Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas

Simpósio Doutoral da Linguateca

Por: Nuno Cardoso

Orientadores: Eugénio de Oliveira (FEUP)
Mário J. Silva (FCUL)

FACULDADE · DE · CIÊNCIAS · UNIVERSIDADE · DE · LISBOA
DEPARTAMENTO · DE · INFORMÁTICA

Linguateca

Universidade do Porto
FEUP Faculdade de Engenharia

Estrutura da Apresentação

- Introdução (Proposta da Tese)
- Motivação (Proposta da Tese)
- Objectivos (Proposta da Tese)
- Metodologia HAREM
  - Colecção Dourada, Directivas
  - Trabalho em progresso
- Plataforma HAREM
  - Arquitectura, Medidas
  - Trabalho em progresso
- Iniciativa HAREM
  - HAREM, MiniHAREM
  - Trabalho em progresso
- Resultados
- Progresso da tese

O que é REM?

- REM (Reconhecimento de Entidades Mencionadas) é uma tarefa da área de PLN
- Objectivo: delimitar, desambiguar e atribuir um significado semântico a Entidades Mencionadas (EMs) importantes na mensagem.
- Exemplo:

```
Eça de Queirós nasceu na Póvoa de Varzim
em 1845, e faleceu 1900, em Paris.
Estudou na Universidade de Coimbra.
```

O que é REM?

- Identificação de EMs

Eça de Queirós nasceu na ***Póvoa de Varzim*** em ***1845***, e faleceu em ***1900***, em ***Paris***. Estudou na ***Universidade de Coimbra***.

- Classificação (semântica) de EMs

Eça de Queirós nasceu na ***Póvoa de Varzim*** em ***1845***, e faleceu ***1900***, em ***Paris***. Estudou na ***Universidade de Coimbra***.

Categorias semânticas:

Cidade, **Ano**, **Pessoa**, **Universidade**

Aplicações de REM

- Tarefa essencial para as diversas áreas de PLN.
- Diversas aplicações (ver [Mota, 06]):
  - Tradução automática

    (evita tradução de Castelo Branco para White Castle)
  - Desambigação de pesquisas com âmbito geográfico

    (pesquisas de romances de Castelo Branco *vs* turismo em Castelo Branco)

Aplicações de REM

– Sistemas de Resposta a Perguntas

(Quem é Castelo Branco? ≠ Onde fica Castelo Branco?)

– Análise sintáctica

(castelo/nome branco/adj vs castelo branco/nome)

– Geração de Texto Automático

– Sumarização de textos

– Reconhecimento de Fala

– ...

Estrutura da Apresentação

- Introdução (Proposta da Tese)
- Motivação (Proposta da Tese)
- Objectivos (Proposta da Tese)
- Metodologia HAREM
  - Colecção Dourada, Directivas
  - Trabalho em progresso
- Plataforma HAREM
  - Arquitectura, Medidas
  - Trabalho em progresso
- Iniciativa HAREM
  - HAREM, MiniHAREM
  - Trabalho em progresso
- Resultados
- Progresso da tese

Motivação (Proposta da Tese)

- Avaliação sempre presente na evolução das áreas de Informática
- Progresso na área de IA sempre medido por avaliação (teste de Turing em 1950)
- Condição importante: avaliação (experimental) só é possível se todos os factores forem idênticos (tarefa, objectivos, colecção, medidas)
- Citando [Gaizauskas, 98]:

"if objective measures can be agreed, winning techniques will come to the fore and better technology will emerge more efficiently"

Motivação (Proposta da Tese)

- REM é uma tarefa complexa (sub-tarefa de interpretação semântica)
- Avaliações anteriores (MUC, COnLL, ACE, MET)
  - Tarefa de REM em evidência...
  - ...mas não focavam em detalhe de REM
- Metodologia adequada?
  - Anotação em contexto? Vagueza das EMs?
  - Só pessoas, locais e organizações?
  - Nenhum dos eventos construído de raíz

Motivação (Proposta da Tese)

- REM em português
  - Não existia um plano organizado para acompanhar os sistemas de REM em português
  - Até que ponto REM em português é diferente de REM noutros idiomas?
  - Português é uma língua diferente, com EMs diferentes? contextos diferentes?
  - Como estava organizada a comunidade que investiga REM em PT? E os seus sistemas?

