

Capítulo 19

A arquitectura dos programas de avaliação do HAREM

Nuno Seco, Nuno Cardoso, Rui Vilela e Diana Santos

A plataforma de avaliação do HAREM consiste num conjunto de módulos utilizado nas avaliações conjuntas realizadas pela Linguatca para medir o desempenho dos sistemas de reconhecimento de entidades mencionadas (REM) participantes no HAREM. Estes programas foram concebidos de acordo com as directivas de avaliação do HAREM, aprovadas pela organização e pelos participantes, e que republicámos no capítulo 18.

A plataforma foi implementada segundo uma arquitectura modular, onde cada módulo executa uma tarefa simples e específica. O resultado final da avaliação é obtido através da sua execução numa determinada sequência.

A opção por uma arquitectura modular, desenvolvida por quatro autores em locais diferentes, foi motivada pelas seguintes considerações:

- a modularização facilita a depuração dos módulos, assim como a verificação de que o seu funcionamento cumpre as directivas de avaliação do HAREM;
- permite o desenvolvimento descentralizado e cooperativo dos programas, com os vários módulos a serem desenvolvidos por diferentes programadores;
- permite o desenvolvimento dos módulos na linguagem de programação em que o programador se sente mais confortável, visto que os módulos podem ser implementados em linguagens diferentes.

Este documento descreve detalhadamente cada um dos programas que compõem a plataforma de avaliação, já apresentada e motivada em Seco et al. (2006). Começamos por apresentar a arquitectura em termos globais, fornecendo depois a descrição pormenorizada de cada módulo.

19.1 Sinopse da arquitectura

A figura 19.1 apresenta o esquema da arquitectura da plataforma de avaliação do HAREM, indicando os módulos que a compõem, e a forma como interagem. A avaliação do HAREM pode ser dividida em quatro fases:

Fase 1: Extração e alinhamento

A sintaxe das saídas dos sistemas é verificada e corrigida através de um **Validador**. O subconjunto de documentos da saída que também estão presentes na colecção dourada (CD, ver Santos e Cardoso (2006)), é extraído pelo **Extractor de CD**. As EM desse subconjunto são posteriormente alinhadas com as respectivas EM da CD pelo **AlinhEM**, gerando uma lista de *alinhamentos*. O **AvalIDA** processa os alinhamentos e produz os primeiros resultados para a tarefa de identificação.

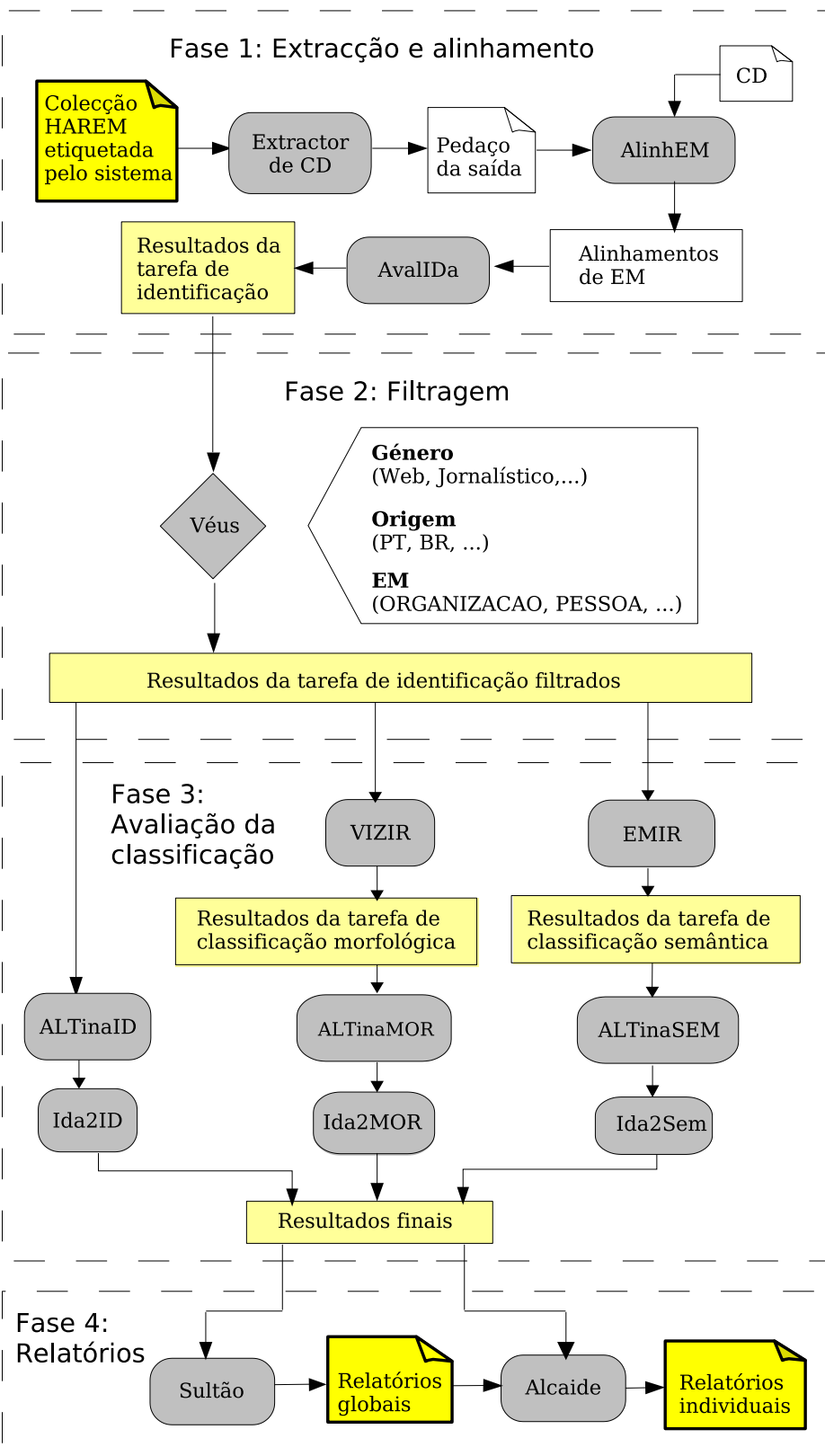


Figura 19.1: Esquema da plataforma de avaliação do HAREM.

Fase 2: Filtragem

A filtragem selectiva dos alinhamentos permite a avaliação parcial segundo diversos cenários específicos. O módulo **Véus** é responsável pela filtragem dos alinhamentos, a partir de uma lista de restrições, que pode incluir um conjunto de categorias e de tipos, um género textual, uma variante ou o resultado da avaliação na tarefa de identificação.

