

FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



**AVALIAÇÃO DE SISTEMAS DE
RECONHECIMENTO DE ENTIDADES
MENCIONADAS**

Nuno Francisco Pereira Freire Cardoso

MESTRADO EM INTELIGÊNCIA ARTIFICIAL E SISTEMAS
INTELIGENTES

2006

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

**AVALIAÇÃO DE SISTEMAS DE
RECONHECIMENTO DE ENTIDADES
MENCIONADAS**

Nuno Francisco Pereira Freire Cardoso

Dissertação submetida para obtenção do grau de MESTRE
em Inteligência Artificial e Sistemas Inteligentes

Orientador:

Eugénio da Costa Oliveira

Co-Orientador:

Mário Jorge Costa Gaspar da Silva

2006

Resumo

O HAREM, organizado pela Linguateca, representa a primeira avaliação conjunta de sistemas de Reconhecimento de Entidades Mencionadas (REM) na língua portuguesa. Os participantes nesta iniciativa desenvolveram uma nova metodologia para a avaliação de sistemas de REM.

O HAREM contou com a participação de 10 grupos de investigação oriundos de 6 países, cujos sistemas de REM produziram 38 saídas. Os resultados da avaliação foram divulgados e os relatórios de desempenho distribuídos aos participantes. Este trabalho descreve o HAREM, valida a metodologia desenvolvida, e apresenta estimativas da exactidão das métricas usadas para comparar a eficácia das saídas produzidas por cada sistema.

PALAVRAS-CHAVE: Reconhecimento de Entidades Mencionadas, Avaliação Conjunta, Metodologia, Validação.

Abstract

HAREM, organised by Linguateca, was the first joint Named Entity Recognition (NER) evaluation initiative for Portuguese. The participants developed a new methodology for evaluation of NER systems.

HAREM involved 10 research groups from 6 countries and their NER systems produced 38 outputs. The global results were published and the detailed performance reports delivered to the participants. This work describes HAREM, validates its methodology, and gives estimates of the accuracy of the metrics used to compare the systems' outputs.

KEYWORDS: Named Entity Recognition, Evaluation Contest, Methodology, Validation.

Agradecimentos

O meu primeiro agradecimento vai para Diana Santos, mentora e principal responsável pelo HAREM, e para Cristina Mota, que iniciou em 2003 um estudo que inspirou o HAREM. Agradeço ao Nuno Seco e Rui Vilela, co-organizadores do HAREM, à Susana Afonso e Anabela Barreiro pela ajuda na anotação das colecções douradas, e a Marília Antunes pela contribuição dada na análise estatística apresentada nesta tese. Agradeço também aos participantes do HAREM, pela permissão dada em reproduzir os resultados dos seus sistemas, e pela contribuição dada na descrição dos seus sistemas.

O meu último agradecimento aos elementos do Grupo XLDB pelo apoio e ajuda prestada durante o trabalho, em particular a Bruno Martins, Daniel Gomes, Leonardo Andrade, Marcirio Chaves, Francisco Couto e Sérgio Freitas.

Este trabalho foi suportado pela Linguatca FCT/ POSI (POSI/-PLP/43931/2001)

Porto, Outubro de 2006

Nuno Francisco Pereira Freire Cardoso

Aos meus familiares e amigos.

Conteúdo

Conteúdo	i
Lista de Figuras	v
Lista de Tabelas	vii
1 Introdução	1
1.1 Contexto	3
1.2 Motivação	5
1.3 Objectivos e contribuições	7
1.4 Metodologia	8
1.5 Organização da tese	10
2 Trabalho relacionado	11
2.1 Metodologia de avaliação	12
2.2 Iniciativas de avaliação	15
2.2.1 Avaliação em REM	15
2.2.2 Avaliação em processamento da fala	16
2.2.3 Avaliação em análise morfossintáctica	17
2.2.4 Avaliação em análise morfológica	17
2.3 Estudo de REM da Linguateca	18

2.4	Sumário	19
3	Introdução à análise estatística	21
3.1	Testes paramétricos	22
3.2	Testes não-paramétricos	23
3.2.1	O método <i>bootstrap</i>	24
3.2.2	Os testes de permutação	26
3.2.3	Exemplos do teste de aleatorização parcial	29
3.2.4	Factores de variação do valor p	31
3.3	Sumário	33
4	HAREM	35
4.1	Criação da metodologia HAREM	38
4.1.1	Directivas de etiquetagem	38
4.2	Colecção HAREM	39
4.2.1	Colecções douradas	41
4.2.2	Comparação entre colecções douradas	44
4.3	Desenvolvimento da plataforma HAREM	51
4.3.1	Pontuações usadas no HAREM	51
4.3.2	Métricas de avaliação	55
4.3.3	Arquitectura da plataforma	56
4.3.4	Cenários de avaliação	59
4.4	Organização dos eventos HAREM	60
4.5	Sumário	62
5	Análise de resultados	63
5.1	Realização da análise estatística	64
5.1.1	Adaptação do teste de permutação ao HAREM	65
5.1.2	Métricas de avaliação	70

5.2	Análise estatística à colecção dourada	71
5.2.1	Variação do número de observações	72
5.2.2	Variação do número de re-amostragens	75
5.3	Resultados dos eventos HAREM	80
5.3.1	Evolução dos sistemas entre eventos	81
5.3.2	Panorama em REM	82
5.3.3	Comparação entre eventos de avaliação	85
5.4	Sumário	87
6	Breve análise aos sistemas de REM	89
6.1	Visão geral dos sistemas de REM	90
6.2	Descrição dos sistemas participantes	91
6.3	Sumário	96
7	Conclusões e trabalho futuro	99
7.1	Conclusões	100
7.2	Futuro do HAREM	101
7.3	Trabalho futuro	101
A	Acrónimos e abreviaturas	103
B	Tabelas de valores p	105
	Bibliografia	111

Lista de Figuras

1.1	Um excerto de texto sem EM marcadas, com EM identificadas, e com EM classificadas na sua semântica.	2
1.2	Exemplos de EM difíceis de identificar e de classificar.	3
1.3	Esquema de avaliação de sistemas inteligentes.	4
1.4	Esquema de avaliação HAREM.	6
3.1	O significado do valor p , no teste de permutação.	28
4.1	Distribuição das categorias semânticas pelas colecções douradas.	45
4.2	Distribuição dos géneros textuais pelas colecções douradas, por contagem de termos.	45
4.3	Distribuição dos géneros textuais pelas colecções douradas, por contagem de EM.	45
4.4	Densidade de EM por géneros textuais.	47
4.5	Distribuição de categorias semânticas por géneros textuais. .	48
4.6	Distribuição de géneros textuais por categorias semânticas. .	48
4.7	Esquema de avaliação da plataforma HAREM.	58
4.8	Cenários de avaliação usados nos eventos HAREM.	59
5.1	Excerto de texto marcado com EM.	66

5.2	Lista de alinhamentos gerados pela plataforma HAREM, para o exemplo da Figura 5.1.	66
5.3	Permutações de termos para o exemplo da Figura 5.1.	67
5.4	Permutações de termos com classificações semânticas diferentes.	68
5.5	Permutações de blocos para o exemplo da Figura 5.1.	69
5.6	A influência do número de blocos na média e desvio padrão da diferença entre re-amostragens.	75
5.7	A influência do número de re-amostragens na média e desvio padrão da diferença entre re-amostragens.	75
5.8	Desempenho dos sistemas para a tarefa de identificação no evento de 2005.	76
5.9	Desempenho dos sistemas para a tarefa de classificação semântica no evento de 2005.	77
5.10	Desempenho dos sistemas para a tarefa de identificação no evento de 2006.	78
5.11	Desempenho dos sistemas para a tarefa de classificação semântica no evento de 2006.	79
5.12	Medida F para os melhores sistemas na tarefa de classificação semântica, discriminada por categorias.	83
5.13	Precisão e abrangência para os melhores sistemas na tarefa de classificação semântica, discriminada por categorias.	83
5.14	Medida F para os melhores sistemas na tarefa de classificação semântica, discriminada por gênero textual.	84
5.15	Precisão e abrangência para os melhores sistemas na tarefa de classificação semântica, discriminada por gênero textual.	84

Lista de Tabelas

4.1	Categorização usada nos eventos HAREM.	40
4.2	Distribuição do género textual e de variante de português na colecção HAREM.	42
4.3	Comparação entre a colecção HAREM e as colecções douradas.	43
4.4	Distribuição das variantes de português pelas colecções douradas.	46
4.5	Tamanho das EM por contagem de termos.	49
4.6	Análise ao teor de EM comuns entre eventos HAREM. . . .	50
4.7	Cenários escolhidos pelos participantes.	61
4.8	Resumo da participação nos eventos HAREM.	62
5.1	Exemplo de uma tabela de contingência usada em sistemas de previsão.	70
5.2	Resultados da tarefa de identificação para duas saídas do evento de 2006.	73
5.3	Médias e desvios-padrão para as re-amostragens da saída <i>A</i> e <i>B</i> , para vários sub-conjuntos de blocos de tamanho de- crescente.	74

5.4	Valores p , médias e desvios-padrão das diferenças entre re-amostragens da saída A e B , para vários sub-conjuntos de blocos de tamanho decrescente.	74
5.5	Comparação da medida F para os eventos de 2005 e de 2006, nas tarefas de identificação e de classificação semântica. . . .	81
5.6	Estatísticas das colecções douradas do HAREM, MUC e CoNLL.	86
5.7	Medida F dos melhores resultados observados no HAREM, MUC e CoNLL.	87
B.1	Valores p para a tarefa de identificação do evento de 2005. .	106
B.2	Valores p para a tarefa de classificação semântica do evento de 2005.	107
B.3	Valores p para a tarefa de identificação do evento de 2006. .	108
B.4	Valores p para a tarefa de classificação semântica do evento de 2006.	109

Capítulo 1

Introdução

Designa-se por Reconhecimento de Entidades Mencionadas (REM) a tarefa de marcação de Entidades Mencionadas (EM) existentes no texto e de interpretação do seu significado semântico. Constituem exemplos de EM referências a nomes próprios de pessoas, organizações, ou cidades no texto.

O REM pode ser conceptualizado em duas sub-tarefas distintas:

Identificação: selecciona os termos que compõem cada uma das EM.

Classificação: determina propriedades linguísticas das EM, como por exemplo o seu significado semântico ou a sua morfologia.

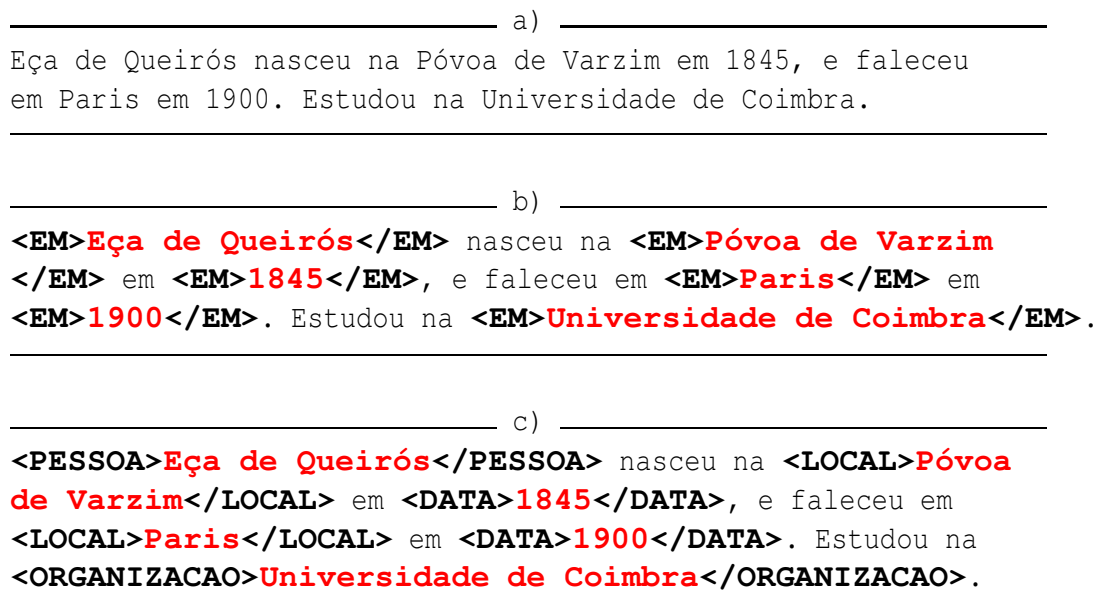


Figura 1.1: Um excerto de texto, **a)** sem EM marcadas, **b)** com EM identificadas, e **c)** com EM classificadas na sua semântica.

A Figura 1.1 ilustra os passos da tarefa de REM. Para o excerto de texto representado em **a)**, a identificação das EM é mostrada em **b)**, e a classificação semântica em **c)**. À primeira vista, a tarefa de REM parece fácil

de concretizar; no entanto, é frequente encontrarmos casos difíceis, onde os próprios humanos não conseguem identificar e classificar as EM sem dúvidas ou discordâncias entre si. As principais dificuldades de REM, como por exemplo o tratamento da vagueza e da ambiguidade, são intrínsecas ao processamento computacional da língua (Santos, 1997; Santos e Barreiro, 2004).

-
- | | |
|---|--|
| 1 | Portugal e Espanha são dois países. |
| 2 | São Tomé e Príncipe é um país. |
| 3 | São Tomé e Príncipe são duas ilhas. |
| 4 | Portugal votou 'sim' no referendo. |
-

Figura 1.2: Exemplos de EM difíceis de identificar e de classificar.

Considere-se o texto da Figura 1.2. Na linha 1 há duas EM referentes a dois países, *Portugal* e *Espanha*, e na linha 2 há uma EM, *São Tomé e Príncipe*, também referente a um país. Porém, na linha 3, *São Tomé* e *Príncipe* são duas EM distintas no contexto dado.

Em termos de semântica, *Portugal* na linha 1 refere-se a um território, enquanto que na linha 4 o contexto dado pelo verbo *votar* sugere que *Portugal* é uma menção a um grupo de pessoas, não a um território.

Estes exemplos mostram que a tarefa de REM pode ser difícil de realizar, sendo por vezes necessário interpretar a mensagem para reconhecer correctamente as EM.

1.1 Contexto

O Processamento de Linguagem Natural (PLN) representa um dos maiores e mais antigos desafios da Inteligência Artificial (IA). Turing

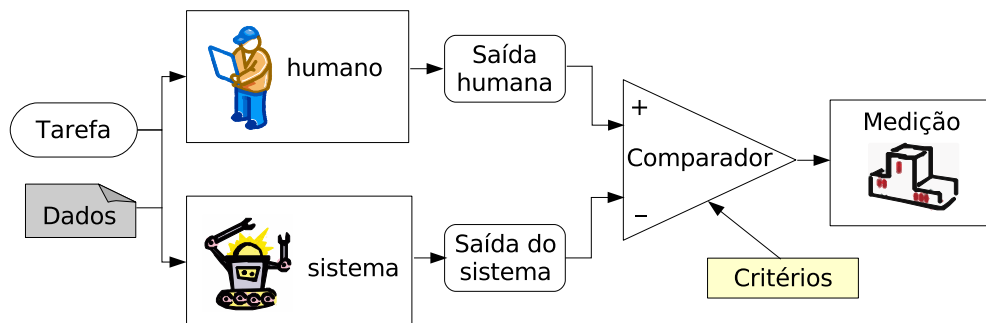


Figura 1.3: Esquema de avaliação de sistemas inteligentes.

(1950) propôs um teste pioneiro para avaliar o desempenho de sistemas inteligentes no domínio da compreensão e geração de linguagem natural, ao comparar o desempenho dos sistemas em relação ao desempenho humano. O teste de Turing, apesar de simples, estabeleceu as bases da avaliação de sistemas inteligentes: medir a diferença de eficácia entre um humano e um sistema na realização de uma tarefa, segundo o mesmo ambiente de avaliação e o mesmo conjunto de critérios (ver Figura 1.3).

O progresso observado em PLN deve muito a várias conferências dedicadas à avaliação de sistemas inteligentes que têm surgido para diversas tarefas envolvidas na compreensão da língua. Estas conferências desenvolvem as suas próprias metodologias e recursos de avaliação, e organizaram periodicamente eventos de avaliação para medir e comparar as saídas produzidas pelos sistemas participantes.

O REM é uma tarefa que pode beneficiar com a realização de eventos de avaliação específicos. Em 1995, os participantes no 6º evento de avaliação do MUC, uma conferência dedicada à avaliação em Extração de Informação (EI, ver Gaizauskas e Wilks, 1998), dividiram a tarefa de avaliação em sub-tarefas independentes susceptíveis de serem avaliadas independentemente. O REM foi uma das sub-tarefas selec-

cionadas (Hirschman, 1998). Após o MUC, outras conferências organizaram eventos de avaliação em REM com metodologias semelhantes.

1.2 Motivação

Em sistemas inteligentes que processam e interpretam a língua, como os sistemas de extracção de informação, resposta a perguntas (Voorhees, 2005), tradução automática (Hutchins e Somers, 1992) ou sumarização de textos (Mani, 1999), é necessária a existência de um componente de *software* que realize o reconhecimento de EM.

A organização de eventos de avaliação em REM representa um estímulo importante para a melhoria de sistemas inteligentes em PLN nas várias línguas. No entanto, não havia até 2003 nenhum plano de avaliação para o processamento da língua portuguesa.

A Linguateca, um centro de recursos (distribuído) para o processamento computacional da língua portuguesa, tem como um dos seus principais objectivos a organização de eventos de avaliação que envolvam a comunidade científica em PLN e que respondam aos interesses desta (Santos, 2000; 2002; Santos et al., 2004).

Em 2003 a Linguateca organizou o primeiro evento de avaliação para sistemas inteligentes em PLN, no caso concreto para sistemas de análise morfológica, denominado *Morfolimpíadas* (Santos et al., 2003). O evento adoptou um modelo de *avaliação conjunta* que envolveu os participantes na definição do evento de avaliação, e foi marcado pelo grande interesse dos participantes em torno de avaliações nas suas áreas de investigação.

O HAREM – Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas (www.linguateca.pt/HAREM/), iniciado em 2005, constitui a

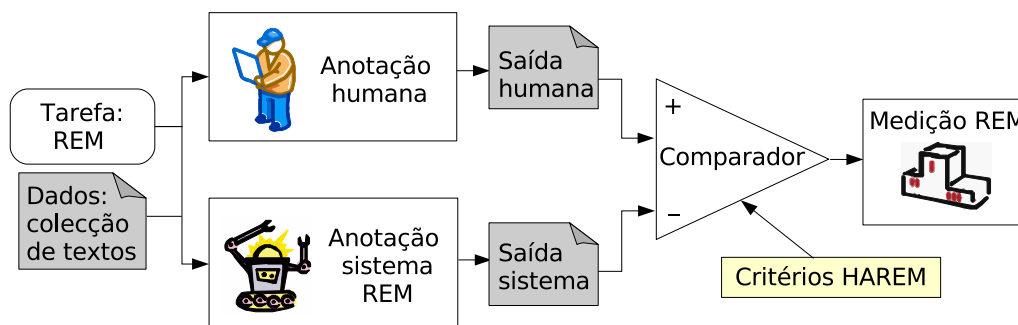


Figura 1.4: Esquema de avaliação HAREM.

primeira avaliação (conjunta) de sistemas de REM em português (Santos et al., 2006). A iniciativa foi motivada por se sentir que os eventos de avaliação de REM anteriores não tinham abordado a tarefa com profundidade suficiente, e com o objectivo de reunir a comunidade científica em torno de outro evento de avaliação dentro do processamento do português.

O HAREM concentrou esforços no desenvolvimento de uma nova metodologia de avaliação específica para REM, na criação de uma plataforma de avaliação de sistemas de REM, e na criação de uma colecção de textos anotada. Os participantes tiveram um papel activo na organização do HAREM, tendo colaborado na criação das directivas e na anotação das colecções. O HAREM segue o modelo de avaliação de sistemas inteligentes atrás descrito (ver Figura 1.4).

No entanto, até à realização do trabalho aqui apresentado, desconhecia-se até que ponto o HAREM avalia adequadamente os sistemas de REM, e qual o nível de confiança que se pode ter na sua avaliação comparativa. Houve para tal que responder às seguintes questões:

- Qual é o nível de confiança que se pode ter nos resultados dos eventos de avaliação do HAREM? Qual é a probabilidade de se cometer

o erro de concluir que duas saídas são diferentes, quando não o são de forma significativa, e vice-versa?

- O tamanho da colecção de textos usada é suficiente para que as diferenças de desempenho observadas entre saídas dos sistemas sejam significativas, com um certo grau de confiança? Até que ponto os desempenhos verificados podem ser extrapolados para outros ambientes?
- As métricas usadas para aferir as saídas dos sistemas são adequadas e conseguem discriminar dois sistemas diferentes?

1.3 Objectivos e contribuições

O trabalho desenvolvido nesta dissertação respondeu aos seguintes objectivos:

- Criar uma metodologia para a avaliação da tarefa de REM – a **Metodologia HAREM** – em conjunto com a comunidade científica interessada em REM.
- Desenvolver um ambiente de avaliação específico para REM – a **Plataforma HAREM** –, uma bancada de ensaios para medir o desempenho dos sistemas.
- Organizar eventos de avaliação conjunta em REM – os **Eventos HAREM** –, aplicando a metodologia HAREM.
- Validar o HAREM através de uma análise estatística dos resultados obtidos nos eventos HAREM.

O trabalho desenvolvido nesta dissertação produziu as seguintes contribuições:

- Uma metodologia nova para avaliação em REM, validada pela comunidade científica.
- Uma colecção de textos ricamente anotada em EM.
- Uma plataforma de avaliação específica para REM, disponível gratuitamente a todos os grupos de investigação.
- Uma caracterização do estado da arte na tarefa de REM em textos de língua portuguesa.

1.4 Metodologia

Para concretizar os objectivos apresentados, o trabalho foi dividido em quatro actividades separadas.

A. Criação da Metodologia HAREM

1. Elaboração de uma proposta inicial da metodologia, e disseminação pelos participantes do HAREM para apreciação.
2. Discussão conjunta das alterações sugeridas pela comunidade e pela equipa organizadora.
3. Revisão e alteração da proposta, de acordo com as conclusões obtidas no ponto anterior. Nova iteração no processo de melhoramento e disseminação da proposta, até a comunidade aprovar a metodologia.

B. Desenvolvimento da Plataforma HAREM

1. Desenho da arquitectura do *software* da plataforma de avaliação.
2. Especificação de cada módulo do *software*.
3. Desenvolvimento do *software* segundo a metodologia HAREM.
4. Teste do *software*.

C. Organização de Eventos HAREM

1. Calendarização dos eventos.
2. Chamada internacional à participação no HAREM.
3. Anotação manual das colecções de texto.
4. Distribuição da colecção de texto sem anotações pelos participantes, e recepção das saídas dos seus sistemas de REM.
5. Medição das saídas dos sistemas participantes.
6. Geração de resultados globais e de relatórios individuais de desempenho.

D. Validação estatística do HAREM

1. Selecção do teste estatístico.
2. Adaptação da análise estatística aos requisitos do HAREM.
3. Desenvolvimento do *software* de análise estatística.
4. Validação estatística das saídas dos eventos HAREM.

As actividades foram realizadas por uma equipa de colaboradores da Linguateca constituída por Nuno Seco, Rui Vilela e o autor da presente tese, e coordenada por Diana Santos. O autor teve uma participação activa na realização das tarefas A, B e C, e realizou a actividade D.

As actividades A, B e C foram executadas em paralelo, devido à grande interdependência e interacção que há entre elas. A actividade D teve início após a conclusão das actividades anteriores. A tarefa de documentação do HAREM foi realizada em paralelo com todas as actividades.

1.5 Organização da tese

Esta tese está estruturada em sete capítulos. No Capítulo 2 descrevem-se eventos de avaliação anteriores relacionados com o tema da tese. No Capítulo 3 faz-se uma breve introdução aos testes estatísticos usados na validação do HAREM. No Capítulo 4 detalham-se a metodologia, a plataforma e os eventos HAREM. No Capítulo 5 apresenta-se uma análise estatística aos eventos HAREM e às colecções de texto usadas. No Capítulo 6 analisam-se as estratégias adoptadas pelos sistemas participantes no HAREM. No Capítulo 7 faz-se uma síntese das conclusões da tese e referem-se trabalho futuro e directrizes para novos eventos de avaliação em REM.

Capítulo 2

Trabalho relacionado

A evolução da engenharia de sistemas deve muito à selecção criteriosa das abordagens mais adequadas, de entre as disponíveis. Essa selecção é feita por uma avaliação imparcial e quantitativa das saídas dos sistemas, segundo uma metodologia comum de comparação. Essa metodologia define um ambiente de avaliação controlado, onde os diversos factores que influenciam a medição são fixados, controlados ou mesmo eliminados, permitindo retirar conclusões objectivas sobre cada uma das abordagens propostas para desempenhar a tarefa de avaliação. Segundo **Gaizauskas et al. (1998)**, *"if objective measures can be agreed, winning techniques will come to the fore and better technology will emerge more efficiently"*.

Nos últimos tempos assistiu-se a um aumento do número de conferências de avaliação que incluem a tarefa de REM nos seus eventos de avaliação. Contudo, a tarefa de REM possui características e requisitos distintos das outras tarefas de PLN, necessitando de uma metodologia de avaliação específica.

Neste capítulo referem-se os primeiros passos na criação de metodologias de avaliação, e descrevem-se alguns eventos de avaliação anteriores. No final, faz-se um sumário do estudo de REM em português realizada pela Linguatca, que precedeu o HAREM.

2.1 Metodologia de avaliação

Cleverdon (1967) definiu a primeira metodologia de avaliação para sistemas de recuperação e processamento de texto. As suas experiências, conhecidas como *experiências de Cranfield*, introduziram vários conceitos que ainda hoje são aplicados nos maiores eventos de avaliação em Extracção

de Informação (EI) e em Recuperação de Informação (RI). A metodologia de Cleverdon postula os seguintes princípios:

Colecção comum de textos: a avaliação deve realizar-se segundo uma colecção de textos comum para todos os sistemas, permitindo uma avaliação comparativa.