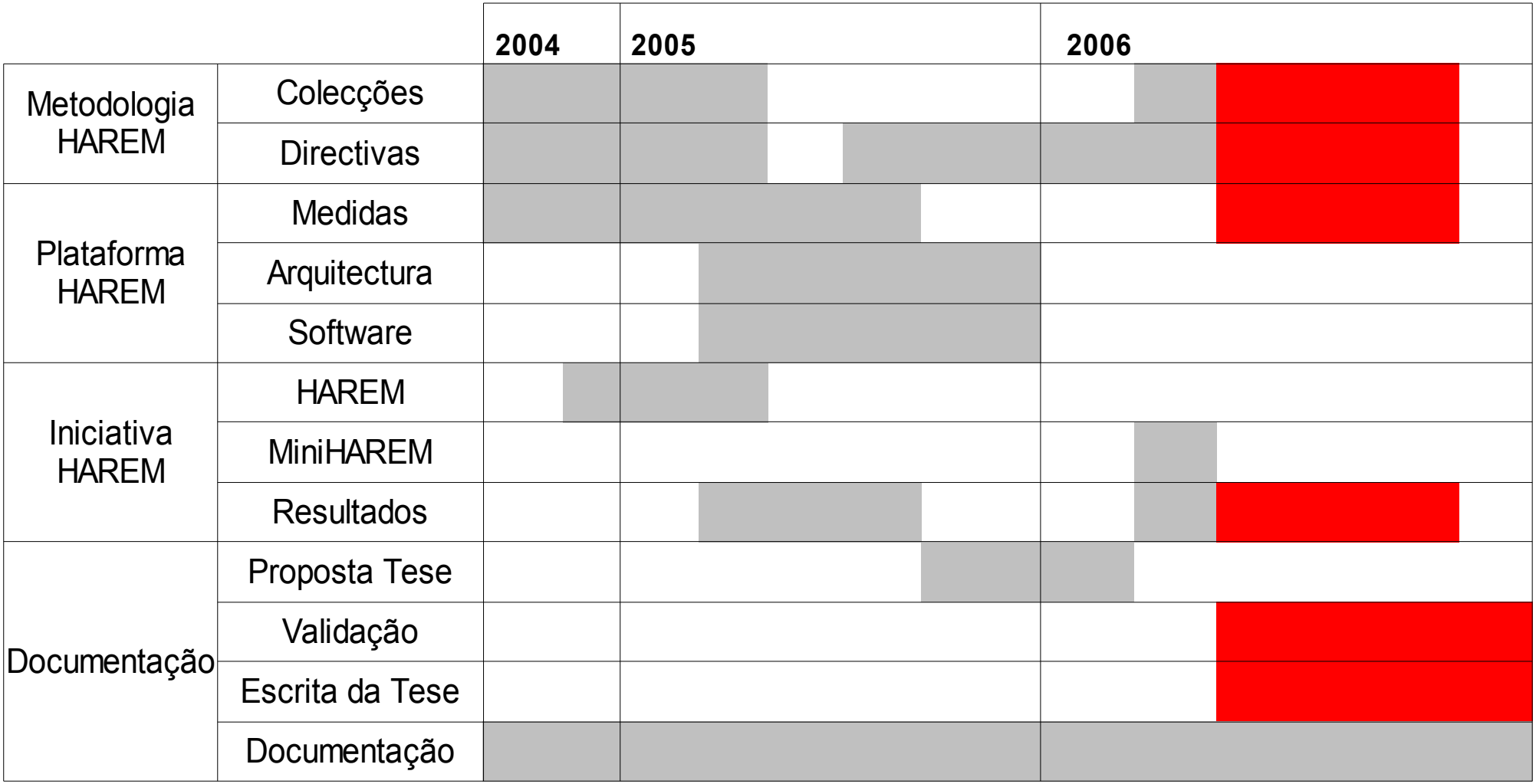
Estrutura da Apresentação

- Introdução (Proposta da Tese)
- Motivação (Proposta da Tese)
- Objectivos (Proposta da Tese)
- Metodologia HAREM
  - Colecção Dourada, Directivas
  - Trabalho em progresso
- Plataforma HAREM
  - Arquitectura, Medidas
  - Trabalho em progresso
- Iniciativa HAREM
  - HAREM, MiniHAREM
  - Trabalho em progresso
- Resultados
- Progresso da tese

