Apêndice D

Documentação técnica da plataforma de avaliação

D.1 Instalação e configuração

Os módulos foram desenvolvidos por Nuno Seco, Nuno Cardoso e Rui Vilela, e encontram-se disponíveis no sítio do HAREM, em `http://poloxldb.linguateca.pt/harem.php?l=programas`. Qualquer investigador tem acesso livre a estes programas e pode usá-los para avaliar o desempenho do seu sistema de REM, e compará-lo com os resultados obtidos pelos outros sistemas em avaliações conjuntas passadas. Dado que o código fonte também foi incluído nos pacotes de distribuição, qualquer utilizador pode estender e melhorar os programas.

Visto que alguns módulos foram programados em Perl, e outros em Java, a plataforma está disponível através de dois pacotes:

`ferramentas_HAREM_java.jar`, o pacote de módulos programados em Java, nomeadamente os módulos AlinhEM, AvalIDa, Véus, Emir, AltinaID, AltinaSEM, Ida2ID, Ida2SEM e Sultão.

`ferramentas_HAREM_perl.tar.gz`, o pacote de módulos programados em Perl, nomeadamente os módulos Extractor, Vizir, AltinaMOR, Ida2MOR e Alcaide.

A versão 1.5 do Java e a versão 5.8 do Perl foram usadas no desenvolvimento dos módulos, em ambiente Linux, e segundo a codificação de caracteres `ISO-8859-1`. Não é necessário nenhum procedimento de instalação para executar os módulos desenvolvidos em Java, sendo contudo necessária a presença da *Java Virtual Machine (JVM)* para a sua execução. Para executar os módulos desenvolvidos em Perl, é primeiro necessário instalar os módulos. Para tal, executa-se os seguinte comando:

```
tar xfz ACMorf.tar.gz
perl Makefile.PL
make
make install
```

Na mesma directoria onde se encontra o ficheiro `ferramentas_HAREM_java.jar`, é obrigatório existir um ficheiro chamado `harem.conf`, que descreve os géneros textuais, variantes, categorias e tipos válidos para a avaliação. O apêndice D.3 inclui o ficheiro `harem.conf` usado no Mini-HAREM.

Para a execução de módulos programados em Java, é necessário especificar na linha de comandos o parâmetro `-Dfile.encoding=ISO-8859-1`, de modo a garantir que os ficheiros sejam processados utilizando codificação de caracteres correcta. Na execução de módulos programados em Perl, é necessário verificar se o ambiente de execução é de codificação `ISO-8859-1`. O Alcaide requer, além disso para a geração dos gráficos, os módulos Perl GD-2.28, GDGraph-1.43 e GDTextUtil-0.86 (as versões dos módulos referidas são as versões utilizadas e testadas).

Dentro do programa Alcaide, é também necessário configurar os seguintes parâmetros, antes da sua execução:

$directoria_identificacao - directoria com os relatórios do SultãoID

$directoria_morfologia - directoria com os relatórios do SultãoMOR

$directoria_semantica - directoria com os relatórios do SultãoSEM

$directoria_ida - directoria com os relatórios dos programas ida2ID, ida2MOR e ida2SEM. Esta directoria deverá manter a estrutura de directorias, ou seja, uma directoria com o nome da saída, e sobre esta uma directoria para cada tarefa (`identificacao`, `morfologia` ou `semantica`), e debaixo das directorias `morfologia` e `semantica`, directorias `absoluto` e `relativo`.

D.2 Utilização

D.2.1 Extractor

Para executar o Extractor, usa-se o seguinte comando:

```
perl extrairCDdasSubmissoes.pl -in FICHEIRO_ENTRADA
-out FICHEIRO_SAIDA -cdids FICHEIRO_CDIDS
```

`FICHEIRO_ENTRADA` corresponde ao ficheiro da saída do sistema REM, a partir do qual serão extraídos os documentos correspondentes à CD para um novo ficheiro,

FICHEIRO_SAIDA. Os identificadores dos documentos a retirar (que, normalmente, correspondem aos identificadores dos documentos da CD) são lidos do ficheiro FICHEIRO_CDIDS, que deve conter uma lista com os últimos cinco números de cada DOCID, um por cada linha (no exemplo HAREM-87J-07845, o valor a colocar seria 07845).

