Abertura científica: processamento computacional da língua portuguesa

Diana Santos
Primeira Escola de Verão da Linguateca

Interdisciplinaridade

- Uns vêm de Letras
- Outros de informática
- Mas o caminho a seguir é o mesmo, na área do processamento computacional da língua
- Não interessa onde começaram
- Professores dos 2 lados

Aulas teóricas e práticas

- Igualmente importantes!
- Exercícios práticos é que contam
- Feitos de forma a ajustar-se ao desenvolvimento de cada um
- Não há **uma** solução, mas sim problemas sobre os quais é preciso pensar, recursos que é preciso experimentar, dúvidas que é preciso ter

Nível esperado

- Alunos de doutoramento ou de mestrado
- Sabem pensar pela sua própria cabeça
- Sabem trabalhar de forma autónoma
- Sabem questionar os professores e o que aqui for dito
- Sabem argumentar
- Sabem citar e reconhecer de onde lhes vieram as ideias

Objectivos da Escola

- Pô-los em contacto com os problemas e aquilo sobre que não há consenso
- Levá-los a pensar
- Levá-los a ter uma postura crítica
- Não vamos falar em uníssono
- Vamos tentar não falar ora para linguistas ora para informáticos – tentar falar para uma audiência de alunos inteligentes

Aviso: Algumas dicotomias que podem não fazer sentido

- sintaxe vs. semântica
- knowledge-lite vs. knowledge-heavy
- automático vs. humano
- teoria vs. prática
- quantitativo vs. qualitativo
- estatístico vs. linguístico
- conhecimento do mundo vs. conhecimento linguístico

Ciência ou engenharia?

- A ciência descreve a realidade
- A engenharia modifica-a, de acordo com o que a ciência ensina
- A ciência produz hipóteses
- A engenharia faz artefactos de acordo com essas hipóteses

Estruturação do PLN

- Através das aplicações
- Através dos objectivos
- Através das áreas teóricas
  - sintaxe, semântica, fonologia, semiótica, aprendizagem automática, lógica modal, estilística, ontologia e epistemologia, estudos literários, ...

Tipos de aplicações em PLN

- LN->LN: Extrair algo
- LN->X->LN: Arrumar algo para subsequente organização
- X->LN: Ajudar a criar algo
- Extra: descobrir plágio, descobrir um autor, autentificaçãoetc.
- tarefas com sentido
  - RAP, EI, TA, RI
- tecnologias
  - análise sintáctica, POS tagging, ...
- recursos e métodos
  - dicionários, ontologias, aprendizagem automática, prospecção (data mining), indexação, ...

Aplicações concretas

- Sugerir sinónimos, no contexto de ajudar a escrever
- Encontrar contextos (exemplos) em que uma palavra é usada para se decidir se se deve aplicar ou não, escrevendo em língua estrangeira
- Confirmar uma data ou uma referência num artigo que se está a escrever

Mais aplicações concretas

- Quem é... ?
- Como é que se vai para... ?
- Que informação há sobre... ?
- Quero ver cenas de encontros ao luar

Panorâmica das áreas...

- Não há nenhum sistema de TA que traduza para português de Portugal
- Não há nenhum sistema que faça raciocínio baseado em textos em português

+ Há sistemas que respondem a perguntas e ajudam à redacção (correctores ortográficos, sintácticos e estilísticos)

+ Há sistemas de indexação para o português

Panorâmica

- A maior parte da I&D concentra-se em criar recursos ou ferramentas de aplicação laboratorial ou didáctica
- Muitas vezes apenas prova de que é possível ou corroborando uma hipótese
- A maior parte das empresas não partilham ou publicam a investigação
- O que as pessoas precisam não interessa

Dois mundos

- Resolver um problema prático não é visto como academicamente útil
- O financiamento de projectos úteis segue as normas (ou falta delas) de uma comunidade académica, cujos valores são
  - publicações
  - número de alunos

É preciso uma motivação

- Mesmo a investigação fundamental pode ter uma origem prática
- É preciso avaliar de acordo com critérios externos: pés na terra
- Inspiração e transpiração são irmãs

- Fazer alguma coisa com a língua

A Linguateca

- tenta lutar contra esta(s) corrente(s)
- produzindo recursos públicos
  - para empresas e universidades
- organizando avaliações conjuntas de sistemas à volta de uma tarefa concreta – idealmente com uma aplicação definida
- produzindo documentação em português

