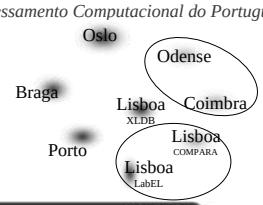
A Linguateca, e em especial a Floresta, o PAPEL e o HAREM

Diana Santos
Universidade de Coimbra, 4 de Abril de 2007

Linguateca, um projeto para o português

- Centro de recursos -- distribuído -- para o processamento computacional da língua portuguesa
- Projecto financiado pela FCT/POSI (2000-2006), POSC (2007-2008)
- Primeiro pólo em Oslo desde 2000 (actividade no SINTEF começou em 1998 com o projeto *Processamento Computacional do Português*)

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



Linguateca num relance

- > 1000 links Mais de 2.500.000 visitas ao sítio
- AC/DC, CETEMPúblico, COMPARA … Recursos consideráveis para o português
- *Morfolimpíadas* A primeira avaliação conjunta para português, seguida pelo CLEF e pelo HAREM

- Recursos públicos
- Incentivar a investigação e colaboração
- Medida e comparação formal
- Uma língua, muitas culturas
- Cooperação usando a Web
- Não à adaptação directa das aplicações para o inglês

A origem da Linguateca

- Resultado da participação no Livro Branco, que identificou
- Problemas: falta de ...
  - recursos públicos
  - cooperação entre os grupos, Brasil e Portugal
  - avaliação
  - esforço na manutenção e disponibilização de recursos
- Soluções: Projeto piloto dedicado à
  - Criação de recursos públicos (desenvolvimento, questões legais, etc.)
  - Organização de avaliações conjuntas
  - Criação de um portal dedicado à área
- Em rede (juntando mão-de-obra a grupos de investigação de acordo com os pressupostos da Linguateca)

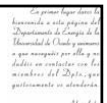
Alguns objectivos da Linguateca

- Fazer com que o PLN do português seja tão qualificado como o das outras línguas
- Impedir que as pessoas continuassem a trabalhar em PLN do inglês com a desculpa de que não havia recursos para o português
- Evitar que os grupos deitassem fora (ou guardassem secretamente) os seus recursos em vez de os disponibilizar, ajudando-os e contribuindo para essa tarefa
- Conseguir colaboração entre os vários países de língua portuguesa para tratarem todas as variantes e não só a "sua"
- Medir o progresso em várias áreas, cimentando e incrementando a colaboração entre os vários actores (avaliações conjuntas)

Alguns resultados: Informação

- Portal constantemente actualizado, www.linguateca.pt
- Catálogo de recursos, actores e publicações
- Resposta a todos os utilizadores
- Manutenção de um repositório
- Documentação sobre os recursos criados pela Linguateca
- Informação sobre as avaliações conjuntas
- Publicações no âmbito da Linguateca

Alguns resultados: Recursos



- Serviços na Web para dar acesso a corpora e ferramentas
  - AC/DC (Acesso a corpora/Disponibilização de corpora)
  - COMPARA
  - Esfinge
  - SIEMÊS
- Várias ferramentas
  - Atomizadores e separadores de frases
  - Sistemas de REM
  - Alinhadores à palavra
- Criação de corpora, colecções, ou dados para distribuição
  - CETEMPúblico, CETENFolha, WPT03
  - GKB (*Geographic Knowledge Base*) e Geo-Net-PT01
  - REPENTINO (REPositório para reconhecimento de ENTidades com NOme)
  - Colecção douradas: CHAVE, Morfolimpíadas e HAREM

Alguns resultados: Avaliações conjuntas

- Selecionar uma área
- Criar recursos para a avaliar, em consenso com os participantes
- Criar programas de avaliação
- Organizar um evento
- Publicar os resultados

- Morfolimpíadas (análise morfológica sem contexto)
- CLEF (RI cruzada e Resposta Automática a Perguntas - RAP)
- HAREM (Reconhecimento de Entidades Mencionadas - REM)

Quem é o público alvo da Linguateca?

- Pessoas envolvidas no desenvolvimento de aplicações de processamento de linguagem natural (PLN)
- Consumidores de dados (linguistas)
- Utilizadores de programas que envolvem PLN

HAREM é **A**valiação de **R**econhecimento de **E**ntidades **M**encionadas

- Problema: Identificar e classificar nomes próprios em contexto em texto em português, dada uma tabela inicial e quanto à morfologia
- A forma mais básica de semântica
- Três tarefas:
  - Identificar uma EM
  - Classificá-la morfologicamente
  - Classificá-la pelo tipo de entidade a que se refere
- Organização de uma avaliação conjunta
  - Criar uma colecção dourada anotada com as soluções
  - Fornecer aos sistemas participantes grandes quantidades de texto (coleção HAREM)
  - Avaliar (através da comparação automática com as soluções)

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

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** ?*

Resultados do HAREM

- Colecção dourada pública
- Arquitectura pública (programas em Perl e Java)
- Dez sistemas prontos a atacar o problema de REM em português (quantos haveria sem o HAREM?)
- Uma primeira medida do estado da técnica em português
- Objectivos científicos
  - Medir a dificuldade do problema para o português
  - Pôr em relevo as especificidades do português
  - Verificar se as EMs podiam ser discriminadoras de gênero textual

Porquê usar os nossos recursos

- Já existem (e deram trabalho)
- Sendo públicos, permitem comparar abordagens, estudos e dados
- Ao usá-los e dando sugestões, está a melhorar um recurso para todos
- Tem apoio na sua utilização
- Ainda há muito a fazer até esgotá-los

Porquê participar numa avaliação conjunta?