Nota: Os ficheiros de identificadores das CD de 2005 e de 2006 (FICHEIRO_CDIDS) estão incluídos no pacote ferramentas_HAREM_perl.tar.gz.

D.2.2 AlinhEM

Para executar o AlinhEM, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.Aligner -submissao FICHEIRO_SUBMISSAO
-cd FICHEIRO_CD [-etiquetas sim|nao] [-ignorar FICHEIRO_ATOMOS]
> FICHEIRO_ALINHEM
```

FICHEIRO_SUBMISSAO corresponde ao nome do ficheiro pré-processado pelo Extractor, e FICHEIRO_CD corresponde ao ficheiro da CD. O resultado do alinhamento é enviado para o *standard output*, pelo que se recomenda o redireccionamento da saída para um ficheiro. Esse ficheiro, o FICHEIRO_ALINHEM, será usado pelo AvalIDa.

O AlinhEM possui dois parâmetros adicionais que podem ser usados na linha de comandos:

etiquetas, que pode ter os valores sim ou nao. A sintaxe é -etiquetas [sim|nao]. A opção nao é usada por defeito. Ao especificar o valor sim, o AlinhEM produz as etiquetas numéricas para identificar os átomos.

ignorar, que recebe como valor o nome de um ficheiro que contém uma lista de átomos que serão ignorados pelo AlinhEM. A sintaxe é -ignorar FICHEIRO_ATOMOS. O ficheiro FICHEIRO_ATOMOS deve ser composto por uma lista de átomos, um por linha.

D.2.3 AvalIDa

Para executar o AvalIDa, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.IndividualAlignmentEvaluator -alinhamento
FICHEIRO_ALINHEM > FICHEIRO_AVALIDA
```

O ficheiro FICHEIRO_ALINHEM corresponde ao nome do ficheiro gerado pelo AlinhEM, que contém os alinhamentos com as etiquetas numéricas. O resultado é enviado para o *standard output*, pelo que se recomenda o redireccionamento da saída para um ficheiro. Esse ficheiro, o FICHEIRO_AVALIDA, será usado pelos módulos Véus, AltinaID, Vizir e Emir.

O AvalIDa requer obrigatoriamente a opção `-alinhamento`, para especificar o ficheiro gerado pelo AlinhEM, o `FICHEIRO_ALINHEM`.

D.2.4 Véus

Para executar o Véus, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.AlignmentFilter -alinhamento FICHEIRO_AVALIDA
[-categoria CATEGORIAS] [-genero GENERO_TEXTUAL] [-origem VARIANTE]
[-estilo muc|relax|harem] > FICHEIRO_VEUS
```

`FICHEIRO_AVALIDA` corresponde ao ficheiro gerado pelo AvalIDa. O Véus escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_VEUS`.

O Véus pode receber até cinco parâmetros de entrada. Só o parâmetro `-alinhamento` é obrigatório, sendo os restantes parâmetros facultativos. Estes parâmetros podem ser combinados de várias formas, de modo a obter o filtro desejado.

-alinhamento, que deve vir acompanhado do nome do ficheiro gerado pelo AvalIDa, `FICHEIRO_AVALIDA`.

-categoria, que especifica as categorias e/ou tipos que devem ser filtradas. O argumento do parâmetro, `CATEGORIAS`, é uma lista de categorias separadas por ':'. Por exemplo, a lista `'PESSOA:ORGANIZACAO:ABSTRACCAO'` faz com que o Véus escreva para o *standard output* todos os alinhamentos que contêm EM de qualquer uma das categorias `PESSOA`, `ORGANIZACAO` ou `ABSTRACCAO`. Note-se que basta existir apenas uma referência à categoria e/ou tipo num dado alinhamento (ou seja, tanto nas EM da CD como nas EM da saída) para que este seja considerado e escrito.

A restrição nos tipos é representada por uma lista de tipos entre parênteses imediatamente a seguir à respectiva categoria. Por exemplo, a lista `'PESSOA(CARGO,GRUPOMEMBRO):ORGANIZACAO'` filtra os alinhamentos para procurar EM de categorias `ORGANIZACAO` e `PESSOA`, sendo que só tipos `CARGO` e `GRUPOMEMBRO` é que são tidos em conta para a categoria `PESSOA`.

-genero, que especifica o(s) género(s) textual(is) a filtrar. Recebe uma lista de géneros separados por ':', ou então um único género textual. Os valores da lista devem estar mencionados na lista `GENEROS` do ficheiro `harem.conf`. Por exemplo, ao especificar `-genero Web`, o Véus escreve todos os alinhamentos de documentos de género textual Web.

-origem, que especifica a(s) variante(s) a filtrar. Recebe uma lista de variantes separadas por ':', ou então uma variante. Os valores da lista devem estar mencionados na lista `ORIGENS` do ficheiro `harem.conf`. Por exemplo, ao especificar `-origem PT`, o Véus filtra e escreve todos os alinhamentos de documentos da variante portuguesa.

-estilo, que pode ter um dos três valores seguintes: **muc**, **relax** e **harem**. Com o valor **muc**, o Véus retira todos os alinhamentos que geraram uma pontuação `parcialmente_correcto`, o que simula o cenário da avaliação dos MUC-6 e MUC-7, que não reconhecia este tipo de pontuação.Com o valor **relax**, o Véus aceita apenas no máximo uma pontuação `parcialmente_correcto` por cada de alinhamento a uma EM na CD. Ou seja, nos casos em que a EM na CD alinhe com várias EM da saída, ou uma EM da saída alinhe com várias EM da CD (gerando várias pontuações `parcialmente_correcto`), só o primeiro alinhamento é pontuado com `parcialmente_correcto`, enquanto que os restantes serão classificadas como `espurio` ou `em_falta`). Esta opção pode ser vista como uma restrição aos alinhamentos múltiplos. Finalmente, com a opção `harem`, todos os alinhamentos parcialmente correctos são considerados para avaliação.

D.2.5 AltinaID

Para executar o AltinaID, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.IdentificationAltAlignmentSelector -alinhamento
FICHEIRO_VEUS > FICHEIRO_ALTINAID
```

