Reconhecimento de entidades mencionadas

Diana Santos
Linguateca
www.linguateca.pt
Palestra na PUC Rio 18 de Maio de 2006

Estrutura

- Qual o problema?
- História
- Aplicações
- Uma digressão sobre anotação gramatical
- Exemplos de ataque
- Avaliação conjunta
- Linguateca
- HAREM
- Futuro

REM, *reconhecimento de entidades mencionadas*

- Identificação e classificação de nomes próprios (e expressões numéricas) em texto -- em português

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

↓

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

O que é?

- É uma espécie de primeira passagem num texto para ter ideia do seu conteúdo...
- Semântica "light"
- Um pré-processamento dos textos com informação que os "agarra" ao mundo
- Uma ajuda a toda e qualquer tarefa de PLN...

Para que serve? Aplicações em que dá jeito:

- IR: indexar e buscar, visualizar
- TA: traduzir como deve ser
  - Rio de Janeiro
  - Prestes
- Análise sintáctica
  - foi a Lisboa de TGV
  - foi a Maria de Adidas para a festa
- Síntese e reconhecimento de fala
  - PUCi, TAP, IPO, Universidade de Aveiro
- Sumarização

História

- Iniciada em 1995 na MUC 6
- Subtarefa de Extracção de Informação
- MET 1996, MUC-7, MET-2
- CoNNL 2002 e 2003
- Reformulada no ACE (entidades e não nomes)
- Estendida e especificada no TERN (expressões temporais)
- Vários sistemas para outras línguas, e ontologias/almanaques multilingues
- HAREM

Digressão sobre "POS tagging" e REM

- Desambiguação da classificação gramatical, ou da categoria semântica
- Motivados pela engenharia, pelo processamento, são um primeiro passo
- Simplificação de um problema maior: análise sintáctica, semântica
- Avaliação dependente da aplicação maior em que estão inseridos

- Léxicos/almanaques que têm várias possibilidades -> aplicados ao contexto, para escolher uma... ou várias
  - *canto* é N ou V
  - *Porto* é equipa ou local
- Limitações intrínsecas nas próprias metodologias

Avaliação conjunta

- Concordar numa tarefa e discutir os promenores em conjunto
- Criar um cenário de avaliação
  - medidas
  - recursos
  - procedimento
- Comparar o desempenho dos vários sistemas, obtendo o estado da arte
- Tornar públicos os recursos, programas e as saídas dos sistemas para
  - validação externa
  - pesquisa tanto sobre a tarefa como sobre a metodologia de avaliação
  - organização de avaliações conjuntas futuras
  - treino de novos actores/participantes

Vantagens de uma avaliação conjunta

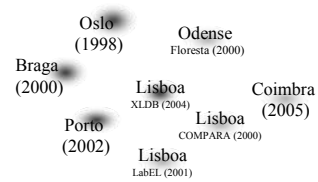
- Acordo sobre detalhes que geralmente tornam as medidas individuais de avaliação incomensuráveis
- Aumentar a visibilidade de uma tarefa particular, dos seus problemas e soluções: construção de comunidade
  - vários novos sistemas nasceram com o HAREM
- Produzir muita documentação que de outra forma nunca seria produzida
  - HAREM directivas; discussão de problemas morfológicos; discussão sobre questões de RAP no CLEF
- Pode dar origem a "baselines" (mínimos garantidos) e recursos (sistemas, almanaques) para trabalho futuro

Linguateca, um projecto para o português

- Um centro de recursos distribuído para o processamento do português
- Projecto POSI (2000-2006) proposto pela FCCN
- Primeiro pólo no SINTEF ICT, Oslo, 2000 (após o projecto *Processamento Computacional do Português*, 1998-2000)

modelo IRA
- **I**nformação
- **R**ecursos
- **A**valiação

www.linguateca.pt



A *Linguateca* em poucas palavras

- > 1000 atalhos Mais de 2 milhões de visitas ao nosso sítio Web
- AC/DC, CETEMPúblico, COMPARA Floresta… Recursos consideráveis para o processamento da língua portuguesa
- ***Morfolimpíadas*** A primeira avaliação conjunta para o português (2002-2003) seguida pelo **CLEF** (2004, 2005, 2006) e pelo **HAREM**

