

III Simpósio Doutoral da Linguateca

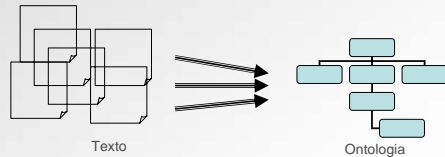
Proposta de tese de doutorado: Uma metodologia para construção de ontologias e integração de conhecimento geográfico

Marcirio Silveira Chaves

3 de outubro de 2006

Motivação

- Informação geográfica
 - é pervasiva na web
 - é transversal a vários domínios de conhecimento
 - pública não é processável (inteligível) por máquina
- Pelo menos alguma informação sobre qualquer local particular já existe em algum lugar na web



03-Out-2006

III Simpósio Doutoral da Linguateca

2

Motivação

O rio Douro (Duero, em castelhano) é um rio que nasce em Espanha, na província de Sória, nos picos da Serra de Urbión (Sierra de Urbión), a 2.080 metros de altitude e atravessa o norte de Portugal. A foz do Douro é junto à cidade do Porto. Tem 850 km de comprimento. Afluentes: Rio Paiva, Rio Sousa, Rio Tua.

wikipedia

03-Out-2006

III Simpósio Doutoral da Linguateca

3

Motivação

O rio Douro (Duero, em castelhano) é um rio que nasce em Espanha, na província de Sória, nos picos da Serra de Urbión (Sierra de Urbión), a 2.080 metros de altitude e atravessa o norte de Portugal. A foz do Douro é junto à cidade do Porto. Tem 850 km de comprimento.

Afluentes: Rio Paiva, Rio Sousa, Rio Tua.

Fatos

nome_rio(R).	comprimento_rio(R,C).
nome_rio(Douro).	comprimento_rio(Douro,850 km).
nome_rio(Paiva).	
nome_rio(Sousa).	afluente(R1,R2).
nome_rio(Tua).	afluente(Douro,Paiva).
nome_rio_es(Douro,Duero)	afluente(Douro,Sousa).
	afluente(Douro,Tua).
nascente(R,L).	foz_rio(R,L).
nascente(Douro,Sória).	foz_rio(Douro,Porto).
nome_serra(Serra de Urbión).	
nome_serra_es(Serra de Urbión, Sierra de Urbión).	
altitude_serra(Serra de Urbión, 2.080 m.).	

03-Out-2006

III Simpósio Doutoral da Linguateca

4

Agenda

- Desafios
- Trabalho realizado
- Objetivos
- Questões de pesquisa
- Contribuições
- Arquitetura do sistema de gerenciamento de conhecimento geográfico
- Sistema de extração e integração de conhecimento geográfico – SEI-Geo
- Avaliação da metodologia proposta

03-Out-2006

III Simpósio Doutoral da Linguateca

5

Desafios

- Reconhecer o conhecimento geográfico nos textos (além da tarefa de REM)
- Representar esse conhecimento em uma ontologia geográfica
- Integrar o conhecimento geográfico distribuído através dos textos ao conhecimento existente

03-Out-2006

III Simpósio Doutoral da Linguateca

6

Complexidade do problema

- Fontes de informação geográfica publicamente disponíveis são raras e a qualidade dos dados é frequentemente baixa
- A linguagem natural é vaga e ambígua
- Os **atributos** dos conceitos geográficos variam bastante. Identificá-los e classificá-los corretamente em textos é uma tarefa complexa
 - podem estar descritos de inúmeras formas
 - integração desses atributos em ontologias
- O estado da arte dos sistemas de **REM** que trabalham com textos em português nas duas edições (2005, 2006) do HAREM para a categoria Local é medida F: 0,6520 (2005) e 0,5304 (2006).

03-Out-2006

III Simpósio Doutoral da Linguatca

7

Trabalho realizado

- GKB – *Geographic Knowledge Base*
- Geo-Net-PT01
- GKB-ML
- Uso de sistemas REM
- Dimensionamento inicial da presença de informação geográfica na web portuguesa

03-Out-2006

III Simpósio Doutoral da Linguatca

8

GKB – *Geographic Knowledge Base*

- Informação geográfica administrativa integrada
- +
- Conceitos e ocorrências históricas integradas
 - Âmbitos geográficos atribuídos a sítios e domínios web
 - GOG – *Geographic Ontology Generator*

03-Out-2006

III Simpósio Doutoral da Linguatca

9

Geo-Net-PT01

- Ontologia gerada a partir da GKB
- Mais de 400.000 features geográficas
- Utilizada por
 - Motor de busca geográfico (Geo-Tumba)
 - Sistema de REM – CAGE

03-Out-2006

III Simpósio Doutoral da Linguatca

10

GKB-ML

- Ontologia geográfica composta por nomes de
 - Países
 - Cidades (> 100.000 habitantes)
 - Regiões
 - Oceanos
 - Rios
 - ...
- Mais de 12.000 nomes em 4 línguas
 - EN, PT, ES, DE
- Relacionamentos geográficos
 - Parte de, adjacência
- Bounding box e centróide
- População
- Número de ocorrências de cada nome na web