`FICHEIRO_VEUS` corresponde ao ficheiro gerado pelo Véus (ou, no caso de não se querer filtrar alinhamentos, pode-se usar o ficheiro gerado pelo AvalIDa). O AltinaID escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_ALTINAID`.

D.2.6 Ida2ID

Para executar o Ida2ID, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.GlobalIdentificationSelector -alinhamento
FICHEIRO_ALTINAID > FICHEIRO_IDA2ID
```

`FICHEIRO_ALTINAID` corresponde ao ficheiro gerado pelo AltinaID, ou seja, sem nenhuma alternativa `<ALT>`. O Ida2ID escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_IDA2ID`.

D.2.7 Emir

Para executar o Emir, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.SemanticAlignmentEvaluator -alinhamento
FICHEIRO_ALTINAID [-relativo sim] > FICHEIRO_EMIR
```

`FICHEIRO_ALTINAID` corresponde ao ficheiro gerado pelo AltinaID, ou seja, já sem nenhuma etiqueta `<ALT>`. O Emir escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_EMIR`.

O Emir aceita o parâmetro opcional `-relativo` com o valor **sim**, para assinalar ao Emir que a avaliação deve ser realizada segundo o cenário relativo (iso é, considerando apenas as EM identificadas como correctas ou parcialmente correctas pela saída). Se nada for especificado, o Emir avalia segundo um cenário absoluto (ou seja, considerando todas as EM da CD, incluindo as que não foram identificadas como correctas ou parcialmente correctas pelo sistema).

D.2.8 AltinaSEM

Para executar o AltinaSEM, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.SemanticAltAlignmentSelector -alinhamento
FICHEIRO_EMIR > FICHEIRO_ALTINASEM
```

`FICHEIRO_EMIR` corresponde ao ficheiro gerado pelo Emir. O AltinaSEM escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_ALTINASEM`.

D.2.9 Ida2SEM

Para executar o Ida2SEM, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.GlobalSemanticSelector -alinhamento
FICHEIRO_ALTINASEM > FICHEIRO_IDA2SEM
```

`FICHEIRO_ALTINASEM` corresponde ao ficheiro gerado pelo AltinaSEM. O Ida2SEM escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_IDA2SEM`.