Um projecto para o português

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

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

Oslo 2.5
Odense 0.5
Braga 2
Lisboa Coimbra 1
XLDB 2.5
Lisboa
COMPARA 1.5
Porto 3
Lisboa
LabEL 1

www.linguateca.pt num relance

- > 1000 links Mais de 2.400.000 visitas ao site
- AC/DC, CETEMPúblico, COMPARA … Recursos valiosos para o processamento do português
- *Morfolimpíadas, parte portuguesa do CLEF, HAREM* Avaliação conjunta para o português

- Recursos públicos
- Incentivar a investigação e colaboração
- Medida e comparação formal

- Uma língua, muitas culturas
- Cooperação usando a Web
- Não à adaptação direta das aplicações para o inglês

Contacto: Diana Santos@sintef.no

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 objetivos 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 jogassem 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 atores

Resultados: Informação

- Portal constantemente actualizado
- Catálogo de recursos, atores 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
- Gestão de um fórum sobre a área

Resultados: Recursos

- Serviços na Web para dar acesso a corpora e ferramentas
  - AC/DC
  - COMPARA
  - Esfinge
  - AnELL
  - 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
  - Floresta sintá(c)tica
  - GKB (*Geographic Knowledge Base*) e Geo-Net-PT01
  - REPENTINO
  - Colecção douradas: CHAVE, Morfolimpíadas e HAREM

Resultados: Avaliações conjuntas

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

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

Quem é o público da Linguateca?

- Pessoas envolvidas no desenvolvimento de aplicações de PLN

- Consumidores de dados (linguistas)
- Utilizadores de programas que envolvem PLN

Quem é a equipe?

- Sêniores: *Diana Santos, José João Dias de Almeida, Elisabete Ranchhod, Eckhard Bick, Belinda Maia, Ana Frankenberg Garcia, Mário J. Silva, Paulo Gomes*
- Contratados para as tarefas "básicas" da Linguateca (um por pólo): *Luís Costa, Nuno Cardoso, Rui Vilela, Luís Miguel Cabral, António Silva*
- Contratados à tarefa: *Paulo Rocha, Susana Afonso, Raquel Marchi, Rosário Silva*
- Doutorandos: *Marcirio Chaves, Alberto Simões, Nuno Seco*
- Bolsistas para tarefas mais curtas: *Susana Inácio, Ana Sofia Pinto*
- Amigos/associados: *Rachel Aires, Cristina Mota, Anabela Barreiro, Luís Sarmento*

Estrutura da Linguateca

- Pólo XLDB de Lisboa: Web portuguesa, RI e RI geográfica
- Pólo do Porto: terminologia, corpora especializados, avaliação de tradução automática
- Pólo de Braga: ferramentas, tradução automática, gestão e validação de recursos
- Pólo de Oslo: organização, portal, avaliações conjuntas, RAP
- Pólo COMPARA: corpora paralelos
- Pólo Odense: floresta sintática
- Pólo Coimbra: ontologias lexicais

Investigação (para doutoramento)

- Rachel Aires: É possível categorizar os textos da Web segundo as necessidades de informação dos usuários? **(concluída Agosto 2005)**
- Marcirio Chaves: É possível gerar ontologias geográficas úteis a partir da análise de textos em português?
- Alberto Simões: É possível aumentar significativamente os exemplos usados na Tradução Automática baseada em exemplos (TABE) com o processamento inteligente de corpora comparáveis?
- Nuno Seco: Métodos de criação e de avaliação de uma ontologia lexical para o português

Investigação (para mestrado)

- Alberto Simões: Alinhador à palavra **(concluída Setembro 2004)**
- Luís Miguel Cabral: Extrator de informação na rede para o catálogo de publicações
- Nuno Cardoso: Avaliação de REM
- Rui Vilela: Extração de informação

Outra investigação na Linguateca

- Métodos de criação de uma floresta sintática (treebank)
- Métodos de RAP
- Anotação sintática do português
- Usabilidade com base nos diários (logs) de serviços na Web
- Métodos de avaliação
- Detecção de entidades mencionadas (EM)
- Criação de serviços na rede
- Avaliação de recursos
- Terminologia para indexação

Onde estamos?

- No princípio!
- Na primeira Escola de Verão
- Lançando as primeiras sementes para fora
- Começando a colher os primeiros frutos

- Alguma visibilidade a nível internacional
- Uso generalizado dos nossos recursos
- Mas ainda não chegámos ao cidadão comum

Resumindo

- Quase tudo está por fazer no processamento computacional da língua portuguesa
- Precisamos de formar muitas pessoas e transmitir as nossas mensagens
- Gozem a Escola de Verão!

Dados sobre a organização

- 604 mails sobre a Escola
- 529 mails para ou da Ana Sofia (90%)
- 5 listas de correio-electrónico
- ?? documentos envolvidos
- 400? slides criados
- 20? documentos HTML