um participante apenas escolher uma categoria para concorrer, aplicando-se um véu que retirava todas as outras categorias (ver secção 19.2.5).

O que não foi compreendido pela maioria dos participantes foi que isso não significava filtrar apenas os casos em que a CD continha EM classificadas como PESSOA, mas sim entrar em conta, também, com todos os casos erradamente marcados como PESSOA pelos sistemas (ou seja, EM espúrias), o que significa que, ao contrário dos casos da variante ou do género textual, usados por todos os sistemas (e discriminados depois nos relatórios de desempenho), as medições por categoria dependem da saída de cada sistema e podem portanto não ser uma forma fácil de comparar os sistemas entre si.

A Tabela 7.1 exemplifica, usando a categoria PESSOA, todos os casos que são levados em conta para as várias pontuações por categoria. Para a tarefa de **identificação** em relação à categoria PESSOA, os casos 1, 2 e 3 são considerados correctos, enquanto que os casos 6, 7 e 8 são considerados parcialmente correctos. Para a tarefa de **classificação** da categoria PESSOA, já apenas o caso 1 e (parcialmente) o caso 6 são correctos. Além disso, a diferença entre os cenários relativo e absoluto (como sempre), é que o primeiro não considera no denominador casos espúrios e em falta, como por exemplo os casos 4 e 5 (veja-se a explicação detalhada dos diferentes valores destas medidas no capítulo 18).

[1] Por exemplo, na primeira CD, num universo de 5086 EM, há 7 casos de COISA com o padrão acima, 4 distintas. Para VALOR, há 132 ocorrências tal como o padrão de cima para unidades tais como metros, kg, escudos ou bits, sendo 106 dessas ocorrências unidades temporais (anos, meses, dias ou minutos).

Caso	CD	Sistema	Comentário
1	PESSOA	PESSOA	identificação e classificação correctas
2	X	PESSOA	o sistema identifica uma EM como PESSOA que na CD é diferente
3	PESSOA	X	o sistema identifica uma EM PESSOA como outro tipo de EM
4		PESSOA	o sistema identifica uma EM espúria como PESSOA
5	PESSOA		o sistema não identifica como EM uma PESSOA na CD
6	PESSOA	PESSOA	apenas parcialmente identificada, e class. semântica correcta
7	X	PESSOA	apenas parcialmente identificada, e class. semântica espúria
8	PESSOA	X	apenas parcialmente identificada, e class. semântica em falta

Tabela 7.1: Todos os casos relacionados com a avaliação da identificação por categoria PESSOA. X significa o nome de uma categoria diferente de PESSOA.

7.1.3 Cenários relativos vistos por outra perspectiva

Outra questão pode ser levantada em geral em relação à pertinência de definir um cenário relativo: se, de facto, como constatámos acima, em alguns casos as duas tarefas (identificação e classificação) não são independentes, isso retira (pelo menos nesses casos) o sentido a tal cenário. Parece ser portanto mais correcto usar apenas o cenário absoluto para avaliar os sistemas, dado que as medidas relativas são de certa forma virtuais, e os sistemas na prática têm de efectuar ambas as decisões até à marcação final da EM (ou melhor, as decisões não são independentes).

Note-se, aliás, que se tornam aparentes mais duas desvantagens do cenário relativo: uma, foi talvez ter induzido os sistemas em erro devido à aparente independência conceptual entre as duas tarefas. Outra, foi a possibilidade de introduzir um elemento de "adaptação ao HAREM": um sistema com dúvida numa dada categoria teria melhores resultados no HAREM (cenário relativo) não a reconhecendo do que tentando classificá-la. Pensamos que ninguém se aproveitou desta característica, mas é uma indicação de que não há vantagem em definir artificialmente um cenário que não representa (e consequentemente mede) uma tarefa independente.

7.1.4 Inconsistência nas medidas usadas

Outra questão refere-se às medidas: Embora nos tenhamos concentrado na capacidade de discriminação dentro de cada categoria, entrando em conta com a quantidade de informação que cada **tipo** (ou conjunto de tipos) implicava, ficou por fazer uma medida que entrasse em conta com a capacidade de discriminação entre **categorias**, e que é claramente mais interessante do ponto de vista de medir a dificuldade da tarefa de REM em português.