D.2.10 Vizir

Para executar o Vizir, usa-se o seguinte comando:

```
vizir.pl [-abs|-rel] -i FICHEIRO_VEUS|AVALIDA -o FICHEIRO_VIZIR
```

O parâmetro `-i` é obrigatório e especifica o ficheiro gerado pelo Véus ou pelo AvalIDa, `FICHEIRO_VEUS|AVALIDA`. O parâmetro `-o` especifica o ficheiro de escrita do Vizir, `FICHEIRO_VIZIR`. Caso esta opção não seja preenchida, é usado o nome do ficheiro `FICHEIRO_VEUS|AVALIDA`, acrescido da extensão `.vizir`.

O Vizir obriga a especificar o tipo de cenário a usar na avaliação. Para tal, é necessário optar por um dos seguintes parâmetros: `-abs`, para cenário absoluto que considera todas as EM para avaliação, ou `-rel`, para cenário relativo, que não considera as EM espúrias nem com classificação morfológica espúria.

D.2.11 AltinaMOR

Para executar o AltinaMOR, usa-se o seguinte comando:

```
altinamor.pl [-abs|-rel] -i FICHEIRO_VIZIR -o FICHEIRO_ALTINAMOR
```

O parâmetro `-i` é obrigatório e especifica o ficheiro gerado pelo Vizir, `FICHEIRO_VIZIR`. O parâmetro `-o` especifica o ficheiro de escrita do AltinaMOR, `FICHEIRO_ALTINAMOR`. Caso esta opção não seja especificada, é usado o nome do `FICHEIRO_VIZIR`, mais a extensão `.altmor`.

D.2.12 Ida2MOR

Para executar o Ida2MOR, usa-se o seguinte comando:

```
ida2mor.pl [-abs|-rel] -i FICHEIRO_ALTINAMOR -o FICHEIRO_IDA2MOR
```

O parâmetro `-i` é obrigatório e especifica o ficheiro gerado pelo AltinaMOR, `FICHEIRO_ALTINAMOR`. O parâmetro `-o` especifica o ficheiro criado pelo Ida2MOR, `FICHEIRO_IDA2MOR`. Caso esta opção não seja preenchida, é usado o nome do `FICHEIRO_ALTINAMOR`, acrescido da extensão `.ida2mor`.

D.2.13 Sultão

Para executar os três módulos do Sultão, omeadamente SultãoID, SultãoMOR e SultãoSEM, usam-se os seguintes comandos, respectivamente:

```
java -Dfile.encoding=ISO-8859-1 -jar ferramentas_HAREM_java.jar
pt.linguateca.harem.GlobalIdentificationReporter [-filtro FILTRO]
[-naooficiais LISTA_NAOOFICIAIS] [-depurar sim|nao]
[-saidas oficiais|naooficiais] > FICHEIRO_SULTAOID

java -Dfile.encoding=ISO-8859-1 -jar ferramentas_HAREM_java.jar
pt.linguateca.harem.GlobalMorphologyReporter [-filtro FILTRO]
[-naooficiais LISTA_NAOOFICIAIS] [-depurar sim|nao]
[-saidas oficiais|naooficiais] > FICHEIRO_SULTAOMOR

java -Dfile.encoding=ISO-8859-1 -jar ferramentas_HAREM_java.jar
pt.linguateca.harem.GlobalSemanticReporter [-filtro FILTRO]
[-naooficiais LISTA_NAOOFICIAIS] [-depurar sim|nao]
[-saidas oficiais|naooficiais] [-tipos sim|nao] > FICHEIRO_SULTAOSEM
```

O Sultão é executado com os seguintes parâmetros opcionais, que podem ser combinados entre si:

-filtro, que diz respeito aos ficheiros que deverão ser utilizados na geração dos relatórios, e recebe como valor o sufixo do ficheiro. Por exemplo, se usar no `FILTRO` o valor '`total.altid.ida2id`', o Sultão processa todos os ficheiros terminados com a extensão `total.altid.ida2id`. Se se pretende mais do que um padrão de ficheiros, pode-se utilizar uma lista de extensões separadas por ':', como por exemplo em `total.local.altid.ida2id:total.organizacao.altid.ida2id`.

-naooficiais, que indica ao Sultão quais os ficheiros que correspondem a saídas não oficiais entregues pelos participantes. O parâmetro recebe como valor o prefixo do ficheiro, que deve ter o nome da saída, como no seguinte exemplo:

```
-naooficiais sistema1_nao_oficial:sistema4
```

O exemplo indica que os ficheiros cujos nomes começam por `sistema1_nao_oficial` ou `sistema4` são para ser considerados não oficiais, e a sua entrada na tabela de resultados não vai ter o pseudónimo a negrito, mas sim a itálico.

