Capítulo 8

PorTexTO: sistema de anotação/extracção de expressões temporais

Olga Craveiro, Joaquim Macedo e Henrique Madeira

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 8, p. 159–170.

As técnicas de recolha de informação assumiram um papel de grande relevo nos últimos anos, em virtude da importância assumida pelos motores de busca na Internet. No entanto, a utilização de informação temporal para melhorar os resultados das pesquisas tem sido pouco explorada, apesar de existir um grande potencial para conseguir esse melhoramento. De facto, a noção de tempo é essencial para muitas das pesquisas efectuadas num sistema de recolha de informação, como por exemplo, na área da saúde onde será pertinente reconstruir o historial clínico dos pacientes com a capacidade de encontrar eventos e apresentá-los num espaço temporal, permitindo, desta forma, dar maior exactidão ao relatório (Alonso et al., 2007). Outro dos exemplos da aplicação da dimensão temporal nos sistemas de recolha de informação é dado pelo trabalho desenvolvido por Uehara e Sato (2005) na implementação de arquivos www.

No entanto, nem sempre o tempo surge de forma explícita nos documentos, mas as referências temporais podem ajudar a identificar a relevância dos documentos encontrados. Beigbeder (2004) apresenta um estudo de como os diferentes aspectos temporais podem intervir nas diversas etapas da recolha de informação. O interesse no processamento de informação temporal tem crescido nos últimos anos e tem-se intensificado nas mais diversas áreas de investigação (Mani et al., 2004; Pustejovsky et al., 2005).

O objectivo do sistema que desenvolvemos é o de identificar informação temporal existente em documentos, para posteriormente ser utilizada, como papel importante na ordenação da lista de resultados obtida pelas pesquisas efectuadas em sistemas de recolha de informação.

Como existe ainda pouco trabalho desenvolvido no processamento de informação temporal da língua portuguesa, decidimos criar um sistema de raiz que seguisse um algoritmo simples e rápido. O processo de anotação/extracção de informação num sistema de recolha de informação terá de ser rápido para que não comprometa todo o sistema.

Pretende-se que o sistema PorTexTO, designado por *PORtuguese Temporal EXpressions Tool*, seja um sistema simples e com baixo tempo de processamento. Para que o desempenho não seja comprometido, o sistema poderá não encontrar todas as expressões temporais existentes nos documentos que processar, mas deverá encontrar as que ocorram mais vezes nos documentos em português e que são definidas através de estudo estatístico.

O PorTexTO faz um processamento frase a frase dos documentos, e utiliza padrões de expressões temporais para o reconhecimento das entidades mencionadas, ao contrário de outros sistemas de processamento de linguagem natural onde o processamento é feito termo a termo, identificando cada termo segundo as suas características linguísticas (Mani, 2004).

Na detecção de expressões temporais na língua inglesa existem diversas abordagens, embora não seja do nosso conhecimento que alguma utilize padrões de expressões criados com recurso às co-ocorrências das palavras temporais. Uma das abordagens apresentada por Mani e Wilson (2000) utiliza uma anotação manual num conjunto de teste e um conjunto de regras obtidas por aprendizagem automática. Makkonen e Ahonen-myka (2003) fazem uma divisão dos termos temporais em categorias e utilizam autómatos de estados finitos no reconhecimento das expressões. Outra abordagem que também utiliza autómatos de estados finitos foi apresentada por Schilder e Habel (2001), mas que introduziram preposições nos seus autómatos. Esta abordagem tem por base o trabalho de Allen (1983) na detecção de intervalos temporais.

A implementação do sistema PorTexTO seguiu as directivas gerais e as directivas do TEMPO (Hagège et al., 2008) publicadas para o Segundo HAREM.

Neste capítulo é descrito o sistema PorTexTO, sendo apresentadas em pormenor as várias etapas de processamento dos dois módulos do sistema (`Anotador` e `Processador de co-ocorrências`), a sua participação no Segundo HAREM e é ainda efectuada uma análise aos resultados obtidos. Por fim, são apresentadas as conclusões e o trabalho futuro.

8.1 Descrição do sistema

O sistema PorTexTO é um sistema de reconhecimento de entidades mencionadas temporais em textos na língua portuguesa. Neste sistema, o processamento dos documentos é efectuado frase a frase, ou seja, o texto é previamente dividido em frases, com a ajuda do atomizador da Linguateca (o módulo de Perl `Lingua::PT::PLNbase`[1]) passando em seguida cada frase pelas diversas etapas de processamento.