Conjunto comum de tópicos: os tópicos representam as experiências realizadas na avaliação, e traduzem diferentes necessidades de informação que os sistemas tem de satisfazer. Os tópicos são comuns a todos os sistemas, tal como a colecção de textos.

Documentos relevantes identificados: o juízo humano da importância de um determinado documento para um determinado tópico é a base da comparação dos sistemas. Cleverdon refere que todos os documentos relevantes para cada tópico devem ser identificados como tal por revisores humanos, para que a avaliação seja completa.

Métricas de avaliação: a introdução de métricas de avaliação, como a precisão e a abrangência, permite quantificar o desempenho dos sistemas.

Houve várias conferências de avaliação em diversas áreas cujas metodologias de avaliação se inspiraram na metodologia de Cleverdon, como são exemplo o MUC (avaliação em EI, ver [Hirschman, 1998](#)), o Parseval (avaliação em análise morfossintáctica, ver [Black et al., 1991](#)) e as Morfolimpíadas (avaliação em análise morfológica, ver [Santos et al., 2003](#)).

Estas conferências de avaliação usam colecções de texto anotadas manualmente segundo as directivas da tarefa. Contudo, o processo de anotação é muito oneroso, pois é um processo manual e realizado por especialistas. Ao usar colecções de texto de tamanho reduzido com anotações

humanas, esta metodologia tem o inconveniente de introduzir um erro humano, uma vez que os vários anotadores podem divergir entre si, influenciando os resultados das medições.

A partir dos anos 90 surgiram conferências de avaliação específicas para sistemas que processam grandes quantidades de texto. O TREC (*Text REtrieval Conference*) teve início em 1992, precisamente com o objetivo de fomentar a investigação em RI sobre colecções de grandes dimensões (Harman, 1993). As colecções de texto usadas nas tarefas do TREC são de dimensão considerável; a título ilustrativo, a colecção usada na tarefa *Terabyte Track* do TREC, em 2004, contém 25 milhões de documentos (Clarke et al., 2004).

Em 1998 surgiu o NTCIR (*NII-NACSIS Test Collection for Information Retrieval systems*, ver Kando et al., 1999), e em 2001 o CLEF (*Cross Language Evaluation Forum*, ver Peters e Braschler, 2001). Ambos os eventos adoptaram uma metodologia semelhante à do TREC, estando o CLEF focado na avaliação em RI multilingue para línguas europeias, enquanto que o NTCIR se dedica à avaliação em RI para línguas asiáticas.

O tamanho das colecções usadas no TREC, CLEF e NTCIR tornam impraticável conhecer a totalidade dos documentos relevantes para cada tópico, um dos princípios da metodologia de Cleverdon. A solução passa pela selecção automática e criteriosa de sub-conjuntos de documentos que serão sujeitos a juízo humano da sua relevância para cada tópico, e supôr que os restantes documentos são irrelevantes para esse mesmo tópico. A esta técnica dá-se o nome de *pooling* (Voorhees, 2002). Apesar de ser uma aproximação ao princípio de Cleverdon, e de já ser sido mostrado que o *pooling* ignora documentos relevantes, influenciando os valores de desempenho absolutos, Zobel (1998) mostrou que os resultados da avaliação

comparativa com *pooling* permitem retirar conclusões com significado estatístico.

2.2 Iniciativas de avaliação

2.2.1 Avaliação em REM

O MUC (*Message Understanding Conference*) foi uma conferência de avaliação que teve como objectivo reunir os grupos de investigação da área em torno de tarefas comuns de avaliação em EI. O MUC teve início em 1987 e organizou sete eventos de avaliação, o último dos quais em 1998 (Sundheim e Chinchor, 1993). O 6º evento de avaliação do MUC, organizado em 1995, foi o primeiro evento de avaliação a incluir uma tarefa independente de avaliação em REM (Grisham e Sundheim, 1996). O 7º evento repetiu a mesma tarefa de REM, com modificações menores (Chinchor e Robinson, 1998).

A tarefa de REM do MUC consistiu em marcar as EM nas seguintes categorias e tipos:

ENAMEX: categoria nominal composta pelos tipos PERSON (pessoa), ORGANIZATION (organização) e LOCATION (local).

TIMEX: categoria temporal composta pelos tipos TIME (hora) e DATE (data).

NUMEX: categoria numérica composta pelos tipos MONEY (moeda) e PERCENT (percentagem).

O MET (*Multilingual Entity Task*) foi a primeira conferência multilingue de avaliação em REM, decorrendo em paralelo com o MUC entre 1996 e 1998 (Merchant et al., 1996). O MET adoptou a mesma metodologia de

avaliação do MUC nos dois eventos organizados. O primeiro evento do MET utilizou o inglês e o espanhol nas colecções de texto, enquanto que o segundo evento de 1997 usou o chinês, o japonês e o inglês.

O CoNLL (*Conference on Computational Natural Language Learning*) foi uma conferência de avaliação que teve como objectivo promover a avaliação em diversas áreas específicas de PLN. O primeiro evento remonta a 1999, e os eventos de 2002 (Sang, 2002) e de 2003 (Sang e de Meulder, 2003) focaram na tarefa de REM, para encorajar a investigação em sistemas de REM independentes da língua. No evento de 2002 usou-se o espanhol e o flamengo nas colecções de texto, e no evento de 2003 o alemão e o inglês. A metodologia de avaliação não difere muito em relação ao MUC, apresentando quatro categorias de classificação semântica: LOC (local), ORG (organização), PER (pessoa) e MISC (diversos).

O ACE (*Automatic Content Extraction*) é uma conferência que teve início em 1999 com um estudo piloto para determinar quais as tarefas de extracção de conteúdos com mais interesse em avaliar (Doddington et al., 2004). Desde 2000 que o ACE organiza eventos de avaliação que incluem uma tarefa denominada EDT - *Entity Detection and Tracking*, que propõe não só a identificação e classificação das EM, mas também as respectivas referências anafóricas. O ACE incluiu o inglês, chinês e o árabe nas suas colecções de texto, som e imagem, e a sua categorização estende-se ao domínio militar, incluindo categorias semânticas como entidades geo-políticas, armas, veículos ou instalações (*facilities*).

2.2.2 Avaliação em processamento da fala

O ATIS (*Air Travel Information System*) teve como objectivo aplicar a experiência do MUC para a avaliação de sistemas de processamento da

fala (Price, 1990; Hirschman, 1992). O nome da conferência deriva do domínio da colecção usada, que contém mensagens sobre o planeamento de viagens aéreas. Apesar de o ATIS não focar na tarefa de REM, observou-se que os sistemas participantes melhoraram o seu desempenho ao longo dos seus eventos, organizados entre 1990 e 1993 (Hirschman, 1998). Este facto demonstra a importância que os eventos de avaliação possuem na melhoria de sistemas inteligentes numa área específica.

2.2.3 Avaliação em análise morfossintáctica

O Parseval (*Parse Evaluation*, ver Black et al., 1991) é uma conferência de avaliação em análise morfossintáctica de textos que surgiu motivada pela criação do *Penn Treebank*, um corpus de frases anotadas na sua sintaxe (Marcus et al., 1994). O Parseval, apesar de não estar directamente relacionado com REM, avalia os sistemas na sua eficácia de análise gramatical, o que é importante para identificar e atribuir o significado semântico de muitas EM.

2.2.4 Avaliação em análise morfológica

O primeiro evento de avaliação organizado pela Linguatca focou-se na avaliação de sistemas de análise morfológica, e denominou-se Morfolimpíadas (Santos et al., 2003). Optou-se pela análise morfológica por ser uma tarefa frequente em PLN, e desempenhada por vários sistemas existentes na altura. Tem também a vantagem de ser uma tarefa bem definida e objectiva, facilitando a criação das directivas.

As Morfolimpíadas contaram com sete participantes, cujos sistemas processaram uma colecção de textos com cerca de 80.000 termos. As Mor-

folimpíadas adaptaram uma metodologia de *avaliação conjunta*, contando com a colaboração dos participantes na definição da tarefa de avaliação. O evento revelou o interesse que a comunidade científica tem na organização de eventos de avaliação em português, e a experiência obtida com a sua organização foi importante para a organização do HAREM.

2.3 Estudo de REM da Linguateca

Apesar de terem sido organizados vários eventos de avaliação focados na tarefa de REM no passado, Santos e Cardoso (2006a) mostram que há diversos aspectos da tarefa que ainda não foram abordados com profundidade suficiente e com a devida atenção. Adicionalmente, o português nunca foi usado em eventos de avaliação em REM, que é uma tarefa dependente da língua, e o seu uso em eventos de avaliação multilingue não é trivial (Santos e Cardoso, 2006b).

A Linguateca promoveu em 2003 um estudo para analisar o problema de REM, recolhendo dados importantes para a organização de uma futura avaliação conjunta em REM para a língua portuguesa. O estudo foi realizado por Mota et al. (2006) e reuniu nove investigadores, que anotaram livremente 20 extractos das colecções CETEMPúblico e CETENFolha (Santos e Rocha, 2001). No final, os investigadores identificaram entre 179 a 250 EM, sugerindo um conjunto variado de categorias semânticas. No caso de EM relativas a nomes próprios de pessoas, verificou-se que no universo de EM marcadas pelos investigadores como tal, 46% destas foram marcadas como tal por todos.

Este estudo permitiu retirar conclusões preciosas para a organização do HAREM:

1. Os investigadores usaram um leque de categorias e de tipos semânticos mais vasto do que as categorizações usadas pelo MUC ou pelo CoNLL, por exemplo.
2. A anotação manual das colecções de textos difere consideravelmente entre anotadores, uma vez que a própria tarefa de REM pode ser realizada de maneiras diferentes.

2.4 Sumário

Após a publicação das experiências de Cleverdon, surgiram diversas conferências de avaliação que desempenharam um papel decisivo no desenvolvimento de sistemas inteligentes, cada vez mais eficazes nas respectivas tarefas. Seguindo a mesma linha de raciocínio, a avaliação de tarefas específicas de PLN facilita o progresso dos sistemas em cada problema particular. A EI é o exemplo de uma área que beneficiou significativamente com a organização de sucessivos eventos de avaliação específicos.

A tarefa de REM já foi objecto de vários eventos de avaliação no passado. Contudo, [Mota et al. \(2006\)](#) mostraram que as metodologias adoptadas para a avaliação em REM não consideraram as especificidades da tarefa, que não se resume apenas a marcar pessoas, organizações, locais e números. Uma análise crítica aos eventos de avaliação anteriores é apresentada no Capítulo 5. Uma vez que o objectivo da avaliação é a medição comparativa da eficácia dos sistemas existentes, a sua metodologia deve considerar a tarefa tal como a comunidade interpreta e implementa nos seus sistemas de REM. Adicionalmente, a metodologia de avaliação deve ter em conta as especificidades e limitações da tarefa, como é exemplo a vagueza das EM e as divergências entre anotadores.

Capítulo 3

Introdução à análise estatística

Os métodos estatísticos permitem caracterizar um conjunto de dados, analisar a sua distribuição e inferir conclusões acerca destes. Os testes de significância estatística, por sua vez, aplicam os métodos estatísticos para quantificar a probabilidade de rejeitar ou não a hipótese de uma amostragem de dados ser representativa de uma população.

Os resultados obtidos na avaliação são a base para seleccionar quais das estratégias adoptadas pelos sistemas têm melhor desempenho na tarefa de avaliação proposta. A análise estatística das saídas produzidas por diversos sistemas reveste-se de grande importância, ao quantificar o nível de confiança que se pode ter nas conclusões sobre o seu desempenho.

Este Capítulo faz uma breve introdução teórica aos métodos estatísticos usados na validação do HAREM, aprofundados no Capítulo 5. Recomenda-se a consulta dos livros de [Good \(2000\)](#), [Sheskin \(2000\)](#) ou [Moore et al. \(2003\)](#), para obter mais detalhe sobre os métodos estatísticos apresentados neste Capítulo.

3.1 Testes paramétricos

Quando é sabido que uma população segue uma distribuição normal (ou Gaussiana), caracterizada por um valor médio μ e um desvio padrão σ , o teste de significância compara os valores da média \bar{x} e do desvio padrão s da amostra contra os parâmetros μ e σ da população. Uma vez que os valores da amostra e da população seguem uma distribuição probabilística conhecida e parametrizável, estes testes são denominados *testes paramétricos*.

Os testes paramétricos requerem que os dados respeitem os seguintes pressupostos:

- A população de onde é extraída a amostra segue uma distribuição probabilística conhecida e parametrizável.
- A amostragem deve ser aleatória e as observações devem ser independentes, ou seja, não deve haver relação directa entre observações.
- Os dados devem ser compatíveis com os pressupostos distribucionais.

Os testes paramétricos são bastantes robustos e mantêm um grau de confiança considerável nos resultados, mesmo quando os pressupostos da distribuição postulada não são cumpridos na sua totalidade. Em casos onde os desvios aos pressupostos são significativos (numa distribuição enviesada, por exemplo), os testes paramétricos deixam de poder ser aplicados.

3.2 Testes não-paramétricos

Quando o fenómeno que se procura medir não segue uma distribuição probabilística conhecida, não se deve aplicar um teste paramétrico. No caso da tarefa de REM, onde as amostras são dadas pelas EM marcadas nas saídas dos sistemas, a distribuição não consegue ser aproximada por uma função conhecida, devido a vários factores:

- A distribuição das EM no texto (e das respectivas categorias semânticas) depende de inúmeros factores linguísticos como o estilo de escrita, o assunto ou o género textual, e o comportamento dos sistemas perante os textos é muito difícil de parametrizar.
- As observações são dependentes entre si, porque quando um sistema identifica e/ou classifica uma dada EM, é provável que a presença

dessa EM influencie a decisão de identificar e/ou classificar outras EM envolventes.

- As observações apresentam diferentes níveis de dificuldade. Existem EM que são relativamente fáceis de reconhecer, enquanto que outras EM são mais difíceis de marcar.
- No teste de hipóteses, as amostras podem conter observações que não podem ser emparelhadas, uma vez que uma observação de uma amostra pode estar relacionada com várias observações da outra amostra.

Os *testes não-paramétricos* são mais adequados nestes casos, pois não necessitam de conhecer à partida a distribuição da população, procurando recriar a distribuição a partir dos dados disponíveis. Estes testes não necessitam de grandes quantidades de dados, e não postulam pressupostos rígidos.

No resto desta Secção analisam-se dois testes não-paramétricos usados na validação estatística de eventos de avaliação, o método *bootstrap* e os testes de permutações.

3.2.1 O método *bootstrap*

O método *bootstrap* é um método não-paramétrico que gera um conjunto n_r de re-amostragens a partir dos dados disponíveis, para recriar a distribuição da população. O método inspira-se no facto de que, se a distribuição normal representa o espectro de valores obtidos a partir de várias amostragens, analogamente, a distribuição gerada pelo *bootstrap* representa o espectro de valores obtidos a partir de n_r re-amostragens (Efron, 1981; Moore et al., 2003).

O método *bootstrap* pressupõe que as observações são independentes, e que as várias re-amostragens geradas são representativas da população. O método é composto pelos seguintes passos:

1. Dada uma amostra inicial de tamanho n_O , são geradas n_r re-amostragens com o mesmo tamanho.
2. Cada re-amostragem *bootstrap* é gerada através da selecção aleatória de n_O observações a partir do conjunto de dados iniciais. Cada observação tem igual probabilidade $\left(\frac{1}{n_O}\right)$ de ser seleccionada (Noreen, 1989). Após a selecção, a observação é reposta no conjunto de observações candidatas; se o valor não fosse reposto, a re-amostragem resumia-se a uma mera re-ordenação dos dados.
3. Para cada uma das n_r re-amostragens, calculam-se os seus parâmetros (como a média, mediana ou o desvio padrão, por exemplo).
4. A distribuição *bootstrap* final, criada a partir dos n_r parâmetros calculados, apresenta uma curvatura gaussiana, com a sua média \bar{x} a apresentar um valor próximo da média μ dos dados originais.

O teste de hipóteses *bootstrap* compara duas amostras distintas a partir das respectivas distribuições *bootstrap*. Se as distribuições *bootstrap* apresentarem parâmetros semelhantes (segundo um determinado intervalo de confiança), a hipótese de que as duas amostras são semelhantes não é rejeitada.

O método *bootstrap* pode suscitar algum cepticismo quanto à sua metodologia, devido à utilização repetida de algumas observações na geração de uma re-amostragem. No entanto, a independência dos dados entre si faz com que as re-amostragens sejam representações válidas dos

dados originais, e existe fundamento estatístico para estimar a distribuição de várias re-amostragens obtidas a partir do mesmo conjunto de dados.

O *bootstrap* é normalmente usado para verificar se certas amostragens seguem uma distribuição normal, ou se são afectadas por algum enviesamento. Savoy (1997) recomenda o método *bootstrap* para determinar a significância em avaliações de sistemas de recuperação de informação. O *bootstrap* também foi usado pelo CoNLL para calcular o intervalo de confiança das saídas, para a tarefa de REM (Sang, 2002; Sang e de Meulder, 2003).

3.2.2 Os testes de permutação

Os testes de permutação são testes não-paramétricos que se baseiam na intuição de que, se a diferença observada entre duas amostras para a medida M (d_M) é significativa, então a permuta aleatória de dados entre as amostras irá alterar consideravelmente os valores de d_M . No caso oposto de a diferença ser ocasional, a permuta de dados não terá um impacto significativo nos valores de d_M .

O teste de permutação pode ser formulado pela seguinte hipótese:

A diferença absoluta entre os valores da métrica M (onde M pode ser a precisão, abrangência, medida F ou outra métrica de desempenho) para as saídas A e B na tarefa T , é aproximadamente igual a zero.

O teste de permutação é composto pelos seguintes passos:

1. Calcular a diferença absoluta d entre os valores da métrica M , para as saídas A e B .

$$d = |M_A - M_B| \quad (3.1)$$

2. Gerar n_r re-amostragens. Para cada re-amostragem:

(a) Percorrer o conjunto de todas as observações de A,

$$O_A = \{O_A^1, O_A^2, \dots, O_A^n\}, \text{ e de B, } O_B = \{O_B^1, O_B^2, \dots, O_B^n\}.$$

(b) Permutar cada par de observações $\{O_A^i, O_B^i\}$, com uma probabilidade θ igual a 0.5.

(c) Calcular a diferença d^* entre os valores da métrica M para as re-amostragens A^* e B^* .

$$d^* = |M_A^* - M_B^*| \quad (3.2)$$

3. Contar o número de vezes (n_m) que o valor de d^* foi igual ou superior a d .

$$n_m = \sum_{i=1}^{n_r} w_i, \quad w_i = \begin{cases} 1 & \text{se } (d^* - d) \geq 0 \\ 0 & \text{se } (d^* - d) < 0 \end{cases} \quad (3.3)$$

4. Calcular o valor p :

$$p = \frac{(n_m + 1)}{(n_r + 1)} \quad (3.4)$$

A hipótese nula formula que as duas amostras são semelhantes, se a diferença d não é significativa. Neste caso, é provável que um certo número n_m de re-amostragens apresente valores de d^* iguais ou superiores a d . Por outro lado, se as duas amostras são diferentes, isso reflecte-se num valor inicial de d elevado. As n_r re-amostragens geradas apresentam uma tendência para obter valores de d menores do que o valor inicial de d , sendo menos frequente observar re-amostragens onde $d^* \geq d$.

O valor p (p -value) representa o rácio entre n_m , o número de re-amostragens onde se observa que $d^* \geq d$, sobre n_r , o número de re-amostragens geradas. Para valores p inferiores a um determinado limite α , pode-se

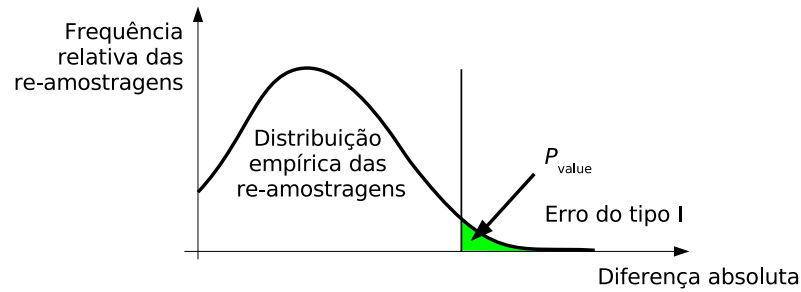


Figura 3.1: Ilustração do significado do valor p , no teste de permutação.

rejeitar a hipótese nula. O α representa a probabilidade de se rejeitar a hipótese nula quando devia ter sido aceite, o denominado Erro do tipo I (ver Figura 3.1).

Quando o número de re-amostragens cobre o universo de todas as permutações possíveis entre amostras, o teste é denominado Aleatorização Completa (*Exact Randomization*). No entanto, para amostras com muitas observações, torna-se impraticável gerar todas as permutações possíveis entre amostras, mesmo para a capacidade computacional actual. O teste de Aleatorização Parcial (*Approximate Randomization*) é uma aproximação ao teste de aleatorização completa, limitado a um determinado número n_r de re-amostragens, e a sua distribuição revela-se uma boa aproximação à distribuição real para números elevados de re-amostragens, podendo ser desprezados os erros derivados da aproximação.

O teste de aleatorização parcial foi aplicado pelo MUC na análise de significância dos seus resultados (Chinchor, 1992; Chinchor et al., 1993). A simplicidade e versatilidade do teste permite adaptá-lo facilmente para a avaliação de outras tarefas de PLN, como a tradução automática ou a análise morfossintáctica (Riezler e Maxwell III, 2005; Morgan, 2006).

3.2.3 Exemplos do teste de aleatorização parcial

Nesta Secção ilustra-se o teste de aleatorização parcial em dois casos:

Duas saídas diferentes

A saída A obteve um valor de medida F de 0,80, enquanto a saída B obteve um valor de medida F de 0,20. A diferença d é igual a 0,60. O número de observações n_O de cada saída é de 1.000. O nível de confiança do teste é de 99%, ou seja, $\alpha = 0,01$.

De um modo geral, pode-se afirmar que a saída A consegue acertar em média 4 em 5 observações, enquanto que a saída B acerta apenas 1 em 5 observações. No caso da saída A , a probabilidade de uma permutação prejudicar o valor final da sua medida F é igual à probabilidade da troca de um alinhamento correcto de A por um alinhamento incorrecto de B , ou seja:

$$\begin{aligned} p_{prejudicado}^A &= p_{correcto}^A \times (1 - p_{correcto}^B) \\ &= 0,8 \times 0,8 \\ &= 0,64. \end{aligned}$$

A probabilidade da permuta não afectar o valor da medida F de A é calculada por:

$$\begin{aligned} p_{indiferente}^A &= p_{correcto}^A \times (p_{correcto}^B) + (1 - p_{correcto}^A) \times (1 - p_{correcto}^B) \\ &= 0,8 \times 0,2 + 0,8 \times 0,2 \\ &= 0,32. \end{aligned}$$

A probabilidade da saída A beneficiar com a permuta é calculada por:

$$\begin{aligned} p_{beneficiado}^A &= (1 - p_{correcto}^A) \times (p_{correcto}^B) \\ &= 0,2 \times 0,2 \\ &= 0,04. \end{aligned}$$

Pode-se concluir que a saída A tem maior probabilidade de sair prejudicada com cada permutação. Ao mesmo tempo, a saída B tem maior probabilidade de sair beneficiada com as permutações. No final, há uma tendência considerável para que d^* seja menor do que 0,60. Para uma média de 500 permutações por re-amostragem, o número esperado de observações correctas para A é determinado por:

$$500p_{correcto}^A + 500p_{beneficiado}^A + 500\frac{p_{indiferente}^A}{2} = 500$$

ou seja, 500 observações correctas contra as cerca de 800 iniciais.

Ao ser muito improvável observar que $d^* \geq d$, n_m terá um valor próximo de 0 e o valor p será inferior a α , o que permite rejeitar a hipótese de que as duas saídas são semelhantes. Este valor foi confirmado num teste *ad-hoc*, que obteve uma média de observações correctas O_A^* próximo de 500, um desvio padrão baixo e um valor p inferior a 0,01.

Duas saídas semelhantes

A saída C obteve um valor de medida F de 0,79. A diferença d em relação à saída A do caso anterior é de 0,01. O número de observações de C é de 1.000.

Um cálculo de probabilidades semelhantes ao do caso anterior mostra que $p_{prejudicado}^A = 0,168$, $p_{indiferente}^A = 0,674$ e $p_{beneficiado}^A = 0,158$, para a

saída A . Neste caso não é possível afirmar que a saída A é significativamente melhor do que C , uma vez que a diferença de probabilidades entre A ficar prejudicado e beneficiado com a permuta é de 0,10, sendo provável que ocorram re-amostragens onde se verifica que $d^* \geq d$. Isso traduz-se em valores de n_m afastados de 0 e em valores p iguais ou superiores a α , o que implica que a hipótese nula não pode ser rejeitada. Um conjunto de testes *ad-hoc* para este caso revelou valores p entre 0,20 e 0,25, provando que a diferença observada entre as saídas A e B não é suficiente para afirmar que são duas saídas distintas.

3.2.4 Factores de variação do valor p

O valor p depende do valor da diferença inicial d , do número de re-amostragens n_r , do número de observações n_O , e da probabilidade de permutação entre observações, θ .

Número de re-amostragens

O número n_r de re-amostragens não afecta o valor absoluto de p , mas afecta a sua resolução. Para valores de n_r elevados, a distribuição gerada aproxima-se mais da distribuição real, o que permite ter maior confiança nos valores p calculados.

Para $n_r = 9.999$, a resolução de p é de 0,0001, o que implica que são precisas 100 ou mais re-amostragens que verifiquem a condição $d^* \geq d$ para que $p \geq \alpha$ (para 99% de confiança). No caso de $n_r = 99$, a resolução de p desce para 0,01, bastando somente 1 re-amostragem que verifique a condição $d^* \geq d$ para que $p \geq \alpha$. Como tal, um número reduzido de re-amostragens n_r torna o teste vulnerável à geração de re-amostragens

excepcionais, e condiciona a confiança que se pode ter no resultado do teste.

Número de observações

Como acontece em todos os métodos estatísticos, o número de observações n_O tem influência directa na margem de erro do teste.