Fase 3: Avaliação da tarefa de classificação

A avaliação das tarefas de classificação morfológica e semântica é realizada em paralelo pelo **Vizir** e pelo **Emir**, respectivamente, a partir dos alinhamentos. O **ALTinaID**, o **ALTinaMOR** e o **ALTinaSEM** analisam as EM vagas em termos de delimitação na CD, e seleccionam as alternativas que conduzem à melhor pontuação para cada saída. Finalmente, o **Ida2ID**, o **Ida2MOR** e o **Ida2SEM** processam os alinhamentos finais e calculam os valores agregados das métricas para as três tarefas, respectivamente.

Fase 4: Geração de relatórios

Os resultados finais da avaliação são compilados em relatórios de desempenho que se desejam facilmente interpretáveis. O **Sultão** gera relatórios globais sobre os resultados de todas as saídas (devidamente anonimizadas), enquanto que o **Alcaide** gera relatórios individuais detalhados para cada saída.

19.2 Descrição pormenorizada de cada módulo**19.2.1 Validador**

O módulo Validador verifica se o formato dos ficheiros de saída enviados durante a análise corresponde ao formato determinado pelas directivas do HAREM. Os documentos incluídos na saída deverão ter a seguinte estrutura, ilustrada abaixo através de uma DTD.

```
<!ELEMENT DOC ( DOCID, GENERO, ORIGEM, TEXTO ) >
<!ELEMENT DOCID ( #PCDATA ) >
<!ELEMENT GENERO ( #PCDATA ) >
<!ELEMENT ORIGEM ( #PCDATA ) >
<!ELEMENT TEXTO ( #PCDATA ) >
```

O formato adoptado pelo HAREM para estruturar os documentos na Colecção HAREM (CH) e nas respectivas CD é o formato SGML.

Veja-se o seguinte exemplo de um documento válido, ilustrando uma saída de um sistema REM que participasse nas tarefas de classificação semântica e morfológica.

```

<DOC>
<DOCID>HAREM-051-00043</DOCID>
<GENERO>Web</GENERO>
<ORIGEM>PT</ORIGEM>
<TEXTO>
<ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S">Lions Clube de Faro</ORGANIZACAO>
DM-115CS
<LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Portugal</LOCAL>
O <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Algarve</LOCAL> , a região mais a sul
do território continental de <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Portugal
</LOCAL>, tem por capital a cidade de <LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">
Faro</LOCAL>.
</TEXTO>
</DOC>

```

O Validador tem em atenção as seguintes questões:

- Um <DOC> deve ser seguido, por esta ordem exacta, pelas etiquetas <DOCID>, <GENERO>, <ORIGEM> e <TEXTO>. Um <DOC> não pode conter outro <DOC>.
- Um <DOCID> deve possuir um único identificador DOCID. Este identificador é usado para identificar os documentos da CH, e é composto pela etiqueta HAREM, seguida de três caracteres alfanuméricos, e terminando por uma sequência de cinco algarismos. Estas três partes distintas são separadas por hífens. Um exemplo de um identificador DOCID válido é HAREM-87J-07845.
- A colecção não pode ter dois ou mais documentos com o mesmo DOCID.
- O texto marcado pelas etiquetas <GENERO> deve ser um dos géneros textuais especificados no ficheiro harem.conf (ver apêndice D.3).
- O texto marcado pelas etiquetas <ORIGEM> deve ser uma das variantes especificadas no ficheiro harem.conf (ver apêndice D.3).
- Dentro da etiqueta <TEXTO>, só são permitidas etiquetas válidas para a marcação de EM no texto.
- Se a saída não se referir à tarefa de classificação semântica, só pode conter etiquetas , que podem incluir o atributo opcional MORF.
- Se a saída se referir à tarefa de classificação semântica, não pode conter etiquetas . As etiquetas devem ter uma ou mais categorias separadas por um '|', e devem possuir obrigatoriamente o atributo TIPO com um ou mais tipos separados por um

'|', em número idêntico. Esses tipos devem corresponder às categorias, pela mesma ordem. A etiqueta pode incluir também o atributo opcional MORF.

- Para o atributo MORF, o formato aceite é “*x,y*”, onde *x* pode tomar os valores M, F ou ?, e *y* os valores S, P ou ?.
- As etiquetas e os atributos devem conter apenas caracteres alfabéticos maiúsculos, além dos caracteres '|' (barra vertical), para especificar mais de uma categoria, e ',' (vírgula), que separa os valores para o género e para o número, dentro do atributo MORF, como explicado acima.
- Todos os atributos dos parâmetros TIPO e MORF devem estar delimitados por aspas.
- Não são aceites etiquetas de abertura quando ainda existe uma etiqueta à espera de ser fechada. Por outras palavras, não são aceites EM marcadas dentro de outras EM.

19.2.2 Extractor

O módulo Extractor extrai o subconjunto dos documentos contidos na CD, da saída do sistema dos participantes. No processo, o Extractor ordena os documentos numericamente pelo seu identificador, o DOCID, e escreve-os sem alterar o seu conteúdo.

19.2.3 AlinhEM

O módulo AlinhEM tem como objectivo produzir uma lista de *alinhamentos* das EM da saída do sistema com as EM da CD. Alinhamentos são linhas de texto que descrevem a correspondência existente entre as EM de dois documentos (no caso da avaliação do HAREM, entre a saída do sistema e a CD).

A tarefa do AlinhEM é muito importante, uma vez que os módulos seguintes baseiam-se nos alinhamentos gerados por este. Um requisito do AlinhEM é que as colecções de textos a alinhar possuam os mesmos documentos, podendo diferir apenas nas etiquetas de EM colocadas nos textos.

Formato de saída

O AlinhEM processa e escreve cada documento no seguinte formato:

```
HAREM_ID ORIGEM GÉNERO
<VERIFICACAO_MANUAL>Informação para o juiz humano</VERIFICACAO_MANUAL>
Alinhamento 1
Alinhamento 2
(...)
Alinhamento n
```

O AlinhEM escreve uma primeira linha com os seus metadados, uma linha (opcional) para depuração manual, seguida de uma lista de alinhamentos. O documento termina com uma ou mais linhas em branco. Os alinhamentos podem ser de cinco tipos:

um para um: uma EM da CD alinha exactamente com uma EM na saída.

um para muitos: uma EM da CD alinha com mais do que uma EM na saída.

muitos para um: mais do que uma EM da CD alinham com uma EM na saída.

nenhum para um: uma EM é identificada na saída mas não há uma EM correspondente na CD.

um para nenhum: uma EM da CD não foi marcada como tal na saída.