- Recursos públicos
- Incentivar a pesquisa e a colaboração
- Medição e comparação formal
- Uma língua, várias culturas
- Cooperação usando a Internet
- Não adaptar aplicações do inglês

Motivação para o HAREM

- Estamos apenas a fazer o mesmo que já se fez, mas agora para português?
- Ou existem também questões científicas e de engenharia válidas a que podemos responder com esta actividade?

- Tentarei convencer a audiência de que
  - É possível fazer ciência e engenharia para o português que sejam melhores do que as que foram feitas para o inglês
  - embora o HAREM tenha sido feito de raiz para o português, como metodologia inovadora pode ser igualmente aplicado ao inglês ou a outra língua qualquer

É a mesma tarefa? "Só" português...

- Uma língua ser diferente é relevante?
- É só mudar os módulos (atomizador, ortografia) e os recursos (almanaques)? Adaptações menores...
- **Ou** uma língua diferente tem desafios diferentes? Assuntos diferentes sobre os quais as pessoas falam, convenções tipográficas diferentes, diferentes conceptualizações do mundo...
- Isto é uma questão que só pode ser resolvida empiricamente... experimentando ver como é para o português e depois comparando

A mesma tarefa? Questões metodológicas

- Qual o conjunto de classificações que nos interessam?
- Como conseguir acordo na sua interpretação?
- É relevante a extensão a outros géneros?
- O conceito de *entidade mencionada* foi delimitado da mesma maneira? Os critérios operacionais são os mesmos?...
  - identificação parcial
  - proximidade ontológica
  - erros ortográficos, variantes diferentes
- A extensão a outros tipos de classificação é relevante?
- Como tratamos da vagueza, e da discordância (efeito de tecto)

Qual a dificuldade de REM?

- O mesmo nome próprio em contextos diferentes...
  O Brasil venceu a Copa (PESSOA $_{GRUPO}$), O Brasil assinou o tratado (ORGANIZACAO $_{ADMINISTRACAO}$), O Brasil tem muitos rios (LOCAL $_{ADMINISTRATIVO}$), Por amor ao Brasil (ABSTRACCAO $_{IDEIA}$), ...
- Ou um nome diferente que inclui um igual... *Camilo Castelo Branco*
- Nem sempre é fácil classificar
  - ***Guimarães** tinha muito poder junto do governo naquele tempo*
  - *Caros amigos dos **Bombeiros***
  - *disse ontem em entrevista à revista **Playboy***
  - *o certificado **ISO-9001** atestou seu nível de qualidade internacional*
  - *o **Brasil** da metade do século XIX não diferia muito da...*
  - *as três repúblicas que surgiriam da divisão da **Bósnia***
  - *Hoje a **Sé** está completamente diferente por dentro*

Qual a dificuldade de REM? (cont.)

- Nem todas as ocorrências são de identificação igualmente fácil
  - *licenciada pelo Ministério da Indústria do Governo cessante*
  - *doação de terras a senhores da nobreza, concretamente com as Honras de Cardoso, de Cantim, de Fonseca ...*
  - *tirada dos Jardins deste Palácio, que era Episcopal, depois passou para Biblioteca Pública e depois para a Universidade do Minho*
  - *Eu não posso deixar de louvar a atitude de V.Exa., prestando assim esses informes à Casa,*
  - *de acordo com as Convenções das Nações Unidas*
  - *para a realização de uma História da Imprensa em Macau*
  - *não herdei a vontade de ser Monárquico*
  - *lutou contra a Ditadura de João Franco*
  - *pegar avião na ponte Rio-São Paulo*