(Buckley e Voorhees, 2000; Voorhees e Buckley, 2002) estudaram a relação que há entre as diferenças observadas entre saídas, o número de observações efectuadas, e o erro associado à conclusão final, para o TREC. Concluiu-se que existe uma relação directa e que esta pode ser determinada empiricamente. Posteriormente, Lin e Hauptmann (2005) conseguiram demonstrar matematicamente o que Voorhees e Buckley tinham calculado empiricamente, e determinaram que a margem de erro pode ser calculada por:

$$\text{Margem de erro} \approx \frac{1}{2} \exp \left(-\frac{2 (\mu_A - \mu_B)^2}{\pi (\sigma_A^2 + \sigma_B^2)} n_O \right) \quad (3.5)$$

onde μ e σ representam a média e desvio padrão das saídas, e n_O representa o número de observações.

A Equação 3.5 mostra que há uma relação exponencial entre o erro da avaliação, a diferença entre valores de métricas e o número de observações efectuadas, onde um aumento do número de observações resulta na diminuição do erro do teste estatístico.

Probabilidade de permutação entre observações

Na literatura, o valor da probabilidade de permutação entre observações θ para os testes de permutações é igual a 0,5. Para valores de θ próximos

de 0, há uma menor tendência para haver permutas entre observações, o que reduz a eficácia do teste.

Por outro lado, com valores de θ próximos de 1, ocorre a permutação quase total das observações entre A e B , produzindo um efeito semelhante ao verificado quando θ tende para 0.

O valor θ de 0,5 é o valor que garante a geração de re-amostragens mais representativas da distribuição real.

3.3 Sumário

Este Capítulo fez uma breve introdução aos testes e métodos estatísticos usados na validação estatística do HAREM, apresentada no Capítulo 5.

Os testes não-paramétricos revelam-se a escolha possível para analisar a significância das saídas do HAREM, dadas as características da tarefa de avaliação de sistemas de REM. Os testes não-paramétricos são mais robustos e precisos do que os testes paramétricos, dado que são menos sensíveis aos desvios dos pressupostos. Estes testes podem ser facilmente adaptados para a validação estatística do HAREM, uma vez que são simples de concretizar e a capacidade de processamento dos computadores actuais permite a sua aplicação (Noreen, 1989; Cohen, 1995).

Capítulo 4

HAREM

O HAREM começou a ser planeado em Junho de 2003 por ocasião do *workshop* AVALON sobre avaliação conjunta organizado pela Linguateca (www.linguateca.pt/avalon2003/), e que decorreu no final da 6ª edição da conferência PROPOR.

A experiência adquirida com a organização das Morfolimpíadas, e o estudo realizado por Mota et al. (2006) foram importantes para o desenvolvimento do HAREM, que culminou com a organização de dois eventos de avaliação, realizados em Fevereiro de 2005 e em Abril de 2006.

O HAREM destaca-se dos eventos de avaliação de sistemas de REM anteriores nos seguintes pontos:

Colecções com diversos tipos de texto: Existem diferenças significativas no teor e na distribuição de EM entre géneros textuais. Uma vez que os sistemas de REM participantes podem ter sido desenvolvidos para processar diferentes tipos de texto, as colecções usadas contêm textos de vários géneros textuais e de várias variantes de português.

A informação sobre o género textual e a variante é incluída nos metadados de cada documento da colecção, permitindo aos sistemas alterar o seu comportamento mediante o tipo de texto, como fizeram Maynard et al. (2001) no seu sistema de REM.

Avaliação independente das tarefas de identificação e de classificação: As tarefas de identificação e de classificação são avaliadas em separado, para diagnosticar detalhadamente o desempenho dos sistemas.

Avaliação selectiva: A avaliação adapta-se às características de cada sistema, medindo o desempenho das saídas segundo um sub-conjunto de categorias e tipos de EM pré-seleccionados pelo sistema participante.

Tarefa de classificação morfológica: A classificação morfológica também é avaliada em separado, uma vez que, nas línguas românicas, a morfologia das EM tem por vezes um papel essencial na desambiguação do seu significado semântico.

Categorização a partir do texto: O leque de categorias usadas para a tarefa de classificação semântica – a Categorização HAREM – foi delineado em conjunto com os participantes do HAREM a partir da análise dos textos da colecção. A categorização HAREM é significativamente diferente das categorizações usadas em eventos de avaliação em REM anteriores.

Anotação em contexto: A anotação manual das colecções tem em consideração o contexto onde se insere a EM, e a classificação semântica é feita de uma maneira mais detalhada (Santos e Cardoso, 2006a).

Suporte à vagueza: A vagueza é uma propriedade intrínseca da linguagem, e no caso de REM existem casos de EM vagas que não permitem desambiguar o seu significado semântico (Santos, 1997). Um exemplo ilustrativo é a frase “*Ajudem os Bombeiros!*”, onde não é possível afirmar com certeza se *Bombeiros* é uma menção a um grupo de pessoas, ou a uma instituição. O tema da vagueza não é mencionado em nenhum evento de avaliação em REM passado. Contudo, os sistemas de REM lidam com EM vagas no texto, e o HAREM suporta a vagueza das EM na anotação manual das suas colecções de texto, ao possibilitar a atribuição de várias categorias semânticas para cada EM.

Métricas específicas para REM: As tarefas de identificação, de classificação morfológica e de classificação semântica apresentam novas

métricas, a seguir descritas, que complementam as métricas de precisão, abrangência e medida F, usadas na aferição dos sistemas.

Este Capítulo descreve as três primeiras das quatro actividades propostas na metodologia da tese: a criação da metodologia, o desenvolvimento da plataforma e a organização dos eventos.

4.1 Criação da metodologia HAREM

A metodologia HAREM foi desenvolvida segundo os seguintes objectivos:

- Abranger as especificidades da tarefa de REM ainda não abordadas com profundidade em eventos de avaliação de sistemas de REM anteriores.
- Representar a tarefa tal como a comunidade científica a interpreta e aplica nos seus sistemas.

A metodologia inclui a definição das directivas de etiquetagem dos textos, a especificação das tarefas de avaliação e o processo de criação das colecções de texto. A Linguatca atribuiu aos participantes do HAREM um papel activo no seu desenvolvimento, para produzir uma metodologia aceite por todos e que responde aos interesses da comunidade.

4.1.1 Directivas de etiquetagem

As directivas de etiquetagem do HAREM representam o conjunto de regras das tarefas de avaliação realizadas num determinado evento HAREM. Estas directivas são seguidas pelos participantes no desenvolvimento dos sistemas, e são usadas na anotação manual da colecção de textos.

As directivas de etiquetagem definem os seguintes aspectos (Cardoso et al., 2006c; Cardoso e Santos, 2006):

- A sintaxe das etiquetas delimitadoras de EM.
- Os critérios de definição de uma EM.
- As regras de aplicação de etiquetas no texto.
- A categorização HAREM, com a descrição dos âmbitos semânticos de cada categoria.

A categorização HAREM é composta por uma hierarquia de dois níveis, denominadas *categorias* e *tipos*, respectivamente. As *categorias* representam as classes semânticas principais das EM e são compostas por vários *tipos*, sub-classes especializadas de cada categoria. Cada tipo pertence a uma única categoria apenas, e cada EM é classificada por uma categoria e por um tipo, no mínimo.

A Tabela 4.1 apresenta as 10 categorias e os 41 tipos usados nos eventos HAREM. As categorias e os tipos não possuem cedilhas nem acentos, para evitar problemas no processamento das saídas relacionados com a codificação de caracteres. As directivas foram revistas e actualizadas para o evento de 2006. Dentro das alterações mais importantes, o tipo `PRODUTO` da categoria `OBRA` foi substituído pelo tipo `MEMBROCLASSE` da categoria `COISA`, e foram revistos alguns âmbitos semânticos.

4.2 Colecção HAREM

O HAREM reuniu 1.202 excertos de textos provenientes de vários géneros textuais e de variantes de português numa única colecção, denominada *Colecção HAREM* (CH). Os textos foram recolhidos das seguintes fontes:

Categoria	Tipos (2005)		Tipos (2006)	
PESSOA (6)	INDIVIDUAL	GRUPOIND	INDIVIDUAL	GRUPOIND
	CARGO	GRUPOCARGO	CARGO	GRUPOCARGO
	MEMBRO	GRUPOMEMBRO	MEMBRO	GRUPOMEMBRO
ORGANIZACAO (4)	ADMINISTRACAO	SUB	ADMINISTRACAO	SUB
	EMPRESA	INSTITUICAO	EMPRESA	INSTITUICAO
LOCAL (5)	ADMINISTRATIVO	GEOGRAFICO	ADMINISTRATIVO	GEOGRAFICO
	CORREIO	VIRTUAL	CORREIO	VIRTUAL
	ALARGADO		ALARGADO	
OBRA (3-4)	ARTE	REPRODUZIDA	ARTE	REPRODUZIDA
	PUBLICACAO	PRODUTO	PUBLICACAO	
ABSTRACCAO (8)	DISCIPLINA	NOME	DISCIPLINA	NOME
	ESTADO	ESCOLA	ESTADO	ESCOLA
	OBRA	IDEIA	OBRA	IDEIA
	MARCA	PLANO	MARCA	PLANO
TEMPO (4)	DATA	HORA	DATA	HORA
	PERIODO	CICLICO	PERIODO	CICLICO
ACONTECIMENTO (3)	EFEMERIDE	EVENTO	EFEMERIDE	EVENTO
	ORGANIZADO		ORGANIZADO	
COISA (3-4)	OBJECTO	SUBSTANCIA	OBJECTO	SUBSTANCIA
	CLASSE		CLASSE	MEMBROCLASSE
VALOR (3)	MOEDA	CLASSIFICACAO	MOEDA	CLASSIFICACAO
	QUANTIDADE		QUANTIDADE	
VARIADO (1)	OUTRO		OUTRO	

Tabela 4.1: Categorização HAREM usada nos eventos de 2005 e de 2006. O número de tipos de cada categoria encontra-se entre parênteses, e as principais alterações entre eventos estão assinaladas a negrito.

Web: textos extraídos de páginas HTML da recolha da *web* portuguesa WPT 03 (Cardoso et al., 2006a) e da recolha da *web* brasileira WBR-99 (Calado, 1999).

Jornalístico: textos retirados dos corpora jornalísticos CETEMPúblico, CETENFolha, Avante!, Viseu Diário, Diário do Minho e Jornal de Macau. Estes corpora são disponibilizados pela Linguatca (www.linguatca.pt/corpora_info.html).

Entrevista: textos transcritos de entrevistas orais cedidas pelo Museu da Pessoa de Portugal (alfarrabio.di.uminho.pt/mp/) e do Brasil (www.museudapessoa.com.br).

Técnico: textos técnicos e científicos extraídos a partir de relatórios contidos no WPT 03 e tratados no Corpógrafo (Sarmiento et al., 2004).

Correio Electrónico: excertos de mensagens da *mailing-list* brasileira da ANCIB (www.ancib.org.br), e do corpus de mensagens CONE (www.linguateca.pt/corpora_info.html).

Expositivo: textos retirados de várias fontes de informação da *web*, como a Wikipedia (pt.wikipedia.org).

Literário: extractos de obras literárias de diversos autores portugueses, brasileiros, angolanos e moçambicanos.

Político: extractos dos corpora EuroParl (people.csail.mit.edu/koehn/publications/europarl/), ECI-EBR (www.linguateca.pt/corpora_info.html) e de discursos de origem timorense.

A Tabela 4.2 apresenta a distribuição dos géneros textuais e das variantes de português pela CH. A colecção possui teores diferentes de géneros textuais segundo a quantidade de texto disponível. O teor baixo de textos de variante africana ou asiática devem-se à dificuldade em encontrar, em tempo útil, textos dessas proveniências.

4.2.1 Colecções douradas

A CH contém 600.086 termos, uma dimensão que torna a sua anotação manual impraticável. Para contornar este problema, seleccionaram-se sub-conjuntos da colecção com o mesmo teor de géneros textuais e de variantes da CH original, com aproximadamente 1/8 do seu tamanho. Estes sub-conjuntos são denominados *Colecções Douradas* (CD), e o seu tamanho já permite a sua anotação manual completa por diversos anotadores (Santos e Cardoso, 2006a).

	Jornalístico	Email	Web	Entrevista
Portugal	101.240 (67,9%)	1.631 (3,3%)	62.474 (44,5%)	52.656 (55,5%)
Brasil	40.889 (27,4%)	48.249 (96,7%)	77.817 (55,5%)	42.288 (44,5%)
Timor-Leste	0	0	0	0
Moçambique	4.650 (3,1%)	0	0	0
Angola	204 (0,1%)	0	0	0
Macau	1.556 (1,0%)	0	0	0
Cabo Verde	527 (0,4%)	0	0	0
Índia	0	0	0	0
Total	149.066 (24,8%)	49.880 (8,3%)	140.291 (23,4%)	94.944 (15,8%)

	Expositivo	Literário	Político	Técnico
Portugal	28.269 (70,1%)	28.776 (56,2%)	54.745 (84,1%)	5.870 (63,1%)
Brasil	8.918 (22,1%)	21.493 (42,0%)	1.262 (1,9%)	3.440 (36,9%)
Timor-Leste	0	0	9.091 (14,0%)	0
Moçambique	0	271 (0,5%)	0	0
Angola	2.076 (5,2%)	656 (1,3%)	0	0
Macau	0	0	0	0
Cabo Verde	509 (1,3%)	0	0	0
Índia	529 (1,3%)	0	0	0
Total	40.301 (6,7%)	51.196 (8,5%)	65.098 (10,4%)	9.310 (1,6%)

Tabela 4.2: Distribuição do género textual e de variante de português na colecção HAREM. Os valores representam a contagem de termos.

Para cada evento HAREM usou-se uma CD distinta. A primeira CD, de 2005, foi usada no evento de 2005, e para o evento de 2006 usou-se uma segunda CD, criada em 2006. A Tabela 4.3 compara a colecção HAREM com as duas CD, e observa-se que a CD de 2006 é ligeiramente menor do que a CD de 2005.

Anotação manual das colecções

Em vez de definir uma categorização à priori, como [Sekine et al. \(2002\)](#) propõem, o HAREM definiu a categorização das EM a partir da análise dos textos das CD.

A anotação manual da CD de 2005 foi realizada em conjunto com os

	Colecção HAREM	Colecção Dourada 2005	Colecção Dourada 2006
Termos	600.086	92.830	62.461
Documentos	1.202	129	128
EM	≈ 40.000	5.270	3.858
EM vagas (classificação)	≈ 1.000	133	142
EM vagas (identificação)	≈ 500	71	56

Tabela 4.3: Comparação entre número de termos, de documentos e de EM, para a colecção HAREM e para as colecções douradas. Os valores de EM para a colecção HAREM são estimativas.

participantes do HAREM, que receberam um pequeno pedaço da CD para anotar livremente e para sugerir novas categorias e alterações às directivas iniciais de etiquetagem. Desta forma, os participantes familiarizaram-se com os problemas reais da tarefa de EM, permitindo desenvolver a metodologia HAREM de uma maneira mais eficaz.

A CD representa o que a comunidade entende ser o resultado ideal da tarefa de REM, mas estão longe de representar o que se espera que os sistemas de REM actuais consigam realizar. A CD permite ter uma perspectiva real da dificuldade da tarefa de REM, estabelecendo um patamar superior para a tarefa a partir do qual se pode observar a evolução dos sistemas inteligentes ao longo do tempo.

Discordância entre anotadores

A CD também mostra que há um limite para os desempenhos da tarefa de REM, imposto pela própria língua. Uma vez que os humanos discordam entre si na marcação de certas EM, não faz sentido exigir aos sistemas de REM que consigam marcar as EM nesses casos (Calzolari e Corazzari, 2000).

Cleverdon (1984) constata que a concordância entre humanos não ultrapassa os 60%, para juízos de relevância de documentos em avaliações de RI. Para a tarefa de REM, o rácio de discordância também é considerável (Mota et al., 2006). Durante a anotação das CD, foi frequente encontrar diferentes interpretações do sentido de várias EM por parte dos anotadores, e leituras diferentes do âmbito semântico dado pela categorização HAREM.

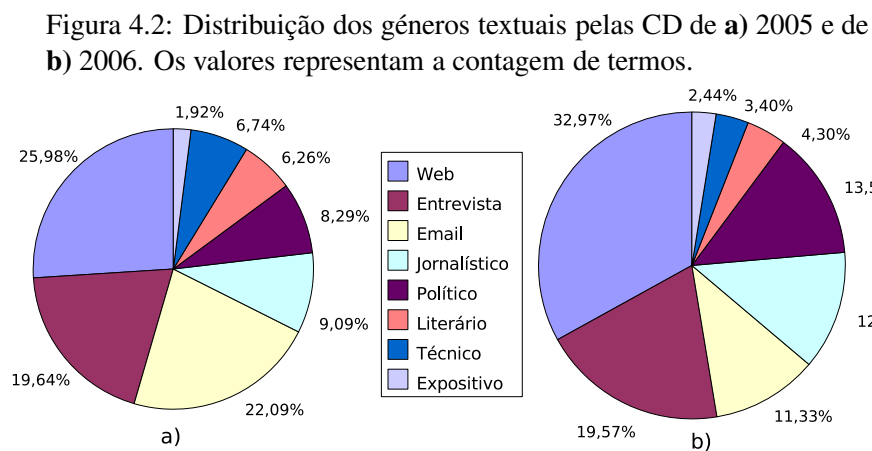
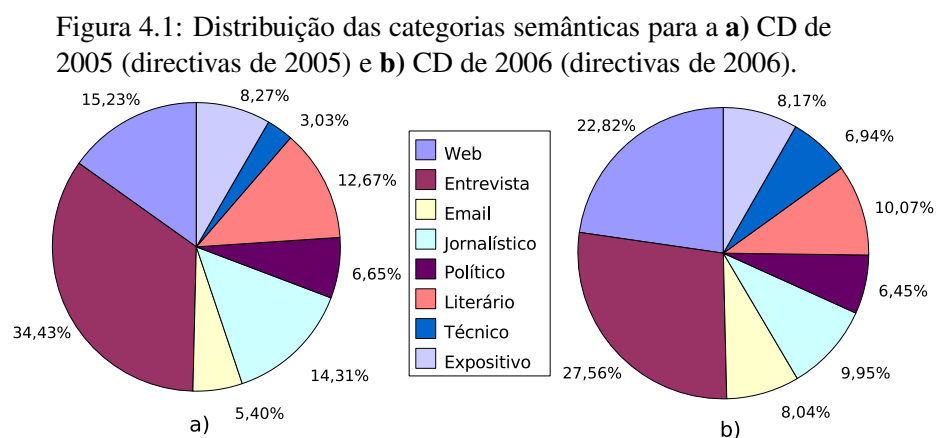
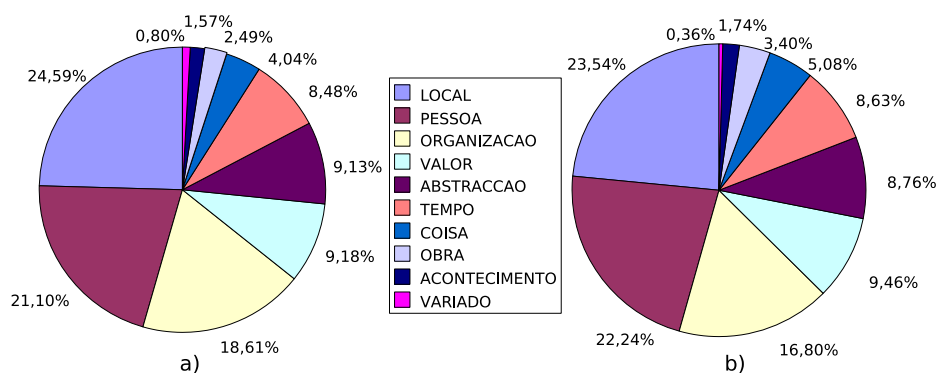
As directivas foram continuamente debatidas, revistas e melhoradas durante a actividade de desenvolvimento da metodologia, para abranger os casos mais difíceis de anotação. A resolução dos casos difíceis envolveram desde a revisão do âmbito semântico das categorias e tipos do HAREM, à marcação de EM vagas nos casos onde não havia consenso numa única opção.

4.2.2 Comparação entre colecções douradas

Ao usar duas CD semelhantes em tamanho e conteúdo, os eventos HAREM procuraram manter o mesmo ambiente de avaliação, possibilitando a medição dos sistemas ao longo do tempo. O resto desta Secção compara a constituição das duas CD.

A Figura 4.1 apresenta a distribuição das categorias semânticas para as duas CD. A distribuição dos géneros textuais pelas duas CD está representada na Figura 4.2 (valores obtidos por contagem de termos), e na Figura 4.3 (valores obtidos por contagem de EM).

A Tabela 4.4 apresenta a distribuição das variantes de português pelas duas CD. Observa-se que as variantes africana e asiática juntas representam aproximadamente 5% da CD de 2005. A CD de 2006 não incluiu documentos africanos ou asiáticos porque os documentos disponíveis dessas



variantes eram poucos, e a escolha aleatória de documentos para a CD não abrangeu nenhum documentos dessas variantes.

Nº de termos	CD de 2005	CD de 2006
Portugal	38.472 (41,44%)	29.864 (47,81%)
Brasil	49.737 (53,58%)	32.597 (52,19%)
África	1.435 (1,55%)	-
Ásia	3.186 (3,43%)	-

Nº de EM	Directivas de 2005	Directivas de 2006
Portugal	2.633 (49,96%)	1.695 (43,93%)
Brasil	2.324 (44,10%)	2.163 (56,07%)
África	76 (1,44%)	-
Ásia	237 (4,50%)	-

Tabela 4.4: Distribuição das variantes de português pelas duas CD, por contagem de termos e por contagem de EM.

As Figuras 4.1, 4.2 e 4.3 e a Tabela 4.4 mostram que os teores de categorias, de géneros textuais e de variantes são semelhantes para as duas CD.

Densidade de EM

As Figuras 4.2 e 4.3 mostram teores diferentes de géneros textuais das CD quando se contam os termos e as EM no texto, respectivamente. Este facto é explicado pelas diferentes *densidades de EM* que cada género textual apresenta nas CD.

A densidade de EM representa o teor de texto que faz parte de EM, e é representada pelo quociente:

$$\text{Densidade de EM} = \frac{\text{Nº de termos que pertencem a EM}}{\text{Nº total de termos}}$$

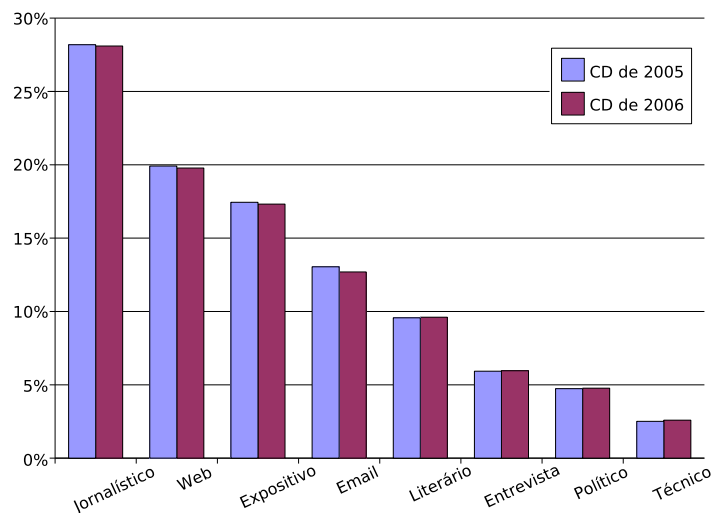


Figura 4.4: Densidade de EM por gêneros textuais, para as colecções douradas de 2005 e de 2006.

A Figura 4.4 mostra a densidade de EM por género textual. Os textos jornalísticos e os textos provenientes da *web* apresentam os maiores valores de densidade de EM. No outro extremo, os géneros textuais com menor densidade de EM são o técnico e o político. Observa-se também que a densidade de EM é semelhante para ambas as CD.

Distribuição de categorias por géneros textuais

Os géneros textuais usados no HAREM possuem diferenças significativas entre si, ao abrangerem diferentes assuntos, autores e estilos de escrita, por exemplo.

As Figuras 4.5 e 4.6 mostram que as diferenças entre géneros textuais também se reflectem na distribuição de EM por categorias semânticas. O teor de categorias por género textual foi semelhante para os dois eventos HAREM, observando-se apenas uma diferença significativa ao nível

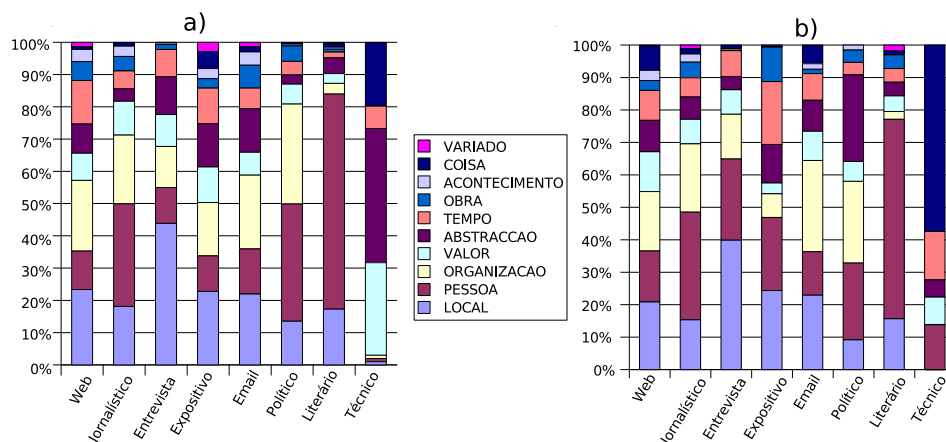


Figura 4.5: Distribuição de categorias semânticas por géneros textuais, para **a)** o evento de 2005, e **b)** o evento de 2006.