-saidas, que indica ao Sultão as saídas que devem ser consideradas. A este parâmetro podem ser atribuídos dois valores: `oficiais` e `naooficiais`. No primeiro caso, só as saídas oficiais é que serão exportadas para o relatório final. No segundo, só as saídas não oficiais é que são consideradas. Se este parâmetro não for utilizado, todas as saídas são consideradas.

-depurar, que pode tomar os valores `sim` ou `nao`. Por defeito, o Sultão assume que a informação para depuração não é para ser colocada no relatório e que a anonimização é

para ser efectuada. Se o parâmetro for fornecido com o valor `sim`, então a anonimização não é efectuada e informação adicional é colocada no relatório final.

-tipos, parâmetro usado apenas no SultãoSEM, e que pode tomar os valores `sim` ou `nao`. Este parâmetro indica ao SultãoSEM se as tabelas referente à avaliação dos tipos devem ou não ser produzidas. Este opção existe uma vez que a avaliação dos tipos é sempre relativa (porque só se avaliam os tipos quando a categoria está correcta), logo os valores destas tabelas seriam sempre iguais na avaliação absoluta e relativa.

D.2.14 Alcaide

para executar o Alcaide, usa-se o seguinte comando:

```
perl alcaide.pl -sistema SISTEMA -run SAIDA -id ID -morf MORF
     -sem SEM -output SAIDA -workingdir DIRECTORIA
```

O Alcaide necessita obrigatoriamente dos seguintes parâmetros:

-sistema, com o nome do sistema que gerou a saída.

-run, com o nome da saída. Este nome deve ser exactamente igual ao nome da directoria que contém os relatórios de entrada, e também ao nome pelo qual começam os nomes dos ficheiros gerados pelos programas Ida2ID, Ida2MOR e Ida2SEM.

-id, que pode tomar o valor de 0 ou 1, assinala ao Alcaide que se pretende gerar tabelas da tarefa de identificação para o relatório individual.

-morf, que pode tomar o valor de 0 ou 1, assinala ao Alcaide que se pretende gerar tabelas da tarefa de classificação morfológica para o relatório individual.

-sem, que pode tomar o valor de 0 ou 1. Diz ao Alcaide que se pretende gerar tabelas da tarefa de classificação semântica para o relatório individual.

-output, que indica a directoria onde o Alcaide irá escrever o relatório. Esta directoria tem de conter uma subdirectoria chamada `images`, para armazenar as imagens que são criadas automaticamente pelo programa.

-workingdir, que designa a directoria raiz com os relatórios do Sultão, Ida2ID, Ida2MOR e Ida2SEM.

D.3 Ficheiro de configuração do HAREM, `harem.conf`

Neste apêndice, apresenta-se o ficheiro `harem.conf` usado no Mini-HAREM para definir as categorias e tipos válidos, bem como os géneros textuais e variantes autorizadas.

```
[ENTIDADES]
PESSOA:INDIVIDUAL,CARGO,GRUPOIND,GRUPOMEMBRO,MEMBRO,GRUPOCARGO
ORGANIZACAO:ADMINISTRACAO,EMPRESA,INSTITUICAO,SUB
TEMPO:DATA,HORA,PERIODO,CICLICO
LOCAL:CORREIO,ADMINISTRATIVO,GEOGRAFICO,VIRTUAL,ALARGADO
OBRA:ARTE,REPRODUZIDA,PUBLICACAO
ACONTECIMENTO:EFEMERIDE,ORGANIZADO,EVENTO
ABSTRACCAO:DISCIPLINA,ESTADO,ESCOLA,MARCA,PLANO,IDEIA,NOME,OBRA
COISA:CLASSE,SUBSTANCIA,OBJECTO,MEMBROCLASSE
VALOR:CLASSIFICACAO,QUANTIDADE,MOEDA
VARIADO:OUTRO

[GENEROS]
CorreioElectrónico
Entrevista
Expositivo
Jornalístico
Literário
Político
Técnico
Web

[ORIGENS]
AO
BR
CV
IN
MO
MZ
PT
TL
```