Critérios de delimitação

- Em abstracto, extrair tudo o que tem um nome, e atribuir-lhe a classificação correcta em contexto
- Primeiro problema: muitos nomes fazem parte de expressões maiores
  - *constante de Planck*
  - *ministro da Defesa*
  - *pasta dos Negócios Estrangeiros*
  - *dona da barraca das farturas da Feira Popular*
- Segundo problema: os nomes podem ser compositionais e como tal referir coisas diferentes simultaneamente
  - *Centro de Lógica e Computação do Departamento de Matemática do Instituto Superior Técnico*

Critérios de delimitação (cont.)

- Terceiro problema: os names não aparecem sempre completos
  - *a Revolução de 30 e a de 33*
  - *o ministro da Educação e a da Ciência*
  - *a Santa Casa*
- Quarto problema: as maiúsculas são quase aleatórias!
  - *que assolam a freguesia de Ferreiró -- um bastião **Socialista** --*
  - *o Pinto Machado que quis fundar a **faculdade de Medicina** e que agora está à frente.*
  - *diz ela. (Do artigo **Fonte da juventude**, publicado em Veja, 25 de julho de 1990*
- Quinto problem: acontecem erros...
  - *cuja verba ronda os **150 ecudos por metro quadrado***
  - *Quantos anos esteve em **Biblau** ?*

HAREM: a primeira avaliação conjunta em REM em português

- Processo
  - Concordar nas categorias e nas tarefas
  - Compilar um recurso dourado (anotado manualmente com EMs)
  - Desenvolver uma arquitectura de avaliação para a comparação automática de sistemas sobre uma colecção grande
  - Produzir resultados divididos por vários critérios
- O acontecimento propriamente dito
  - Três tarefas: identificação, classificação morfológica e semântica
  - A própria AC decorreu a 14-16 Fev. 2005: 10 participantes (5 países), 18 "runs"
  - Vencedores diferentes em diferentes medidas (resultados em Out 2005)
  - O encontro do HAREM vai ser a 15 de Julho de 2006
  - Repetição do HAREM (mini-HAREM) em Abril 2006 para estudar a relevância estatística e apreciar o progresso dos sistemas

Três eixos principais

- Compilar a colecção dourada: a anotação certa, e como exprimi-la
- Desenvolver o ambiente de avaliação (um conjunto de módulos gerais com diversas opções para experimentar várias maneiras de ordenar sistemas e de lidar com estes problemas, etc.)
- Perceber e interpretar os resultados

- As três coisas estão obviamente ligadas!

As categorias contempladas pelo HAREM

A partir da observação dos vários textos e de outras fontes

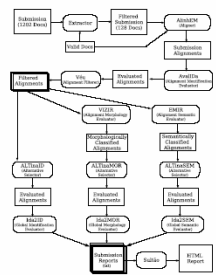
- PESSOA
- ORGANIZACAO
- LOCAL
- TEMPO
- OBRA
- ABSTRACCAO
- ACONTECIMENTO
- COISA
- QUANTIDADE, VARIADO

Os tipos

- Tentamos apenas distinguir subcategorias com motivação linguística
- PESSOA
  - INDIVIDUAL
  - CARGO
  - MEMBRO
  - GRUPOIND
  - GRUPOCARGO
  - GRUPOMEMBRO
- OBRA
  - (PRODUTO)
  - ARTE
  - PUBLICACAO
  - REPRODUZIDA
- ABSTRACCAO
  - DISCIPLINA
  - ESTADO
  - ESCOLA
  - OBRA
  - MARCA
  - PLANO
  - IDEIA
  - NOME