A identificação das expressões temporais é feita usando padrões de expressões, criados a partir de co-ocorrências existentes em referências temporais. Os padrões encontram-se armazenados num ficheiro para que facilmente possam ser acrescentados novos padrões ou alterados os existentes. Este ficheiro foi designado por `REGEX` e encontra-se representado na figura 8.1.

O PorTexTO permite o processamento de documentos tanto em formato de texto simples não estruturado como em formato estruturado em XML. O resultado produzido pelo sistema pode ser um ficheiro no seu formato original, mas com as devidas anotações nas expressões temporais encontradas ou então um ficheiro com todas as expressões temporais encontradas e a sua posição relativamente ao texto original. Como o formato pretendido pelo Segundo HAREM era o do ficheiro original devidamente anotado, só será abordado neste capítulo o módulo que produz este resultado e que foi designado por módulo `Anotador`. A figura 8.1 apresenta a arquitectura deste módulo.

Este módulo tem como entrada o texto original e os padrões de expressões que são previamente criados por outro dos módulos do PorTexTO, o módulo `Processador de co-ocorrências` (ver descrição detalhada na secção 8.1.2).

Além da colecção e dos padrões de expressões, o sistema tem ainda como entrada uma lista de palavras-chave temporais usadas para unicamente excluir do processamento frases que não contenham expressões temporais e assim conseguir diminuir o tempo final de processamento dos documentos. Esta lista funciona no sistema como um filtro das frases a serem processadas. As frases excluídas são todas as que não têm nem datas, nem nenhuma das palavras-chave temporais. As palavras-chave temporais são definidas consoante a lista de expressões temporais que o sistema deverá identificar e classificar, de modo a atingir os objectivos de uma determinada tarefa. Por exemplo, se existirem padrões de expressões temporais com a palavra temporal *ano*, então *ano* deverá existir na lista de palavras-chave temporais.

De seguida, são apresentados com maior detalhe os módulos `Anotador` e `Processador de co-ocorrências`.

8.1.1 Módulo `Anotador`

O módulo `Anotador` é responsável por identificar as expressões temporais, mediante os padrões definidos pelo módulo `Processador de co-ocorrências`, fazer a sua classificação

[1] Disponível em http://search.cpan.org/~ambs/Lingua-PT-PLNbase-0.20.

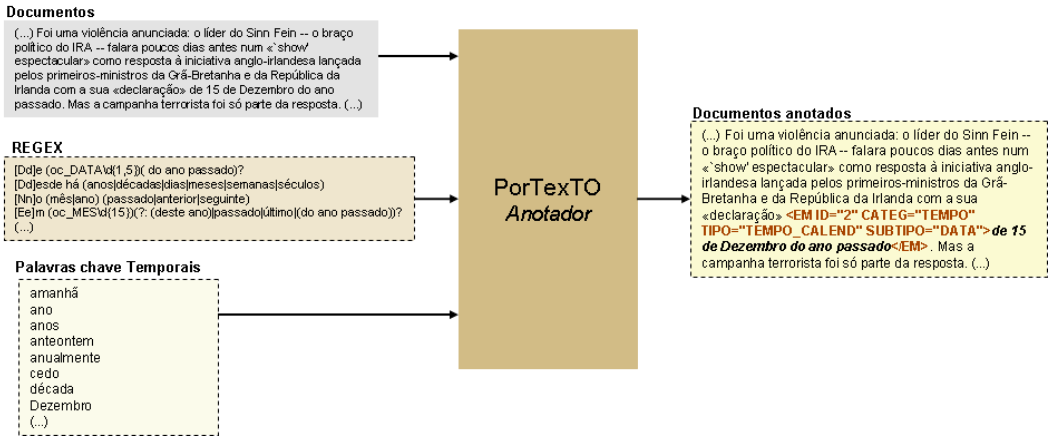


Figura 8.1: Arquitectura do módulo `Anotador` do sistema PorTexTO.

e, posteriormente, proceder à anotação no texto original.