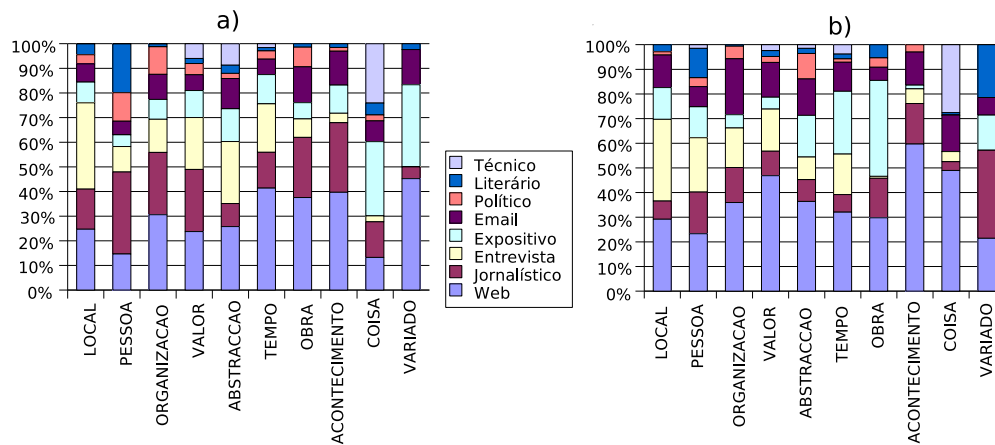


Figura 4.6: Distribuição de géneros textuais por categorias semânticas, para **a)** o evento de 2005, e **b)** o evento de 2006.

Categoria	Evento de 2005			Evento de 2006		
	Média	Mediana	Desv.Pad.	Média	Mediana	Desv.Pad.
ACONTECIMENTO	3,27	3	2,51	3,99	3	4,17
OBRA	3,41	2	3,00	3,52	3	2,94
ABSTRACCAO	1,94	1	1,27	2,64	1	3,51
VARIADO	2,21	1	2,57	2,21	1	2,15
ORGANIZACAO	2,23	1	1,89	2,17	1	2,17
PESSOA	1,91	2	1,13	1,89	2	1,06
TEMPO	1,84	1	1,32	1,79	1	1,36
VALOR	1,76	2	0,94	1,75	2	0,87
LOCAL	1,66	1	1,47	1,66	1	1,43
COISA	1,24	1	0,6	1,53	1	0,89
TOTAL	1,98	1	1,62	1,99	1	1,94

Tabela 4.5: Tamanho das EM por contagem de termos.

da categoria *COISA* dos textos técnicos, devido à revisão da categorização HAREM de 2005 para 2006 (ver Tabela 4.1).

As figuras mostram também que a percentagem de EM de categorias semânticas não abrangidas por eventos de avaliação em REM anteriores, como *ABSTRACCAO* ou *COISA*, é significativa em certos géneros textuais.

Tamanho médio das EM

A Tabela 4.5 analisa o tamanho das EM para os eventos de 2005 e de 2006, discriminadas por categoria semântica. A EM média é composta por dois termos, e a mediana das EM é de um termo. O tamanho das EM por categoria semântica manteve-se semelhante em ambos os eventos HAREM.

EM comuns entre colecções douradas

A colecção HAREM contém alguns documentos que foram criados a partir do mesmo extracto. Apesar de as duas CD não possuírem documentos em comum entre si, podem existir EM comuns entre diferentes documentos,

	2005	2006
Número total de EM	5.132	3.712
Número de EM distintas	3.060	2.434
Rácio	59,63%	65,57%
<hr/>		
Número total de EM comuns	623	
Número de EM distintas comuns	380	
Rácio de total de EM	12,14%	16,78%
Rácio de EM distintas	12,42%	15,61%

Tabela 4.6: Análise ao teor de EM comuns entre eventos HAREM.

uma vez que é provável que as duas CD possuam documentos criados sobre o mesmo extracto.

Em sistemas de REM que operam com base em técnicas de aprendizagem automática, é frequente treinar-se o sistema com colecções anotadas, para aprender a identificar e classificar as EM com base nos exemplos de EM analisadas, aplicando-se o conhecimento adquirido na marcação de novas colecções de textos. Assim sendo, um teor elevado de EM comuns entre as duas CD poderia beneficiar os sistemas de REM baseados em técnicas de aprendizagem automática em relação aos restantes sistemas, se usassem a CD de 2005 como colecção de treino, podendo afectar os resultados do evento de 2006 e as conclusões sobre as aproximações usadas pelos sistemas participantes.

A Tabela 4.6 analisa o teor de EM comuns entre os dois eventos. Os valores totais consideram todas as ocorrências de EM, e os valores distintos só consideram uma vez cada EM repetida. Observa-se que o teor de EM comum a ambos os eventos HAREM não é significativo.

4.3 Desenvolvimento da plataforma HAREM

A plataforma HAREM é uma bancada de avaliação onde os grupos de investigação podem comparar as CD com os textos anotados automaticamente pelos seus sistemas. A avaliação segue um conjunto de directivas estabelecidas conjuntamente com os participantes do HAREM, denominadas *Directivas de Avaliação do HAREM*.

As directivas de avaliação representam o conjunto de *pontuações*, *regras* e *medidas* usadas para comparar as saídas dos sistemas em relação às CD (Cardoso et al., 2006b). As *pontuações* são os valores atribuídos a cada EM marcada, para a respectiva tarefa. As pontuações são calculadas segundo um conjunto de *regras* de pontuação. As *medidas* combinam as várias pontuações de cada tarefa, para representar diferentes componentes da avaliação.

A plataforma é constituída por um conjunto de módulos de *software* que aplicam as directivas de avaliação, estabelecidas conjuntamente pela comunidade durante a actividade de desenvolvimento da metodologia.

4.3.1 Pontuações usadas no HAREM

Pontuações para a tarefa de identificação

Existem cinco pontuações possíveis para a tarefa de identificação.

correcto: todos os termos da EM marcada correspondem aos termos da respectiva EM da CD.

parcialmente_correcto (por Excesso): pelo menos um termo corresponde a outro termo da respectiva EM da CD, e o número de termos da EM da saída é superior à EM da CD.

parcialmente_correcto (por Omissão): pelo menos um termo corresponde a outro termo da respectiva EM da CD, e o número de termos da EM da saída é inferior à EM da CD.

espurio: a EM da saída não tem uma EM correspondente na CD.

em_falta: a EM da CD não tem uma EM correspondente na saída.

O valor atribuído a uma pontuação `correcto` é igual a 1. Para as pontuações `espurio` e `em_falta`, o valor atribuído é igual a 0. Para as pontuações `parcialmente_correcto`, o valor é dado pela seguinte equação:

$$p = 0,5 \frac{n_c}{n_d}$$

onde n_c representa o número de termos em comum, e n_d o número de termos distintos entre as EM da saída e da CD. A pontuação máxima é limitada a 0,5 para garantir que várias EM pontuadas como `parcialmente_correcto` em relação à mesma EM da CD não totalizem o mesmo valor de uma EM pontuada como `correcto`.

Pontuações para a tarefa de classificação morfológica

Na tarefa de classificação morfológica há certas categorias e tipos de EM que não possuem morfologia, como é exemplo o caso da categoria VALOR (Cardoso et al., 2006c). Para outras categorias, a morfologia de certas EM pode ser indefinida, sendo registado nas etiquetas de marcação. Assim sendo, as pontuações possíveis para o género ou o número da tarefa de classificação morfológica são as seguintes:

correcto: o género (ou número) da EM da saída corresponde ao género (ou número) da EM da CD, e a pontuação na tarefa de identificação for `correcto`.

parcialmente_correcto: o género (ou número) da EM da saída corresponde ao género (ou número) da EM da CD, e a pontuação na tarefa de identificação for `parcialmente_correcto`.

incorrecto: se o género (ou número) da EM da saída não corresponde ao género (ou número) da respectiva EM da CD.

espurio: existe um género (ou número) na EM da saída, mas não há classificação morfológica na EM da CD.

em_falta: não existe um género (ou número) na EM da saída, mas há classificação morfológica na EM da CD.

sobre_especificado: existe um género (ou número) na EM da saída, mas a respectiva classificação morfológica na EM da CD é indefinida.

O valor da pontuação `correcto` é igual a 1. A pontuação `parcialmente_correcto` é igual ao valor obtido para a tarefa de identificação. As pontuações restantes têm valor igual a 0.

A tarefa de classificação morfológica é medida através de três medidas:

Género: só é considerada a pontuação para o género.

Número: só é considerada a pontuação para o número.

Combinada: combina as pontuações anteriores, atribuindo o menor valor das duas.

Pontuações para a tarefa de classificação semântica

A tarefa de classificação semântica é pontuada através da comparação das categorias e dos tipos marcados nas etiquetas das EM. Existem três pontuações possíveis para a categoria (ou o tipo):

correcto: a categoria (ou tipo) da EM da saída corresponde à categoria (ou tipo) da EM da CD.

em_falta: a categoria (ou tipo) da EM da CD não está presente na EM da saída.

espurio: a categoria (ou tipo) da EM da saída não está presente na EM da CD.

As pontuações para as categorias e para os tipos podem ser combinadas segundo quatro medidas:

Categoria apenas: apenas as pontuações das categorias são consideradas, e os tipos são ignorados.

Tipos apenas: considera a pontuação dos tipos, se a pontuação da categoria for `correcto`.

Combinada: combina a pontuação da categoria e do tipo num único valor, segundo a seguinte fórmula:

$$p_{CSC} = \begin{cases} 0 & \text{para categoria incorrecta.} \\ 1 & \text{para categoria correcta e tipo incorrecto.} \\ 1 + \left(1 - \frac{n_c}{n_t}\right) - \frac{n_e}{n_t} & \text{para categoria e tipo correcto.} \end{cases}$$

onde n_c representa o número de tipos pontuados como `correcto`, n_e o número de tipos pontuados como `espurio` e n_t o número total de tipos válidos para a respectiva categoria.

Plana: atribui um valor de 1 se as pontuações para a categoria e para o tipo forem ambas `correcto`.

4.3.2 Métricas de avaliação

As directivas de avaliação do HAREM inspiram-se nas directivas usadas pelo MUC (Douthat, 1998), que apresentaram novas métricas para a avaliação das suas tarefas: *overgeneration*, *undergeneration*, *substitution* e *error per response fill*. Assim, além das métricas de precisão, abrangência e de medida F, o HAREM desenvolveu as métricas de *Sobre-geração*, *Sub-geração* e *Erro Combinado* para aferir dos desempenhos dos sistemas.

A *precisão* mede o teor de acertos que o sistema obtém em relação ao número de EM que marcou, e é calculada pela Equação 4.1.

$$\text{Precisão} = \frac{\sum \text{Pontuação do sistema}}{\sum \text{Pontuação máxima do sistema}} \quad (4.1)$$

A *abrangência* mede o teor de acertos que o sistema obtém em relação ao número de EM possíveis de marcar, e é calculada pela Equação 4.2.

$$\text{Abrangência} = \frac{\sum \text{Pontuação do sistema}}{\sum \text{Pontuação máxima da CD}} \quad (4.2)$$

A *medida F* combina a precisão e a abrangência segundo a Equação 4.3 (van Rijsbergen, 1979):

$$\text{Medida F} = \frac{2 \times \text{precisão} \times \text{abrangência}}{(\text{precisão} + \text{abrangência})} \quad (4.3)$$

A *sobre-geração* (*sobre-especificação*, na tarefa de classificação morfológica) mede o teor de classificações que o sistema faz em excesso, resultando em pontuações espúrio (ou sobre-especificado). A sobre-geração e a sobre-especificação são calculadas segundo as Equações 4.4 e 4.5, respectivamente.

$$\text{Sobre-geração} = \frac{\sum \text{Pontuações espúrio}}{\sum \text{Pontuação máxima do sistema}} \quad (4.4)$$

$$\text{Sobre-especificação} = \frac{\sum \text{Pontuações sobre-especificado}}{\sum \text{Pontuação máxima do sistema}} \quad (4.5)$$

A *sub-geração* mede o teor de marcações que o sistema não efectuou, em relação ao número de EM possíveis de marcar, e é calculada pela Equação 4.6.

$$\text{Sub-geração} = \frac{\sum \text{Pontuações em_falta}}{\sum \text{Pontuação máxima da CD}} \quad (4.6)$$

O *erro combinado* combina a sobre-geração e a sub-geração num único valor que quantifica o teor de erros cometido pelo sistema na tarefa que realizou, e é dado pela Equação 4.7.

$$\text{Erro combinado} = \frac{\sum \text{em_falta} + \sum \text{espúrio} + \sum (1\text{-par.cor.})}{\sum (\text{Pont. máx. sistema} \cup \text{Pont. máx. CD})} \quad (4.7)$$

4.3.3 Arquitectura da plataforma

A arquitectura da plataforma HAREM é composta por um *pipeline* de módulos de *software* que executam tarefas simples e específicas (Seco et al., 2006). A opção por esta arquitectura flexível permite realizar a avaliação segundo vários cenários. O esquema da avaliação está representada na

Figura 4.7. A avaliação das saídas dos sistemas é dividida em quatro fases, a seguir descritas:

Fase 1: Extracção e alinhamento. As saídas dos sistemas são verificadas e corrigidas na sua sintaxe. O sub-conjunto de documentos relativos à CD é retirado da saída pelo *Extractor de CD*. As EM do sub-conjunto são posteriormente extraídas e alinhadas com as respectivas EM da CD pelo módulo *AlinhEM*, gerando informação sobre os emparelhamentos das EM, denominados *alinhamentos*. O *AvalIDA* processa os alinhamentos e gera os primeiros resultados para a tarefa de identificação.

Fase 2: Filtragem. Nesta fase realiza-se uma filtragem selectiva dos alinhamentos, para permitir a avaliação parcial segundo diversos cenários específicos. O *Véus* é o módulo responsável pela filtragem dos alinhamentos, a partir 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 pelos módulos *Vizir* e *Emir*, respectivamente, a partir dos alinhamentos. Os módulos *ALTinaID*, *ALTinaMor* e *ALTinaSem* analisam as EM vagas nos resultados para as três tarefas respectivas, e seleccionam as alternativas escolhidas pelo sistema. Finalmente, os módulos *Ida2ID*, *Ida2Mor* e *Ida2Sem* processam os alinhamentos finais e calculam os valores das métricas da saída para as três tarefas, respectivamente.

Fase 4: Relatórios. Nesta fase final, os resultados finais da avaliação são processados para gerar relatórios de desempenho que possam ser

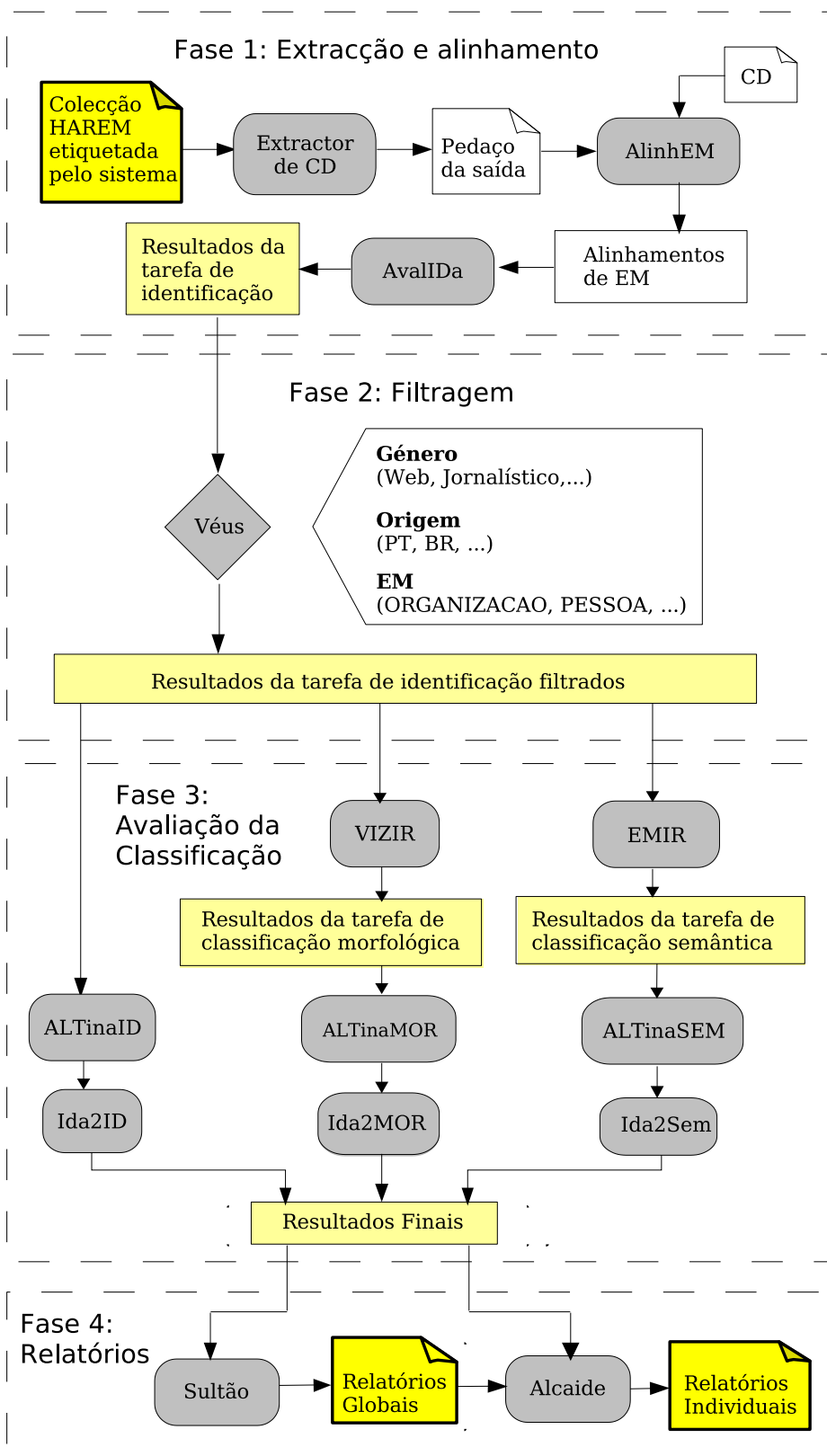


Figura 4.7: Esquema de avaliação da plataforma HAREM.

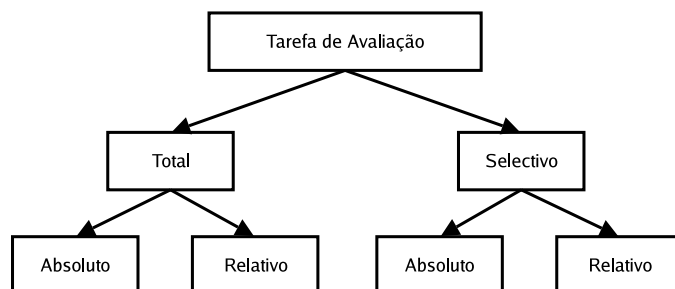


Figura 4.8: Cenários de avaliação usados nos eventos HAREM.

facilmente interpretados pelos humanos. 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.

4.3.4 Cenários de avaliação

Os grupos de investigação que participaram no HAREM desenvolveram os seus sistemas de REM com objectivos variados, que vão desde a atribuição de âmbitos geográficos em páginas da *web* à análise morfossintáctica de textos, extracção terminológica ou a criação de bases de conhecimento.

A plataforma HAREM, através do módulo *Véus*, permite realizar *cenários de avaliação* para a medição parcial das saídas dos sistemas, ajustando a avaliação às características de cada sistema participante.

A avaliação dos eventos HAREM realizou-se segundo dois *eixos* de cenários: o eixo de cenários *absoluto – relativo*, e o eixo de cenários *total – selectivo* (ver Figura 4.8).

Cenário absoluto–relativo: O *cenário absoluto* avalia o desempenho do sistema em relação à totalidade das EM na CD para a tarefa de REM

completa, ou seja, a identificação e a classificação de EM. O *cenário relativo*, por seu lado, restringe a avaliação às EM pontuadas como *correcto* ou *parcialmente_correcto* na tarefa de identificação. Este cenário permite avaliar o desempenho do sistema apenas na tarefa de classificação (semântica ou morfológica), independentemente do desempenho na tarefa de identificação.

Cenário total-selectivo: O *cenário total* abrange todas as categorias de EM da CD, avaliando a tarefa de classificação morfológica ou semântica em relação à tarefa tal como foi proposta pelo HAREM. No *cenário selectivo*, o participante escolhe previamente um sub-conjunto de categorias e de tipos da categorização HAREM que o seu sistema consegue processar. Assim, a tarefa da classificação (semântica ou morfológica) é avaliada segundo esse sub-conjunto de categorias e de tipos.

As comparações entre saídas de sistemas participantes apresentadas nesta tese são realizadas segundo o cenário total e absoluto, ou seja, a tarefa inicial proposta pelo HAREM. Nestas comparações, os sistemas de REM mais específicos e que desempenham a tarefa de REM segundo um cenário selectivo são penalizados, uma vez que focam num sub-conjunto de categorias de EM. Contudo, os relatórios de desempenho gerados permitem comparar parcialmente os sistemas de REM por categoria, género textual ou por variante.

4.4 Organização dos eventos HAREM

A avaliação dos eventos HAREM segue o modelo de avaliação de sistemas inteligentes descrito na Figura 1.4. Os eventos HAREM tiveram início com

Categoria	2005	2006
Total de saídas	18	20
Cenário total	10	9
Cenário selectivo	8	11
LOCAL	8	11
PESSOA	6	9
ORGANIZACAO	6	9
VALOR	5	8
TEMPO	4	9
ACONTECIMENTO	1	5
ABSTRACCAO	0	0
COISA	0	0
OBRA	0	0
VARIADO	0	0

Tabela 4.7: Cenários escolhidos pelos participantes, e categorias escolhidas nos cenários selectivos.

o envio da CH sem anotações para os participantes. Num prazo máximo de 48 horas, os participantes devolveram a CH anotada automaticamente pelos seus sistemas. Após a recepção de todas as saídas, a plataforma HAREM mediu o desempenho dos sistemas e gerou relatórios de desempenho globais e individuais. Os relatórios de desempenho globais podem ser consultados no sítio do HAREM.

A Tabela 4.7 apresenta os cenários escolhidos pelos participantes, e para os cenários selectivos as categorias escolhidas. Em 2005, quatro participantes escolheram cenários selectivos, num total de 8 saídas enviadas segundo um cenário selectivo. Em 2006, quatro participantes voltaram a escolher cenários selectivos, num total de 11 saídas segundo um cenário selectivo. Pode-se observar que a categoria `LOCAL` é a única que todos os sistemas de REM se propõem reconhecer, enquanto que as categorias `ABSTRACCAO`, `COISA`, `OBRA` e `VARIADO` não foram escolhidas por nenhum sistema nos seus cenários selectivos.

Evento	2005	2006
Número de participantes	10	5
Número de países representados	6	2
Número de saídas	18	20
Saídas para classificação morfológica	4	4
Saídas para classificação semântica	15	19

Tabela 4.8: Resumo da participação nos eventos HAREM.

4.5 Sumário

Este Capítulo descreveu os aspectos mais importantes da metodologia e da plataforma HAREM e que foram aplicados em dois eventos HAREM, tendo o primeiro sido realizado em Fevereiro de 2005, e o segundo em Abril de 2006 (também conhecido como *MiniHAREM*).

O evento de 2006 usou uma CD ligeiramente menor do que a CD usada no evento de 2005, e aplicou uma categorização melhorada. No entanto, estas alterações não têm um impacto significativo na distribuição das categorias de EM nos textos, como tal os dois eventos usaram CD com características semelhantes.

A Tabela 4.8 apresenta um resumo do número de participantes e de saídas enviadas aos eventos HAREM. O evento de 2006 foi restringido aos grupos de investigação que participaram no evento de 2005, para permitir observar e medir a melhoria dos sistemas em pouco mais de um ano e obter mais dados de cada sistema participante para a validação estatística do HAREM.

No próximo Capítulo apresenta-se os resultados obtidos nos dois eventos, para as tarefas de identificação e de classificação semântica, em conjunto com a análise estatística.

Capítulo 5

Análise de resultados

O presente Capítulo apresenta a análise estatística realizada ao HAREM para validar a sua metodologia. Inicia-se com a selecção e justificação da escolha do teste estatístico usado, e descreve a seguir a adaptação do teste aos requisitos do HAREM. Apresenta-se também uma análise estatística completa às saídas dos eventos HAREM. No final, faz-se uma comparação entre os resultados obtidos e os de eventos de avaliação anteriores a sistemas de REM.

5.1 Realização da análise estatística

O Capítulo 3 mostrou que apenas os testes não-paramétricos são adequados à análise estatística de avaliações em PLN. Entre os testes não-paramétricos mais usados, destacam-se o teste *bootstrap* e os testes de permutação. O teste seleccionado para a análise estatística do HAREM foi o teste de permutação de aleatorização parcial, pelas razões a seguir expostas.

Riezler e Maxwell III (2005) compararam os dois testes estatísticos para um conjunto de métricas de avaliação e concluíram que o teste de permutação obtém valores p mais conservadores do que o método *bootstrap*. O *bootstrap* é mais sensível à qualidade dos dados iniciais, o que pode originar re-amostragens enviesadas se os dados iniciais são insuficientes ou pouco representativos da população. Noreen (1989) afirma que o enviesamento da distribuição *bootstrap* pode originar a rejeição da hipótese nula, quando na verdade a diferença entre os parâmetros em avaliação não sugerem a sua rejeição. Este facto limita a confiança que se pode ter nos resultados do teste de hipóteses *bootstrap*.

Na adaptação do método *bootstrap* aos requisitos do HAREM, existem dúvidas se a aplicação seria apropriada, já que:

- A metodologia de geração de re-amostragens do *bootstrap* não tem em consideração as fortes dependências que existem entre EM.
- No método *bootstrap*, não há garantias de que todas as EM serão usadas nas re-amostragens, ao contrário dos testes de permutação. Assim, não há certezas de que as re-amostragens geradas sejam representativas da população (ou da saída do sistema, no caso particular do HAREM).

O método de permutação revela-se mais adequado para a validação estatística do HAREM, pois usa todos os alinhamentos das saídas no teste, e não repete observações durante a geração de re-amostragens.

5.1.1 Adaptação do teste de permutação ao HAREM

Os testes de permutação pressupõem que as observações são permutáveis entre si. No entanto, este pressuposto não é satisfeito na saídas geradas pela plataforma HAREM:

- É frequente encontrar observações com pontuação *espurio* ou *em_falta* da saída *A* que não têm correspondência na saída *B* e vice-versa, comprometendo o pressuposto de permutabilidade entre observações.
- As alternativas das EM vagas na tarefa de identificação podem totalizar números diferentes de observações para as saídas *A* e *B*.
- As observações da saída *A* podem depender de várias observações da saída *B*, e vice-versa.