- **ABSTRACCAO**
  - DISCIPLINA
  - ESTADO
  - ESCOLA
  - OBRA
  - PLANO
  - IDEIA
  - NOME
- **OBRA**
  - ARTE
  - REPRODUZIDA
  - PRODUTO
  - PUBLICACAO
- **ACONTECIMENTO**
  - EFEMERIDE
  - ORGANIZADO
  - EVENTO
- **COISA**
  - OBJECTO
  - SUBSTANCIA
  - CLASSE
- **ORGANIZACAO**
  - INSTITUICAO
  - ADMINISTRACAO
  - EMPRESA
  - SUB
- **TEMPO**
  - DATA
  - HORA
  - PERIODO
  - CICLICO
- **PESSOA**
  - INDIVIDUAL
  - GRUPOIND
  - CARGO
  - GRUPOCARGO
  - MEMBRO
  - GRUPOMEMBRO
- **LOCAL**
  - GEOGRAFICO
  - ADMINISTRATIVO
  - VIRTUAL
  - ALARGADO
  - CORREIO
- **VALOR**
  - MOEDA
  - CLASSIFICACAO
  - QUANTIDADE
- **VARIADO**
  - OUTRO

A arquitectura de avaliação



Medidas usadas no HAREM

- Para a tarefa de identificação
  - precisão: (número de EMs correctas + $\Sigma_i 0.5*(nc_i / nd_i)$)/EMs identificados
  - abrangência (recall): (número de NEs correctas + $\Sigma_i 0.5*(nc_i / nd_i)$)/número de EMs na CD escolhendo ALT that maximizem a medida-F
- Para as tarefas de classificação
  - Dois cenários: relativo (só contando as EMs correctamente identificadas) e absoluto
  - Escolhendo independentemente a ALT que maximiza a medida-F
  - Lidando com casos A|B e ?
  - 4 tipos de classificação semântica: plana, só categorias, ou tipos, e combinada
  - 3 escalas para a classificação morfológica: número, gênero, combinada
  - Peso para parcialmente identificadas: nc_i / nd_i (sem), 0.5 (morf) se no início

Exemplo de cálculo das medidas

<ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S">Departamento de Cultura Científica do Centro Acadêmico Pedro Nunes</ORGANIZACAO>

- alinhado com

<ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S">Departamento de Cultura </ORGANIZACAO>

<ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S">Científica do Centro Acadêmico Pedro Nunes </ORGANIZACAO>

- Identificação: 0.17 e 0.33
- Classificação semântica: 0.34 e 0.66
- Classificação morfológica: 0.5 e 0

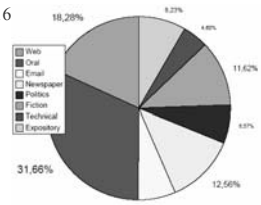
Comparação com as Morfolimpíadas e CLEF

- Texto seguido: Todos os casos são classificados, nas Morfolimpíadas escolhemos casos morfologicamente interessantes
- O HAREM permite uma avaliação quantitativa melhor do desempenho real dos sistemas
- Mas: muitos casos "estranhos" tiveram de ser tratados
- É mais fácil correlacionar EMs e gênero textual do que morfologia e gênero; mas é mais difícil comparar variantes

- Comparação com QA@CLEF: só um gênero (jornalístico), grande variedade de saída, 200 perguntas são menos representativas
- Comparação com CLEF adhoc: amontoar ("pooling")

Tamanho do recurso e sua constituição

- Se quiserem usar a CD para treinar algoritmos de aprendizagem automática (machine learning)...
- Colecção HAREM: 520 mil palavras; aprox. 40 mil EMs
- Colecção dourada 2005+2006: 16
- Cada texto está marcado com variante (PT, BR, outra) e gênero textual



O futuro do HAREM, Maio 2006

- Estamos a organizar o **Primeiro Encontro** a 15 de Julho de 2006

- apresentar os resultados estatísticos (mini-HAREM)
- apresentar os sistemas e seus comentários ao HAREM
- decidir, em conjunto, o formato do próximo HAREM
- novas tarefas: sem maiúsculas, co-referência, RAP/RI?
- novos participantes: de GIR, de extracção de ontologias, semântica, ...

- tratamento dos dados públicos para pesquisa matemática