O funcionamento deste módulo é apresentado na figura 8.2. Os documentos de entrada são trabalhados um de cada vez e o processamento de cada documento é feito numa frase de cada vez. Cada frase será submetida às quatro etapas de processamento do `Anotador` do PorTexTO. A frase só será dividida nos seus termos caso haja necessidade de reconhecer datas com o mês por extenso, ou datas que tenham também o dia da semana. Por exemplo, a frase *Domingo, 7 de Setembro de 2008* é dividida nos seguintes termos: *Domingo*, *7*, *de*, *Setembro*, *de*, *2008*. Com estes termos, é verificado se os que estão à esquerda e à direita do mês podem fazer ou não parte de uma data. No caso desses termos poderem constar de uma data então serão incluídos na expressão que vai ser marcada como `DATA`. No caso contrário, só o mês é que será marcado, mas com o marcador `MES`. No exemplo apresentado, a frase inicial terá a marcação `DATA`.

Na primeira etapa é decidido se a frase vai ser ou não processada. Uma frase só será processada caso possa conter, pelo menos, uma expressão temporal. Isto é, a frase deverá ter termos numéricos ou, pelo menos, uma das palavras definidas na lista de palavras-chave temporais (ver figura 8.1). As frases que não têm nenhuma expressão temporal ficam excluídas do processamento.

Vamos considerar a frase (8.1), extraída da colecção do Segundo HAREM, para exemplificar as próximas etapas do processamento.

(8.1) A missão científica da nave foi concluída *em 30 de abril de 2002*.

Na segunda etapa são geradas expressões candidatas a entidades mencionadas temporais. No caso da frase conter dígitos, o sistema aplica regras para reconhecer horas, datas completas ou incompletas (datas constituídas só por dia e mês ou mês e ano) e anos.

De seguida, as frases são analisadas quanto à existência de meses (palavras por extenso ou abreviaturas) e dias da semana, sendo posteriormente aplicadas regras para reconhecimento de datas constituídas por termos numéricos e palavras. Depois de identificadas, as expressões são marcadas e passam a ser expressões candidatas. Nesta etapa, a frase do exemplo passaria a:

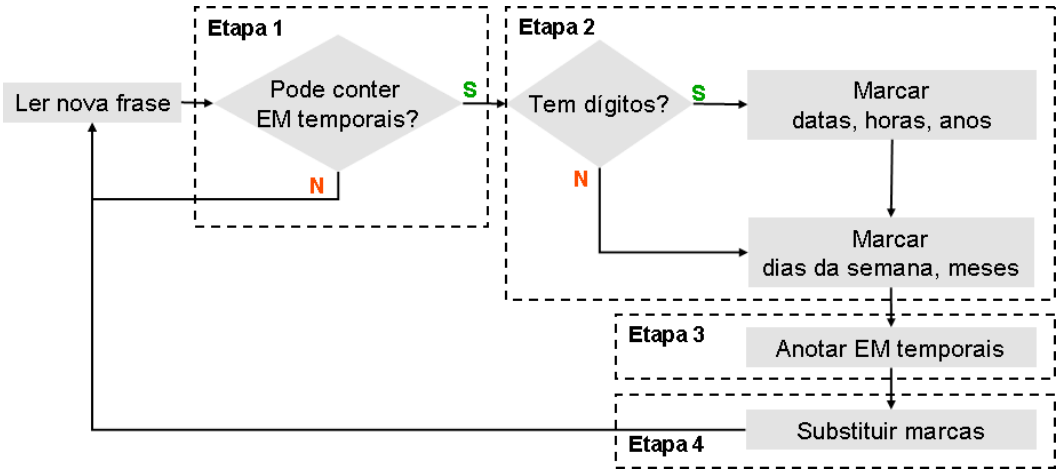


Figura 8.2: Funcionamento do módulo `Anotador` do sistema PorTexTO.

(8.2) A missão científica da nave foi concluída *em* `oc_DATA12`.

Na terceira etapa, verifica-se a correspondência entre a frase resultante das anteriores etapas (frase (8.2)) e os padrões definidos para o actual processamento e que foram anteriormente criados pelo módulo `Processador de co-ocorrências` (ver secção 8.1.2). A frase (8.2) tem correspondência com o padrão, definido através da seguinte expressão regular:

```
[Ee]m (oc_DATA\d{1,5})(?: (deste ano)|passado|último)?
```

A anotação da expressão será efectuada caso ocorra correspondência e a classificação atribuída está associada ao padrão responsável pela correspondência. A frase (8.2) resultaria na frase (8.3).