Para cada tipo de alinhamento, o AlinhEM representa cada uma destas situações num formato específico, para facilitar o processamento dos módulos seguintes. Todos os formatos exibem primeiro a correspondência na CD, seguido de um separador '--->' e a(s) correspondências na saída, entre parênteses rectos. Existem cinco formatos diferentes de alinhamentos, um para cada tipo:

1. No caso de um alinhamento do tipo **um para um**, a lista de entidades da saída contém uma EM:

```
<EM>17:00<EM> ---> [<EM>17:00</EM>]
```

2. No caso de um alinhamento do tipo **um para muitos**, onde múltiplas EM da saída alinham com uma EM da CD, o alinhamento apresenta as várias EM da saída separadas por vírgulas, como é ilustrado a seguir:

```
<EM>17:00<EM> ---> [<EM>17</EM>, <EM>00</EM>]
```

3. No caso de um alinhamento do tipo **muitos para um**, cada EM da CD alinhada é representada numa linha distinta:

```
<EM>17<EM> ---- [ <EM>17:00</EM> ]
<EM>00<EM> ---- [ <EM>17:00</EM> ]
```

4. No caso de um alinhamento do tipo **nenhum para um**, ou seja, EM espúrias na saída, esta é marcada com a etiqueta <ESPURIO>:

```
<ESPURIO>Ontem</ESPURIO> ---- [ <EM>Ontem</EM> ]
```

5. No caso de um alinhamento do tipo **um para nenhum**, ou seja, EM que não foram identificadas na saída, a EM da CD aponta para uma lista com o termo null.

```
<EM>Departamento de Informática</EM> ---- [ null ]
```

Etiquetas <ALT>

Nas situações em que as etiquetas <ALT> foram usadas na CD, o AlinhEM faz o alinhamento para cada alternativa, e marca cada uma das alternativas com uma etiqueta <ALT n >, com n a ser o número incremental da alternativa. De seguida pode-se ver exemplos de alternativas escritas pelo AlinhEM. A selecção da melhor alternativa é posteriormente realizada pelos módulos AltinaID, AltinaMOR e AltinaSEM.

Segue-se um exemplo de alternativas para um alinhamento do tipo **um para um**, com uma EM vaga na CD, para o caso em que na CD esteja <ALT> 98 anos e meio | 98 anos e meio </ALT> e a saída do sistema tenha sido 98 anos :

```
<ALT>
<ALT1>
<VALOR TIPO="QUANTIDADE">98 anos e meio</VALOR> ---- [ <VALOR TIPO="QUANTIDADE">98 anos</VALOR> ]
</ALT1>
<ALT2>
<VALOR TIPO="QUANTIDADE">98 anos</VALOR> ---- [ <VALOR TIPO="QUANTIDADE">98 anos</VALOR> ]
<ALT2>
</ALT>
```

O próximo é um exemplo de alternativas para um alinhamento do tipo **um para um** ou do tipo **muitos para um**, uma ou mais EM vagas na CD, para o caso em que na CD esteja <ALT> Aves-Campomaiorense | Aves - Campomaiorense </ALT> e a saída do sistema tenha sido Aves-Campomaiorense :

```
<ALT>
<ALT1>
```



```

<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorens</ACONTECIMENTO> --->
[<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorens</ACONTECIMENTO>]
</ALT1>
<ALT2>
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Aves</PESSOA> --->
[<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorens</ACONTECIMENTO>]
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Campomaiorens</PESSOA> --->
[<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorens</ACONTECIMENTO>]
</ALT2>
</ALT>

```

Apresentamos agora um exemplo de alternativas para um alinhamento do tipo **nenhum para nenhum** ou do tipo **um para nenhum**, uma ou nenhuma EM na CD, para o caso em que na CD esteja `<ALT> Monárquico | Monárquico <ALT>` e a saída do sistema tenha sido Monárquico:

```

<ALT>
<ALT1>
</ALT1>
<ALT2>
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA> ---> [null]
</ALT2>
</ALT>

```

Finalmente, eis um exemplo de alternativas para um alinhamento do tipo **nenhum para um** ou do tipo **um para um**, uma ou nenhuma EM na CD, para o caso em que na CD esteja `<ALT> Monárquico | Monárquico <ALT>` e a saída do sistema tenha sido ` Monárquico `:

```

<ALT>
<ALT1>
<ESPURIO>Monárquico</ESPURIO> --->
[<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA>]
</ALT1>
<ALT2>
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA> --->
[<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA>]
</ALT2>
</ALT>

```

Etiquetas <OMITIDO>

A etiqueta <OMITIDO> foi introduzida na versão 2.1 da CD de 2005, em plena avaliação do HAREM, por se ter achado necessário ignorar certos excertos de texto sem qualquer interesse do ponto de vista linguístico, sem interferir com a avaliação do HAREM. Assim, as

etiquetas <OMITIDO> identificam esses excertos de texto, alertando os módulos de avaliação para ignorarem o conteúdo. Apresentamos abaixo um exemplo contido num documento oriundo da Web, e que, do ponto de vista da tarefa de REM em português, é inadequado para avaliar o desempenho dos sistemas.

```
<OMITIDO>
Sorry, your browser doesn't support <OBRA TIPO="PRODUTO">Java</OBRA>.
</OMITIDO>
```

Numeração distinta de átomos

O AlinhEM, ao ser executado com a opção `-etiquetas sim`, regista todos os átomos presentes nos alinhamentos de cada documento, e depois numera-os sequencialmente por ordem de aparição. Desta forma, impede-se que haja emparelhamentos de EM com átomos em comum, mas que estão localizados em partes diferentes do documento.

Para ilustrar tais situações, considere-se o seguinte extracto de texto, marcado como uma CD (só para a categoria ORGANIZACAO):

```
<DOC>
<DOCID>HAREM-051-00043</DOCID>
<GENERO>Web</GENERO>
<ORIGEM>PT</ORIGEM>
<TEXTO>
<ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S"><1>Lions</1> <1>Clube</1> de <1>Faro</1>
</ORGANIZACAO>

É no Hotel Eva, situado na lateral da marina, que se reúne o <ORGANIZACAO TIPO="INSTITUICAO"
MORF="M,S"><2>Clube</2> <2>Lions</2></ORGANIZACAO>, nas primeiras quartas-feiras de cada mês.
</TEXTO>
</DOC>
```

E a respectiva (e hipotética) saída de um sistema participante:

```
<DOC>
<DOCID>HAREM-051-00043</DOCID>
<GENERO>Web</GENERO>
<ORIGEM>PT</ORIGEM>
<TEXTO>
<1>Lions</1> <1>Clube</1> de <1>Faro</1>

É no Hotel Eva, situado na lateral da marina, que se reúne o <ORGANIZACAO TIPO="INSTITUICAO"
MORF="M,S"><2>Clube</2> <2>Lions</2></ORGANIZACAO>, nas primeiras quartas-feiras de cada mês.
</TEXTO>
</DOC>
```

O documento da CD tem duas EM, e ambas incluem o átomo *Lions*. Contudo, a saída do sistema apresenta apenas uma EM, com o átomo *Lions*. Se os textos não fossem marcados com etiquetas numéricas, o AlinhEM não tinha informação suficiente para saber qual das EM da CD é que vai alinhar com a EM da saída.

