

Segundo HAREM

Balanço final e perspectivas de futuro

Diana Santos,
Hugo Oliveira, Cláudia Freitas, Cristina Mota e Paula Carvalho

Encontro do Segundo HAREM
Universidade de Aveiro, 7 de Setembro de 2008

Mapa da apresentação

- O que correu mal
- O que correu bem
- O futuro... a quem pertence?

O que correu mal

- Do ponto de vista científico
 - Delimitação das EM
 - Utilização de dois modelos de avaliação com filosofias e objectivos distintos
- Do ponto de vista da organização
 - Falta de comunicação/coordenação entre as duas equipas na questão de avaliação do TEMPO
- Do ponto de vista dos participantes
 - Pouca interacção com a organização na avaliação conjunta

Em pormenor: ident. vs. class

- A identificação ainda teve um peso demasiado grande em relação à classificação, fazendo com que sistemas sem qualquer classificação fossem superiores aos que tentaram classificar
- Alguns dos participantes ao identificarem só um subconjunto de categorias implicitamente estavam a classificar
- Muito possivelmente, deveríamos remover a identificação simples ou garantir que era ínfima comparada com a classificação

Em pormenor: minúsculas

- Para simultaneamente
 - reduzir a importância das diferentes estratégias de identificação
 - garantir que as EM na colecção dourada estivessem bem delimitadas (ao contrário da CD do Primeiro HAREM)
- Procurámos todos os casos da CD em que havia minúsculas que faziam parte, e listámos esses casos
- Dissemos que todos os outros casos não deviam ser marcados (o que provocou muita confusão)

Em pormenor: modelos semânticos incompatíveis

- Modelo do HAREM clássico: é o contexto que decide, a análise é a da pessoa que anota a CD
- Modelo do TEMPO: baseado em critérios fundamentalmente sintácticos, ignorando em muitos casos o uso das entidades em contexto
- Resultado: uma CD com anotação de categorias seguindo filosofias diferentes ☹

Em pormenor: falta de coordenação no TEMPO

- Embora o grupo do TEMPO tenha fornecido material de treino e exemplo, não podia naturalmente ser contactado para resolver os problemas da anotação na CD (visto que eram participantes)
- Como é impossível especificar todos os pormenores antes de deparar com o texto real, muitas vezes tivemos de fazer escolhas que – embora com boa vontade – podem ser consideradas como desvirtuando ou discordando com a intenção da pista

Novo formato XML

- Criou mais problemas do que resolveu
 - Reformatação das antigas CD e dos programas
 - UTF-8 por omissão quando pedimos ISO
 - Fez-nos descobrir o maravilhoso mundo da padronização: há várias versões dos padrões, incompatíveis entre si ☺
 - Não levámos suficientemente longe a proposta de novo formato para poder utilizar cabalmente as capacidades do XML
- ```
<alt id=x><em categ="obra">|<em categ="local"></alt>
```

## ReReEM: primeiro balanço

- Tarefa demasiado ambiciosa
- Carregando com as complexidades do HAREM
  - ALT
  - Vagueza
  - Cenários selectivos do HAREM
- Participantes muito divergentes
  - Um que seguiu à risca o que esperávamos
  - Dois que divergiram inesperada e substancialmente
    - 1 sem classificação !
    - 1 sem identidade !

## ReReEM: primeiro balanço (2)

- Vários becos sem saída: separação de identidade e das outras relações
  - Agrupamento através da identidade
  - Medidas de agrupamento
  - Com emparelhar os grupos
- Expansão da participação ou não?
- Como comparar o incomparável?
- O que fazer aos ALT?
- O que fazer a participações inconsistentes?

## O que correu bem

- As CD foram muito melhor revistas
- As opções foram incomparavelmente melhor documentadas
- Houve mais retorno dos participantes
- A questão dos cenários e dos véus foi levada às últimas consequências
- Tivemos vários novos participantes ou interessados
- Claro progresso na definição da tarefa e nos desafios

## Construção de recursos mais robustos e melhor pensados

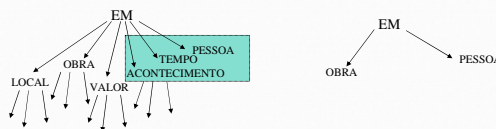
- Muito mais discussão e fundamentação, numa equipa maior, de todas as opções linguísticas tomadas
- Muito mais revisão e consideração das divergências, linguísticas e de interpretação
- Mais informação
  - SUBTIPOs em LOCAL e TEMPO
  - Relações semânticas (entre facetas)
  - Identificação única das EM
  - Outra informação para futuro estudo (dúvidas, discordâncias, casos problemáticos – OMITIDOS)

## Recursos mais variados

- Além de um recurso valioso para REM, a constituição da nova colecção HAREM e dos resultados dos sistemas permite efectuar trabalhos interessantes em
  - recolha de informação geográfica
  - resposta automática a perguntas
  - normalização temporal (graças a Hagege et al.)
  - co-referência
  - relações semânticas de inclusão e localização
  - outras relações semânticas entre EM

## Tratamento de cenários como ontologias distintas

- Foi clarificado e cabalmente implementado o tratamento de cenários de participação que permitam comparar melhor os vários sistemas entre si
- Não só comparar cada sistema segundo as suas próprias condições



## ALT linguisticamente motivados

- Foi aumentada a semântica dos ALT, que passaram a identificar consistentemente todas as EM possíveis e não apenas a maior
- A avaliação dos ALT deixou de ser feita por critérios quantitativos em termos de número de palavras, para passar a sê-lo em termos do conteúdo
- Foram identificadas uma série de regras de construção de EM complexas, estruturalmente sistemáticas

## Futuro: que futuro?

- Agora que a Linguateca termina...
  - Existe uma comunidade de REM que pode continuar?
  - Algum participante ou grupo de participantes que quer continuar a organizar um Terceiro HAREM?
- Ou devemos tentar tornar o HAREM multilingue
  - por exemplo no CLEF, GeoCLEF, GikiP, ARE...
  - ou independentemente?
- Faz mais sentido agora atacar outras áreas?
  - Discussão para o *Encontro Linguateca: 10 anos* ?

## Discussão: a palavra aos outros

- O que é que podia ter sido feito melhor
- O que é que pode ser melhorado já nas actas e no futuro
- Para que é que este encontro pode contribuir

## Agradecimentos

- A Linguateca e o HAREM são financiados através do contrato nº 339/1.3/C/NAC, financiado pelo governo português e pela União Europeia, e executado pela FCCN.