(8.3) A missão científica da nave foi concluída <EM ID="41" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**em oc_DATA12**</EM>.

A quarta e última etapa é responsável por substituir pelo texto original, todas as marcas que foram colocadas na frase durante a execução da segunda etapa de processamento. A frase de exemplo depois de terminado o processamento ficará como em (8.4).

(8.4) A missão científica da nave foi concluída <EM ID="41" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**em 30 de Abril de 2002**</EM>.

8.1.2 Módulo `Processador de co-ocorrências`

O módulo `Processador de co-ocorrências` só é executado quando ainda não existem padrões de expressões temporais ou os que existem são insuficientes para a tarefa a desempenhar pelo PorTexTO. O objectivo principal deste módulo é determinar as expressões temporais mais utilizadas numa determinada colecção segundo uma abordagem estatística e

com estas expressões criar os padrões que vão ser posteriormente utilizados no módulo `Anotador`. Este módulo produz como resultado um ficheiro com os padrões definidos através de expressões regulares e a respectiva classificação. A arquitectura deste módulo é apresentada na figura 8.3.

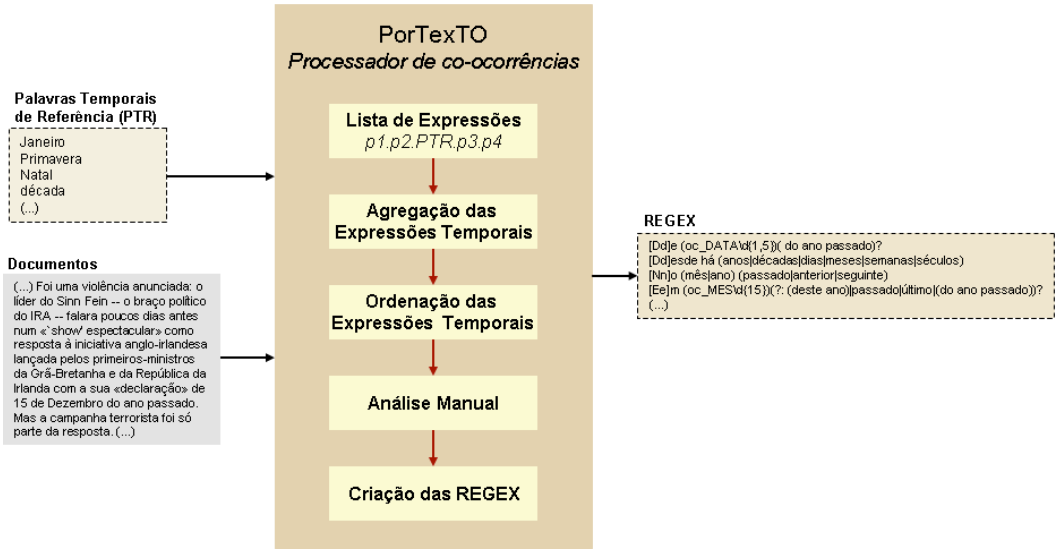


Figura 8.3: Arquitectura do módulo `Processador de co-ocorrências` do sistema PorTexTO.

A determinação das co-ocorrências é realizada utilizando um conjunto de palavras, palavras essas que são referidas neste capítulo como palavras temporais de referência (PTR). A lista das PTR deve ser constituída por todas as palavras temporais que apareçam em expressões com no mínimo duas palavras, como por exemplo, meses do ano, estações do ano, dias da semana, festividades e unidades de medida temporal, para que realmente haja a necessidade de determinar as palavras que ocorrem conjuntamente. Assim, as palavras que sozinhas formam uma expressão temporal devem ser excluídas desta lista, como por exemplo, advérbios temporais com terminação *-mente* (e.g., *diariamente*).

O funcionamento deste módulo está dividido em cinco etapas de processamento e tal como no módulo `Anotador` os documentos de entrada também são processados frase a frase.

O objectivo da primeira etapa é obter uma lista com as expressões encontradas onde as co-ocorrências detectadas têm uma distância máxima de *n* palavras antes e/ou *n* palavras depois da palavra temporal de referência. Por exemplo, considerando a palavra temporal de referência *ano* podemos obter as seguintes expressões *No* ***ano*** *passado*, *No último* ***ano***, *No próximo* ***ano*** *de 2009*. A lista obtida, para além das expressões encontradas, tem também o respectivo número de ocorrências.