Nos processos de atomização e de etiquetagem numérica, o AlinhEM pode ignorar ocorrências de um dado conjunto de átomos. Esta opção permite não só ignorar termos muito frequentes, como também permite ultrapassar situações em que os textos originais das saídas são alterados, especialmente nas suas EM.

O AlinhEM possui uma lista interna de termos a ignorar nas avaliações conjuntas do HAREM, que apresentamos abaixo. Esta lista pode ser complementada com outra lista, segundo a opção `-ignorar`, descrita no apêndice D.2.2.

a, A, à, À, ao, AO, Ao, as, AS, As, com, COM, Com, como, COMO, Como, da, DA, Da, das, DAS, Das, de, DE, De, do, DO, Do, dos, DOS, Dos, e, E, é, É, em, EM, Em, for, FOR, For, mais, MAIS, Mais, na, NA, Na, não, NÃO, Não, no, NO, No, nos, NOS, Nos, o, O, os, OS, Os, ou, OU, Ou, para, PARA, Para, pela, PELA, Pela, pelo, PELO, Pelo, por, POR, por, que, QUE, Que, se, SE, Se, um, UM, Um, uma, UMA, Uma.

O processo de atomização do AlinhEM não se preocupa em garantir que cada átomo gerado corresponda a algo que faça parte do léxico, uma vez que a preocupação principal é o alinhamento correcto das EM. O AlinhEM pode mesmo partir palavras e números em locais que os atomizadores para a língua portuguesa não o fariam. O AlinhEM utiliza as seguintes regras de atomização:

1. Todos os caracteres não alfa-numéricos são considerados delimitadores de átomos.

alguem@algures.com -> <1>alguem</1> @ <1>algures</1> . <1>com</1>

2. Todos os números são atomizados ao nível do dígito.

1979 -> <1>1</1> <1>9</1> <1>7</1> <2>9</2>

1.975 -> <1>1</1> . <1><9/1> <1><7/1> <1><5/1>

3. A transição de um carácter numérico para um alfabético (ou vice-versa) delimita átomos.

NBR6028 -> <1>NBR</1> <1>6</1> <1>0</1> <1>2</1> <1>8</1>

Etiquetas <VERIFICACAO_MANUAL>

A etiqueta <VERIFICACAO_MANUAL> é gerada quando o AlinhEM é executado com a opção `-etiquetas sim`, e no final da etiquetagem numérica aos átomos do mesmo documento

na CD e na saída, os números finais não coincidem. Isto normalmente sugere que o texto original da saída foi alterado, o que pode impedir o alinhamento correcto das EM. Quando tal acontece, os alinhamentos com as etiquetas numéricas discordantes são envolvidas em etiquetas <VERIFICACAO_MANUAL>, para que sejam inspeccionados manualmente de forma a que a origem do problema seja identificada. Estas etiquetas são ignoradas pelos módulos seguintes.

19.2.4 AvalIDa

O módulo AvalIDa avalia e pontua os alinhamentos produzidos pelo AlinhEM, segundo as directivas de avaliação para a tarefa de identificação. Para tal, o AvalIDa acrescenta no final de cada alinhamento a respectiva pontuação dentro de parênteses rectos, com um carácter de dois pontos como separador, como é exemplificado abaixo:

```
<EM>17:00<EM> ---> [<EM>17:00</EM>]:[Correcto]
```

No caso de um alinhamento do tipo **um para muitos**, as várias pontuações são separadas por vírgulas, como é mostrado no exemplo abaixo. Este caso é sintomático de pontuações parcialmente correctas, que é complementado com a informação do valor do *factor de correcção* e do *factor de erro* (ver a secção 18.2.1):

```
<EM>17:00<EM> ---> [<EM>17</EM>, <EM>00</EM>] :[Parcialmente_Correcto_
por_Defeito(0.25; 0.75), Parcialmente_Correcto_por_Defeito(0.25; 0.75)]
```

Existem, no entanto, casos que requerem um processamento mais cuidado, como o caso exemplificado abaixo:

```
<EM>Gabinete do Instituto</EM> ---> [<EM>Gabinete do Instituto da Juventude
em Lisboa</EM>]:[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]
<EM>Juventude em Lisboa</EM> ---> [<EM>Gabinete do Instituto da Juventude
em Lisboa</EM>]:[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]
```

Este exemplo apresenta uma EM da saída (Gabinete do Instituto da Juventude em Lisboa) alinhada com duas EM da CD (Gabinete do Instituto e Juventude em Lisboa). Como o alinhamento é representado em duas linhas, os módulos seguintes (como por exemplo, o Ida2ID) precisam de saber se as duas linhas se referem a um único alinhamento (uma situação **muitos para um**) ou a dois alinhamentos (duas situações **um para um**), evitando cair no erro de contar mais de uma vez a mesma EM. O AvalIDa distingue entre as duas situações usando as etiquetas numéricas produzidas pelo AlinhEM.

19.2.5 Véus

O módulo Véus permite seleccionar criteriosamente grupos de documentos com determinadas características, tais como o seu género textual (Web, Jornalístico, etc) ou a sua variante (PT, BR, etc), ou filtrar os alinhamentos segundo as classificações semânticas das etiquetas das EM, permitindo a avaliação do desempenho do sistema segundo um determinado leque de categorias/tipos.

É dessa forma que o HAREM permite avaliar os sistemas segundo um cenário *selectivo*, ou seja, comparando a saída sobre a CD segundo o universo das EM de categoria/tipo que o sistema se propõe tentar identificar/classificar, e não segundo o universo total das EM. Além disso, o Véus ainda permite parametrizar as avaliações em três estilos: Além do do HAREM, descrito no presente capítulo e volume, também permite uma avaliação “relaxada” em que apenas o primeiro valor de um alinhamento com EM parcialmente correctas é contabilizado, e uma avaliação estilo “muc” em que nenhum caso parcialmente correcto é contabilizado (são todos considerados errados, veja-se Douthat (1998)).

Filtragem por género textual ou por variante

Quando o Véus é executado apenas com um filtro por género textual ou variante, apenas os cabeçalhos dos documentos são analisados, para decidir se o documento é ignorado ou se é copiado para a saída.