Objectivos (Proposta da Tese)

- **Metodologia HAREM** - Criar uma nova metodologia para a avaliação em REM, em conjunto com a comunidade científica interessada em REM
- **Plataforma HAREM** - Desenvolver um ambiente de avaliação específico para REMs, que aplica a metodologia e permite a qualquer grupo de investigação a avaliação comparativa do seu sistema
- **Iniciativa HAREM** – Aplicar e validar a Metodologia HAREM em eventos de avaliação conjunta, usando a Plataforma HAREM.

Plano da Tese



Estrutura da Apresentação

- Introdução (Proposta da Tese)
- Motivação (Proposta da Tese)
- Objectivos (Proposta da Tese)
- Metodologia HAREM
  - Colecção Dourada, Directivas
  - Trabalho em progresso
- Plataforma HAREM
  - Arquitectura, Medidas
  - Trabalho em progresso
- Iniciativa HAREM
  - HAREM, MiniHAREM
  - Trabalho em progresso
- Resultados
- Progresso da tese

Características da Metodologia HAREM

- Anotação em contexto
- Suporta indefinição das EMs como característica inerente à tarefa
- Avalia a classificação morfológica, essencial na tarefa de REM
- Avaliação compatível (sistemas com objectivos diferentes, para propósitos diferentes)
- Colecção, Directivas e Medidas feitas e aprovadas pela comunidade/participantes

“ *everyone debates, everyone contributes, everyone participates, everybody wins!”*

Colecções de texto do HAREM

- Colecção HAREM: conjunto de textos não anotados de diversos géneros de texto e várias variantes de português
- Colecção Dourada (CD): fracção da Colecção HAREM, manualmente anotada
- HAREM: CD de 2005
- MiniHAREM: CD de 2006 (ainda por rever)

Tamanho	Colecção HAREM	CD 2005	CD 2006
Palavras	520752	92761	75664
Extractos	1202	129	128
EMs	cerca de 40000	5132	3714
EMs vagas (class.)	cerca de 1000	131	142
EMs vagas (ident.)	cerca de 500	65	58

Validação da Metodologia HAREM

- Anotação manual das CDs
  - Discordância entre anotadores – influencia os resultados? Testes com anotações unipessoais...
  - Estimar a % de concordância entre humanos. Influencia os valores absolutos de desempenho? E relativos?
- Tamanho da CD
  - Número de extractos suficientes?
  - CDs 2005, 2006: Mesmo número de extractos (mas menos 18% de palavras, menos 27% de EMs). Importante?

Validação da Metodologia HAREM

- Composição da CD
  - As CDs são representativas?
  - Variantes: mais documentos de PALOPs e ex-colónias asiáticas?
  - Origem reflete a variância?
  - Géneros representativos? Teores de cada género são realistas?
- Textos da CD
  - Textos web difíceis de segmentar. Deviam ter etiquetas HTML? Sistemas penalizados?
  - Texto técnico quase não tem EMs...

Validação da CD Metodologia HAREM

- Género das CDs
  - Analisar a distribuição de EMs por géneros
  - Será possível estimar o género através das suas EMs?
  - Será que a informação do género pode ajudar os sistemas (identificação / classificação)?
  - Qual a distribuição de EMs vagas por categoria? Género? Há correlações?
  - Medir a dificuldade dos sistemas com as categorias, géneros e variantes

Validação da Metodologia HAREM

- Directivas de Etiquetagem do HAREM
  - Adequadas? Aspectos a melhorar?
  - Definição da EM nada pacífica...
    - Delimitação, contextos, etc.
  - Alterações 2005->2006
    - Qual a “cobertura” das melhorias?
    - Deviam ter sido mais profundas?
- Encontro HAREM: Importante para receber sugestões / críticas dos participantes para:
  - as colecções usadas
  - as directivas do HAREM e MiniHAREM

Estrutura da Apresentação

- Introdução (Proposta da Tese)
- Motivação (Proposta da Tese)
- Objectivos (Proposta da Tese)
- Metodologia HAREM
  - Colecção Dourada, Directivas
  - Trabalho em progresso
- Plataforma HAREM
  - Arquitectura, Medidas
  - Trabalho em progresso
- Iniciativa HAREM
  - HAREM, MiniHAREM
  - Trabalho em progresso
- Resultados
- Progresso da tese

