Universidade de Lisboa
Faculdade de Ciências
Departamento de Informática

Simpósio Doutoral - Linguateca

Plano de doutoramento

Marcirio Silveira Chaves

Estrutura

- Motivação
- Introdução
- Problemas de Pesquisa
- Objetivo
- Plano
- Expectativa
- Critérios de Sucesso
- Trabalho em desenvolvimento
  - Experimentos
- Considerações Finais

Motivação

- Informação geográfica legível por máquina
- Detecção de relações entre entidades mencionadas geográficas em corpora português
- Interconexão do conteúdo em textos em linguagem natural.
- Formalização das interconexões encontradas, de forma a tornar o conteúdo legível por máquina.

Problemas de Pesquisa

- Nomes não são únicos
  - Uma mesma entidade geográfica pode receber mais de um nome

  Ex.:Vila Nova de Gaia e Gaia
- Variações na forma de escrita

  Ex.: Freixo de Espada à Cinta =
  Freixo de Espada Cinta,
  Freisco de Espada a Cinta,
  Freixo de Espado a Cinta,
  Freisco de Espada à Cinta,
  Freixo-de-Espada-à-Cinta

Problemas de Pesquisa

- Mesmo nome geográfico pode identificar diferentes entidades
  - Localidade **Sobral de Monte Agraço:**
    - parte do concelho de *Lisboa*
    - nome alternativo da localidade de **Sobral** no concelho de *Lourinhã*
- REM
  - **Castelo Branco**
    - distrito, concelho, freguesia, localidade, ou ainda rua

Objetivos

- Criar um modelo conceitual de uma ontologia geográfica sobre Portugal
- Descobrir de padrões geográficos para extrair ontologias na língua portuguesa
- Propor de uma ontologia geográfica subjacente à língua portuguesa na web portuguesa

Plano

- Construção de uma base de conhecimento geográfico a partir de fontes de informação de autoriadades
- Realização e validação de experimentos em texto utilizando técnicas de PLN
- Avaliação dos resultados

Cronograma

	2005							
Tarefa	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Estimativa da "geograficidade" de termos geográficos	X							
Estimativa da presença da GKB na Web	X							
Implementação de gramáticas específicas para expressões geográficas	X	X						
Aplicação dos padrões Hearst para o português	X	X						
Replicação parcial do KnowItAll		X	X	X				
Uso de padrões geográficos em corpora anotados (CETEMPúblico/CETENFolha)	X	X	X	X				
Escrita capítulo ontologias	X	X	X	X	X	X	X	X
Escrita capítulo detecção automática de relações semânticas a partir de texto	X	X	X	X	X	X	X	X
Descrição dos experimentos	X	X	X	X	X	X	X	X

Expectativas

- Fornecer suporte a um motor de RI
- Aumentar o nosso conhecimento da linguagem geográfica utilizada na web
- Criar de ontologias geográficas para o português (início de uma biblioteca de ontologias)

Critérios de Sucesso

- A web portuguesa antes e depois dos resultados apresentados na tese
- Aplicações utilizando as idéias, algoritmos, bases de conhecimento, etc. produzidos durante a tese
- Publicações

Experimentos

- Experimento para expansão de instâncias
  - "FeatureType" ta[l|is] como "List"
  - "FeatureType" incluindo "List"
  - "FeatureType" especificamente "List"
  - "FeatureType" principalmente "List"
  - "FeatureType" nomeadamente "List"
  - "FeatureType" como por exemplo "List"
  - "FeatureType" particularmente "List"
- Plural?
- Limite da "List"?
- **Objectivo:** Expandir o conhecimento fornecido pelas fontes de informação das autoridades geográficas

Experimentos

- Padrões Hearst
  - "tais como" no Google

Classe		Número de ocorrências
• Cidades	-	30
• Freguesias	-	9
• Concelhos	-	6
• Distritos	-	1
• Ruas	-	2

Experimentos

- **Estimativa da presença da GKB na Web**
  - WPT 03 completo
  - ~500MB
  - 50 mil páginas
- **Nomes simples**

Lisboa	**1.162.441**
Porto	699.641
Centro	**445.243**
Coimbra	314.677
Aveiro	**194.554**

- **Nomes compostos**

Diogo Dias	**7.237**
São Bento	1.297
Vila Nova	**1.122**
Santa Maria	842
Quarta Feira	**764**
Vila Real	754
São João	**747**
Viana do Castelo	634
Castelo Branco	**541**
São Paulo	530

Experimentos

- Nano-gramáticas geográficas Unitex
  - Padrões Hearst
    - Raras ocorrências no contexto geográfico

Experimentos

- Verificação da "geograficidade" em textos na web
- Comparação com as informações da Geo-Net
- Suporte à anotação geográfica do WPT03
- Suporte à atribuição de âmbitos

Considerações Finais

- Plano de doutoramento
- Trabalho em desenvolvimento