Nesse caso, a primeira linha escrita pelo Véus contém a informação sobre todas as categorias e tipos utilizadas na avaliação (ou seja, a repetição das categorias e tipos especificados no ficheiro `harem.conf`). A linha é ilustrada abaixo (o exemplo está abreviado para facilitar a leitura):

```
#PESSOA=[ "MEMBRO", "GRUPOIND", "CARGO", "GRUPOCARGO", "INDIVIDUAL",
"GRUPOMEMBRO" ]; LOCAL=[ "GEOGRAFICO", "ALARGADO", "ADMINISTRATIVO",
"VIRTUAL", "CORREIO" ]; (...)
```

Filtragem por categorias e tipos semânticos

Quando o Véus é executado com um filtro por categorias e/ou tipos, a primeira linha da saída do Véus reproduz todas as categorias e tipos aceites, para que não se perca a informação sobre o tipo de filtro aplicado e que originou o resultado do Véus.

Se, por exemplo, o Véus fosse executado com um filtro para obter apenas alinhamentos contendo a categoria ORGANIZACAO com todos os seus quatro tipos, e LOCAL nos seus tipos GEOGRAFICO, ADMINISTRATIVO, CORREIO e ALARGADO (ou seja, todos excepto o VIRTUAL), como é ilustrado na Figura 19.2, a primeira linha da saída do Véus seria:

```
#ORGANIZACAO=[ "INSTITUICAO", "ADMINISTRACAO", "SUB", "EMPRESA" ];
LOCAL=[ "GEOGRAFICO", "ADMINISTRATIVO", "CORREIO", "ALARGADO" ]
```

O símbolo '#' no início de cada ficheiro gerado pelo Véus indica aos módulos de avaliação seguintes qual o *cenário* de avaliação especificado, para efeitos de avaliação semântica e morfológica.

De seguida, o Véus filtra todos os alinhamentos previamente identificados, extraindo o subconjunto de alinhamentos que interessa considerar. Note-se que a filtragem por categorias só faz sentido quando o sistema em causa efectuou a respectiva classificação semântica, ou seja, quando a etiqueta genérica não é usada.

19.2.6 ALTinaID

O módulo ALTinaID analisa as alternativas na tarefa de identificação, marcadas com <ALT>, e selecciona a alternativa segundo os critérios descritos no capítulo 18. A alternativa escolhida é a única escrita como resultado do programa. As etiquetas <ALT> e <ALT*n*> também são eliminadas.

19.2.7 Ida2ID

O módulo Ida2ID calcula os valores das métricas de avaliação para a tarefa de identificação, fornecendo dados para aferir o desempenho do sistema REM participante.

O funcionamento do Ida2ID pode ser dividido em dois passos: em primeiro lugar, o Ida2ID percorre todos os alinhamentos do ficheiro fornecido, realizando várias contagens. No segundo passo, usa os valores finais dos contadores para chegar aos valores das métricas de avaliação.

À primeira vista, a tarefa do Ida2ID parece simples. Contudo, o formato usado para representar os alinhamentos pode induzir o Ida2ID à contagem errada de EM. Estes casos potencialmente problemáticos normalmente verificam as seguintes condições:

1. O alinhamento em consideração foi pontuado como `parcialmente_correcto`;
2. O alinhamento imediatamente anterior ao que está a ser considerado também foi pontuado como `parcialmente_correcto`;
3. A EM da saída identificada no alinhamento imediatamente anterior é idêntica à EM da saída identificada no alinhamento que está a ser considerado.

Quando estas três condições se verificam, o Ida2ID precisa de decidir se está na presença de uma EM nova, ou se está na presença da mesma ocorrência da EM anterior. Para decidir, o Ida2ID averigua se existe alguma sobreposição das EM da CD, com o auxílio das etiquetas numéricas.

Considere-se o seguinte exemplo (hipotético) de um alinhamento do tipo **muitos para um**:

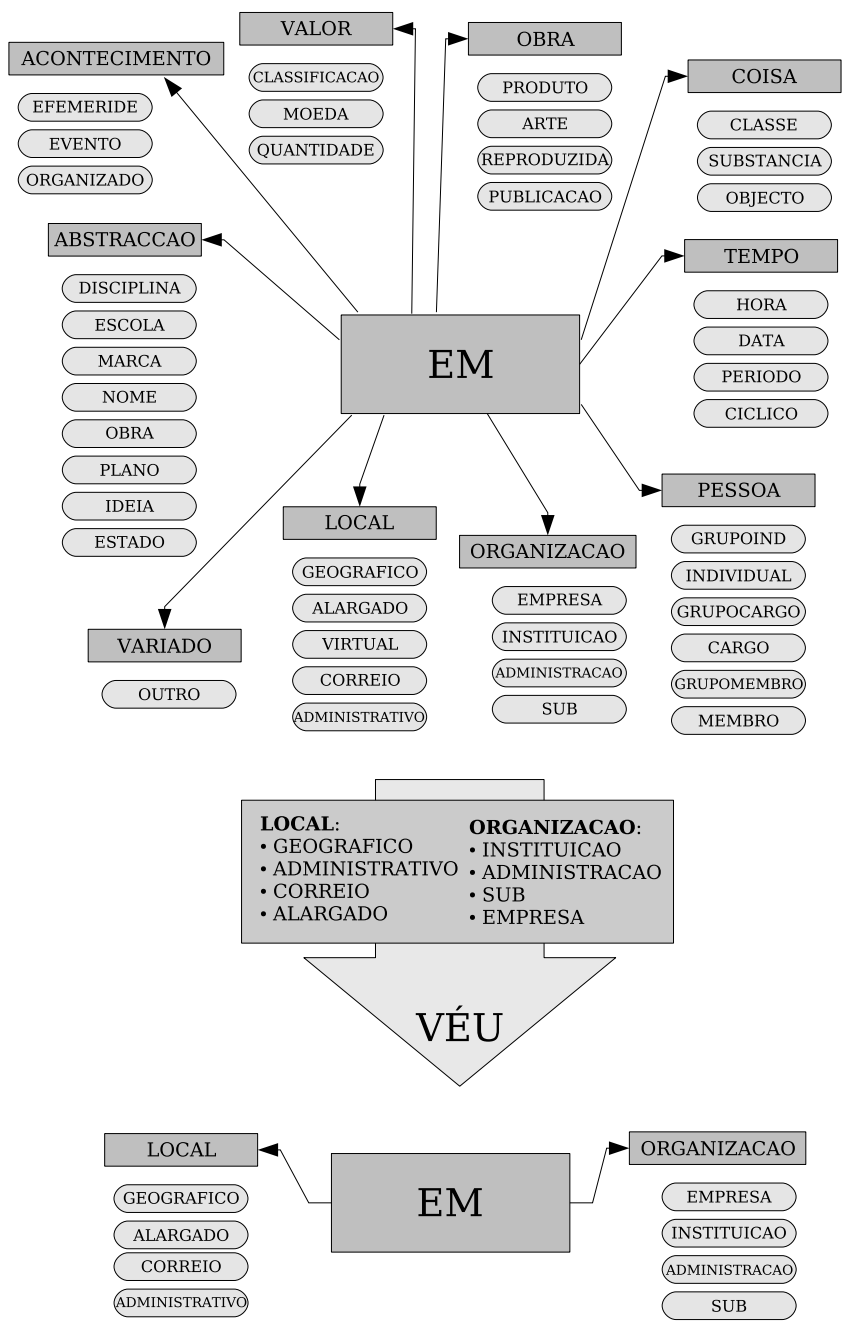


Figura 19.2: Esquema de um exemplo do processamento efectuado pelo Véus.

```

<EM><3>Gabinete</3> do <2>Instituto</2></EM> ---> [<EM><3>Gabinete
</3> do <2>Instituto<2> da <1>Juventude</1> em <5>Lisboa</5></EM>]:
[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]
<EM><1>Juventude</1> em <5>Lisboa</5></EM> ---> [<EM><3>Gabinete
</3> do <2>Instituto<2> da <1>Juventude</1> em <5>Lisboa</5></EM>]:
[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]

```

Com a ajuda das etiquetas numéricas, o Ida2ID consegue determinar que as duas linhas referem-se à mesma EM da saída, uma vez que essa EM, tal como está representada, refere-se à EM que contém a 3ª ocorrência do átomo 'Gabinete', ou a 5ª ocorrência do átomo 'Lisboa'. Como tal, o contador das EM de saídas do Ida2ID faz uma correcção e conta apenas uma EM na saída.

Agora, considere-se também o seguinte exemplo de dois alinhamentos do tipo **um para um**:

```

<EM><3>Gabinete</3> do <2>Instituto</2></EM> ---> [<EM><3>Gabinete
</3> do <2>Instituto<2> da <1>Juventude</1> em <5>Lisboa</5></EM>]:
[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]
<EM><2>Juventude</2> em <6>Lisboa</6></EM> ---> [<EM><4>Gabinete
</4> do <3>Instituto<3> da <2>Juventude</2> em <6>Lisboa</6></EM>]:
[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]

```

Neste exemplo, há duas EM da saída alinhadas respectivamente com outras duas EM na CD. As etiquetas numéricas mostram que na saída há duas ocorrências de uma EM e, como tal, o Ida2ID conta duas EM na saída.

Um exemplo de um relatório (fictício) produzido pelo Ida2ID para um sistema, contendo as várias contagens e avaliações a levar em conta é apresentado em seguida:

```

Total na CD: 4995
Total Identificadas: 2558
Total Correctos: 1927
Total Ocorrências Parcialmente Correctos: 601
Soma Parcialmente Correctos: 128.57140579578655
Soma Parcialmente Incorrectos: 472.42859420421337
Espúrios: 73
Em Falta: 2545
Precisão: 0.8035853814682512
Abrangência: 0.41152580696612345

```


Medida F: 0.5443059461924498
 Sobre-geração: 0.028537920250195466
 Sub-geração: 0.5095095095095095
 Erro Combinado: 0.600549668520057

É de notar que estes cálculos só podem ser efectuados após a escolha da alternativa mais favorável ao sistema, realizada pelo AltinaID. Esta escolha influencia o número total de entidades encontradas na CD, o que também implica que saídas diferentes podem ser avaliadas segundo diferentes conjuntos de EM da CD. Contudo, estas diferenças saldaram-se sempre no favorecimento de cada sistema.

19.2.8 Emir

O módulo Emir pode ser considerado o homólogo do AvalIDa e do Vizir, mas para a avaliação da tarefa de classificação semântica, ao pontuar cada alinhamento segundo a classificação semântica das EM.

O Emir recebe os resultados gerados pelo AvalIDa, filtrados pelo Véus. A primeira linha desses resultados, que contém a informação sobre as categorias e/ou tipos usados no cenário da avaliação, é usada para efectuar o cálculo das várias medidas de avaliação correspondentes à classificação semântica.

Depois de avaliar o alinhamento em relação à classificação semântica, o Emir concatena o resultado no fim do alinhamento, usando um formato semelhante ao do AvalIDa. Considere-se o seguinte alinhamento hipotético gerado pelo AvalIDa:

```
<LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL> ---> [<LOCAL TIPO="GEOGRAFICO">
Lisboa e Porto</LOCAL>]:[Parcialmente_Correcto_por_Excesso(0,6666; 0,3333)]
```

No seu processamento, o Emir retira a pontuação respeitante à tarefa de identificação (no exemplo dado, [Correcto]), e substitui-a por um novo resultado referente à tarefa de classificação semântica. Esse novo resultado é colocado no final do alinhamento, com dois pontos (:) como separador. Este resultado pode conter uma ou mais avaliações (uma por cada EM), e cada avaliação fica envolvida entre chavetas ({}).

O resultado da avaliação da classificação semântica, adicionado no final do alinhamento, contém quatro campos:

1. a lista de categorias que foram pontuadas como correcto, espurio ou em_falta;
2. a lista de tipos que foram pontuadas como correcto, espurio ou em_falta;
3. o valor da classificação semântica combinada (CSC), uma das quatro medidas de classificação semânticas adoptadas pelo HAREM (consulte-se a secção 18.3.2 para mais detalhes sobre a CSC e o seu cálculo);

4. o valor do peso da EM. Para mais informação sobre o cálculo deste, consulte-se o capítulo 18.

No final, o alinhamento processado pelo Emir pode apresentar o seguinte aspecto:

```
<LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL> ---> [<LOCAL TIPO="GEOGRAFICO">
Lisboa e Porto</LOCAL>]:[{Categoria(Correcto:[LOCAL] Espúrio:[ ] Em_Falta:[ ])
Tipo(Correcto:[ ] Espúrio:[GEOGRAFICO] Em_Falta:[ADMINISTRATIVO]) CSC(1.0)
Peso(0.66)}]
```

No caso de alinhamentos **um para muitos**, o Emir escreve os vários resultados da avaliação da forma que se apresenta no exemplo abaixo, separados por vírgulas (,):

```
<LOCAL TIPO="ADMINISTRATIVO">Lisboa e Porto</LOCAL> --->
[<LOCAL TIPO="GEOGRAFICO">Lisboa</LOCAL>, <LOCAL TIPO="GEOGRAFICO">
Porto</LOCAL>]:[{Categoria(Correcto:[LOCAL] Espúrio:[ ] Em_Falta:[ ])
Tipo(Correcto:[ ] Espúrio:[GEOGRAFICO] Em_Falta:[ADMINISTRATIVO])
CSC(1.0) Peso(0.33)}, {Categoria(Correcto:[LOCAL] Espúrio:[ ]
Em_Falta:[ ]) Tipo(Correcto:[ ] Espúrio:[GEOGRAFICO] Em_Falta:
[ADMINISTRATIVO]) CSC(1.0) Peso(0.33)}]
```

