Apresentação do SIEMÊS2

Encontro HAREM

Porto, 15 de Julho de 2006

Luís Sarmento

O SIEMÊS2 num relance

- Camada de identificação de candidatos, usando pistas morfológicas
  - classificação de entidades numéricas: datas, quantidades, numerário…
- Cadeia de classificação:
  - 5 níveis
  - Nível de desambiguação
- Recorre a um léxico-semântico em construção

A Classificação no SIEMÊS2

- Objectivo: gerar hipóteses de classificação para depois desambiguar
- Obter uma ou mais hipóteses de classificações de entre as quais se escolherá uma num processo de "desambiguação"
- Desambiguação feita (se necessária) tendo em conta informação contextual

Geração de hipóteses no SIEMÊS2

- Cadeia com 5 níveis:
  - regras "simples" de *grande precisão*
  - regras de pesquisa imediata no REPENTINO
  - regras de emparelhamento de prefixo sobre o REPENTINO
  - regras de posteriores usadas quando não foi possível obter nenhuma hipótese anteriormente (camada reduzida)
  - regras de semelhança sobre REPENTINO (2)
- Aplicadas por esta ordem,,,

Regras "simples"

- Exemplo:

```
## o imperdador Hiroito
{{ -1:@cargo =>
      meta(-1,CLASSE=SER); meta(-1,SUBCLASSE=CARGO);
      meta(CLASSE=SER); meta(SUBCLASSE=HUM);
      sai();
}}
```

- Vários conjuntos de regras para algumas entidades
- Acesso ao léxico-semântico: @cargo

Participação no Harem

- Foi realizada com o SIEMÊS1
  - O SIEMÊS2 mantém a "filosofia" do SIEMÊS1
- Do ponto de vista de engenharia tinha vários problemas que levaram à sua re-implementação
  - Estruturas de dados mais poderosas
  - Motor de interpretação de regras
  - Bancos de regras externos
  - Léxico-semântico independente
- SIEMÊS1: acabou por ser abandonado…

Participação no mini-HAREM

- Com SIEMÊS2:
  - Conjunto de regras aumentado
  - Léxico-semântico maior
- Mas… ainda não completo:
  - Certo tipo de entidades não estava coberto
  - A camada de desambiguação estava incompleta
- Avaliação de Componentes:
  - Quanto contribui cada uma dos 5 níveis de classificação

Avaliação de Componentes

- Testar os resultados obtidos activando **apenas 1** dos 5 níveis de geração de hipóteses
- Perceber quanto é que contribui cada uma das várias estratégias:
  - Uso das regras
  - Uso do almanaque nas suas várias possibilidades de "semelhança"

Submissão de 9 ensaios

Camadas / Ensaio	Regras Simples	Exacto REP.	Prefixo REP.	Regras Post.	Sem 1 REP.	Sem 2 REP,
.siemes_simples	X					
siemes_exacto		X				
siemes_prefixo2			X (2)			
siemes_prefixo4			X (4)			
siemes_posterior				X		
siemes_difuso1					X	
siemes_difuso2						X
siemes_total1	X	X	X	X	X	
siemes_total2	X	X	X	X		X

Como foram feitas as submissões?

- Remoção de tudo o que não se conseguiu marcar, incluindo a própria identificação
  - só manter o que se tem a "certeza"
- **Objectivo:** testar a precisão e abrangência da classificação (e não da identificação)
  - Regras so HAREM corresponde ao ***Cenário Relativo***
- Todas as submissões incluiam a análise às EM "numéricas" (data, numerário…)
  - O que em rigor não deveria ter sido feito

Todos no cenário absoluto

- Os resultados globais…

Saída	Precisão Máxima do Sistema (%)	Abrangência Máxima na CD (%)	Medida F
siemes_total2	53.02	**51.38**	0.5219
siemes_total1	52.63	51.01	0.5181
siemes_prefixo4	57.25	46.08	0.5106
siemes_prefixo2	55.17	46.92	0.5071
siemes_difuso2	45.88	42.31	0.4403
siemes_exact	66.00	32.99	0.4399
siemes_posterior	58.12	25.33	0.3528
siemes_difuso1	35.48	32.33	0.3383
siemes_simples	68.79	14.97	0.2458