Uma outra área com clara potencialidade de melhoria refere-se à classificação de EM com alternativas de delimitação e/ou de encaixe, com a respectiva classificação de parcialmente correcto. Embora tenhamos argumentado em Santos et al. (2006) a favor da existên-

cia da classificação parcialmente correcta em vez de um "tudo ou nada" como preconizado pelo MUC, é claro que há casos em que tal faz mais sentido do que outros. Ou seja, pode haver EM disparatadas que recebem no HAREM uma gratificação que não merecem, enquanto que outras são desvalorizadas (pelo tamanho) embora com muito mais significado intrínseco. Apresentamos um exemplo hipotético apenas para ilustrar esta questão:

> As Actas do ETNR do Departamento de Informática do Rio Azul/Brasil e as do PROPOR foram publicadas pela Springer.

Segundo as directivas do HAREM, o exemplo seria anotado da seguinte forma:

> As **<OBRA TIPO="REPRODUZIDA"> Actas do ETNR do Departamento de Informática do Rio Azul/Brasil </OBRA>** e as do **<ACONTECIMENTO TIPO="ORGANIZADO"> PROPOR </ACONTECIMENTO>** foram publicadas pela **<ORGANIZACAO TIPO="EMPRESA"> Springer </ORGANIZACAO>**.

Neste caso, os sistemas que produzissem EM como <EM> Azul/Brasil </EM>, <EM> Informática do Rio </EM> ou <EM> As Actas </EM> não deveriam receber qualquer pontuação, enquanto que aqueles que marcassem <EM> Actas do ETNR </EM> ou <EM> Departamento de Informática do Rio Azul/Brasil </EM> já nos parecem merecer uma pontuação parcial.

7.1.5 Tratamento dos problemas incluídos em texto real

Finalmente, uma questão muitas vezes referida mas que não foi tratada convenientemente refere-se à inclusão de texto real (por exemplo, com erros ortográficos ou com uso indevido de maiúsculas) na Colecção HAREM e na CD. Esses casos deveriam estar marcados, de forma a poderem ser automaticamente ignorados pelos módulos da avaliação. É muito importante sublinhar que consideramos que os sistemas devem ser alimentados com texto real; contudo, nos casos em que não é possível obter um consenso, não se deve favorecer ou prejudicar os sistemas através de uma decisão arbitrária, e por isso a avaliação destes não deve incluir erros ou problemas não resolvidos. Embora tal já tenha sido parcialmente feito através da etiqueta <OMITIDO> na CD (ver capítulo 19), ainda muitos casos ficaram por tratar.

7.2 Receitas para uma nova avaliação conjunta fundamentada

Antes de nos abalançarmos a organizar um novo HAREM, há vários estudos que precisam de ser realizados, de forma a que todo o processo possa ser melhor avaliado, e sabermos que escolhas vale a pena manter e quais as que podemos abandonar ou mudar.

No que se refere à validação estatística do método, já foi feito um trabalho importante (veja-se o capítulo 5 e Cardoso (2006a)); contudo, é ainda preciso esclarecer algumas outras questões conceptuais.

Em alguns casos, isto requer o enriquecimento ou verificação adicional da CD, por isso principiamos por listar o que pretendemos fazer como uma continuação lógica do trabalho de investigação sobre o REM em português:

- marcação da CD por mais investigadores independentes, de forma a medir a concordância inter-anotadores e refinar também a compreensão (e documentação) das directivas. A determinação da concordância inter-anotador permitirá calcular o erro da medição inerente ao erro humano (Will, 1993), e determinar com maior rigor o nível de confiança nos resultados das avaliações (Maynard et al. (2003a) comparam o MUC e o ACE a esse respeito).
- marcação sistemática dos casos problemáticos e com erros, de forma a não serem contados pela arquitectura de avaliação;
- marcação de todas as EM encaixadas;
- marcação com o tipo semântico pormenorizado (país, cidade, jornal, etc) e eventualmente traduzi-lo para um esquema MUC, em que, por exemplo, país e cidade são `LOCAL` independentemente do seu contexto, ou menções a jornais classificadas como `ORGANIZACAO` (ver os capítulos 4 e 3 para explicação detalhada das diferenças entre os tipos semânticos empregues);
- marcação da CD segundo as directivas do ACE;
- marcação de dependências anafóricas.

Talvez a tarefa mais importante que se nos depara é a medição da dificuldade das tarefas, quer através do recurso a um almanaque "ideal", quer através da simplicidade da atribuição de uma dada classificação – e para isto teremos não só que classificar os contextos sintácticos como a possibilidade de encaixe e/ou de ambiguidade das várias EM.