Características da Plataforma HAREM

- Bancada de comparação de sistemas composto por diversos módulos independentes (permite flexibilidade e facilita depuração)
- Disponível gratuitamente (licença GPL)
- Permite a avaliação parcial de um subconjunto de categorias (cenário absoluto/relativo)
- Permite a avaliação somente às tarefas de classificação (cenário total/selectivo)
- Novas medidas para além de Precisão, Abrangência e Medida F: Sobre-geração, Sub-geração, Sobre-especificação

Pontuação no HAREM

- **Tarefa de Identificação**:

  "O <EM>**João António**</EM> falou hoje":

Pontuação	Saída
Correcto	O <EM>**João António**</EM> falou hoje.
Parcialmente correcto por excesso	O <EM>**João António falou**</EM> hoje
Parcialmente correcto por defeito	O **João** <EM>**António**</EM> falou hoje.
Espúrio	O João António falou <EM>**hoje**</EM>.
Em Falta	O **João António** falou hoje.

Pontuação no HAREM

- Correcto: $p = 1$
- Parcialmente correcto (excesso ou defeito): $p = 0{,}5 \frac{n_c}{n_d}$

n_c - termos em comum. n_d – termos distintos.

- Espúrios, Em Falta, Outros: 0

Exemplo: <EM>**João António Santos**</EM>:

<EM>João António Santos</EM>: Correcto (p=**1**)

João <EM>António</EM> Santos: Par.Cor. (p=0,5(1/3)=**1/6**)

<EM>João António</EM> <EM>Santos</EM>: Par.Cor. (p=0,5(2/3)+0,5(1/3)=**0,5**)

p_{max} Parcialmente Correcto, por cada EM: **0,5** (e nunca 1).

Pontuação no HAREM

- **Tarefa de Classificação Morfológica**:

"O <EM MORF="?,S">**João António**</EM> falou":

Pontuação	Saída
Correcto	O <EM MORF="?,S">**João António**</EM> falou.
Parcialmente correcto	O <EM MORF="?,S">**João António falou**</EM>.
Incorrecto	O <EM MORF="?,P">**João António**</EM> falou.
Em Falta	O <EM >**João António**</EM> falou.
Espúrio	O João António <EM MORF="M,S">**falou**</EM>.
Sobre-especificado	O <EM MORF="M,S">**João António**</EM> faltou.

- Medidas: Género, Número, Combinado
  - Combinado: correcto se Género e Número correcto

Métricas no HAREM

- **Tarefa de Classificação Semântica**:

  Há quatro medidas:

  - Categorias apenas – só são avaliadas as categorias. Os tipos são ignorados.
  - Tipos apenas – só são avaliados os tipos, no universo de EMs com categorias correctas.
  - Plana – só é considerado correcto se a categoria e o tipo são correctos.
  - Combinada:

$$\begin{cases} 0 & \text{se categoria incorrecta} \\ 1 & \text{se categoria correcta e tipo incorrecto} \\ 2-\frac{1}{n} & \text{se categoria correcta e pelo menos um tipo correcto} \end{cases}$$

Métricas no HAREM

- Significado das quatro medidas
  - Categorias apenas – avaliação semântica a nível de categoria (mais geral)
  - Tipos apenas – avaliação semântica a nível de tipo (mais detalhado)
  - Plana – avalia o desempenho do sistema em conseguir definir por completo a categoria e tipo semântica da EM (mais restrito)
  - Combinada – Medida que combina a categoria e tipo, tendo em conta as diversas opções de tipo para cada categoria (mais 'fuzzy')

Métricas no HAREM

- Classificação Semântica:
  - **Caso 1**: categoria correcta, tipo correcto.
- Saída:

“<PESSOA TIPO=”INDIVIDUAL”>**João**</PESSOA>”

- CD:

“<PESSOA TIPO=”INDIVIDUAL”>**João**</PESSOA>”

Medida	Pontuação
Categorias apenas	1 (Correcto)
Tipos apenas	1 (Correcto)
Plana	1 (Correcto)
Combinada	2 – 1/6 = 1,833(3)