- Não foram excelentes. Pior que o SIEMÊS1

Todos no cenário relativo

- É nestes resultados que vamos prestar mais atenção (só os correctamente identificados)

Saída	Precisão Máxima do Sistema (%)	Abrangência Máxima na CD (%)	Medida F
siemes_simples	**77.21**	**78.20**	**0.7770**
siemes_exact	72.11	72.54	0.7233
siemes_prefixo4	64.87	66.28	0.6557
siemes_prefixo2	62.29	63.90	0.6308
siemes_total2	61.14	63.25	0.6217
siemes_posterior	62.45	61.66	0.6205
siemes_total1	60.67	62.79	0.6171
siemes_difuso2	52.99	53.45	0.5322
siemes_difuso1	40.98	40.83	0.4091

siemes_simples

- Método: cerca de 20 regras simples
  - com operadores de disjunção seriam apenas 5

Avaliação global da classificação semântica combinada

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F
VALOR	11º	94,13	95,69	0,9490
TEMPO	16º	91,35	91,75	0,9155
LOCAL	14º	87,25	87,25	0,8725
PESSOA	17º	75,43	75,43	0,7543

siemes_exacto

- Método: olha para o REPENTINO, se existir exactamente igual, marca com a categoria correspondente

Avaliação global da classificação semântica combinada

ABSTRACCAO	1º	**85,32**	**85,32**	**0,8532**
ACONTECIMENTO	1º	**80,00**	**80,00**	**0,8000**
COISA	6º	92,59	92,59	0,9259
VALOR	8º	94,13	95,69	0,9490
TEMPO	10º	91,72	92,11	0,9192
LOCAL	1º	**95,27**	**95,87**	**0,9557**
PESSOA	5º	88,24	88,85	0,8854
ORGANIZACAO	1º	91,60	**90,60**	**0,9109**
OBRA	3º	88,72	88,72	0,8872

siemes_prefixo2/4

- Método:
  - olha para a estrutura da entidade e procura quais as instâncias no REPENTINO que possuem o mesmo prefixo com 2 ou 4 palavras
  - Em cada iteração reduz uma palavra no emparelhamento.
  - Para com a maior sequência cuja cobertura seja maior que limite (40%)
  - Pretende-se explorar heuristicamente a informação que se encontra no "prefixo" de uma referência

siemes_prefixo2/4

Avaliação global da classificação semântica combinada (2 & 4)

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F
ABSTRACCAO	7º	74,02	75,05	0,7453
ACONTECIMENTO	4º	70,47	70,47	0,7047
COISA	10º	85,82	85,82	0,8582
VALOR	10º	94,13	95,69	0,9490
TEMPO	11º	91,59	91,98	0,9178
LOCAL	13º	87,53	88,55	0,8804
PESSOA	4º	89,05	89,43	0,8924
ORGANIZACAO	3º	80,89	83,33	0,8209
OBRA	4º	84,17	84,17	0,8417

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F
ABSTRACCAO	5º	74,77	75,82	0,7529
ACONTECIMENTO	5º	70,47	70,47	0,7047
COISA	9º	85,82	85,82	0,8582
VALOR	5º	94,13	95,69	0,9490
TEMPO	12º	91,51	91,91	0,9171
LOCAL	12º	88,11	88,99	0,8855
PESSOA	6º	87,91	88,29	0,8810
ORGANIZACAO	2º	82,09	83,54	0,8281
OBRA	7º	82,03	82,03	0,8203

siemes_posterior

- Descrição: Um conjunto de regras muito simples sobre o contexto…
- Tenta apenas resolver alguns casos
  - Na verdade pouco interessante…

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F
VALOR	6º	94,13	95,69	0,9490
TEMPO	13º	91,50	91,92	0,9171
LOCAL	15º	87,25	87,25	0,8725
PESSOA	3º	89,41	89,80	0,8960
ORGANIZACAO	16º	**94,48**	53,99	0,6871
OBRA	10º	97,92	58,75	0,7344

siemes_difuso1/2

- Descrição: tenta encontrar a entidade mais semelhante no REPENTINO e atribui uma medida de mérito:
  - Difuso1: para cada palavra do candidato ver qual a sua "cobertura" por categorias. Fazer uma média ponderada das categorias.
  - Difuso2: Cada palavra do candidato contribui com as suas potênciais categorias com um peso inversamente proporcional ao número de categoria em que ela "ocorre" (poder descrimintativo de cada palavra). Fazer uma média pondera das categorias.