Na segunda etapa é feita a agregação das expressões temporais contidas na lista criada na etapa anterior. As expressões são agregadas quando têm mais do que uma palavra que ocorre com uma determinada palavra temporal de referência, na mesma posição. Quando ocorre a agregação de expressões o número total de ocorrências passa a ser a soma das ocorrências de cada uma das expressões agregadas. Por exemplo, as expres-

sões *No **ano** passado* e *No **ano** seguinte* são agregadas e passamos a ter o padrão `No` **`ano`** `passado|seguinte`.

A agregação também é feita para as referências temporais consideradas como datas, os meses do ano e as horas. Por exemplo, as expressões *Em **Janeiro*** e *Em **Fevereiro*** são agregadas no padrão `Em` **`oc_MES`**. No caso da expressão *No dia 25 de Janeiro* o padrão criado seria `No dia` **`oc_DATA`**.

A terceira etapa ordena a lista de expressões, já com expressões que foram agregadas, por ordem decrescente do número total de ocorrências.

Na quarta etapa é efectuada uma análise manual necessária para excluir todas as expressões que embora tenham uma unidade lexical que representa um elemento temporal, não sejam na realidade uma expressão temporal e outras que não façam sentido, pelo senso comum de quem tem domínio da língua. Por exemplo, as expressões *Feliz Natal* e *Bom Ano Novo* são excluídas da lista, embora de acordo com as directivas do Segundo HAREM estas expressões temporais devessem ter como classificação `GENERICO` no atributo `TIPO` (Hagège et al., 2008). No entanto o PorTexTO não faz esta classificação.

Esta é a etapa mais complexa do módulo `Processador de co-ocorrências`, pois mesmo com o conhecimento da língua portuguesa e seguindo as directivas do TEMPO definidas para o Segundo HAREM aparecem sempre situações dúbias tornando-se difícil decidir se realmente uma determinada expressão deve ser ou não considerada uma expressão temporal.

A quinta etapa cria as expressões regulares que definem os padrões que foram considerados após a análise manual e associa a respectiva classificação, segundo as directivas do Segundo HAREM (Hagège et al., 2008). Como exemplo, apresentamos de seguida um dos padrões criados:

```
[Nn]o (mês|ano) (passado|anterior|seguinte)
```

8.2 Participação no Segundo HAREM

A participação do PorTexTO no Segundo HAREM foi efectuada com a sua versão 1.0 (PorTexTO 1.0). Este sistema só foi submetido à avaliação nas tarefas de identificação e classificação de entidades mencionadas no cenário `TEMPO`, uma vez que só processa expressões temporais.

Na classificação das entidades mencionadas temporais o sistema PorTexTO só utilizou o `TIPO` e o `SUBTIPO` da categoria `TEMPO`, não tendo utilizado nem a classificação de `TEMPO` estendido, nem a normalização. Além disso, ficaram excluídos da classificação o subtipo `INTERVALO` do tipo `TEMPO_CALEND` e o tipo `GENERICO`.

Os padrões utilizados no sistema PorTexTO 1.0 só representam expressões temporais simples que se iniciam com os termos *a*, *às*, *de*, *em*, *há*, *durante*, *desde*, *pelas*, e os termos *no*, *naquele*, *neste*, *nesse*, *este*, *esse* também no género feminino e no plural. Por expressões temporais simples entendem-se as expressões linguísticas compostas por uma só unidade lexical que denote um elemento temporal.

As expressões temporais compostas, como por exemplo, *no **dia** 10 do **mês** passado*, no PorTexTO 1.0 são tratadas como duas expressões simples, não obtendo a correcta classificação segundo as directivas do TEMPO do Segundo HAREM (Hagège et al., 2008).

Os padrões das expressões temporais foram criados pelo módulo `Processador de co-ocorrências` utilizando a lista das palavras temporais de referência composta pelos

meses do ano, estações do ano, dias da semana, festividades (*Natal*, *Páscoa*, *Carnaval* e *Entrudo*), unidades de medida temporal (*década*, *período*, *século*, *ano*, *mês*, etc.) e outras (*altura*, *instante*, *momento*, *tempo*). Na determinação das co-ocorrências a distância máxima considerada foi de *n*=2 palavras antes e/ou *n*=2 palavras depois da palavra temporal de referência.