Métricas no HAREM

- Classificação Semântica:
  - **Caso 2**: categoria correcta, tipo incorrecto.
- Saída:

“<PESSOA TIPO=”CARGO”>**João**</PESSOA>”

- CD:

“<PESSOA TIPO=”INDIVIDUAL”>**João**</PESSOA>”

Medida	Pontuação
Categorias apenas	1 (Correcto)
Tipos apenas	0
Plana	0
Combinada	1 (Cat. correcta)

Medidas de HAREM

Nota: as medidas variam:

- Consoante os cenários
- Consoante as tarefas de classificação

- **Precisão**: $\frac{\sum p_{Correctas} + \sum p_{Parc.Correctas}}{\sum EMs_{saída\ \ do\ \ sistema}}$
- **Abrangência**: $\frac{\sum p_{Correctas} + \sum p_{Parc.Correctas}}{\sum EMs_{na\ \ CD}}$
- **Medida F**: $\frac{2\mathrm{PR}}{(P+R)}$

Medidas de HAREM

- **Sobre-geração**
  - excesso de espúrios que o sistema produz (só em cenários absolutos).

$$Sobre-geração = \frac{\sum EMs_{espúrios}}{\sum EMs_{saída\ \ do\ \ sistema}}$$

  - Para Morfologia: **Sobre-especificação**

$$Sobre-especificação = \frac{\sum p_{Correcto\ ,sobre-esp.} + \sum p_{ParCorr\ ,sobre-esp.}}{\sum EMs_{com\ \ class.\ \ morfológica}}$$

Medidas de HAREM

- **Sub-geração**
  - mede o faltou ao sistema analisar / atribuir, em relação à Colecção Dourada.

$$Subgeração = \frac{\sum EMs_{em\ \ falta}}{\sum EMs_{CD}}$$

Erro Combinado (já não usada)

- Objectivo: Combinar a sobre-geração e sub-geração, numa medida de teor de erros do sistema

$$Erro\ \ Combinado = \frac{\sum EMs_{em\ \ falta} + \sum EMs_{espúrias} + \sum (1 - p_{Par.Cor})}{\sum (EMs_{saída\ \ do\ \ sistema} \cup EMs_{saída\ \ do\ \ sistema})}$$

Cenários de HAREM

- **Absoluto** – Pontuações de classificação são calculadas em relação a todas as EMs da CD.
- **Relativo** – Pontuações de classificação são calculadas em relação às EMs identificadas total ou parcialmente correctas pelo sistema.

Significado

Absoluto: avalia o sistema em relação ao universo das EMs na CD que seriam possíveis de classificar, identificadas ou não identificadas.

Relativo: avalia o sistema em relação ao conjunto de EMs que o sistema realmente "viu".

Cenários de HAREM

- **Total** – são consideradas todas as categorias de EMs da CD.
- **Selectivo** – são consideradas parte das categorias de EMs da CD.

Significado

Total: avalia o sistema em relação a todas as EMs incluídas na CD.

Selectivo: avalia o sistema em relação apenas às categorias de EMs que este se propõe classificar. Ignora as EMs que o sistema não pretende classificar

Validação da Plataforma HAREM

- Usabilidade da Plataforma
  - Fácil de usar? Permite aos participantes fazer as suas próprias avaliações?
- Pontuações HAREM:
  - Adequadas em cada tarefa?
  - Situações de EMs com erro na CD (atributo META=ERRO)
    - como pontuar?
  - Sobre-geração, Sub-geração, Sobre-especificação...
    - Fornecem informação útil? P, A e F bastam?

Validação da Plataforma HAREM

- Pontuações HAREM
  - F é boa medida? Adequada para comparar sistemas?
- Medidas HAREM (Classificação Semântica):
  - Intuitivas? Complicadas?
  - Contribuem para mostrar os pontos fortes / fracos dos sistemas na classificação?
- Cenários HAREM:
  - Opinião dos participantes
  - Total/selectivo: Sistemas selectivos portaram-se melhor nas 'suas' categorias?