Saídas	Texto / EMs
A	<div> <div>①</div> <div>Segundo</div> <div>o</div> <div>②</div> <div>presidente da Fundação</div> <div>para o</div> <div>③</div> <div>Desenvolvimento da Produção, Costa e Silva, ...</div> </div>
CD	<div> <div>Segundo</div> <div>o</div> <div>presidente</div> <div>da</div> <div>Fundação</div> <div>para o</div> <div>Desenvolvimento da Produção,</div> <div>Costa e Silva, ...</div> </div>
B	<div> <div>Segundo</div> <div>o</div> <div>①</div> <div>presidente</div> <div>da</div> <div>②</div> <div>Fundação</div> <div>para o</div> <div>Desenvolvimento da Produção,</div> <div>③</div> <div>Costa</div> <div>e</div> <div>④</div> <div>Silva,</div> <div>⑤</div> <div>...</div> </div>

Figura 5.1: Excerto de texto marcado com EM nas saídas A e B, e respectivos alinhamentos com a CD representados por setas.

CD ---> Saída A
<ESPÚRIO> ---> <A1>Segundo</A1> (ESP)
<CD1>Fund. Des. Produção</CD1> ---> <A2>presidente da Fundação</A2> (PC)
<CD1>Fund. Des. Produção</CD1> ---> <A3>Des. Produção, Costa Silva</A3> (PC)
<CD2>Costa e Silva</CD2> ---> <A3>Des. Produção, Costa Silva</A3> (PC)
CD ---> Saída B
<ESPÚRIO> ---> <B1>presidente</B1> (ESP)
<CD1>Fundação Des. Produção</CD1> ---> <B2>Fundação</B2> (PC)
<CD1>Fundação Des. Produção</CD1> ---> <B3>Des. Produção</B3> (PC)
<CD2>Costa e Silva</CD2> ---> <B4>Costa</B4> (PC)
<CD2>Costa e Silva</CD2> ---> <B5>Silva</B5> (PC)

Figura 5.2: Lista de alinhamentos gerados pela plataforma HAREM para as saídas A e B do exemplo da Figura 5.1. Entre parênteses apresenta-se as pontuações dos alinhamentos, ESP (espurio) e PC (parcialmente_correcto).

O exemplo de texto das Figuras 5.1 e 5.2 ilustra o problema. Pode-se observar que a CD identifica 2 EM, a saída A identifica 3 EM e produz 4 alinhamentos, e a saída B identifica 5 EM e produz 5 alinhamentos. A diferença entre o número de alinhamentos para as saídas A e B viola o pressuposto de permutabilidade dos testes de permutações. Outra situação relevante é ilustrada nos alinhamentos respeitantes à EM “*presidente da Fundação*”, onde se pode verificar que a observação A2 da saída A depende das observações B1 e B2 da saída B. A permutação das observações A2, B1

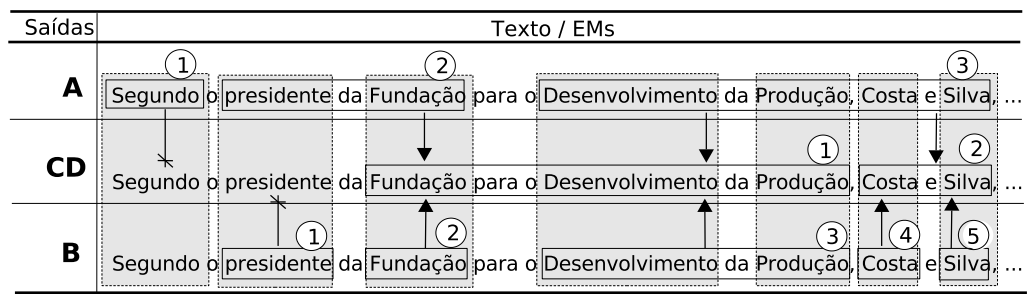


Figura 5.3: Permutações de termos para o exemplo da Figura 5.1.

ou *B2* não pode violar o pressuposto de independência entre observações permutadas.

Apontam-se duas estratégias para resolver os problemas encontrados, que são a seguir dissecadas:

- 1. Reduzir as observações ao seu elemento mínimo comum, ou seja, permutar os termos do texto.
- 2. Agrupar as observações ao seu elemento máximo comum, ou seja, permutar blocos de observações do texto.

Permutação de termos

A Figura 5.3 ilustra o exemplo da Figura 5.1 com as possíveis permutações segundo a estratégia de permutação de termos. As *stop-words* não são consideradas na permutação. A permutação de termos procura reproduzir uma conhecida estratégia de REM, no qual o sistema processa sequencialmente os termos do texto, a estratégia BIO (**Sang e de Meulder, 2003**). Segundo esta estratégia, usada nas colecções de texto da tarefa de REM do CoNLL, os termos são etiquetados com os seguintes marcadores:

- B** (*Begin*), se o termo está no início de uma EM.
- I** (*Inside*), se o termo pertence a uma EM, mas não a inicia.

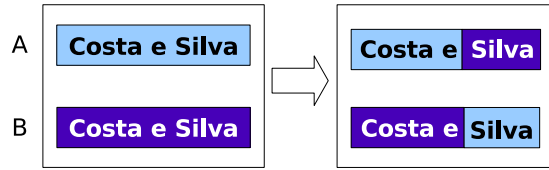


Figura 5.4: Permutações de termos com classificações semânticas diferentes. As saídas A e B marcaram a EM “Costa e Silva” com categorias diferentes, representadas na figura por tons diferentes.

O (*Outside*), se o termo não pertence a nenhuma EM.

Contudo, a permutação de termos traz as seguintes dificuldades:

- A permutação de termos pode partir as EM em pedaços, o que altera a sua pontuação original. Uma vez que os alinhamentos com pontuação `correcto` e um valor de 1 podem ser partidos em vários pedaços com pontuações `parcialmente_correcto`, cujo somatório tem um valor máximo limitado a 0,5, é muito provável que o valor absoluto da métrica final para as saídas A e B seja prejudicado pelas quebras de EM. A alteração no valor absoluto das métricas pode ter consequências nefastas na decisão de rejeição da hipótese nula.
- Após a quebra das EM e a permuta dos termos, é necessário unir os termos para restaurar as respectivas EM originais. No entanto, no caso da classificação semântica, a reconstrução pode gerar EM com diferentes categorias semânticas (ver Figura 5.4).
- A quebra das EM implica recalcular as pontuações de cada saída. Para tal, é necessário re-avaliar as EM em relação à CD para cada re-amostragem. O processo de avaliação das saídas é algo demorado, e a necessidade de gerar um número considerável de re-amostragens torna impraticável a estratégia de permutação de termos.

Permutação de blocos

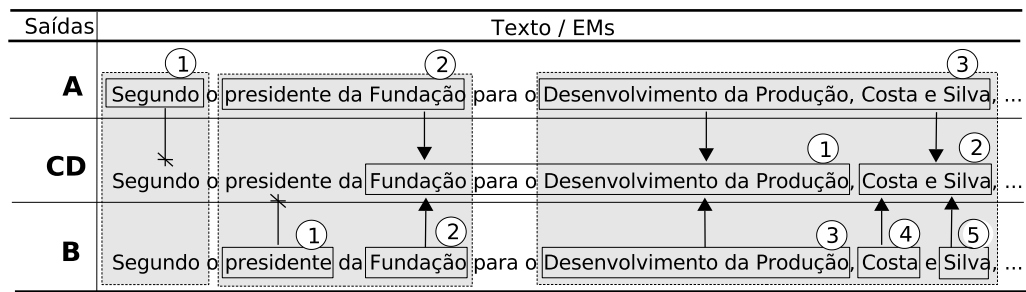


Figura 5.5: Permutações de blocos para o exemplo da Figura 5.1.

A Figura 5.5 ilustra o exemplo da Figura 5.1 com as possíveis permutações segundo a estratégia de permutação de blocos de EM. A permutação de blocos apresenta as seguintes vantagens:

- a permutação por blocos não quebra as EM, evitando os inconvenientes da permutação de termos.
- A pontuação de cada alinhamento não sofre alterações com a permuta, não sendo necessário recalcular as pontuações para cada re-amostragem.
- Para alinhamentos sobre EM vagas na sua identificação, a permuta por blocos não é afectada pelo número diferente de alinhamentos que pode existir entre saídas.

A permutação de blocos de EM pode ser interpretada como uma permutação ao nível de determinadas unidades de texto, como expressões multi-palavra, frases ou mesmo parágrafos. Apesar do agrupamento de EM em blocos diminuir o número de observações permutáveis, a estratégia mantém a independência entre observações e adequa-se aos objectivos apontados para a validação estatística do HAREM.

	Esperado Correcto	Esperado Incorrecto	Total
Observado Correcto	Verdadeiros Positivos (VP)	Falsos Positivos (FP)	VP + FP
Observado Incorrecto	Falsos Negativos (FN)	Verdadeiros Negativos (VN)	FN + VN
Total	VP + FN	FP + VN	VP+FP+FN+VN

Tabela 5.1: Exemplo de uma tabela de contingência usada em sistemas de previsão.

Com base nesta análise, adoptou-se a estratégia de permutação por blocos para os testes de aleatorização parcial usados na análise estatística ao HAREM.

5.1.2 Métricas de avaliação

No HAREM, as métricas de sobre-geração, sub-geração e erro combinado medem o teor de erros que os sistemas de REM cometem nas tarefas. Estas métricas servem para o diagnóstico individual de cada sistema, e não são adequadas para a comparação de sistemas.

Para que o teste de permutação consiga distinguir os sistemas, é necessário usar métricas que representem adequadamente as diferenças de desempenho dos mesmos (Riezler e Maxwell III, 2005). Esta Secção descreve o processo de selecção das métricas utilizadas na análise estatística do HAREM.

A medição do desempenho dos sistemas participantes baseia-se na criação de *tabelas de contingência* para as classes de decisão que o sistema pode realizar. Estas tabelas de contingência agrupam os resultados esperados e os resultados observados em duas classificações: *correctos* e *incorrectos* (ver Tabela 5.1).

A precisão e a abrangência são duas métricas muito usadas em even-

tos de avaliação em PLN. A precisão representa a proporção de resultados correctos em relação ao número de resultados observados ($\frac{VP}{VP+FP}$), enquanto que a abrangência representa a proporção de resultados correctos em relação ao número de resultados esperados ($\frac{VP}{VP+FN}$).

A medida F é uma métrica proposta por **van Rijsbergen (1979)** e que combina a precisão e abrangência num único valor (ver Equação 4.3). Um sistema inteligente eficaz procura obter bons desempenhos tanto em termos de precisão como em termos de abrangência, logo procura maximizar os valores de medida F . No entanto, a medida F sozinha não é suficiente para comparar os sistemas; por exemplo, duas saídas A e B podem apresentar valores de medida F semelhantes, mas apresentar também valores de precisão e de abrangência distintos, evidenciando que as duas saídas são diferentes.

Assim sendo, a análise estatística ao HAREM usa as métricas de precisão, abrangência e medida F para comparar as saídas dos sistemas.

5.2 Análise estatística à colecção dourada

Muitas das reticências à metodologia de avaliação usada em eventos como o HAREM dizem respeito à colecção de textos usada:

- Uma colecção de textos como a *web* é muito difícil de representar numa colecção estática, não só devido à diversidade de assuntos, formatos, autores e estilos de escrita, mas também à volatilidade dos seus conteúdos (**Gomes e Silva, 2006**). A colecção de textos usada é uma amostra representativa da colecção real de textos?

- Qual é o tamanho mínimo da colecção para poder ser considerada uma amostra válida da colecção real que se pretende representar? Como se pode determinar esse tamanho mínimo?
- Os resultados destes eventos de avaliação podem ser extrapolados para a colecção real? Se o sistema *A* revela-se superior ao sistema *B* no evento de avaliação, será que o mesmo se sucede fora da avaliação?

No caso do HAREM, é possível verificar se os dois eventos HAREM produzem os mesmos resultados, uma vez que usaram a mesma CH que serviu de base às duas CD. Observou-se que os resultados dos eventos HAREM não sofrem alterações significativas, usando a CD de 2005, a CD de 2006 ou ambas as CD. Este facto mostra que o tamanho das CD usadas não influencia os resultados, e que estes traduzem as diferenças reais entre os sistemas de REM participantes. O passo seguinte é determinar o tamanho mínimo da CD que permitiria manter os mesmos resultados observados nos eventos.

5.2.1 Variação do número de observações

A experiência foi realizada sobre duas saídas reais do evento de 2006. Adoptando o critério (subjectivo) de Sparck-Jones (1974), "*differences of 5% are noticeable, and differences of 10% are material*", pode-se concluir que a saída *A* é melhor do que a saída *B* com base nos valores das métricas apresentadas na Tabela 5.2.

Na Tabela 5.2 observa-se que há uma diferença de 2 EM no número de EM da CD entre as duas saídas. Esta diferença explica-se pela opção feita por diferentes alternativas em dois casos de EM vagas na sua iden-

	Saída A	Saída B	Diferença
Número de EM na saída	4.086	4.191	105
Número de EM na CD	3.663	3.661	2
Número de blocos	4.312	4.312	-
Precisão	79,77%	72,84%	6,93%
Abrangência	87,00%	69,58%	17,42%
Medida F	0,8323	0,7117	0,1206

Tabela 5.2: Resultados da tarefa de identificação para duas saídas do evento de 2006.

tificação, por cada saída. Observa-se também que o número de blocos é aproximadamente 4% maior do que o número de EM marcadas nas saídas. Esta discrepância deriva do número de alinhamentos de cada saída com pontuação `espurio` e `em_falta` que não têm contrapartida na saída oposta, gerando blocos semelhantes ao primeiro bloco do exemplo da Figura 5.5.

A Tabela 5.3 mostra que as médias das re-amostragens das saídas *A* e *B* se mantêm constantes para os vários sub-conjuntos de blocos usados. Contudo, a Tabela 5.4 mostra que a média da diferença entre re-amostragens tem tendência a aumentar com a diminuição do número de blocos, assim como os respectivos desvios-padrão. A precisão é a primeira métrica a registar valores p acima de α para um intervalo de confiança de 99%, uma vez que apresenta a diferença inicial mais baixa entre as três métricas.

Esta experiência mostra que quando se diminui o número de blocos no teste de permutação, a média e o desvio padrão da distribuição empírica aumentam até se atingir um ponto em que o valor p excede o valor de α (este comportamento é ilustrado na Figura 5.6). Como a significância dos resultados depende da métrica escolhida e da diferença inicial de valores entre as saídas, não é possível determinar um tamanho mínimo absoluto para a CD.

Re-amostragens de A						
NºBlocos	Média			Desvio Padrão		
	Precisão	Abrang.	Medida F	Precisão	Abrang.	Medida F
Todos	0,7653	0,7831	0,7741	0,0035	0,0041	0,0032
2.000	0,7654	0,7717	0,7685	0,0080	0,0092	0,0073
1.000	0,7654	0,7653	0,7653	0,0128	0,0146	0,0117
500	0,7657	0,7614	0,7634	0,0188	0,0216	0,0174
250	0,7654	0,7592	0,7620	0,0272	0,0312	0,0254
200	0,7651	0,7589	0,7616	0,0302	0,0341	0,0278
100	0,7656	0,7588	0,7614	0,0432	0,0490	0,0402
75	0,7667	0,7590	0,7619	0,0503	0,0570	0,0471
50	0,7657	0,7593	0,7611	0,0619	0,0689	0,0576
25	0,7672	0,7624	0,7622	0,0865	0,0928	0,0793

Re-amostragens de B						
NºBlocos	Média			Desvio Padrão		
	Precisão	Abrang.	Medida F	Precisão	Abrang.	Medida F
Todos	0,7654	0,7831	0,7741	0,0035	0,0041	0,0032
2.000	0,7655	0,7718	0,7686	0,0080	0,0091	0,0072
1.000	0,7652	0,7648	0,7650	0,0127	0,0145	0,0116
500	0,7654	0,7610	0,7631	0,0185	0,0213	0,0171
250	0,7659	0,7600	0,7627	0,0268	0,0314	0,0253
200	0,7650	0,7592	0,7617	0,0304	0,0346	0,0283
100	0,7651	0,7581	0,7609	0,0437	0,0493	0,0408
75	0,7665	0,7591	0,7619	0,0500	0,0566	0,0468
50	0,7658	0,7597	0,7613	0,0618	0,0685	0,0574
25	0,7660	0,7596	0,7602	0,0864	0,0928	0,0794

Tabela 5.3: Médias e desvios-padrão para as re-amostragens da saída A e B, para vários sub-conjuntos de blocos de tamanho decrescente ($n_r=9.999$).

NºBlocos	Valor p			Média			Desvio Padrão		
	Prec.	Abr.	Med.F	Prec.	Abr.	Med.F	Prec.	Abr.	Med.F
Todos	0,0001	0,0001	0,0001	0,006	0,006	0,005	0,004	0,005	0,004
2.000	0,0001	0,0001	0,0001	0,008	0,009	0,008	0,006	0,007	0,006
1.000	0,0001	0,0001	0,0001	0,012	0,013	0,011	0,009	0,010	0,008
500	0,0004	0,0001	0,0001	0,017	0,018	0,015	0,013	0,014	0,011
250	0,0181	0,0001	0,0001	0,024	0,026	0,021	0,018	0,020	0,016
200	0,0351	0,0001	0,0001	0,026	0,029	0,024	0,020	0,022	0,018
100	0,1391	0,0009	0,0047	0,037	0,041	0,034	0,028	0,031	0,026
75	0,1912	0,0034	0,0123	0,043	0,048	0,039	0,032	0,036	0,029
50	0,2900	0,0181	0,0453	0,053	0,058	0,048	0,040	0,045	0,036
25	0,4488	0,0843	0,1505	0,073	0,081	0,066	0,056	0,061	0,051

Tabela 5.4: Valores p , médias e desvios-padrão das diferenças entre re-amostragens da saída A e B, para vários sub-conjuntos de blocos de tamanho decrescente ($n_r=9.999$).

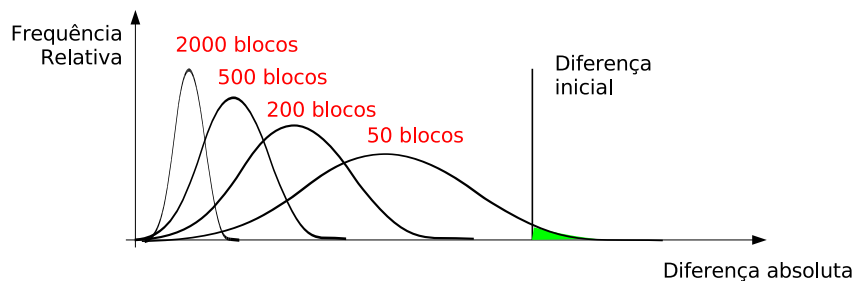


Figura 5.6: A influência do número de blocos na média e desvio padrão da diferença entre re-amostragens.

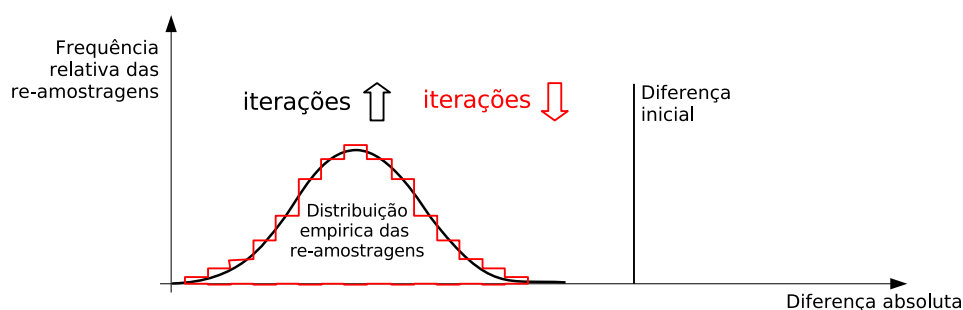


Figura 5.7: A influência do número de re-amostragens na média e desvio padrão da diferença entre re-amostragens.

5.2.2 Variação do número de re-amostragens

A experiência anterior gerou 9.999 re-amostragens para obter os resultados. Desta feita, repetiu-se a experiência para os sub-conjuntos de 2.000, 200 e 25 blocos, usando valores de 999 e de 99 re-amostragens. Os resultados mostram que o número de re-amostragens não têm influência nos valores de média e desvio padrão das re-amostragens. Este comportamento comprova a análise realizada na Secção 3.2.4, segundo o qual o número de re-amostragens influencia a resolução da distribuição empírica, mas não altera os seus parâmetros (ver Figura 5.7).

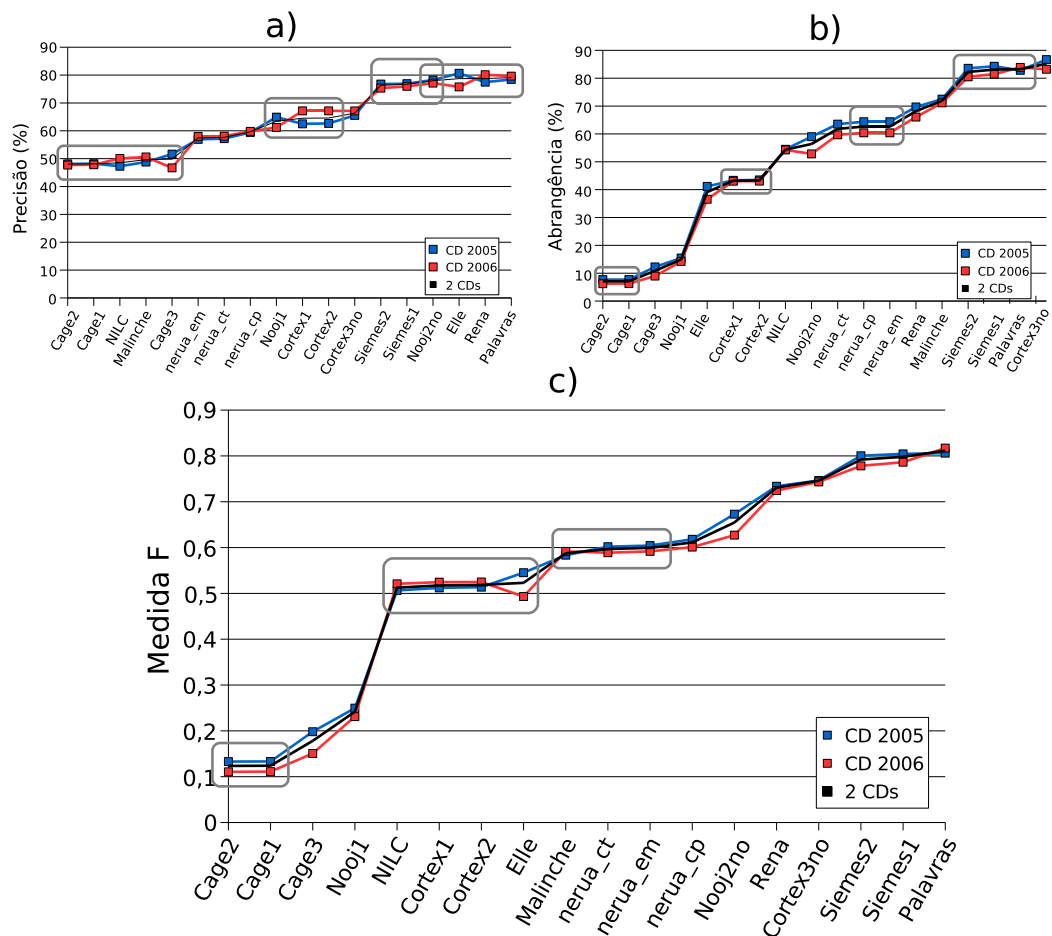


Figura 5.8: Desempenho dos sistemas para a tarefa de identificação no evento de 2005, para a **a)** precisão, **b)** abrangência e **c)** medida F.

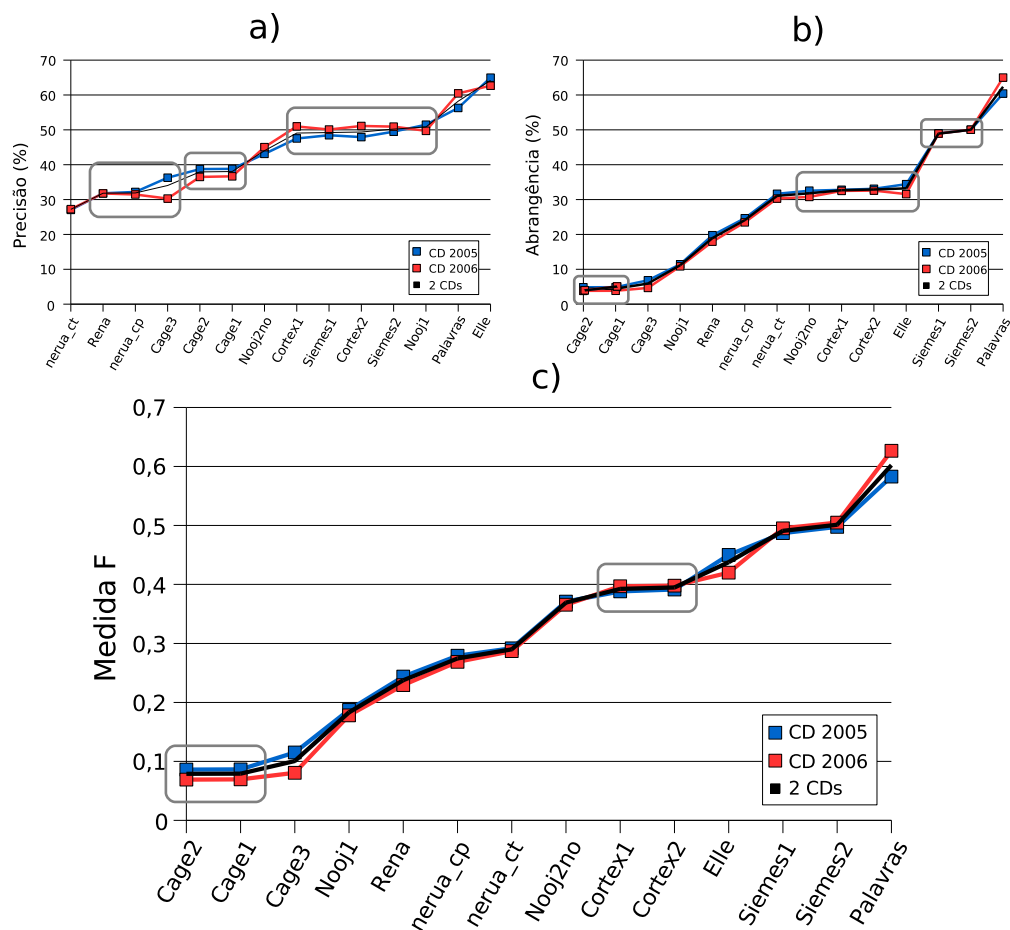


Figura 5.9: Desempenho dos sistemas para a tarefa de classificação semântica (na medida combinada) no evento de 2005, para a **a)** precisão, **b)** abrangência e **c)** medida F.