O módulo `Processador de co-ocorrências` foi aplicado à colecção do HAREM versão 2.0[2] utilizada nos dois eventos de avaliação do Primeiro HAREM. Esta colecção apresenta cerca de 520 mil palavras distribuídas por cerca de 40 mil linhas e provenientes de 1202 documentos de diferentes géneros textuais (textos técnicos, políticos, literários, expositivos, jornalísticos, entrevistas, mensagens de correio electrónico e páginas web) (Santos e Cardoso, 2007a).

Os padrões de expressões utilizados no Segundo HAREM só foram criados depois da colecção do Segundo HAREM também ter sido submetida ao módulo `Processador de co-ocorrências`. Seguindo uma abordagem estatística, os resultados obtidos neste módulo foram posteriormente incluídos na lista de expressões que já possuíamos aquando do processamento dos outros textos e verificou-se o aparecimento de novas expressões temporais e alteração da ordem de outras expressões na lista de frequência.

O PorTexTO 1.0 participou com quatro corridas. A corrida PorTexTO_1 serviu só para validar o envio de resultados ao Segundo HAREM e verificar se existiam algumas incoerências na anotação das entidades mencionadas. Após o envio desta corrida verificámos que estava tudo correcto.

As outras corridas (PorTexTO_2, PorTexTO_3 e PorTexTO_4) têm pequenas diferenças ao nível da definição das expressões regulares utilizadas – mais precisas e menos abrangentes ou menos precisas e mais abrangentes, como por exemplo, `[Nn]o (passado|último) ano` e `[Nn]o \w+ ano`, respectivamente.

O envio destas três corridas teve por objectivo podermos avaliar qual a penalização da precisão quando se aumentou a abrangência na definição das expressões regulares usadas nos padrões.

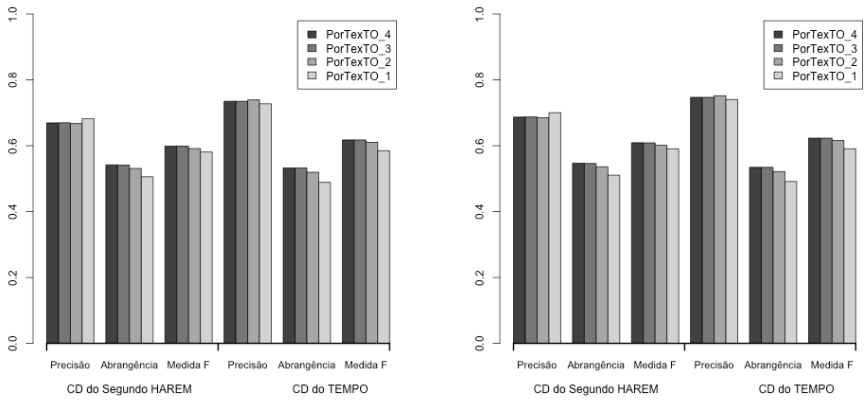
8.3 Resultados da participação no Segundo HAREM

Os resultados obtidos pelo sistema PorTexTO 1.0 na sua primeira participação em avaliações conjuntas excederam as nossas expectativas porque esta versão apresentava muitas limitações, tal como foi referido na secção 8.2. O sistema foi criado de raiz para a participação no Segundo HAREM e não houve tempo suficiente até à avaliação para tudo estar implementado e devidamente testado.

O sistema utilizou para o processamento um computador pessoal com 1GB de memória RAM, processador Intel Core 2 E6600 a 2.4 GHz e sistema operativo Microsoft Windows XP Professional, versão 2002, SP2. Em termos do desempenho computacional, o sistema PorTexTO 1.0 anotou a colecção do Segundo HAREM a um débito de aproximadamente 22KB por segundo, tendo processado as 33 mil linhas com cerca de 675 mil palavras da colecção do Segundo HAREM em cerca de três minutos e vinte segundos. O tempo de processamento conseguido é bastante aceitável para os objectivos traçados para o sistema, como por exemplo a sua incorporação numa aplicação de recolha de informação (em inglês, *ad-hoc retrieval)*.

[2] Disponível no sítio do HAREM (http://www.linguateca.pt/HAREM), na secção dedicada ao Primeiro HAREM.