No caso de haver vagueza na classificação semântica, ou seja, a EM possuir mais do que uma categoria ou tipo, estas são tratadas como se fossem uma classificação única, como exemplificado abaixo:

```
<ORGANIZACAO|ABSTRACCAO TIPO="SUB|IDEIA">Lisboa</ORGANIZACAO|
ABSTRACCAO> ---> [<LOCAL TIPO="ADMINISTRATIVO">Lisboa e Porto
e Faro e Braga</LOCAL>]:[{Categoria(Correcto:[ ] Espúrio:[LOCAL]
Em_Falta:[ORGANIZACAO|ABSTRACCAO]) Tipo(Correcto:[ ] Espúrio:[ ]
Em_Falta:[ ]) CSC(0.0) Peso(0.142)}]
```

Quando o Emir é executado **sem** a opção de cenário relativo, os alinhamentos espúrios são contabilizados pelo Emir, que considera todas as categorias e tipos como espúrio. Um alinhamento como este:

```
<ESPURIO>DM-115CS</ESPURIO> ---> [<ABSTRACCAO TIPO="MARCA"
MORF="F,S">DM-115CS</ABSTRACCAO>]:[Espúrio]
```

é convertido pelo Emir (se não se optar pelo cenário relativo) para:

```
<ESPURIO>DM-115CS</ESPURIO> ---> [<ABSTRACCAO TIPO="MARCA"
MORF="F,S">DM-115CS</ABSTRACCAO>]:[{Categoria(Correcto:[ ]
Espúrio:[ABSTRACCAO] Em_Falta:[ ]) Tipo(Correcto:[ ] Espúrio:[ ]
Em_Falta:[ ]) CSC(0.0) Peso(0.0)}]
```

Da mesma forma que acontece com alinhamentos espúrios quando o Emir é executado **sem** a opção de cenário relativo, o Emir também considera e escreve as categorias e tipos `em_falta` quando as EM não foram identificadas, como se pode ver no seguinte exemplo:

```
<LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">Pinheiros</LOCAL> ----> [null]:[Em_Falta]
```

o alinhamento é convertido (se não se optar pelo cenário relativo) para:

```
<LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">Pinheiros</LOCAL> ---->
[null]:[{Categoria(Correcto:[ ] Espúrio:[ ] Em_Falta:[LOCAL])
Tipo(Correcto:[ ] Espúrio:[ ] Em_Falta:[ ] CSC(0.0) Peso(0.0))}]
```

No apêndice E.1 apresentam-se mais exemplos do processamento do Emir.

19.2.9 AltinaSEM

O módulo `AltinaSEM`, de um modo análogo aos módulos `AltinaID` e `AltinaMOR`, recebe os resultados do Emir e processa os alinhamentos marcados com etiquetas `<ALT>`, escolhendo as melhores alternativas para cada saída. Os critérios tomados em consideração na escolha da melhor alternativa estão descritos na página 18.2.4 do capítulo 18, e ao contrário do `AltinaID`, tomam em consideração os valores calculados pelo Emir para a tarefa de classificação semântica, no processo de selecção da melhor alternativa. A alternativa escolhida é escrita, enquanto que as restantes alternativas são eliminadas, tal como as etiquetas `<ALTn>` e `<ALTn>`.

19.2.10 Ida2SEM

O módulo `Ida2SEM` é o avaliador global da tarefa de classificação semântica, ao calcular os valores das métricas, fornecendo dados para aferir o desempenho do sistema. Tal como o `Ida2ID` e `Ida2MOR`, a execução do `Ida2SEM` pode ser dividida em duas fases: i) todos os alinhamentos avaliados relativamente à classificação semântica são processados, procedendo-se a várias contagens; ii) os contadores são usados para calcular as métricas e gerar um relatório.

De seguida, reproduz-se um exemplo hipotético de um relatório gerado pelo `Ida2SEM`, que possui as seguintes informações:

1. O domínio da avaliação: quais as categorias e tipos a avaliar;
2. A avaliação referente à classificação semântica por categorias;
3. A avaliação referente à classificação semântica por tipos;
4. A avaliação referente à classificação semântica combinada;

302 CAPÍTULO 19. A ARQUITECTURA DOS PROGRAMAS DE AVALIAÇÃO DO HAREM

5. A avaliação referente à classificação semântica plana.

Avaliação Global - Classificação Semântica por Categorias

Total de EMs classificadas na CD: 5004
Total de EMs classificadas pelo sistema: 5269
Total Correctos: 3120
Espúrios: 1866
Em Falta: 1832
Precisão: 0.5922527110682176
Abrangência: 0.6236170133130373
Medida F: 0.6075303289435293
Sobre-geração: 0.34937277663358923
Sub-geração: 0.36610711430855314

Avaliação Global - Classificação Semântica por Tipos

Total de EMs classificadas na CD: 3440
Total de EMs classificadas pelo sistema: 3448
Total Correctos: 2641
Espúrios: 599
Em Falta: 631
Precisão: 0.7660720776326169
Abrangência: 0.7678536406038555
Medida F: 0.7669618245288219
Sobre-geração: 0.17237410071942447
Sub-geração: 0.18343023255813953

Avaliação Global - Classificação Semântica Combinada

Valor máximo possível para a Classificação Semântica Combinada
na CD: 8987.450000000072
Valor máximo possível para a Classificação Semântica Combinada
do sistema: 7309.8648131094515
Precisão Máxima do Sistema: 0.7081119047925152
Abrangência Máxima na CD: 0.5759367002416341
Medida F: 0.6352214896681005

Avaliação Global - Classificação Semântica Plana

Total de EMs classificadas na CD: 23
Total de EMs classificadas pelo sistema: 23
Total Correctos: 1.8403361344537812
Espúrios: 18
Em Falta: 17
Precisão: 0.08001461454146874

Abrangência: 0.08001461454146874
 Medida F: 0.08001461454146874
 Sobre-geração: 0.782608695652174
 Sub-geração: 0.7391304347826086

19.2.11 Vizir

O módulo Vizir faz a avaliação da tarefa de classificação morfológica, de uma forma análoga ao Emir na tarefa de classificação semântica, e ao AvalIDa na tarefa de identificação. Para tal, o Vizir pontua os alinhamentos cujas EM possuem atributos MORF.