Estrutura da Apresentação

- Introdução (Proposta da Tese)
- Motivação (Proposta da Tese)
- Objectivos (Proposta da Tese)
- Metodologia HAREM
  - Colecção Dourada, Directivas
  - Trabalho em progresso
- Plataforma HAREM
  - Arquitectura, Medidas
  - Trabalho em progresso
- Iniciativa HAREM
  - HAREM, MiniHAREM
  - Trabalho em progresso
- Resultados
- Progresso da tese

Iniciativas HAREM

- **HAREM**: 14 a 16 de Fevereiro de 2005
  - 10 participantes de 6 países
  - 15 saídas oficiais + 3 extra-oficiais
- **MiniHAREM**: 3 a 5 de Abril de 2006
  - 5 participantes de 2 países
  - 20 saídas oficiais
- Como foi realizada:
  - Distribuição da Colecção HAREM, sem EMs
  - 48 horas para devolver a colecção, anotada automaticamente

Análise às Iniciativas HAREM

- Desempenho dos sistemas...
  - Quais os géneros / variantes mais fáceis / mais difíceis de processar?
  - Parece-me que sistemas BR são melhores em BR, sistemas PT melhores em PT. Porquê?
  - Quais as categorias / tipos EMs mais difíceis de identificar / classificar morfologicamente / classificar semanticamente?
  - Como se portaram os sistemas com as EMs de categorias mais difíceis / mais vagas?