siemes_difuso1/2

Avaliação global da classificação semântica combinada (2 & 4)

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F
ABSTRACCAO	9º	72,27	75,56	0,7388
ACONTECIMENTO	10º	62,97	66,28	0,6458
COISA	11º	85,54	85,54	0,8554
VALOR	12º	94,13	95,69	0,9490
TEMPO	17º	91,36	91,75	0,9155
LOCAL	16º	78,84	79,47	0,7915
PESSOA	10º	83,99	84,28	0,8414
ORGANIZACAO	13º	73,67	77,70	0,7563
OBRA	9º	75,90	79,00	0,7742

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F
ABSTRACCAO	4º	77,28	77,28	0,7728
ACONTECIMENTO	11º	51,85	59,26	0,5530
COISA	5º	95,00	95,00	0,9500
VALOR	9º	94,13	95,69	0,9490
TEMPO	9º	91,73	92,13	0,9193
LOCAL	17º	77,16	78,38	0,7776
PESSOA	7º	85,13	86,24	0,8568
ORGANIZACAO	12º	74,74	77,13	0,7591
OBRA	5º	82,78	82,78	0,8278

siemes_total1/2

- Método: os vários andares seguidos

siemes_total1/2

Avaliação global da classificação semântica combinada (Total 1)

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F
ABSTRACCAO	6º	73,46	76,12	0,7477
ACONTECIMENTO	7º	67,72	70,67	0,6917
COISA	7º	87,53	87,53	0,8753
VALOR	7º	94,13	95,69	0,9490
TEMPO	14º	91,44	91,82	0,9163
LOCAL	8º	89,26	90,38	0,8982
PESSOA	9º	84,38	84,83	0,8460
ORGANIZACAO	4º	81,50	82,48	0,8199
OBRA	8º	81,95	81,10	0,8152

Avaliação global da classificação semântica combinada (Total 2)

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F
ABSTRACCAO	8º	72,66	75,16	0,7389
ACONTECIMENTO	6º	70,02	70,02	0,7002
COISA	8º	87,53	87,53	0,8753
VALOR	4º	94,13	95,69	0,9490
TEMPO	15º	91,44	91,82	0,9163
LOCAL	9º	89,02	90,14	0,8958
PESSOA	8º	84,53	85,12	0,8483
ORGANIZACAO	5º	81,15	82,26	0,8170
OBRA	6º	82,52	81,58	0,8205

Os melhores em cada categoria

- Cenário Relativo (ignorando NUMEX)

Categoria	Sistema	Precisão (%)	Abrangência (%)	Medida F
ABSTRACCAO	**siemes_exact**	**85.32**	**85.32**	**0.8532**
ACONTECIMENTO	**siemes_exact**	**80.00**	**80.00**	**0.8000**
COISA	**siemes_difuso2**	95.00	95.00	0.9500
LOCAL	**siemes_exact**	**95.27**	**95.87**	**0.9557**
PESSOA	**siemes_posterior**	89.41	89.80	0.8960
ORGANIZACAO	**siemes_exact**	91.60	**90.60**	**0.9109**
OBRA	**siemes_exact**	88.72	88.72	0.8872

- Cenário Absoluto (ignorando NUMEX)

Categoria	Sistema	Precisão (%)	Abrangência (%)	Medida F
ABSTRACCAO	**siemes_total2**	43.02	**19.82**	**0.2713**
ACONTECIMENTO	**siemes_prefixo2**	36.91	25.42	0.3010
COISA	**siemes_prefixo2**	41.05	10.43	**0.1664**
LOCAL	**siemes_total2**	61.29	56.69	0.5890
PESSOA	**siemes_total2**	59.78	57.49	0.5861
ORGANIZACAO	**siemes_total2**	40.25	46.95	0.4334
OBRA	**siemes_total1**	15.85	**36.46**	**0.2209**

SIEMÊS 2 : Breves conclusões

- O sistema foi completamente re-implementado embora mantendo a filosofia da primeira versão
- As múltiplas opções de marcação permitem obter resultados diferentes, alguns aparentemente supreendentes (?)
- Os resultados obtidos terão de ser analisados cuidadosamente no futuro para formularmos novas hipóteses de desenvolvimento do SIEMÊS