Na apresentação dos resultados obtidos iremos focar-nos no único cenário que se adequa ao nosso sistema, o cenário constituído pela categoria TEMPO, uma vez que nos restantes cenários existem outras categorias que estão também a ser avaliadas e que o PorTexTO não tentou reconhecer. Faremos a análise tanto na CD do Segundo HAREM, como na CD do TEMPO. Neste último caso, apenas estamos interessados no modo de avaliação do HAREM clássico, ou seja, sem ter em conta os atributos específicos da categoria TEMPO, pois não tentámos atribuí-los. Não fizemos distinção entre os dois tipos de avaliação de ALT, já que nas directivas do TEMPO não foi considerada a etiqueta ALT e, consequentemente, no cenário da categoria TEMPO a avaliação estrita ou relaxada de ALT produz os mesmos valores.



(a) Classificação (b) Identificação

Figura 8.4: Resultados das quatro corridas do PorTexTO obtidos na pista do TEMPO na CD do Segundo HAREM e na CD do TEMPO.

Como se pode constatar pelos gráficos da figura 8.4, as quatro corridas apresentam resultados bastante similares com diferenças pouco significativas. No entanto, a corrida PorTexTO_1 teve uma maior diferença relativamente às outra corridas, como era esperado, já que quando esta corrida foi submetida a avaliação o sistema ainda não estava devidamente configurado.

Nos resultados obtidos nas corridas PorTexTO_2, PorTexTO_3 e PorTexTO_4 verificámos que a definição dos padrões de forma mais abrangente não foi muito penalizadora para a precisão, em ambas as colecções douradas e tanto na identificação como na classificação.

Na CD do TEMPO, as corridas PorTexTO_3 e PorTexTO_4 chegam a ter os mesmos resultados.

Os resultados obtidos pelo PorTexTO na categoria TEMPO na CD do Segundo HAREM, relativamente às métricas abrangência e precisão para a classificação e para a identificação não têm diferenças significativas (ver tabela 8.1), ou seja, as entidades temporais identificadas pelo PorTexTO como TEMPO estão a ser bem classificadas quanto ao seu tipo e subtipo. Aliás, este comportamento é seguido na CD do TEMPO, como se pode constatar pelos resultados apresentados na tabela 8.2. Como as três corridas (PorTexTO_2, PorTexTO_3 e

Tabela 8.1: Resultados do PorTexTO na pista do TEMPO na CD do Segundo HAREM.

Corrida	Classificação			Identificação		
	Precisão	Abrangência	Medida F	Precisão	Abrangência	Medida F
PorTexTO_4	0,6694	**0,5419**	**0,5990**	0,6871	**0,5470**	**0,6091**
PorTexTO_3	0,6698	0,5410	0,5986	0,6875	0,5462	0,6087
PorTexTO_2	0,6674	0,5310	0,5915	0,6849	0,5360	0,6014
PorTexTO_1	**0,6825**	0,5058	0,5810	**0,7002**	0,5106	0,5905

Tabela 8.2: Resultados da corrida PorTexTO_4 na pista do TEMPO na CD do Segundo HAREM (CDSH) e na CD do TEMPO (CDT).

CD	Classificação			Identificação		
	Precisão	Abrangência	Medida F	Precisão	Abrangência	Medida F
CDSH	0,6694	**0,5419**	0,5990	0,6871	**0,5470**	0,6091
CDT	**0,7350**	0,5327	**0,6177**	**0,7470**	0,5345	**0,6231**

PorTexTO_4) tiveram resultados bastante idênticos, nesta tabela só apresentamos a corrida PorTexTO_4.

O PorTexTO conseguiu uma pequena melhoria na precisão quando a avaliação utilizou o subconjunto da CD – CD do TEMPO, mas que não é significativa. A diferença foi de aproximadamente 6%.

Nesta CD, o sistema PorTexTO na corrida PorTexTO_4 obteve a 5.ª posição, mas a diferença relativamente à corrida XIP-L2F/Xerox_3 do sistema que obteve a 1.ª posição verificou-se somente na abrangência. O sistema XIP-L2F/Xerox foi o melhor sistema com uma precisão de aproximadamente 75% para uma abrangência de 78%. A figura 8.5 faz a comparação entre os resultados obtidos pelos dois sistemas.