- Para ajudar na especificação do resultado final
- Para garantir um nível de qualidade
- Para conhecer de perto os vários sistemas
- Para participar na comunidade em torno de um dado problema
- Para saber o nível de dificuldade envolvido

PAPEL

- Palavras Associadas Porto Editora Linguateca
- Uma ontologia lexical criada a partir da versão electrónica do Dicionário da Porto Editora
- A ser disponibilizada gratuitamente à comunidade científica pela Linguateca

- objectivo principal/primordial do pólo de Coimbra da Linguateca
- inspirado pela WordNet e pela MindNet e por muitas outras "ontologias" provenientes de dicionários

Competência e desempenho

- O dicionário representa uma sistemarização da competência linguística
- Os corpora representam uma amostra do desempenho de uma comunidade linguística

- *gosto / amo*
- *não desgosto / * não desamo*
- *não me importo / importo*
- *cair em desuso*
- *não conseguir - desconseguir*

Objectivos de um dicionário

- Tudo o que serve para alguma coisa tem um determinado objectivo
- Há vários tipos de dicionários muito diferentes (com públicos muito diferentes)
  - dicionário escolar
  - dicionário de ensino de português língua estrangeira
  - dicionário de língua geral
  - dicionário etimológico, ...
- Um dicionário organiza-se pelas palavras (forma)
- Um tesauro organiza-se pelos sentidos
- Uma ontologia lexical é uma tentativa de definir/representar sentidos através de um conjunto de palavras (formas)

Dicionário como objecto matemático

- constitui um grafo: todas as palavras são definidas por outras palavras

o que significa que:

- se podem agrupar (cluster)
- podem definir-se relações hierárquicas
- há palavras "aristocráticas" (Wilks)
- pode atribuir-se uma importância baseada no seu uso (pagerank)
- pode definir-se "proximidade lexical"

- sintaxe "simples"

Problemas: homonímia e polissemia

- uma palavra tem apenas uma definição **muito raramente**
- o que fazer quando tem mais do que um

- homonímia: coxa, junta, venda, revista, banco
- polissemia: quando estão relacionadas...

- 1.a questão: como distinguir automaticamente?
- 2.a questão: quantos sentidos relacionados e como? a que nível é que os lexicógrafos desceram/subiram? quantas meta-relações encontraram?

Problema: o que queremos no PAPEL?

- Definir as especificações de forma a guiar o processo de construção
- Conhecendo as aplicações para as quais são usadas as outras ontologias
  - semelhança entre palavras
  - relações semânticas: antonímia, hiponímia, meronímia, ...
  - distância entre palavras
  - relacionamento com outras palavras
- Resposta automática a perguntas (RAP)
- Análise sintáctica e semântica de um texto
- Recolha de informação (restruturação da pesquisa)
- Melhoria dos próprios dicionários

Frequências PAPEL

acto	5783	qualidade	2250
pessoa	4207	fazer	2215
efeito	4205	parte	2180
família	3052	onde	2015
muito	2594	grande	1878
forma	2530	conjunto	1776

Exemplos de polissemia sistemática

- acto/colocação [enquadramento, distribuição, enquadramento, relegação, substituição, sobreposição, imposição, seriação, acentuação, marcação, marcação, anteposição, internamento]
- acto/declaração [confissão, licitação, manifesto, assentamento, contradeclaração, preconização, ressalva, imputação, depoimento, protesto, confissão, aclamação, profissão, protesto]
- fruto/planta [marubá, pimento, quivi, murtinho, andu, mastruço, beringela, noz-vómica, carriço, doce-amarga, pimenta, chuchu, zacum, baunilha, framboesa, ...]
- atitude/doutrina [iconoclastia, objectivismo, ateísmo, dogmatismo, moralismo, estetismo, idealismo, cepticismo, liberalismo, moralismo, ...]

Floresta Sintá(c)tica

- Um corpo de frases analisadas sintacticamente, revistas intelectualmente por linguistas
- Gramática empírica (desempenho e não competência)
- Para estudos quantitativos
- Para treino de analisadores sintácticos
- Para avaliação

- Talvez o mais rico de todos os recursos criados pela Linguateca
- Decididamente o menos usado

Problemas da Floresta

- Muita informação
- Difícil de usar
- Pequeno demais para generalizações avassaladoras
- Muitas opções controversas (equipa de 4)
- Muitos erros/inconsistências: errar é humano

- Ainda não há muitos sistemas que façam coisas parecidas...
- Mais uso internacional do que na comunidade do PLP
- Serviu para melhorar consideravelmente o PALAVRAS

Melhorias à Floresta

- Em quantidade
- Em qualidade
- Em diversidade
  - Entidades mencionadas
  - Códigos semânticos
  - Formato de gramática categorial
  - Relações anafóricas
- Na usabilidade das várias interfaces
- Na clareza da documentação
- Em exemplos concretos do seu uso

Pólo de Coimbra: intenções

- Desenvolvimento harmonioso do PAPEL, da Floresta, e de um novo HAREM
- Idealmente aproveitando o trabalho e os recursos de cada "subprojecto" para melhorar cada um
- Adicionando bolseiros e projectos relacionados (POP?) para aumentar a massa crítica e para ter testadores e utilizadores directos
- Integrado no projecto maior da Linguateca, que tentará fazer o mesmo com o "resto" dos recursos e pólos: tentar que colaborem e que usem mutuamente os trabalhos desenvolvidos