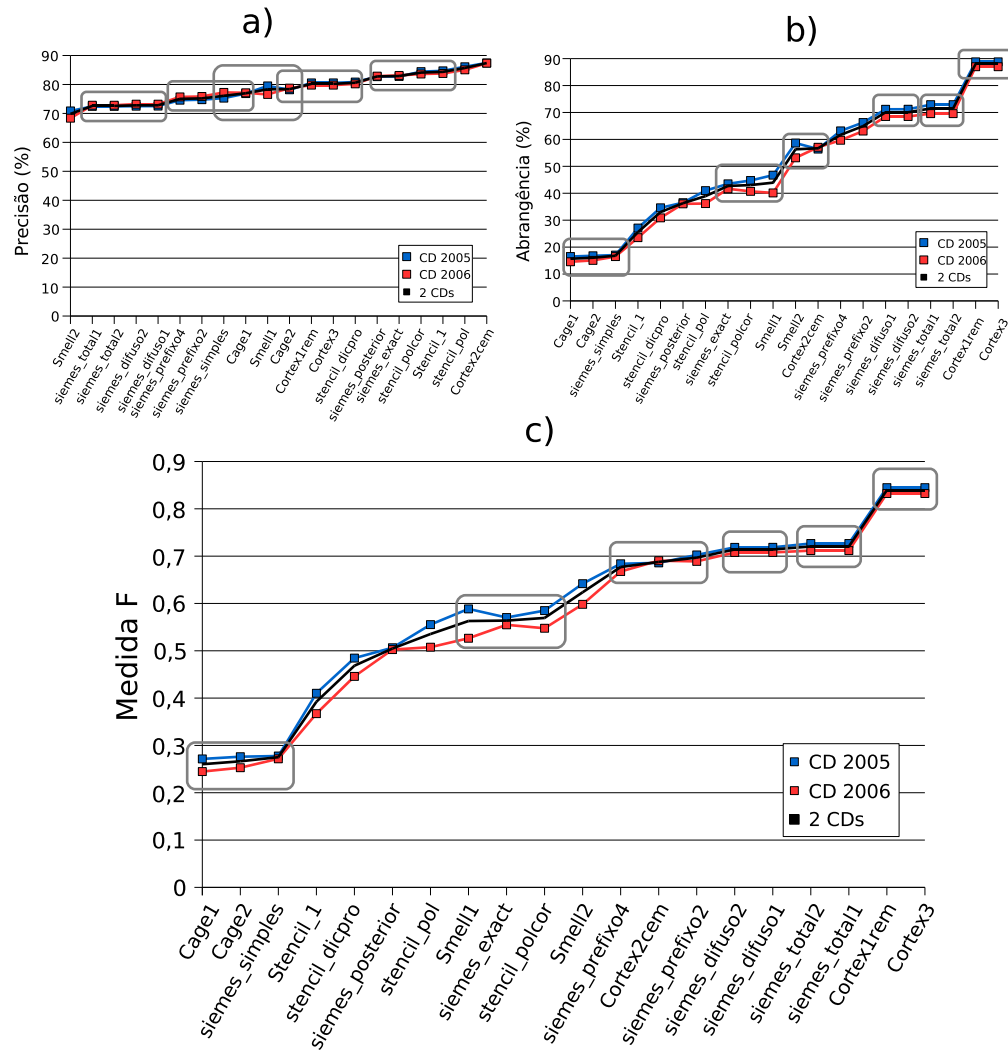


Figura 5.10: Desempenho dos sistemas para a tarefa de identificação no evento de 2006, para a **a)** precisão, **b)** abrangência e **c)** medida F.

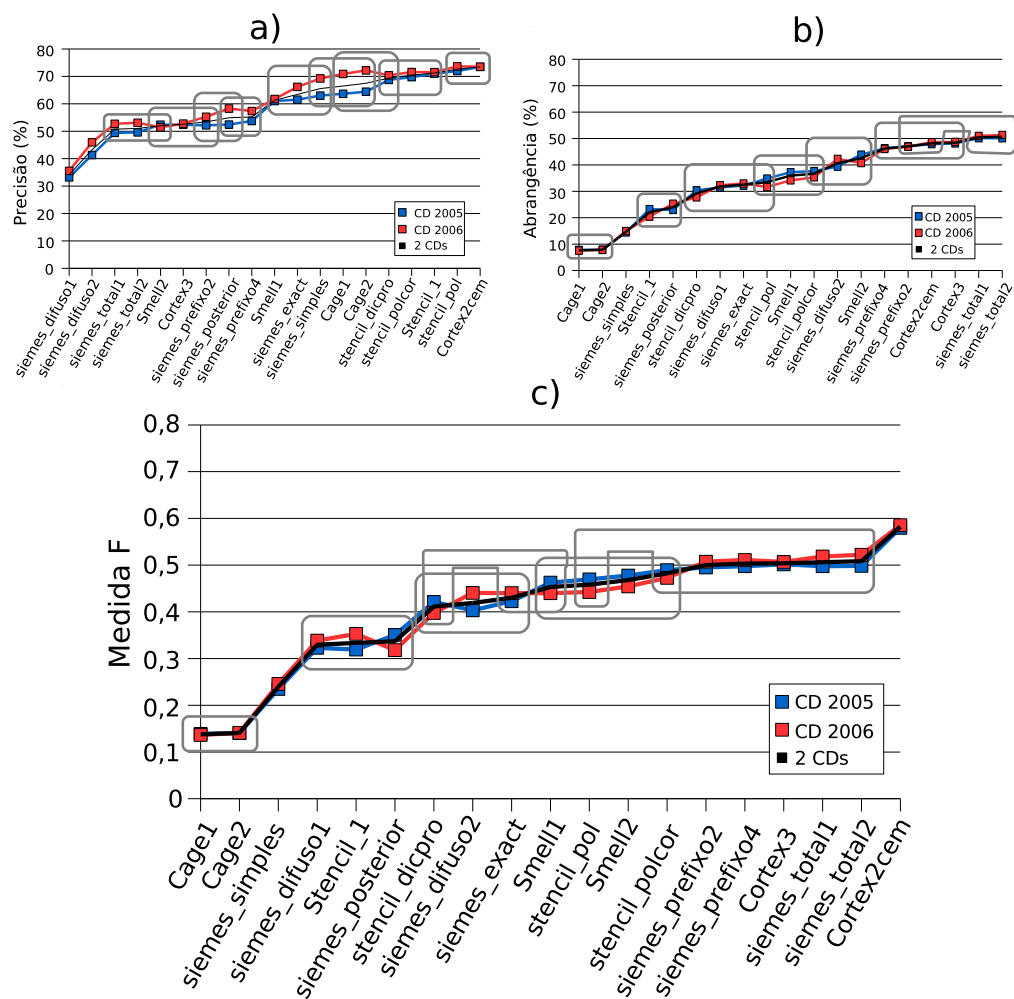


Figura 5.11: Desempenho dos sistemas para a tarefa de classificação semântica (na medida combinada) no evento de 2006, para a **a)** precisão, **b)** abrangência e **c)** medida F.

5.3 Resultados dos eventos HAREM

As Figuras 5.8, 5.9, 5.10 e 5.11 apresentam os resultados dos eventos de 2005 e de 2006, para as tarefas de identificação e de classificação semântica (na medida combinada). A classificação morfológica não foi objecto de análise estatística, porque não é uma etapa essencial de REM.

Pode-se observar nas figuras os desempenhos das saídas em relação à CD de 2005, à CD de 2006 e às duas CD. Contudo, a análise estatística realizada usou apenas os alinhamentos das saídas para as duas CD. O intervalo de confiança foi fixado em 99% ($\alpha = 0,01$), e foram geradas 9.999 re-amostragens para cada teste. Os valores p calculados podem ser consultados no Apêndice B. As saídas agrupadas em caixas nas Figuras apresentam valores p superiores a α entre si.

A análise estatística mostra que, em certos casos, a diferença entre amostras não é directamente proporcional à margem de erro, como afirmam Voorhees e Buckley (2002). Na Figura 5.11 pode-se observar esse comportamento excepcional no agrupamento das saídas em caixas.

Considere-se o seguinte caso de 3 saídas A , B e C com métricas de valor $M_A < M_B < M_C$. Se se observar que $p(M_A, M_C) \geq \alpha$ e que $p(M_B, M_C) \geq \alpha$, pode-se concluir que as saídas A e C são semelhantes, e que o mesmo se passa entre as saídas B e C . No entanto, não é possível afirmar que as saídas A e B são semelhantes a partir destes valores.

Este comportamento observado é consequência da *consistência* com que uma saída produz melhores resultados do que outra saída, sem no entanto se reflectir significativamente nas métricas usadas no teste. As saídas A e B podem representar os resultados de um sistema de REM, cuja saída B resulta de um ligeiro melhoramento ou da correcção de um erro. Assim, a probabilidade da saída B ser prejudicada com a permutação é igual a 0, e

Tarefa de identificação	2005	2006	Diferença
CaGE3 / CaGE2	0,178	0,266	49,4%
ELLE / SMELL2	0,523	0,624	19,2%
Siemes1 / Siemes Total	0,798	0,720	-9,7%
Nooj2no / Stencil_polcor	0,655	0,569	-13,0%
Cortex3no / Cortex3, Cortex1rem	0,746	0,839	12,6%

Tarefa de classificação semântica	2005	2006	Diferença
CaGE3 / CaGE2	0,101	0,141	39,8%
ELLE / SMELL2	0,438	0,468	6,9%
Siemes2 / Siemes Total	0,501	0,508	1,4%
Nooj2no / Stencil_polcor	0,369	0,482	30,7%
Cortex2 / Cortex2cem	0,395	0,582	47,5%

Tabela 5.5: Comparação dos valores de medida F para os eventos de 2005 e de 2006, nas tarefas de identificação e de classificação semântica (na medida combinada), para os sistemas participantes nos dois eventos.

consequentemente, a geração de re-amostragens nunca apresenta valores de diferenças entre A e B maiores, mesmo que o impacto da melhoria nas observações de B fosse ténue.

Assim sendo, a *consistência* com que B é melhor do que A faz com que o teste de permutação conclua que A e B são duas saídas estatisticamente diferentes, mesmo que a diferença entre M_A e M_B seja mínima.

5.3.1 Evolução dos sistemas entre eventos

A Tabela 5.5 apresenta os melhores resultados dos sistemas que participaram em ambos os eventos HAREM, para as tarefas de identificação e de classificação semântica (na medida combinada), em relação às duas CD. No global, os resultados dos participantes melhoraram o seu desempenho em 2006.

5.3.2 Panorama em REM

As Figuras 5.12 e 5.13 apresentam os resultados obtidos pelos melhores sistemas agrupados por categorias semânticas para os eventos de 2005 e 2006, respectivamente. Estas figuras mostram a dificuldade relativa das várias categorias semânticas para a tarefa de REM, com destaque para as novas categorias semânticas propostas pelo HAREM – COISA, OBRA, ACONTECIMENTO e ABSTRACCAO –, as mais difíceis de processar. Na Figura 5.13 observa-se que os melhores desempenhos do evento de 2006 são inferiores aos melhores desempenhos do evento de 2005, porque o melhor sistema de REM de 2005 na tarefa de classificação semântica não participou no evento de 2006.

As Figuras 5.14 e 5.15 apresentam os resultados obtidos pelos melhores sistemas agrupados por género textual, para os eventos de 2005 e de 2006, respectivamente. A dificuldade relativa de cada género varia segundo a CD usada em certos casos. Pode-se observar que os textos técnicos e os textos provenientes da *web* e de correio electrónico representam um desafio maior para os sistemas.

No caso dos textos da *web*, a dificuldade pode estar relacionada com a apresentação dos documentos sem etiquetas HTML, o que retira alguma informação sobre a segmentação dos textos, e prejudica o seu processamento automático. Os textos de mensagens de correio electrónico são ainda menos estruturados do que os textos da *web*, o que dificulta ainda mais a interpretação dos seus conteúdos. Por outro lado, os textos técnicos apresentam uma densidade baixa de EM, ao mesmo tempo que os assuntos abordados e o estilo de escrita tornam difícil a tarefa de REM neste tipo de textos.

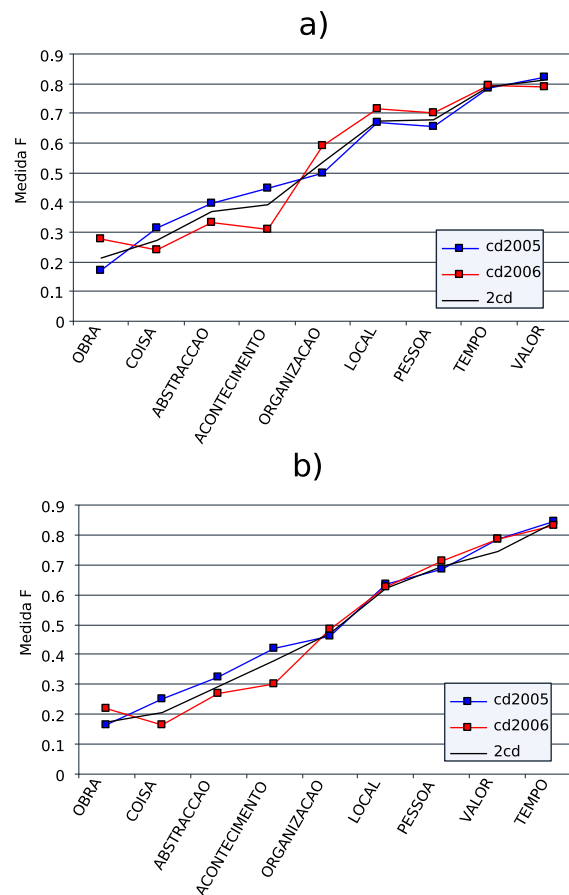


Figura 5.12: Medida F para os melhores sistemas na tarefa de classificação semântica (medida por categorias apenas) discriminada por categorias, para **a)** o evento de 2005, e para **b)** o evento de 2006.

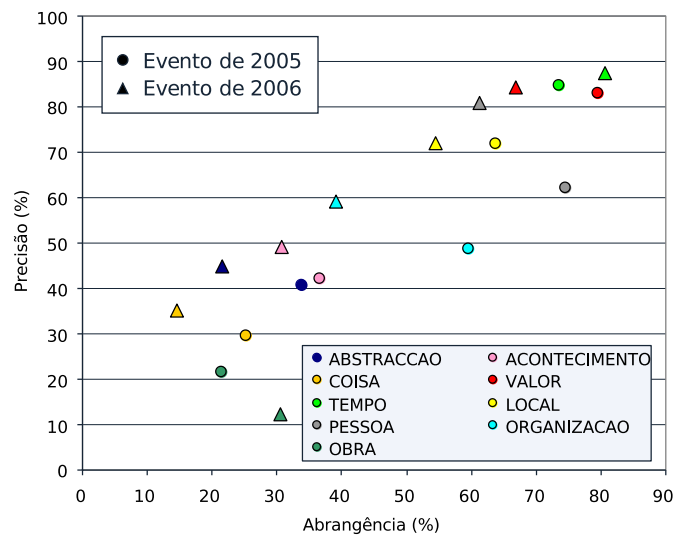


Figura 5.13: Precisão e abrangência para os melhores sistemas na tarefa de classificação semântica (medida por categorias apenas) discriminada por categorias, para ambos os eventos HAREM.

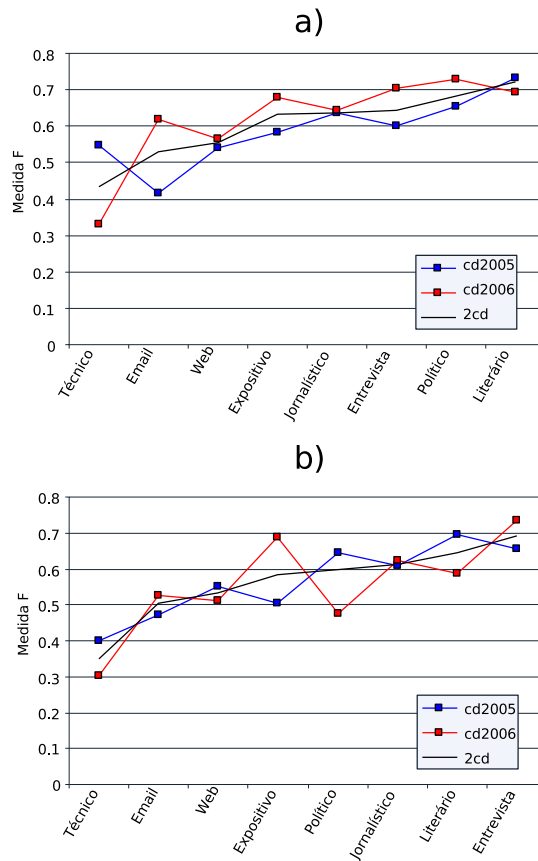


Figura 5.14: Medida F para os melhores sistemas na tarefa de classificação semântica (medida por categorias apenas) discriminada por gênero textual, para **a)** o evento de 2005, e para **b)** o evento de 2006.

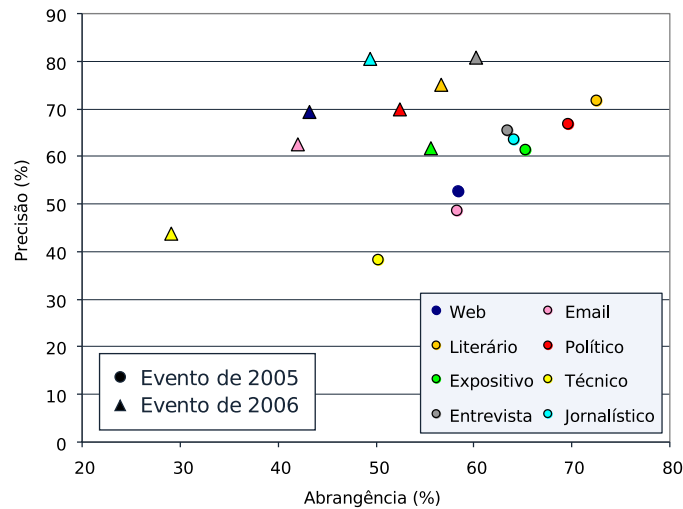


Figura 5.15: Precisão e abrangência para os melhores sistemas na tarefa de classificação semântica (medida por categorias apenas) discriminada por gênero textual, para ambos os eventos HAREM.

5.3.3 Comparação entre eventos de avaliação

A comparação dos resultados obtidos no HAREM em relação aos resultados de anteriores eventos de avaliação de sistemas de REM não tem significado, pelos seguintes motivos:

- O HAREM é o único evento de avaliação que usa textos em português. O REM é uma tarefa que depende da língua, e, como tal, a comparação entre eventos de avaliação só tem significado se estes forem realizados usando a mesma língua.
- O HAREM anota as CD segundo o significado semântico das EM no contexto, de uma maira mais aprofundada (Santos e Cardoso, 2006a).
- As directivas de etiquetagem e as categorizações adoptadas pelos diferentes eventos são consideravelmente diferentes. Por vezes, os critérios de definição de uma EM e os âmbitos semânticos das categorias usadas são divergentes (comparem-se, por exemplo, as directivas do HAREM (Cardoso e Santos, 2006) com as directivas do MUC (Chinchor e Robinson, 1998)).
- O HAREM usa uma CD com textos de diversas proveniências, enquanto que os eventos anteriores criaram as suas colecções a partir de corpora jornalísticos. O MUC usou textos provenientes de 58.000 artigos do *Wall Street Journal* de Janeiro de 1993 a Junho de 1994 (www ldc upenn edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T13), e o CoNLL 2003 usou o corpus da *Reuters* (www.reuters.com/researchandstandards/) na criação da sua colecção em inglês.

Apesar das diferenças significativas entre os eventos de avaliação, no resto desta Secção faz-se uma comparação das CD usadas (ver Tabela 5.6) e dos melhores resultados obtidos por cada evento de avaliação (ver Tabela 5.7).

Os resultados do HAREM são relativos ao sub-conjunto jornalístico das duas CD, o mesmo género que foi adoptado pelo MUC e pelo CoNLL. O cenário usado é selectivo em relação à categorização usada no MUC, e o *software* de avaliação aplica as directivas do MUC, ignorando as marcações de EM identificadas com `parcialmente_correcto`. Os resultados do HAREM referem-se ao melhor sistema na tarefa de classificação semântica, na medida por categorias apenas.

	HAREM		MUC		CoNLL
	2005	2006	MUC-6 (1995)	MUC-7 (1998)	2003
Nº documentos	62		30	100	231
Nº termos	19.503		≈ 14.000	≈ 40.000	46.435
PESSOA	472	469	373	887	1.617
ORGANIZACAO	333	331	447	1.880	1.661
LOCAL	277	227	110	1.324	1.668
TEMPO	77	75	112	1.481	-
VALOR	154	154	93	327	-
MISC	-	-	-	-	702
TOTAL	1.313	1.256	1.135	5.899	5.608

Tabela 5.6: Estatísticas das colecções douradas do HAREM, MUC e CoNLL.

Na Tabela 5.7 observa-se que o MUC-6 apresenta os melhores valores de medida F, e o CoNLL apresenta valores de medida F consideráveis. Sundheim (1995) refere que no MUC-6 metade dos seus participantes obteve valores de medida F acima de 90%. No entanto, Palmer e Day (1997) demonstraram que um sistema REM simples, com regras básicas e o auxílio de uma lista de EM extraída a partir do conjunto de textos de

	HAREM		MUC		CoNLL
	2005	2006	MUC-6 (1995)	MUC-7 (1998)	2003
PESSOA	0,7320	0,7184	0,9811	0,9966	0,9385
ORGANIZACAO	0,6158	0,5461	0,9253	0,9852	0,8467
LOCAL	0,6520	0,5304	0,9689	0,9920	0,9115
TEMPO	0,5839	0,8834	0,9867	0,9726	-
VALOR	0,9164	0,8939	1,000	0,9956	-
MISC	-	-	-	-	0,8044
TOTAL	0,7030	0,6815	0,9602	0,9856	0,8876

Tabela 5.7: Medida F dos melhores resultados observados no HAREM, MUC e CoNLL.

treino, é capaz de obter bons resultados na tarefa de REM proposta pelo MUC. Mikheev et al. (1999) mostram que o seu sistema REM que participou no MUC-7 conseguiu valores elevados de medida F usando regras simples e uma lista de EM de tamanho reduzido.

5.4 Sumário

A análise estatística usada para a validação do HAREM baseou-se nos testes de permutação, que foram adaptados aos requisitos impostos pelas tarefas de avaliação do HAREM. A análise realizada às saídas dos eventos HAREM mostra que o tamanho das CD permitiu comparar satisfatoriamente o desempenho das saídas e retirar conclusões acerca dos seus sistemas. A validação mostra que não há relação directa entre o tamanho da colecção e a margem de erro associada à comparação dos sistemas.

Contudo, é possível calcular o intervalo de confiança para comparações de saídas sobre CD de menor tamanho, usadas em cenários mais específicos. Um exemplo concreto é a comparação do desempenho de sistemas para as EM de categoria *LOCAL* em textos jornalísticos. No conjunto das duas CD, existem 279 EM que correspondem a este cenário. Apesar do

número reduzido de observações, é possível avaliar os sistemas e verificar se os resultados são significativos do ponto de vista estatístico.

A comparação do HAREM com eventos similares anteriores não é conclusiva, e a diferença observada deriva de vários factores como as directivas mais exigentes da tarefa do HAREM, a diferença de línguas usadas nas colecções, ou a comparação de resultados entre sistemas distintos, desenvolvidos para tarefas diferentes.

Capítulo 6

Breve análise aos sistemas de REM

6.1 Visão geral dos sistemas de REM

A avaliação de sistemas de REM conta com mais de uma década de actividade. Ao longo dos diversos eventos de avaliação, os grupos de investigação aplicaram várias estratégias para realizar as tarefas de avaliação. Contudo, o modelo típico de um sistema de REM contém um conjunto de *regras* de reconhecimento de expressões candidatas a EM, que são normalmente auxiliadas por um conjunto de listas de EM, denominados almanaques (ou *gazetteers*).

De um modo geral, as regras procuram detectar evidências internas das EM (como por exemplo a presença de maiúsculas ou de dígitos) e as evidências externas (como por exemplo o seu contexto) (McDonald, 1996). Os almanaques representam uma base de dados com várias instâncias de EM com significados semânticos desambiguados, contribuindo para a identificação e classificação de EM.

O tipo de regras usadas pelos sistemas como os dos participantes no HAREM varia consoante a estratégia adoptada. Vários grupos de investigação aplicam técnicas de aprendizagem automática, para desenvolver sistemas de REM independentes da língua. Estas técnicas não necessitam de regras complexas nem de almanaques extensos, mas requerem colecções de treino anotadas, para poderem aprender as suas próprias regras de detecção de EM e construir almanaques próprios.

Por outro lado, outros grupos de investigação aproveitam todo o conhecimento linguístico disponível para maximizar o desempenho dos seus sistemas numa determinada língua. Em geral, estes sistemas incluem várias regras gramaticais, introduzidas manualmente, para reconhecer as EM. O seu desempenho melhora com a utilização de mais regras, mas isso tem o inconveniente de os tornar mais complexos.