Parece-nos pois interessante estudar meios de realizar uma selecção automática das EM mais difíceis de reconhecer e/ou classificar, e realizar uma nova avaliação (usando os resultados já existentes dos sistemas) segundo este cenário de "elite". A principal intuição subjacente a esta proposta é a de que há tipos de EM (por exemplo, as expressões numéricas) que pouco contribuem para distinguir os sistemas, e que "diluem" os valores dos resultados finais. Ao usar um leque de EM difíceis como um novo Véu (ver secção 19.2.5),será mais fácil distinguir os melhores sistemas, eventualmente para tarefas diferentes.

Outra questão de interesse óbvio é investigar a relação entre a dificuldade de anotação para um sistema automático e para a anotação intelectual. Na pista dessa, e após reanotação da CD, será também preciso comparar, como sugerido no capítulo 4, a dificuldade do esquema MUC com a do esquema HAREM e quantificar, ao mesmo tempo, em quantos casos é que há sobreposição, ou seja, em que a diferença é apenas teórica.

Finalmente, esperamos que a disponibilização pública, quer das CD quer dos resultados dos sistemas, permita estudar métodos de análise sintáctico-semântica que indiquem o tipo ou categoria de forma a podermos compilar semi-automaticamente mais texto, usando por exemplo a Floresta Sintá(c)tica (Afonso et al., 2002; Bick et al., 2007) para texto jornalístico, o COMPARA (Frankenberg-Garcia e Santos, 2002) para texto literário e o BACO (Sarmento, 2006a) (marcado automaticamente com o SIEMÊS (Sarmento, 2006b)) para texto da Web. Estes métodos permitirão não só criar maiores colecções de texto, mais variadas, como também alcançar (se tal for considerado desejável) um determinado balanço entre os vários casos difíceis, em vez de prosseguir uma abordagem cega de apenas mais quantidade de material.

7.3 Alguns futuros possíveis

Esta secção descreve algumas propostas feitas no Encontro do HAREM, dando evidentemente crédito aos seus autores, mas tentando sobretudo fazer um ponto da situação sobre os vários futuros que a comunidade tem à sua frente, convencidos de que o futuro dependerá tanto de nós, organizadores, como da comunidade.

Martins et al. (2006) sugeriram que o significado (ou seja, o resultado da análise semântica), pelo menos das EM geográficas, fosse dado com mais detalhe, ou seja, que além de simplesmente `LOCAL` se indicassem, por exemplo, as coordenadas geográficas. Para uma `PESSOA`, poder-se-ia especificar a data de nascimento, ou até uma pequena biografia; para uma obra, o seu ISBN ou a data da primeira edição; e para uma empresa, o seu número fiscal, por exemplo. Isto tornaria a tarefa mais realista, embora consideravelmente mais específica, e exigiria que os sistemas fizessem uso de almanaques muito maiores.

Sarmento e Mota (2006) sugeriram uma pista robusta, em que as maiúsculas ou minúsculas não importassem (apropriada, por exemplo, à detecção de entidades em texto transcrito automaticamente). De notar que nesse caso estamos a aproximarmo-nos do ACE, em que não só nomes próprios mas quaisquer referências/menções a entidades devem ser marcadas.

Mais uma vez, e embora tal já tenha sido aflorado no capítulo 4, convém relembrar que Mota, Bick, Sarmento e Almeida mencionaram o interesse de fazer algo semelhante ao MUC para poder ser comparável entre línguas – dada a repetição de afirmações como "para o inglês, o problema está resolvido a 95%, para o português ainda vamos a 70%",

afirmações essas que não são rigorosas mas que têm sido repetidamente feitas, como já referido em Cardoso (2006a, p. 85-87).

Pensamos que todos estes futuros (excepto o primeiro) dependem dos resultados das medições mencionadas na secção anterior, que nos permitirão ajuizar: o trabalho necessário, o esforço de anotação envolvido, e a necessidade de reformular ou não a arquitectura de avaliação e de criação de recursos.

Notamos também que, se não nos afastarmos demasiado do que já foi feito, os participantes em edições seguintes de uma avaliação conjunta têm a possibilidade de reutilizar os recursos criados na primeira para o treino dos seus sistemas. Essa é uma consideração que deve ser tida em conta antes de modificações demasiado radicais.

Agradecimentos

Este capítulo foi escrito no âmbito da Linguateca, financiada pela Fundação para a Ciência e Tecnologia através do projecto POSI/PLP/43931/2001, co-financiado pelo POSI, e pelo projecto POSC 339/1.3/C/NAC.