Encontro da Linguateca 10 Anos, 11 Setembro 2008, Curia, Portugal

Microsoft | Development Center Portugal

Os recursos da Linguateca ao serviço do desenvolvimento da tecnologia de voz na Microsoft

Daniela Miguel Sales Dias

Promoted by: Microsoft Language Development Center

http://www.microsoft.com/portugal/mldc

Sumário

Microsoft | Development Center Portugal

- Introdução
- Objectivos
- Linguateca no desenvolvimento de tecnologia de voz na MS
- Conclusões

Introdução

Microsoft | Development Center Portugal

-Posição do Português no conjunto das línguas mais faladas do mundo com 235 milhões de falantes
-língua materna (com cerca de 235 milhões de falantes)
-língua oficial de 9 Estados Independentes (Angola, Brasil, Cabo Verde, Guiné Bissau, Moçambique, Portugal, São Tomé e Príncipe, Timor)

Objectivos

Microsoft | Development Center Portugal

- Mostrar a importância da Linguateca no desenvolvimento de tecnologia de voz em Português Europeu e do Brasil na Microsoft
- Elencar os pontos de aplicabilidade dos recursos disponíveis

Linguateca no desenvolvimento de tecnologia de voz na MS

Microsoft | Development Center Portugal

- Principais recursos utilizados:
  - CETNFolha
  - CETEMPúblico
  - COMPARA
  - Floresta Sintáctica e Floresta Virgem

Linguateca no desenvolvimento de tecnologia de voz na MS

Microsoft | Development Center Portugal

- Tarefas executadas sobre CETEMPublico, CETENFolha e COMPARA:
  - Selecção automática de scripts para a voice font
  - Selecção de casos de teste para validação de algoritmos de:
    - Separador de frases
    - Separador de palavras
    - Normalização de texto
    - Desambiguação de homógrafos
    - Conversão grafema-fone
  - Obtenção e generalização de padrões para criação de regras de normalização de texto
  - Obtenção de listas de frequência de léxico em certos domínios

Linguateca no desenvolvimento de tecnologia de voz na MS

Microsoft | Development Center Portugal

- Tarefas executadas sobre Floresta Sintáctica e Florestas Virgem:
  - Extracção de tags de POS e extrapolação para léxico, após o mapeamento
  - Utilização do corpus da Floresta Sintáctica para treino dos POS tagger automático

Conclusões

Microsoft | Development Center Portugal

- Há 10 anos eram muito escassos os recursos disponíveis para a comunidade científica, sobretudo em formatos inteligíveis para processamento computacional
- A Linguateca veio preencher com sucesso essa lacuna, contribuindo não só para a aproximação entre as comunidades científicas portuguesa e brasileira que trabalham em Processamento da Linguagem Natural, Linguística Computacional e áreas relacionadas, como também para a divulgação de trabalhos académicos e bases de dados que de outra forma se manteriam dispersos e dificilmente acessíveis

Microsoft | Development Center
Portugal

www.microsoft.com/portugal/mldc

Daniela Braga

Program Manager Lead

i-dbraga@microsoft.com

Microsoft®

Your potential. Our passion.™