O tipo e o tamanho dos almanaques a usar nos sistemas de REM variam consoante a estratégia adoptada e o objectivo do sistema (Mikheev et al., 1999).

6.2 Descrição dos sistemas participantes

O HAREM contou com a participação de 10 grupos de investigação, cujos sistemas são a seguir descritos de forma breve.

PALAVRAS

Bick (2006) participou no HAREM com o PALAVRAS_NER, um sistema REM baseado no analisador morfossintáctico para o português, o PALAVRAS (Bick, 2003). O PALAVRAS_NER começou a ser desenvolvido em 2002, possuindo um conjunto considerável de regras e de restrições gramaticais, inseridas manualmente, para analisar as evidências internas e externas das EM.

O PALAVRAS_NER utiliza um dicionário com cerca de 70.000 lexemas, cerca de 6.000 regras gramaticais e aproximadamente 900 regras morfológicas. O almanaque do PALAVRAS_NER contém cerca de 17.000 entradas para ajudar a classificação das EM segundo a sua própria hierarquia composta por 6 categorias e 18 tipos, e que foi estendida para adoptar a categorização do HAREM (Bick, 2004).

SIEMÊS

Sarmento (2006) participou no HAREM com o sistema SIEMÊS. A estratégia do SIEMÊS procura evitar os inconvenientes do uso de regras e heurísticas de reconhecimento de EM, usando apenas cinco regras que procu-

ram semelhanças lexicais entre as EM no texto e as EM contidas no seu repositório de grandes dimensões, o REPENTINO (Sarmiento et al., 2006).

O REPENTINO contém cerca de 450.000 EM recolhidas através de diversas fontes, categorizadas numa hierarquia de dois níveis, com 11 categorias de topo e 102 subtipificações. As regras do SIEMÊS reconhecem as EM que não se encontram no REPENTINO, mas cuja estrutura é semelhante a outras EM. No HAREM, o SIEMÊS mostrou que a sua estratégia permite obter desempenhos significativos.

Cortex

O Cortex (www.cortex-intelligence.com) é um sistema de REM baseado em modelos cognitivos, linguísticos e estatísticos, desenvolvido por Christian Nunes Aranha na PUC-Rio. O Cortex usa técnicas de aprendizagem automática a partir de regras primitivas para reconhecer padrões conhecidos e sugerir novos padrões, mas também inclui módulos dependentes da língua e regras adicionais para adaptar o sistema à metodologia do HAREM.

O processo de aprendizagem não precisa de colecções de texto anotado, nem de gramáticas ou de almanaques. O Cortex usa um conjunto de léxicos no início, e aprende lexemas novos ou novos usos dos lexemas existentes à medida que vai processando textos novos.

A aprendizagem é realizada por vários módulos em paralelo que comunicam entre si, cada um com as suas próprias heurísticas. A eficiência do Cortex depende muito da quantidade de texto usada no processo de aprendizagem (o Cortex já processou cerca de 300 mil notícias nos seus três anos e meio de desenvolvimento).

RENA

O RENA foi um protótipo desenvolvido pelo projecto Natura da Universidade do Minho (natura.di.uminho.pt/natura/natura), que marca e extrai as EM a partir do texto segundo uma hierarquia de categorias configurável. O objectivo do RENA é a extracção de conhecimento a partir do texto, procurando identificar múltiplas referências à mesma entidade e fazendo a fusão com a informação já adquirida anteriormente.

O RENA usa a ferramenta *jspell* (de Almeida e Pinto, 1994) para auxílio na detecção de EM, e recorre a almanaques de carácter geral. Aplica ainda um conjunto de regras simples de desambiguação de nomes e de interpretação de contextos para a classificação das EM. O RENA usa uma categorização detalhada e específica (contém cerca de 100 categorias principais), e de vários níveis com herança múltipla para extrair informação rica a partir do texto. Para o HAREM, o RENA usou a sua própria categorização, a que aplicou no final um conversor para a categorização HAREM nas suas saídas.

CaGE

O CaGE foi desenvolvido pelo Grupo XLDB da Faculdade de Ciências da Universidade de Lisboa, para o motor de busca geográfico Geo-Tumba (local.tumba.pt, ver Silva et al., 2006). O CaGE é um módulo de prospecção de textos que reconhece referências geográficas no texto e atribui âmbitos geográficos a documentos recolhidos da *web* (Martins e Silva, 2005). Combina um conjunto de regras baseadas em padrões para identificar EM geográficas (ex: “cidade de X”), recorrendo a uma ontologia geográfica específica, a Geo-Net-PT01 (Chaves et al., 2005).

No HAREM, o CaGE focou-se na tarefa de identificação e de classificação de EM de categoria `LOCAL`, usando cenários selectivos restritos a EM com informação geográfica, tais como locais ou códigos postais.

NERUA

O NERUA é um sistema de REM desenvolvido pela Universidade de Alicante, que usa dois tipos de algoritmos: baseados em conhecimento, ou em aprendizagem automática. No caso dos algoritmos baseados em aprendizagem automática, o NERUA aplica as técnicas *Memory-based learner*, *Maximum Entropy* e *Hidden Markov Models* para realizar a tarefa de REM independentemente da língua usada nos textos (Ferrández et al., 2005; Kozareva et al., 2005). O NERUA usa uma estratégia de votação pesada entre os algoritmos, para combinar as vantagens de cada estratégia e maximizar os resultados (Kozareva et al.).

O NERUA teve início em Outubro de 2004 e foi desenhado inicialmente para a língua espanhola, mas a estratégia adoptada pelo NERUA e a proximidade que existe entre a língua espanhola e portuguesa permitiram a participação do NERUA no HAREM. A colecção de treino usada incluiu a colecção espanhola do CoNLL 2002 e o pedaço da CD enviada aos participantes para anotação.

O NERUA contou ainda com a ajuda inicial de ferramentas gramáticas dependentes da língua, e de pequenos almanaques criados a partir de informação da *web* para auxiliar a tarefa de classificação semântica.

Stencil-NooJ

O Stencil/NooJ é um conjunto de recursos linguísticos desenvolvidos por Mota (2006) e aplicados no sistema NooJ, uma plataforma genérica de

processamento de linguagem natural (www.nooj4nlp.net). O objectivo do Stencil/NooJ é anotar EM em corpora jornalísticos de forma semi-automática.

Esses recursos, criados manualmente, são constituídos por dicionários de palavras-chave e por gramáticas locais. Estas utilizam as informações dos dicionários para descrever, por um lado, a estrutura do nome próprio, e, por outro, contextos locais onde os nomes próprios podem ocorrer. O Stencil não utiliza nenhum almanaque na tarefa de REM.

Malinche

Solorio (2005) participou no HAREM com o seu sistema, o Malinche. O Malinche usa uma estratégia de aprendizagem automática para realizar a tarefa de REM, independente da língua usada.

O Malinche usa uma versão modificada do algoritmo C4.5 para aprender regras de identificação a partir de uma colecção de treino (**Quinlan, 1993**), dispensando a codificação manual de regras e a manutenção de grandes almanaques de EM.

Elle

O ELLE é um sistema de REM desenvolvido pelo LaBEL (label.ist.utl.pt), especificamente para participar no HAREM (**Marcelino, 2005**). O ELLE baseia-se em gramáticas e dicionários construídos sobre o sistema INTEX (**Silberztein, 1993**).

O ELLE contém uma gramática específica para cada categoria da categorização HAREM. Estas gramáticas são usadas em paralelo, competindo entre si pelo reconhecimento das EM.

NILC

O ReGra começou a ser desenvolvido pelo NILC em 1993 com o objectivo de analisar a sintaxe das frases, verificando automaticamente a gramática (Martins e Nunes, 2006). O NILC participou nas Morfolimpíadas e no HAREM com o seu módulo de reconhecimento lexical do ReGra, que faz a etiquetagem morfossintáctica dos textos, mas não a sua classificação semântica.

Como o módulo não foi desenvolvido especificamente para a tarefa de REM, a identificação das EM é realizada demarcando as fronteiras de unidades lexicais que não constem no dicionário que serve a ferramenta, com mais de 500.000 léxicos. As sequências de unidades lexicais marcadas e de palavras iniciadas por maiúsculas são agrupadas, e o analisador sintáctico classifica as marcações na sua morfologia (Hasegawa et al., 2002).

6.3 Sumário

Ao longo dos vários eventos de avaliação em REM, não houve nenhuma estratégia de abordagem ao problema de REM que se revelasse superior às restantes. No entanto, observa-se que os melhores sistemas adoptam estratégias por vezes consideravelmente diferentes entre eles, em relação ao tamanho de almanaques usados, ou no tipo e número de regras de reconhecimento de EM, entre outros.

Nos eventos HAREM também se verificou que os melhores sistemas participantes usaram diferentes aproximações para a tarefa de REM em português, o que não permite retirar conclusões sobre as melhores estratégias a aplicar em sistemas de REM. No entanto, é necessário que sejam organizados mais eventos de avaliação em REM no futuro, pois sente-se que

os sistemas de REM ainda possuem uma ampla margem de progressão, e eventos de avaliação como o HAREM contribuem para o melhoramento da eficácia dos sistemas.

Capítulo 7

Conclusões e trabalho futuro

7.1 Conclusões

O REM é uma tarefa desempenhada por vários sistemas inteligentes no domínio do processamento da língua. Na última década houve várias conferências de avaliação em PLN que destacaram a importância do REM, ao incluir nos seus eventos tarefas específicas para a sua avaliação.

O HAREM representa a primeira avaliação de REM em textos portugueses, e apresentou uma nova metodologia de avaliação de REM desenvolvida especificamente para a avaliação da tarefa, em conjunto com a comunidade científica. Nos dois eventos organizados em 2005 e 2006, o HAREM mediu e comparou o desempenho dos sistemas participantes, permitindo caracterizar o estado da arte em sistemas inteligentes de REM em português. A validação estatística realizada sobre os resultados mostra que a metodologia usada nos eventos de avaliação e as colecções de textos usadas são adequadas para comparar sistemas inteligentes em REM.

Os sistemas desenvolvidos pelos participantes recorrem a diversas estratégias, que vão desde a aprendizagem automática até à codificação manual das regras de reconhecimento de EM. Não foi possível, no entanto, retirar conclusões sobre as melhores estratégias para REM, segundo os resultados obtidos nos eventos.

Esta tese faz parte do processo de documentação do HAREM. O trabalho desenvolvido nesta tese produziu recursos de valor, como uma colecção de textos ricamente anotada, uma plataforma de avaliação específica para comparação de sistemas, e *software* de análise estatística para medição da significância das comparações entre saídas.

7.2 Futuro do HAREM

No dia 15 de Julho de 2006 decorreu o *Encontro HAREM*, um *workshop* que contou com a presença dos participantes e organizadores do HAREM. O encontro encerrou com um debate sobre o futuro do HAREM, onde os participantes apresentaram algumas propostas concretas para novos eventos do HAREM (www.linguateca.pt/HAREM/harem.php?l=encontroharem). Os participantes foram convidados a escrever um artigo sobre a sua participação e sobre considerações sobre o futuro do HAREM, para uma futura publicação sobre o HAREM (www.linguateca.pt/HAREM/DocActas/).

Mota e Sarmento (2006) sugerem, para uma futuro evento de avaliação do HAREM, a criação de uma *tarefa de identificação robusta* sobre uma CD com letras convertidas em minúsculas, para eliminar parte da evidência interna de uma EM, bem como uma revisão à metodologia de atribuição de categorias semânticas.

7.3 Trabalho futuro

A avaliação de sistemas inteligentes na área de PLN tem como principal objectivo servir a comunidade científica. Como tal, os eventos de avaliação devem acompanhar os interesses da comunidade ao longo do tempo, ao mesmo tempo que permitem a avaliação periódica dos seus sistemas, fornecendo dados úteis sobre as estratégias adoptadas.

O HAREM contou com a participação de sistemas de REM com objectivos e características diferentes, que utilizaram diversas estratégias. Os resultados obtidos ainda não permitem tecer conclusões sobre as abordagens usadas, sendo necessário organizar mais eventos de avaliação em

REM para acompanhar o desenvolvimento dos sistemas, e seleccionar as melhores estratégias para a tarefa de REM.

Assim sendo, os futuros eventos de HAREM deverão apresentar novas tarefas de avaliação desenvolvidos em conjunto com a comunidade científica, respondendo às necessidades comunidade. Ao mesmo tempo, os eventos deverão rever as suas metodologias de avaliação, permitindo uma comparação mais eficaz entre saídas de sistemas. Um exemplo seria aproveitar o facto de as EM apresentarem diferentes níveis de dificuldade nas tarefas propostas, o que pode permitir distinguir os melhores sistemas de REM de uma maneira mais eficaz.

Apêndice A

Acrónimos e abreviaturas

ACE: *Automatic Content Extraction.*

ANCIB: Associação Nacional de Pesquisa e Pós-graduação em Ciência da Informação.

ATIS: *Air Travel Information Systems.*

AVALON: Encontro de avaliação conjunta em sistemas de processamento computacional do português, organizado pela Linguatca.

CD: Colecção Dourada.

CENTEMPúblico: Corpus de Extractos de Textos Electrónicos MCT/Público.

CENTENFolha: Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo.

CH: Colecção HAREM.

CLEF: *Cross Language Evaluation Forum.*

CONE: Corpus de Correio Não-Endereçado.

CoNLL: *Conference on Computational Natural Language Learning.*

ECI-EBR: *European Corpus Initiative - ECI Borba-Ramsey.*

EDT: *Entity Detection and Tracking.*

EI: *Extracção de Informação.*

EM: *Entidade Mencionada.*

HAREM: *HAREM - Avaliação de sistemas de Reconhecimento de Entidades Mencionadas.*

IA: *Inteligência Artificial*

MET: *Multilingual Entity Task.*

MUC: *Message Understanding Conference.*

NTCIR: *NII-NACSIS Test Collection for Information Retrieval systems.*

Parseval: *Parse Evaluation.*

PLN: *Processamento de Linguagem Natural.*

PROPOR: *Conferência de Processamento Computacional da Língua Portuguesa Escrita e Falada.*

REM: *Reconhecimento de Entidades Mencionadas.*

RI: *Recuperação de Informação.*

TREC: *Text REtrieval Conference.*

WPT 03: *Recolha da web portuguesa de 2003.*

WBR-99: *Recolha da web brasileira de 1999.*

Apêndice B

Tabelas de valores p

	palavras	siemes1	siemes2	rena	noo2no	nerua_em	nerua_cp	nerua_ct	malinche	elle	cortex2	cortex1	cortex3no	nilc	noo1	cage3	cage2	cage1
Prec.	79,18%	0,0010	0,0010	0,3150	0,0400	0,0010	0,0010	0,0010	0,0010	0,4670	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Abr.	83,31%	0,6510	0,0230	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,8119	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Prec.	76,75%		0,0130	0,0010	0,0180	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Abr.	83,11%		0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,7981		0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Prec.	76,39%			0,0010	0,0030	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Abr.	82,29%			0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,7923			0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Prec.	78,76%				0,2060	0,0010	0,0010	0,0010	0,0010	0,9840	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Abr.	68,17%				0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,7308				0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Prec.	78,03%					0,0010	0,0010	0,0010	0,0010	0,3190	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Abr.	56,42%					0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0040	0,0010	0,0010	0,0010	0,0010
MedF	0,6549					0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Prec.	57,47%						0,0010	0,0060	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Abr.	62,68%						1,0000	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,5996						0,0010	0,0010	0,0230	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Prec.	59,70%							0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Abr.	62,68%							0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,6115							0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Prec.	57,67%								0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Abr.	61,87%								0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,5970								0,0690	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Prec.	49,68%									0,0010	0,0010	0,0010	0,0010	0,0090	0,0010	0,8440	0,0500	0,0640
Abr.	71,95%									0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,5878									0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Prec.	78,74%										0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
Abr.	39,16%										0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,5231										0,4920	0,4170	0,0010	0,0820	0,0010	0,0010	0,0010	0,0010
Prec.	64,64%											0,0240	0,0010	0,0010	0,2570	0,0010	0,0010	0,0010
Abr.	43,33%											0,0180	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,5188											0,0190	0,0010	0,3540	0,0010	0,0010	0,0010	0,0010
Prec.	64,57%												0,0010	0,0010	0,2960	0,0010	0,0010	0,0010
Abr.	43,22%												0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,5178												0,0010	0,4280	0,0010	0,0010	0,0010	0,0010
Prec.	66,37%													0,0010	0,0010	0,0010	0,0010	0,0010
Abr.	85,21%													0,0010	0,0010	0,0010	0,0010	0,0010
MedF	0,7462													0,0010	0,0010	0,0010	0,0010	0,0010
Prec.	48,52%													0,0010	0,0010	0,0010	0,0010	0,0010
Abr.	54,35%														0,0010	0,0010	0,0010	0,0010
MedF	0,5127														0,0010	0,0010	0,0010	0,0010
Prec.	63,54%														0,0010	0,1110	0,5950	0,6650
Abr.	14,96%														0,0010	0,0010	0,0010	0,0010
MedF	0,2421														0,0010	0,0010	0,0010	0,0010
Prec.	49,84%															0,0230	0,0320	0,0320
Abr.	10,86%															0,0010	0,0010	0,0010
MedF	0,1783															0,0010	0,0010	0,0010
Prec.	48,05%																0,5140	0,5140
Abr.	7,09%																0,3740	0,3740
MedF	0,1236																	0,3740
Prec.	48,14%																	
Abr.	7,12%																	
MedF	0,1240																	

Tabela B.1: Valores p para a tarefa de identificação do evento de 2005.

		palavras	siemes1	siemes2	rena	noj2no	nerua_cp	nerua_ct	elle	cortex2	cortex1	noj1	cage3	cage2	cage1
palavras	Prec.	58,22%													
	Abr.	58,68%	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,5845	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes1	Prec.	49,23%													
	Abr.	45,96%		0,0219	0,0001	0,0001	0,0001	0,0001	0,0001	0,8348	0,8252	0,0714	0,0001	0,0001	0,0001
	MedF	0,4754		0,0005	0,0001	0,0001	0,0001	0,0001	0,0004	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes2	Prec.	50,21%													
	Abr.	46,96%								0,2269	0,1031	0,4985	0,0001	0,0001	0,0001
	MedF	0,4853								0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
rena	Prec.	31,83%													
	Abr.	18,21%					0,9108	0,0001	0,0001	0,0001	0,0001	0,0001	0,0186	0,0001	0,0001
	MedF	0,2316				0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
noj2no	Prec.	44,01%													
	Abr.	30,47%						0,0001	0,0371	0,0812	0,1155	0,0001	0,0001	0,0001	0,0001
	MedF	0,3601					0,0001	0,0001	0,0001	0,0004	0,0012	0,0001	0,0001	0,0001	0,0001
nerua_cp	Prec.	31,90%													
	Abr.	22,21%						0,0001	0,0001	0,0001	0,0001	0,0001	0,0123	0,0001	0,0001
	MedF	0,2619						0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
nerua_ct	Prec.	27,19%													
	Abr.	28,67%							0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,2791							0,0005	0,0010	0,0018	0,0001	0,0001	0,0001	0,0001
elle	Prec.	64,18%													
	Abr.	32,59%								0,7680	0,6374	0,0001	0,0001	0,0001	0,0001
	MedF	0,4323								0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
cortex2	Prec.	49,37%													
	Abr.	32,28%									0,0001	0,1108	0,0001	0,0001	0,0001
	MedF	0,3904									0,8941	0,0001	0,0001	0,0001	0,0001
cortex1	Prec.	49,08%													
	Abr.	32,08%										0,0586	0,0001	0,0001	0,0001
	MedF	0,3880										0,0001	0,0001	0,0001	0,0001
noj1	Prec.	50,82%													
	Abr.	10,92%													
	MedF	0,1797													
cage3	Prec.	34,04%													
	Abr.	5,81%													
	MedF	0,0993												0,0006	0,0008
cage2	Prec.	37,94%													
	Abr.	4,37%													
	MedF	0,0784													
cage1	Prec.	38,04%													
	Abr.	4,39%													
	MedF	0,0787													

Tabela B.2: Valores p para a tarefa de classificação semântica (na medida combinada) do evento de 2005.

	cortex3	cortex1re m	siemes_t cia2	siemes_t cia1	siemes_d fus2	siemes_d fus1	cortex2ce m	siemes_pr efixo2	siemes_pr efixo4	small2	siemes_e xact	stencil_p dcor	small1	stencil_p ol	siemes_p osterior	stencil_d icpro	stencil_1	siemes_si mples	case2	case1
cortex3	Prec. 80.23% Abr. 88.04% MedF 0.8395	1.0000 1.0000 1.0000	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0022 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.5792 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0138 0.0001 0.0001	0.0001 0.0001 0.0001
cortex1rem	Prec. 80.23% Abr. 88.04% MedF 0.8395		0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0012 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.5853 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0143 0.0001 0.0001	0.0002 0.0001 0.0001
siemes_total2	Prec. 72.72% Abr. 71.62% MedF 0.7216		1.0000 1.0000 1.0000	0.0001 0.0001 0.0001	0.0203 0.0001 0.0001	0.0115 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001
siemes_total1	Prec. 72.72% Abr. 71.62% MedF 0.7216				0.0213 0.0001 0.0001	0.0155 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001
Siemes_diffuso2	Prec. 72.91% Abr. 70.18% MedF 0.7152					0.4991 0.4991 0.4991	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001
siemes_diffuso1	Prec. 72.93% Abr. 70.18% MedF 0.7152						0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001
cortex2cam	Prec. 87.39% Abr. 56.69% MedF 0.6877						0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.5324 0.0817	0.0001 0.0001 0.0001 0.0001	0.0001 0.0001 0.0001 0.0001	0.0001 0.0001 0.0001 0.0001	0.0023 0.0001 0.0001 0.0001	0.0001 0.0001 0.0001 0.0001	0.0001 0.0001 0.0001 0.0001	0.0001 0.0001 0.0001 0.0001	0.0001 0.0001 0.0001 0.0001	0.0001 0.0001 0.0001 0.0001	0.0001 0.0001 0.0001 0.0001
Siemes_prefixo2	Prec. 75.33% Abr. 65.04% MedF 0.6981								0.1191 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0670 0.0001 0.0001	0.0003 0.0001 0.0001	0.0587 0.0001 0.0001
Siemes_prefixo4	Prec. 75.16% Abr. 61.81% MedF 0.6783									0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0460 0.0001 0.0001	0.0001 0.0001 0.0001	0.0398 0.0001 0.0001
small2	Prec. 69.84% Abr. 56.34% MedF 0.6236									0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001
Siemes_exact	Prec. 83.14% Abr. 42.83% MedF 0.9654										0.1740 0.7337 0.9051	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0002 0.0001 0.0001	0.8510 0.0001 0.0001	0.0012 0.0001 0.0001	0.1443 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001
stencil_polcor	Prec. 84.10% Abr. 43.04% MedF 0.5694												0.1342 0.0001 0.2708	0.0001 0.0001 0.0001	0.1360 0.0001 0.0001	0.0001 0.0001 0.0001	0.5468 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001
small1	Prec. 78.33% Abr. 43.92% MedF 0.5629													0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0028 0.0001 0.0001	0.0001 0.0001 0.0001	0.9739 0.0001 0.0001	0.0677 0.0001 0.0001	
stencil_pol	Prec. 85.69% Abr. 38.94% MedF 0.5355														0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001
Siemes_posterior	Prec. 83.04% Abr. 36.43% MedF 0.5064															0.0019 0.0001 0.0001	0.1097 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001
stencil_dicpro	Prec. 80.59% Abr. 33.02% MedF 0.4684															0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0001 0.0001 0.0001	0.0211 0.0001 0.0001	0.0001 0.0001 0.0001
stencil_1	Prec. 84.36% Abr. 25.56% MedF 0.3923																		0.1578 0.0370 0.1573	0.7947 0.0370 0.1573
siemes_simples	Prec. 76.62% Abr. 16.91% MedF 0.2771																		0.1578 0.0370 0.1573	0.7947 0.0370 0.1573
case2	Prec. 78.39% Abr. 16.05% MedF 0.2664																		0.1578 0.0370 0.1573	0.7947 0.0370 0.1573
case1	Prec. 76.96% Abr. 15.66% MedF 0.2603																		0.1578 0.0370 0.1573	0.7947 0.0370 0.1573

Tabela B.3: Valores p para a tarefa de identificação do evento de 2006.