A precisão poderá ser melhorada quando o sistema passar a identificar expressões temporais compostas. Mas a sua maior limitação deve-se à abrangência dos padrões utilizados, sendo necessário acrescentar mais padrões para conseguir reconhecer um maior número de entidades mencionadas temporais. Por exemplo, padrões para identificar expressões com classificação de subtipo `INTERVALO` do tipo `TEMPO_CALEND` (*entre 1990 e 2000*, *de 2 a 6 meses*).

8.4 Conclusões e trabalho futuro

O sistema PorTexTO foi desenvolvido com o objectivo de utilizar um algoritmo simples para conseguir obter um bom desempenho computacional, mesmo que não identifique e classifique todas as expressões temporais, conforme foi referido na secção introdutória deste capítulo.

A participação do PorTexTO no Segundo HAREM foi bastante importante, pois para além de permitir saber qual o desempenho do sistema nas suas tarefas de reconhecimento de entidade mencionadas temporais, também facultou o acompanhamento na criação das directivas de classificação e normalização destas entidades.

Os resultados obtidos pelo sistema PorTexTO na sua versão 1.0 são bastante motivadores. Os resultados vieram demonstrar que o algoritmo seguido pelo sistema permite obter

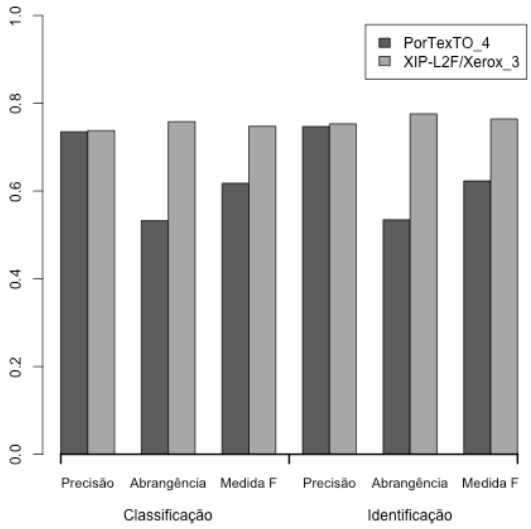


Figura 8.5: Resultados da corrida PorTexTO_4 e da corrida XIP-L2F/Xerox_3 para a pista do TEMPO na CD do TEMPO.

um bom desempenho, embora obviamente ainda precise de várias afinações. O sistema ainda se encontrava numa versão inicial e os padrões de expressões utilizados no reconhecimento das entidades mencionadas temporais foram muito limitados, mas, mesmo assim, o sistema conseguiu uma abrangência de aproximadamente 55%.

Os objectivos foram atingidos nesta participação, pois apesar do sistema não ter conseguido reconhecer todas as entidades mencionadas temporais, cerca de 75% das que o sistema reconheceu estão correctas.

No entanto, o sistema necessita de ultrapassar as limitações verificadas na versão 1.0. Um dos melhoramentos a fazer será criar padrões que representem expressões com um maior número de palavras, especificamente ultrapassar o limite de n=2 na determinação de co-ocorrências. O processamento de expressões também deveria ser alargado a expressões temporais complexas, isto é, criar padrões para expressões com mais de uma unidade temporal. Por exemplo, reconhecer a seguinte expressão, como uma única entidade mencionada: *no **dia** 10 do **mês** passado*.

A tarefa da criação de padrões de expressões poderá ser mais automatizada, para que com mais facilidade se possa acrescentar novos padrões.

Num trabalho futuro será interessante utilizar o sistema PorTexTO já melhorado, no reconhecimento de entidades mencionadas temporais em outras línguas. A adaptação do sistema ao processamento da língua, para além da língua portuguesa, não será uma tarefa difícil, porque os módulos funcionam de forma independente da língua. No entanto será sempre necessário ter algum conhecimento da língua em que se pretende aplicar o sistema. Resumidamente, o PorTexTO necessita de ter a parte importante do reconhecimento de entidades mencionadas temporais que é a lista de PTR na língua em questão, e as pa-

lavras-chave temporais dessa língua (Pustejovsky et al., 2005). Com esta informação, o módulo `Processador de co-ocorrências` cria os padrões de expressões que são necessários para que o módulo `Anotador` possa fazer o reconhecimento de entidades mencionadas temporais nos textos dessa língua.