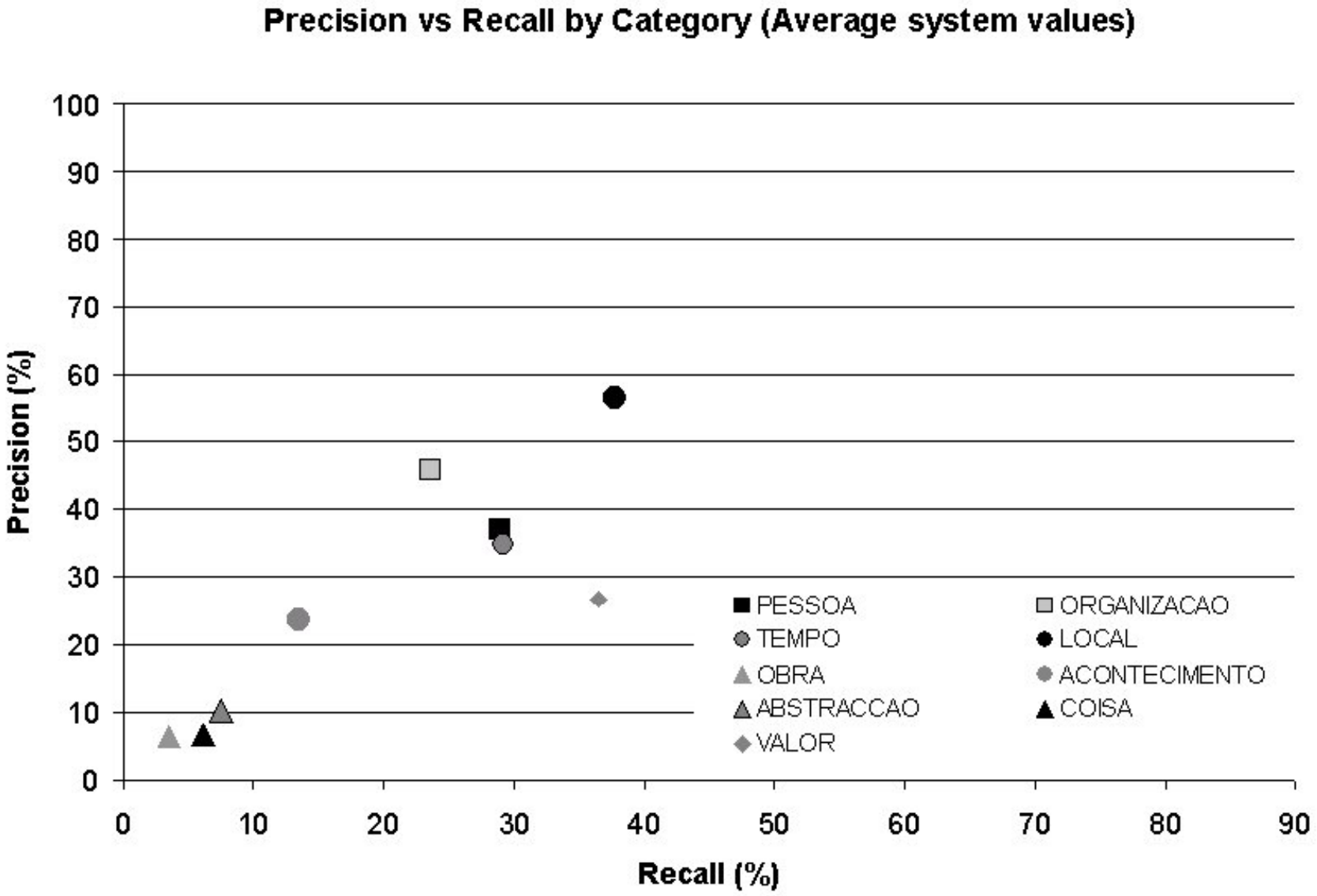
Análise às Iniciativas HAREM

- Comparação de sistemas...
  - Qual o melhor? O melhor foi em todas as categorias? Géneros? Variantes?
  - Sistemas 'selectivos' melhores que sistemas 'gerais'?
  - As CDs são ambas representativas? Calcular o intervalo de confiança para poder dizer que “Sistema A melhor que Sistema B”
  - Comparar com uma 'baseline' (conjunto de EMs que todos identificam / classificam com facilidade)