		cortex3	siemes_t otai2	siemes_t otai1	siemes_d ifuso2	siemes_d fuso1	cortex2ce m	siemes_p refixo2	siemes_pr efixo4	small2	siemes_e xact	stencil_po lcor	small1	stencil_p ol	siemes_p osterior	stencil_d cpo	siemes_si mples	cage2	cage1
cortex3	Prec.	52.56%	0.0086	0.0014	0.0001	0.0001	0.0001	0.1462	0.0001	0.3174	0.0001	0.0001	0.0001	0.0001	0.0001	0.0016	0.0001	0.0001	0.0001
	Abr.	45.63%	0.0010	0.0017	0.0001	0.0001	0.2465	0.9995	0.6016	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	MedF	0.4885	0.1826	0.2973	0.0001	0.0001	0.0001	0.6250	0.3596	0.0009	0.0001	0.4248	0.0005	0.0193	0.0001	0.0001	0.0001	0.0001	0.0001
siemes_total2	Prec.	51.03%	0.0029	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.1605	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	Abr.	48.53%	0.8821	0.0001	0.0001	0.0001	0.3638	0.0130	0.0065	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	MedF	0.4975	0.6915	0.0001	0.0001	0.0001	0.0001	0.4614	0.8512	0.0001	0.0001	0.0694	0.0001	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001
siemes_diffuso2	Prec.	50.78%			0.0001	0.0001	0.0001	0.0736	0.0001	0.0736	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	Abr.	48.38%			0.0001	0.0001	0.4229	0.0198	0.0088	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	MedF	0.4955			0.0001	0.0001	0.0001	0.6506	0.9470	0.0001	0.0001	0.0986	0.0001	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001
siemes_diffuso1	Prec.	43.25%			0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	Abr.	38.97%			0.0001	0.0001	0.0001	0.0001	0.0001	0.0278	0.0001	0.0296	0.0041	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	MedF	0.4100			0.0001	0.0001	0.0001	0.1834	0.0001	0.0001	0.1834	0.0001	0.0001	0.0001	0.0001	0.9309	0.0001	0.0001	0.0001
siemes_diffuso1	Prec.	34.16%				0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	Abr.	30.53%			0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.3063	0.0001	0.0003	0.0360	0.0001	0.2330	0.0001	0.0001	0.0001
	MedF	0.3224			0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.2566	0.3907	0.0001	0.0001	0.0001	0.0001
cortex2cem	Prec.	73.49%					0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.2319	0.0001	0.0001	0.0001	0.0001	0.0001
	Abr.	47.37%					47.37%	0.0131	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	MedF	0.5761					0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
siemes_prefixo2	Prec.	53.43%					0.0001	0.0001	0.0001	0.0294	0.0001	0.0001	0.0001	0.0001	0.0255	0.0001	0.0001	0.0001	0.0001
	Abr.	45.63%					0.0001	0.6309	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	MedF	0.4923					0.0001	0.6203	0.0002	0.0001	0.0001	0.1704	0.0002	0.0011	0.0001	0.0001	0.0001	0.0001	0.0001
siemes_prefixo4	Prec.	55.22%					0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.6679	0.0001	0.0001	0.0001	0.0001
	Abr.	45.02%					0.0001	45.02%	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	MedF	0.4960					0.0001	0.0001	0.0001	0.0001	0.0001	0.0597	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001
small2	Prec.	51.94%					0.0001	0.0001	0.0001		0.0001	0.0001	0.0001	0.0001	0.0005	0.0001	0.0001	0.0001	0.0001
	Abr.	41.05%					0.0001		0.0001		0.0001	0.0003	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	MedF	0.4586					0.0001		0.0001		0.0714	0.0001	0.2937	0.7548	0.0001	0.0001	0.0001	0.0001	0.0001
siemes_exact	Prec.	63.45%										0.0001	0.0012	0.3709	0.0001	0.0041	0.0169	0.0001	0.00076
	Abr.	32.15%										0.0003	0.0030	0.3709	0.0001	0.0041	0.0001	0.0001	0.0001
	MedF	0.4267										0.0001	0.0371	0.0091	0.0001	0.1046	0.0001	0.0001	0.0001
stencil_polcor	Prec.	70.49%								0.0001	0.0001	0.0001	0.0001	0.0552	0.0001	0.1873	0.0001	0.0026	0.0001
	Abr.	36.12%											0.3843	0.0928	0.0001	0.0001	0.0001	0.0001	0.0001
	MedF	0.4776											0.0011	0.1602	0.0001	0.0001	0.0001	0.0001	0.0001
small1	Prec.	61.21%													0.0001	0.0001	0.0001	0.0001	0.0001
	Abr.	35.22%													0.0814	0.0001	0.0001	0.0001	0.0001
	MedF	0.4471													0.4763	0.0001	0.0001	0.0001	0.0001
stencil_pol	Prec.	72.65%													0.0001	0.0045	0.0001	0.0001	0.0001
	Abr.	33.12%													0.0001	0.0028	0.0001	0.0001	0.0001
	MedF	0.4550													0.0001	0.0013	0.0001	0.0001	0.0001
siemes_posterior	Prec.	54.92%															0.0001	0.0001	0.0001
	Abr.	23.72%															0.0001	0.0001	0.0001
	MedF	0.3313															0.0001	0.0001	0.0001
stencil_dicpro	Prec.	69.37%															0.1133	0.1148	0.0001
	Abr.	29.00%															0.0001	0.0001	0.0001
	MedF	0.4090															0.0001	0.0001	0.0001
stencil_1	Prec.	71.23%															0.0001	0.0001	0.0001
	Abr.	21.94%															0.0001	0.0001	0.0001
	MedF	0.3354															0.0001	0.0001	0.0001
siemes_simples	Prec.	65.56%																0.1986	0.5485
	Abr.	14.57%																0.0001	0.0001
	MedF	0.2385																0.0001	0.0001
cage2	Prec.	67.50%																	0.0001
	Abr.	7.81%																	0.0003
	MedF	0.1401																	0.7895
cage1	Prec.	66.49%																	0.7895
	Abr.	7.65%																	0.7895
	MedF	0.1373																	

Tabela B.4: Valores p para a tarefa de classificação semântica (na medida combinada) do evento de 2006.

Bibliografia

Eckhard Bick. “Multi-level NER for Portuguese in a CG Framework”. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso e Maria das Graças Volpe Nunes, editores, *Proceedings of the 6th Computational Processing of the Portuguese Language, PROPOR 2003*, volume 2721 de *Lecture Notes in Computer Science*, págs. 118–125, Faro, Portugal, 26–27 Junho 2003. Springer. 91

Eckhard Bick. “A Named Entity Recognizer for Danish”. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, págs. 305–308, Lisboa, Portugal, 26–28 Maio 2004. ELRA. 91

Eckhard Bick. “Functional Aspects in Portuguese NER”. Em Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Claudia Oliveira e Maria Carmelita Dias, editores, *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language, PROPOR 2006*, volume 3960 de *Lecture Notes in Computer Science*, págs. 80–89, Itatiaia, Brasil, 13–17 Maio 2006. Springer. 91

Ezra Black, Steven Abney, Dan Flickenger, Claudia Gdaniec, Ralph Grishman, Philip Harrison, Donald Hindle, Robert Ingria, Fred Jelinek, Ju-

- dith Klavans, Mark Liberman e Tomek Strzalkowski. "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars". Em Patti Price, editor, *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, págs. 306–311, Pacific Grove, Califórnia, EUA, 19–22 Fevereiro 1991. Morgan Kaufmann. 13, 17
- Chris Buckley e Ellen Voorhees. "Evaluating Evaluation Measure Stability". Em *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2000*, págs. 33–40, Atenas, Grécia, 24–28 Julho 2000. ACM Press. 32
- Pável Calado. "The WBR-99 Collection: Description of the WBR-99 Web Collection Data-Structures and File Format". Relatório técnico, LATIN - Laboratório para o Tratamento de Informação, Departamento de Computação, Universidade Federal de Minas Gerais, Brasil, 1999. 40
- Nicoletta Calzolari e Ornella Corazzari. "Senseval/Romanseva: The Framework for Italian". *Computers and the Humanities*, 34(1–2):61–78, Abril 2000. 43
- Nuno Cardoso, Bruno Martins, Daniel Gomes e Mário J. Silva. "WPT 03: a primeira colecção pública proveniente de uma recolha da web portuguesa". Em Diana Santos, editor, *Avaliação Conjunta: um novo paradigma no processamento computacional da língua portuguesa*, capítulo 23. 2006a. No prelo. ISTPress. 40
- Nuno Cardoso e Diana Santos. "Directivas e categorias para identificação e classificação semântica na colecção dourada do HAREM". Relatório Técnico DI-FCUL TR 06–18, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, Outubro 2006. 39, 85

Nuno Cardoso, Diana Santos e Nuno Seco. “Avaliação no HAREM: Métodos e medidas”. Relatório Técnico DI-FCUL TR 06–20, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, Outubro 2006b. 51

Nuno Cardoso, Diana Santos e Rui Vilela. “Directivas para identificação e classificação morfológica na colecção dourada do HAREM”. Relatório Técnico DI-FCUL TR 06–19, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, Outubro 2006c. 39, 52

Marcirio Silveira Chaves, Mário J. Silva e Bruno Martins. “GKB - Geographic Knowledge Base”. Relatório Técnico DI-FCUL TR 05–12, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, Junho 2005. 93

Nancy Chinchor. “The Statistical Significance of MUC-4 Results”. Em *Proceedings of the 4th Conference on Message Understanding, MUC-4*, págs. 30–50, McLean, Virgínia, EUA, 16–18 Junho 1992. Association for Computational Linguistics. 28

Nancy Chinchor, Lynette Hirschman e David D. Lewis. “Evaluating Message Understanding Systems: An Analysis of the 3rd Message Understanding Conference”. *Computational Linguistics*, 19(3):409–449, 1993. 28

Nancy Chinchor e Patty Robinson. “MUC-7 Named Entity Task Definition (versão 3.5)”. Em *Proceedings of the 7th Conference on Message Understanding, MUC-7*, Washington, D.C., EUA, Abril 1998. 15, 85

Charles Clarke, Nick Craswell e Ian Soboroff. “Overview of the TREC 2004 Terabyte Track”. Em Ellen. M. Voorhees e Lori P. Buckland, editores,

- Proceedings of the 13th Text REtrieval Conference*, Gaithersburg, Maryland, EUA, 16–19 Novembro 2004. NIST. 14
- Cyril W. Cleverdon. “The Cranfield Tests on Index Language Devices”. *Aslib Proceedings*, 19(6):173–193, 1967. 12
- Cyril W. Cleverdon. “Optimizing Convenient Online Access to Bibliographic Databases”. *Information Services and Use*, 4:37–47, 1984. 43
- Paul R. Cohen. *Empirical Methods for Artificial Intelligence*. 1995. MIT Press. 33
- J. J. Dias de Almeida e Ulisses Pinto. “Jspell - um módulo para análise léxica genérica de linguagem natural”. Em *Actas do X Encontro da Associação Portuguesa de Linguística*, págs. 1–15, Évora, Portugal, 6–8 Outubro 1994. 93
- George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel e Ralph Weischedel. “The Automatic Content Extraction (ACE) Program: Tasks, Data and Evaluation”. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, págs. 837–840, Lisboa, Portugal, 26–28 Maio 2004. ELRA. 16
- Aaron Douthett. “The Message Understanding Conference Scoring Software User’s Manual”. Em *Proceedings of the 7th Conference on Message Understanding, MUC-7*, Washington, D.C., EUA, Abril 1998. http://www-nlpir.nist.gov/related_projects/muc/muc_sw/muc_sw_manual.html. 55

Bradley Efron. “Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods”. *Biometrika*, 68:589–599, 1981. 24

Oscar Ferrández, Zornitsa Kozareva, Andres Montoyo e Rafael Muñoz. “NERUA: Sistema de Detección y Clasificación de Entidades Utilizando Aprendizaje Automático”. *Procesamiento del Lenguaje Natural*, 35:37–44, 2005. 94

Robert J. Gaizauskas, Mark Hepple e Chris Huyck. “A Scheme for Comparative Evaluation of Diverse Parsing Systems”. Em Antonio Rubio, Natividad Gallardo, Rosa Castro e Antonio Tejada, editores, *Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC 98*, págs. 143–149, Granada, Espanha, 28–30 Maio 1998. 12

Robert J. Gaizauskas e Yorick Wilks. “Information Extraction: Beyond Document Retrieval”. *Journal of Documentation*, 54(1):70–105, 1998. 4

Daniel Gomes e Mário J. Silva. “Modelling Information Persistence on the Web”. Em *Proceedings of the 6th International Conference on Web Engineering, ICWE 2006*, págs. 193–200, Palo Alto, Califórnia, EUA, 11–14 Julho 2006. ACM Press. 71

Philip Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. 2ª edição. 2000. Springer. 22

Ralph Grisham e Beth Sundheim. “Message Understanding Conference 6: A Brief History”. Em *Proceedings of the 16th International Conference on Computational Linguistics, COLING 96*, págs. 466–471, Copenhaga, Dinamarca, Agosto 1996. 15

- Donna Harman. "Overview of the First TREC Conference". Em Robert Korfhage, Edie M. Rasmussen e Peter Willett, editores, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 93*, págs. 36–47, Pittsburgh, Filadélfia, EUA, 1993. ACM Press. 14
- Ricardo Hasegawa, Ronaldo Teixeira Martins e Maria das Graças Volpe Nunes. "ReGra 2002: Características e Desempenho". Relatório Técnico 02-08, NILC ICMC-USP, Brasil, Junho 2002. 96
- Lynette Hirschman. "Multi-Site Data Collection for a Spoken Language Corpus". Em *Proceedings of the 5th DARPA Speech and Natural Language Workshop, HLT 91*, págs. 7–14, Harriman, Nova Iorque, EUA, 23-26 Fevereiro 1992. 17
- Lynette Hirschman. "The Evolution of Evaluation: Lessons from the Message Understanding Conferences". *Computer Speech and Language*, 12(4): 281–305, 1998. 5, 13, 17
- W. John Hutchins e Harold L. Somers. *An Introduction to Machine Translation*. 1992. Academic Press. 5
- Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato e Souichiro Hidaka. "Overview of IR Tasks at the first NTCIR Workshop". Em *Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, págs. 11–44, Tóquio, Japão, Agosto 1999. 14
- Zornitsa Kozareva, Oscar Ferrández, Andres Montoyo e Rafael Muñoz. "Using Language Resource Independent Detection for Spanish Named

- Entity Recognition". Em Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov e Nikolai Nikolov, editores, *Proceedings of the 5th Conference on Recent Advances in Natural Language Processing, RANLP 2005*, págs. 279–283, Borovets, Bulgária, Setembro 2005. 94
- Zornitsa Kozareva, Óscar Ferrández, Andres Montoyo, Rafael Muñoz e Armando Suárez. "Combining Data-Driven System for Improving Named Entity Recognition". Em *Proceedings of 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005*, págs. 80–90. 94
- Wei-Hao Lin e Alexander Hauptmann. "Revisiting the Effect of Topic Set Size on Retrieval Experiment Error". Em Ricardo Baeza-Yates, Nivio Ziviani, Gary Marchionini, Alistair Moffat e John Tait, editores, *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005*, págs. 637–638, Salvador, Brasil, 15–19 Agosto 2005. ACM Press. 32
- Inderjeet Mani. *Advances in Automatic Text Summarization*. 1999. MIT Press. 5
- Isabel Marcelino. Documentação do ELLE, 2005. http://www.linguateca.pt/Equipa/isabel/Documentacao_ELLE.doc. 95
- Mitchell Marcus, Beatrice Santorini e Mary Ann Marcinkiewicz. "Building a Large Annotated Corpus of English: the Penn Treebank". *Computational Linguistics*, 19(2):313–330, 1994. 17
- Bruno Martins e Mário J. Silva. "A Graph-Ranking Algorithm for Geo-referencing Documents". Em *Proceedings of the 5th IEEE International Con-*

ference on Data Mining, ICDM 2005, págs. 741–744, Houston, Texas, EUA, Novembro 2005. 93

Ronaldo Teixeira Martins e Maria das Graças Volpe Nunes. “Dos processos de individuação e de categorização lexical: sobre a participação do ReGra nas Morfolimpíadas”. Em Diana Santos, editor, *Avaliação Conjunta: um novo paradigma no processamento computacional da língua portuguesa*, capítulo 4. 2006. No prelo. ISTPress. 96

Diana Maynard, Valentin Tablan, Cristian Ursu, Hamish Cunningham e Yorick Wilks. “Named Entity Recognition from Diverse Text Types”. Em Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov e Nikolai Nikolov, editores, *Proceedings of the 3rd Conference on Recent Advances in Natural Language Processing, RANLP 2001*, págs. 257–274, Tzigov Chark, Bulgária, 5–7 Setembro 2001. 36

David McDonald. “Internal and External Evidence in the Identification and Semantic Categorization of Proper Names”. *Corpus processing for lexical acquisition*, págs. 21–39, 1996. 90

Roberta Merchant, Mary Ellen Okurowski e Nancy Chinchor. “The Multilingual Entity Task (MET) Overview”. Em *Proceedings of TIPSTER Text Program (Phase II)*, págs. 445–447, Vienna, EUA, Maio 1996. Association for Computational Linguistics. 15

Andrei Mikheev, Marc Moens e Claire Grover. “Named Entity Recognition without Gazetteers”. Em *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, EACL 99*, págs. 1–8, Bergen, Noruega, 8–12 Junho 1999. 87, 91

- David S. Moore, George P. McCabe, William M. Duckworth e Stanley L. Sclove. *The Practice of Business Statistics: Using Data for Decisions*. 2003. W.H. Freeman and Co. 22, 24
- William Morgan. Statistical Hypothesis Tests for NLP, 2006. <http://www-nlp.stanford.edu/local/talks/sigtest.pdf>. 28
- Cristina Mota. “NooJ as a Corpus Annotator of Named Entities”. Em *Proceedings of the 9th Intex/NooJ Workshop*, Belgrado, Sérvia, 1–3 Junho 2006. 94
- Cristina Mota, Diana Santos e Elisabete Ranchhod. “Avaliação de Reconhecimento de Entidades Mencionadas: Princípio de AREM”. Em Diana Santos, editor, *Avaliação Conjunta: um novo paradigma no processamento computacional da língua portuguesa*, capítulo 14. 2006. No prelo. ISTPress. 18, 19, 36, 44
- Cristina Mota e Luís Sarmento. Proposta concreta para o futuro do HAREM. http://poloxldb.linguatca.pt/harem/publicacoes/harem_2.0.ppt, 2006. 101
- Eric W. Noreen. *Computer Intensive Methods for Testing Hypothesis. An Introduction*. 1989. John Wiley & Sons. 25, 33, 64
- David D. Palmer e David S. Day. “A Statistical Profile of the Named Entity Task”. Em *Proceedings of the 5th ACL Conference for Applied Natural Language Processing, ANLP-97*, págs. 190–193, Washington, D.C., EUA, Março 1997. 86
- Carol Peters e Martin Braschler. “Cross-Language System Evaluation: the CLEF campaigns”. *Journal of the American Society for Information Science and Technology*, 52(12):1067–1072, 2001. 14

- Patti Price. "Evaluation of Spoken Language Systems: The ATIS Domain". Em *Proceedings of the 3rd DARPA Speech and Natural Language Workshop*, págs. 91–95, Hidden Valley, Pensilvânia, EUA, 24–27 Junho 1990. Morgan Kaufmann. 17
- Ross Quinlan. *C4.5: Programs for Machine Learning*. 1993. Morgan Kaufmann. 95
- Stefan Riezler e John T. Maxwell III. "On Some Pitfalls in Automatic Evaluation and Significance Testing for MT". Em *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization, MTSE 2005*, págs. 57–64, Ann Arbor, Michigan, EUA, 29 Junho 2005. 28, 64, 70
- Erik F. Tjong Kim Sang. "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition". Em *Proceedings of the 6th Workshop on Computational Natural Language Learning, CoNLL 2002*, págs. 155–158, Taipei, Formosa, 2002. 16, 26
- Erik F. Tjong Kim Sang e Fien de Meulder. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". Em *Proceedings of the 7th Workshop on Computational Natural Language Learning, CoNLL 2003*, págs. 142–147, Edmonton, Canadá, 2003. 16, 26, 67
- Diana Santos. "The Importance of Vagueness in Translation: Examples from English to Portuguese". *Romansk Forum*, 5:43–69, Junho 1997. 3, 37
- Diana Santos. "O projecto Processamento Computacional do Português: balanço e perspectivas". Em Maria das Graças Volpe Nunes, editor, *Pro-*

ceedings of the 5th International Workshop on Computational Processing of the Portuguese Language, PROPOR 2000, págs. 105–113, Atibaia, Brasil, 19–22 Novembro 2000. 5

Diana Santos. “Um centro de recursos para o processamento computacional do português”. *DataGramaZero - Revista de Ciência da Informação*, 3(1), 2002. 5

Diana Santos e Anabela Barreiro. “On the Problems of Creating a Consensual Golden Standard of Inflected Forms in Portuguese”. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, págs. 483–486, Lisboa, Portugal, 26–28 Maio 2004. ELRA. 3

Diana Santos e Nuno Cardoso. “A Golden Resource for Named Entity Recognition in Portuguese”. Em Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Claudia Oliveira e Maria Carmelita Dias, editores, *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language, PROPOR 2006*, volume 3960 de *Lecture Notes in Computer Science*, págs. 69–79, Itatiaia, Brasil, 13–17 Maio 2006a. Springer. 18, 37, 41, 85

Diana Santos e Nuno Cardoso. “Portuguese at CLEF 2005”. Em Carol Peters, Fredric Gey, Julio Gonzalo, Henning Müller, Gareth Jones, Michael Kluck, Bernardo Magnini e Marteen de Rijke, editores, *Proceedings of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, volume 4022 de *Lecture Notes in Computer Science*, págs. 1007–1010, Viena, Áustria, 21–23 Setembro 2006b. Springer. 18

Diana Santos, Luís Costa e Paulo Rocha. “Cooperatively Evaluating Portuguese Morphology”. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso e Maria das Graças Volpe Nunes, editores, *Proceedings of the 6th International Workshop on Computational Processing of the Portuguese Language, PROPOR 2003*, volume 2721 de *Lecture Notes in Computer Science*, págs. 259–266, Faro, Portugal, Junho 2003. Springer. 5, 13, 17

Diana Santos e Paulo Rocha. “Evaluating CETEMPúblico, a Free Resource for Portuguese”. Em *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, ACL 2001*, págs. 450–457, Toulouse, França, 9–11 Julho 2001. Association for Computational Linguistics. 18

Diana Santos, Nuno Seco, Nuno Cardoso e Rui Vilela. “HAREM: An Advanced NER Evaluation Contest for Portuguese”. Em Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, págs. 1986–1991, Génova, Itália, 22–28 Maio 2006. ELRA. 6

Diana Santos, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, J. J. Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmiento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela e Susana Afonso. “Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa”. Em Guillermo de Ita Luna, Olac Fuentes Chávez e Mauricio Osorio Galindo, editores, *Proceedings of the International Workshop “Taller de Herramientas y Recursos Lingüísticos para el Español y el Por-*

tugués”, IX Iberoamerican Conference on Artificial Intelligence, IBERAMIA 2004, págs. 147–154, Puebla, México, Novembro 2004. 5

Luís Sarmento. “Siemês - a Named-Entity Recognizer for Portuguese Relying on Similarity Rules”. Em Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Claudia Oliveira e Maria Carmelita Dias, editores, *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language, PROPOR 2006*, volume 3960 de *Lecture Notes in Computer Science*, págs. 90–99, Itatiaia, Brasil, 13–17 Maio 2006. Springer. 91

Luis Sarmento, Belinda Maia e Diana Santos. “The Corpógrafo - a Web-based Environment for Corpora Research”. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, págs. 449–452, Lisboa, Portugal, 26–28 Maio 2004. ELRA. 40

Luís Sarmento, Ana Sofia Pinto e Luís Cabral. “REPENTINO - a Wide-Scope Gazetteer for Entity Recognition in Portuguese”. Em Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Claudia Oliveira e Maria Carmelita Dias, editores, *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language, PROPOR 2006*, págs. 31–40, Itatiaia, Brasil, 13–17 Maio 2006. 92

Jacques Savoy. “Statistical Inference in Retrieval Effectiveness Evaluation”. *Information Processing and Management*, 33(4):495–512, 1997. 26

Nuno Seco, Diana Santos, Nuno Cardoso e Rui Vilela. “A Complex Evaluation Architecture for HAREM”. Em Renata Vieira, Paulo

- Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Claudia Oliveira e Maria Carmelita Dias, editores, *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language, PROPOR 2006*, volume 3960 de *Lecture Notes in Computer Science*, págs. 260–263, Itatiaia, Brasil, 13–17 Maio 2006. Springer. 56
- Satoshi Sekine, Kiyoshi Sudo e Chikashi Nobata. “Extended Named Entity Hierarchy”. Em Manuel González Rodríguez e Carmen Paz Suarez Araujo, editores, *Proceedings of the 3rd Language Resource and Evaluation Conference, LREC-2002*, págs. 1818–1824, Las Palmas de Gran Canaria, Espanha, 2002. ELRA. 42
- David J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. 2ª edição. 2000. Chapman and Hall. 22
- Max Silberztein. *Dictionnaires électroniques et analyse lexicale du français. Le système INTEX*. Paris. 1993. Masson. 95
- Mário J. Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso e Nuno Cardoso. “Adding Geographic Scopes to Web Resources”. *CEUS - Computers, Environment and Urban Systems*, 30:378–399, 2006. 93
- Tamar Solorio. “Exploiting Named Entity Taggers in a Second Language”. Em Chris Callison-Burch e Stephen Wan, editores, *Student Research Workshop at the 43rd Annual Meeting of the Association for Computational Linguistics, ACL 2005*, págs. 25–30, Ann Arbor, Michigan, EUA, Junho 2005. Association for Computational Linguistics. 95
- Karen Sparck-Jones. “Automatic Indexing”. *Journal of Documentation*, 30 (4):393–432, 1974. 72

- Beth Sundheim. "Overview of Results of the MUC-6 Evaluation". Em *Proceedings of the 6th Conference on Message Understanding, MUC-6*, págs. 13–31, Columbia, Maryland, EUA, 6–8 Novembro 1995. Morgan Kaufmann. 86
- Beth Sundheim e Nancy Chinchor. "Survey of the Message Understanding Conferences". Em *Proceedings of the Workshop on Human Language Technology, HLT 93*, págs. 56–60, Princeton, Nova Jersey, EUA, Março 1993. Association for Computational Linguistics. 15
- Alan M. Turing. "Computing Machinery and Intelligence". *Mind*, 59: 433–460, 1950. 3
- C. J. van Rijsbergen. *Information Retrieval*. 2ª edição. 1979. Department of Computer Science, University of Glasgow. 55, 71
- Ellen Voorhees. "The Philosophy of Information Retrieval Evaluation". Em Carol Peters, Martin Braschler, Julio Gonzalo e Michael Kluck, editores, *Proceedings of the 2nd Workshop of the Cross-Language Evaluation Forum, CLEF 2001*, volume 2406 de *Lecture Notes in Computer Science*, págs. 355–370, Darmstadt, Alemanha, 3-4 Setembro 2002. Springer. 14
- Ellen Voorhees. "Question Answering in TREC". Em Ellen M. Voorhees e Donna K. Harman, editores, *TREC - Experiment and Evaluation in Information Retrieval*, capítulo 10, págs. 233–257. 2005. MIT Press. 5
- Ellen Voorhees e Chris Buckley. "The Effect of Topic Set Size on Retrieval Experiment Error". Em *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2002*, págs. 316–323, Tampere, Finlândia, 11–15 Agosto 2002. ACM Press. 32, 80

Justin Zobel. "How Reliable are the Results of Large-Scale Information Retrieval Experiments?". Em *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1998*, págs. 307–314, Melbourne, Austrália, Agosto 1998. ACM Press. 14