O Vizir retira toda a informação semântica contida na EM, substituindo as categorias pela etiqueta , e eliminando os atributos TIPO. Esta etapa é ilustrada no seguinte exemplo abaixo, onde o alinhamento:

```
<LOCAL TIPO="ADMINISTRATIVO" MORF="F,S">Rússia</LOCAL> --->
[<LOCAL TIPO="ALARGADO" MORF="F,S">Rússia</EM>]
```

é convertido pelo Vizir na seguinte linha:

```
<EM MORF="F,S">Rússia</EM> ---> [<EM MORF="F,S">Rússia</EM>]
```

Em seguida, o Vizir, tal como o Emir, remove dos alinhamentos os resultados respeitantes à tarefa de identificação, substituindo-os por novos resultados referentes à tarefa de classificação morfológica. Esses resultados detalham as pontuações e valores para as três medidas usadas: Género, Número e Combinada. Os critérios de atribuição de pontuação e do respectivo valor para cada medida encontram-se detalhados na secção 18.4.2 deste livro.

O seguinte caso exemplifica o resultado da avaliação do Vizir:

```
<ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S">Lions Clube de
Faro</ORGANIZACAO> ---> [<PESSOA TIPO="GRUPOMEMBR" MORF="F,S">
Lions Clube de Faro</PESSOA>]:[Correcto]
```

O alinhamento é convertido em:

```
<EM MORF="M,S">Lions Clube de Faro</EM> ---> [<EM MORF="F,S">
Lions Clube de Faro</EM>]:[(Género: Incorrecto 0) (Número:
Correcto 1) (Combinada: Incorrecto 0)]
```

Para mais exemplos do processamento do Vizir, consulte-se o apêndice E.2.

19.2.12 AltinaMOR

O módulo AltinaMOR, de um modo análogo ao AltinaID e ao AltinaSEM, recebe os resultados do Vizir e processa os alinhamentos marcados com etiquetas <ALT>, escolhendo as melhores alternativas para cada saída. Os critérios tomados em consideração na escolha da melhor alternativa estão descritos no capítulo 18, e tomam agora em consideração os valores calculados pelo Vizir para a tarefa de classificação morfológica.

19.2.13 Ida2MOR

O módulo Ida2MOR, de um modo análogo ao Ida2ID e ao Ida2SEM, calcula e gera os resultados globais para a tarefa de classificação morfológica. Para tal, o Ida2MOR processa os alinhamentos gerados pelo AltinaMOR, contando as pontuações calculadas. O relatório produzido pelo Ida2MOR apresenta os valores das métricas para as medidas de avaliação da tarefa de classificação morfológica: género, número, e combinada. Em seguida apresentamos um exemplo de um relatório do Ida2MOR:

RELATÓRIO DA AVALIAÇÃO DA CLASSIFICAÇÃO MORFOLÓGICA

Gerado em: 25 de Maio de 2005

Avaliação Global da Classificação Morfológica - Número

Total de classificações da CD: 111

Total de classificações do sistema : 92

Precisão: 0.940217391304348

Abrangência: 0.779279279279279

Medida F: 0.852216748768473

Sobre-especificação: 0

Sobre-geração: 0

Sub-geração: 0.171171171171171

Avaliação Global da Classificação Morfológica - Género

Total de classificações da CD: 88

Total de classificações do sistema : 92

Precisão: 0.652173913043478

Abrangência: 0.681818181818182

Medida F: 0.666666666666667

Sobre-especificação: 0.25

Sobre-geração: 0
Sub-geração: 0.215909090909091

Avaliação Global da Classificação Morfológica - Combinada

Total de classificações da CD: 111
Total de classificações do sistema : 92

Precisão: 0.652173913043478
Abrangência: 0.540540540540541
Medida F: 0.591133004926108

19.2.14 Sultão

O módulo Sultão tem por objectivo interpretar todos os relatórios globais gerados, e resumir os valores obtidos por todos os participantes na forma de tabelas, de modo a fornecer resultados comparativos da avaliação conjunta do HAREM. O Sultão é composto por três programas dedicados a cada tarefa de avaliação: o SultãoID, o SultãoMOR e o SultãoSEM, respectivamente para as tarefas de identificação, de classificação morfológica e de classificação semântica.

O Sultão precisa de ler os resultados dos vários sistemas segundo vários cenários para poder gerar os relatórios globais, pelo que o seu maior interesse é para os organizadores da avaliação conjunta. Ao resumir os resultados dos participantes, foi implementada no Sultão a opção de substituir o nome das saídas por pseudónimos, gerando também uma chave para poder desvendar os mesmos (Para conservar o anonimato dos resultados, esta chave deve naturalmente ser separada dos ficheiros, antes de serem divulgados).

As tabelas geradas pelo Sultão recorrem aos seguintes estilos:

1. os pseudónimos a **negrito** identificam as saídas consideradas oficiais, ou seja, as saídas enviadas durante a avaliação conjunta dentro do prazo estipulado;
2. os pseudónimos a *itálico* identificam os resultados no cenário selectivo escolhido para a saída em causa;
3. os valores a verde identificam os melhores para a métrica em questão.

A tabela 18.25 da secção 18.5.1 é um exemplo de tabelas geradas pelo Sultão.

19.2.15 Alcaide

O módulo Alcaide tem por objectivo gerar relatórios individuais para cada saída que participou no HAREM. Para tal, o Alcaide lê e processa os relatórios gerados pelo Sultão e

os relatórios gerados pelos módulos *Ida2ID*, *Ida2MOR* e *Ida2SEM*, organizando-os num único relatório composto por tabelas e gráficos, sub-dividido por tarefas, formas de avaliação, cenários, categorias, géneros textuais e variantes.

A saída do *Alcaide* consiste num relatório final em HTML, que resume o desempenho de uma saída, nas tarefas que esta se propôs realizar, nos diversos cenários e formas de avaliação discriminada por categoria, género textual e variante. Tal como o *Sultão*, o *Alcaide* é um módulo vocacionado para ser utilizado pela organização do *HAREM*, uma vez que os seus relatórios são apresentados de uma forma comparativa, que, para ser compilada, exige o acesso aos resultados dos outros participantes.

As tabelas 18.26 a 18.28 e as figuras 18.1 a 18.5 da secção 18.5.2 são exemplos de tabelas e gráficos gerados pelo *Alcaide*.

19.3 Comentários finais

Os programas aqui descritos foram desenvolvidos pelo primeiro autor (*Véus*, *AlinhEM*, *AvalIDa*, *Ida2ID*, *AltinaID*, *Emir*, *AltinaSEM*, *Ida2SEM* e *Sultão*), pelo segundo autor (*Validador*, *Extractor* e *Alcaide*) e pelo terceiro autor (*Vizir*, *AltinaMOR* e *Ida2MOR*), e testados exaustivamente pela quarta autora, com a ajuda dos primeiros.

Estes programas encontram-se acessíveis no sítio do *HAREM*, e a informação técnica para a sua utilização está patente no apêndice D.2.

Congratulamo-nos com o facto de existirem já alguns utilizadores que os usam rotineiramente, e esperamos que possam vir a ser usados, com poucas alterações, em futuras edições do *HAREM*.

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.