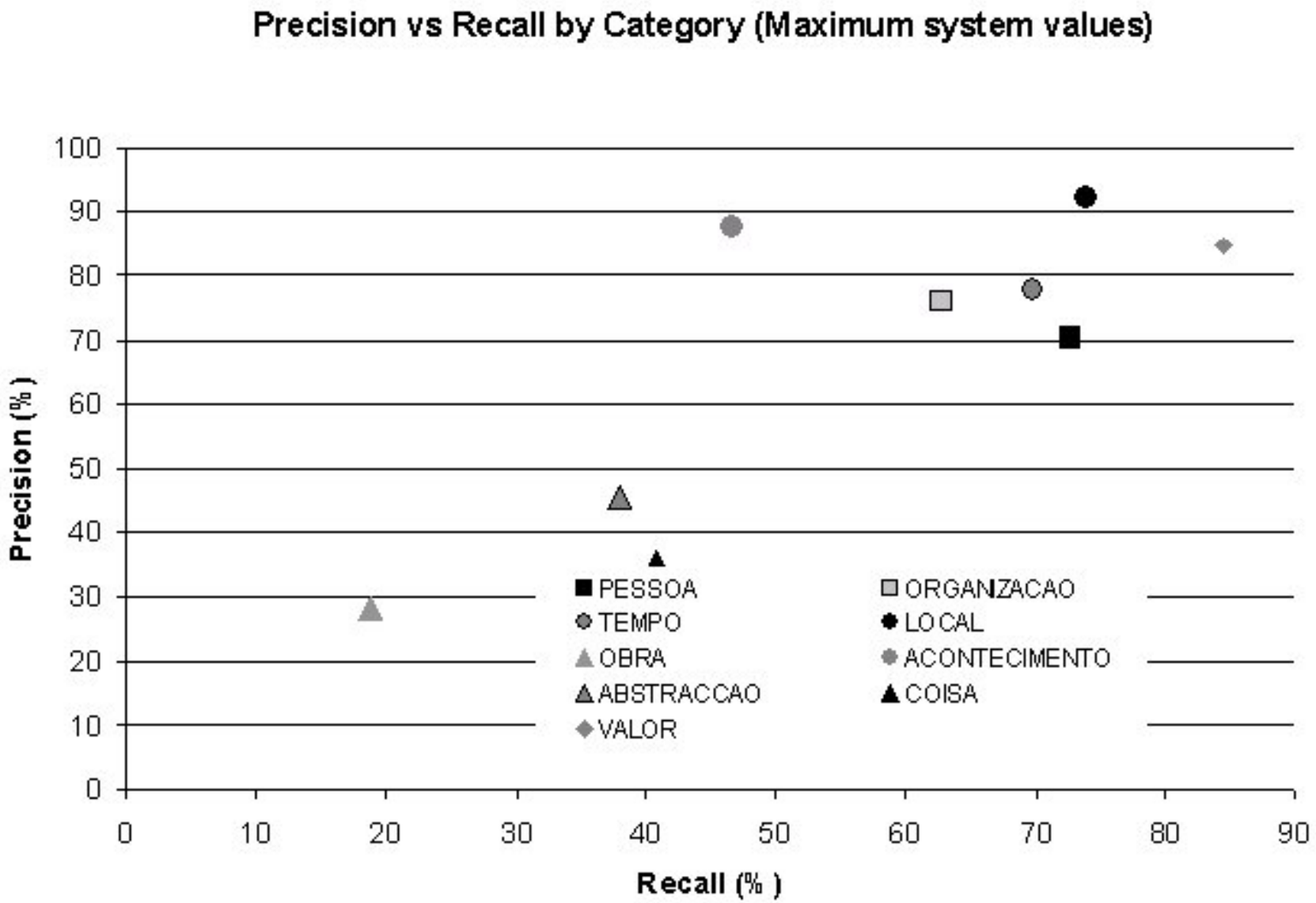
Análise às Iniciativas HAREM

- Evolução dos sistemas...
  - Houve evolução ao longo do tempo?
  - Porquê? Porque não?
- Várias comparações interessantes:
  - Comparar <Sistema n> de <Ano a>, usando <Directivas d>, segundo a <CD c>
  - n=10, a=2, d=2, c=2 (ou +)...
- Verificar impacto da evolução das directivas
- Qual a estratégia dos melhores sistemas? Se quiser construir de raiz um sistemas REM novo, o que posso aprender do HAREM?

Resultados Iniciais (HAREM 2005)



Resultados Iniciais (HAREM 2005)



Estrutura da Apresentação

- Introdução (Proposta da Tese)
- Motivação (Proposta da Tese)
- Objectivos (Proposta da Tese)
- Metodologia HAREM
  - Colecção Dourada, Directivas
  - Trabalho em progresso
- Plataforma HAREM
  - Arquitectura, Medidas
  - Trabalho em progresso
- Iniciativa HAREM
  - HAREM, MiniHAREM
  - Trabalho em progresso
- Resultados
- Progresso da tese

Resultados esperados (Proposta da Tese)

- Uma **metodologia nova**
  - Validada pela comunidade científica
  - Satisfaça os requisitos de REM
  - Base para medição de sistemas REM
  - Inspiração para outras iniciativas semelhantes
- **Colecções** de textos ricamente anotados
  - Importante para a evolução de sistemas REM (avaliação, colecção treino, ...)
  - Não representam o 'objectivo supremo', mas permitem aferir o Δ que falta.

Resultados da Tese

- **Software** de avaliação
  - Acessível, gratuito, bem documentado
  - Permite recrear ambientes de avaliação anteriores, para comparar novos sistemas
- **Caracterização** do estado da arte
  - Análise crítica das controvérsias geradas pelo HAREM
  - Apresentar o estado actual de REM em PT
  - Delinear aspectos a melhorar no futuro de avaliações em REM

Progresso da Tese – Linhas guia

- Tese irá documentar o HAREM...
- ...mas também validar o trabalho, e caracterizar o estado de REM em português
  - Trabalhar com as CDs e as saídas
  - Avaliar a avaliação
    - Colecções, Métricas, Medidas, Tarefas propostas, Cenários, Anotações, Categorias,...
    - Adequados? Realistas? Validados?
    - Permitiu caracterizar REM em PT?
    - **Permitiu melhorar os sistemas?**

Progresso da Tese – Linhas guia

- Ter sempre presente:
  - Onde é que o HAREM foi uma contribuição científica de grande valor para a área onde se insere (REM)?
  - Medir a contribuição da avaliação na evolução da área (no caso, REM em PT)
- Tese deve mostrar claramente esses pontos

Progresso da Tese

- "Material" a trabalhar:
  - 2 avaliações
  - 2 conjuntos de directivas
  - 2 Cds (com várias anotações)
  - 10 participantes
  - 38 saídas
  - 5 sistemas "repetentes"
  - Plataforma de avaliação
  - *n*º relatórios de desempenho dos sistemas
- Comunidade activa, artigos publicados (LREC 2006, PROPOR 2006)
- Encontro HAREM (importante para a tese)

Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas

- Fim
- Obrigado pela atenção.
- Questões?

Simpósio Doutoral da Linguateca

Por: Nuno Cardoso

Orientadores: Eugénio de Oliveira (FEUP)
Mário J. Silva (FCUL)