03-Out-2006

III Simpósio Doutoral da Linguatca

11

GKB-ML

- Desenvolvida especialmente para participação do pólo XLDB da Linguatca em avaliações
 - HAREM
 - Mini-HAREM
 - Geo-CLEF 2005/2006
- Geobase – Interface web para navegação dentro das ontologias

03-Out-2006

III Simpósio Doutoral da Linguatca

12



Dimensionamento inicial da presença de informação geográfica na web portuguesa

- Coleção HAREM – parte PT
 - 30% dos Locais na coleção presentes na Geo-Net-PT01
 - 25% dos Locais na coleção presentes na GKB-ML
- Tipo Administrativo
 - 38% dos Locais do tipo administrativo da coleção presentes na Geo-Net-PT01
 - 39% dos Locais do tipo administrativo da coleção presentes na GKB-ML

Objetivos

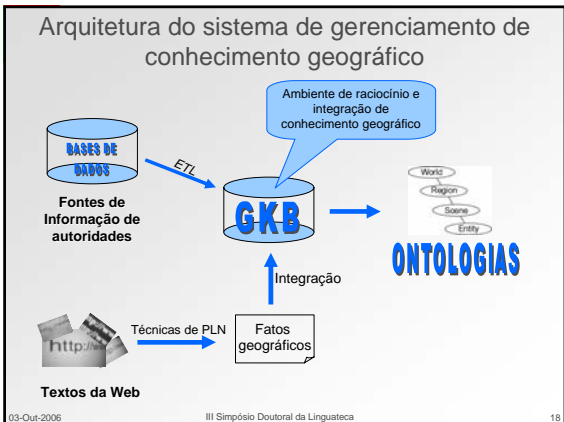
- dimensionar a “geograficidade” presente em textos web escritos em português
- derivar uma ontologia geográfica a partir de textos
- integrar o conhecimento geográfico físico adquirido dos textos em um ambiente de raciocínio geográfico com conhecimento geográfico administrativo existente

Questões de pesquisa

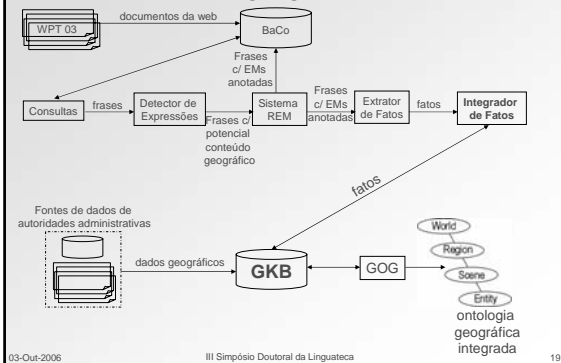
- quais os conceitos, atributos e relacionamentos geográficos presentes nos textos que podem ser representados numa ontologia?
- quais os tipos (rios, serras, ...) de ocorrências geográficas mais frequentes nos textos Web em português?
- quais são os atributos geográficos presentes nos textos Web em português?
- quais os padrões lexicais que mais indicam presença de conteúdo geográfico em textos?

Contribuições

- uma arquitetura para um sistema de gerenciamento de conhecimento geográfico
- uma metodologia para extração de conhecimento geográfico de textos, tendo como estudo de caso textos web em português, sendo o primeiro trabalho nessa língua
- uma metodologia para integração de informação geográfica
- construção e disponibilização gratuita de uma ontologia geográfica com conhecimento integrado de diversas fontes de informação



Sistema de extração e integração de conhecimento geográfico – SEI-Geo



```
<rdf:comment>Conhecimento existente</rdf:comment>
<gn:Geo_Feature rdf:ID="GEO_238">
  <gn:geo_id>238</gn:geo_id>
  <gn:geo_name xml:lang="pt">Porto</gn:geo_name>
  <gn:geo_type_id rdf:resource="#CON"/>
  <gn:info_source_id rdf:resource="#INE"/>
  ...
</gn:Geo_Feature>

<rdf:comment>Novo conhecimento integrado</rdf:comment>
<gn:Geo_Feature rdf:ID="GEO_169">
  <gn:name>
    <rdf:Bag>
      <rdf:li xml:lang="pt">Douro</rdf:li>
      <rdf:li xml:lang="es">Duero</rdf:li>
    </rdf:Bag>
  </gn:name>
  <gn:geo_type_id rdf:resource="#RIO"/>
  <gn:source rdf:resource="#GEO_120"/>
  <gn:outlet rdf:resource="#GEO_238"/>
  <gn:affluent>
    <rdf:Bag>
      <rdf:li rdf:resource="#400"/>
      <rdf:li rdf:resource="#401"/>
      <rdf:li rdf:resource="#402"/>
    </rdf:Bag>
  </gn:affluent>
  <gn:length unit="km">850</gn:length>
  <gn:info_source_id rdf:resource="#texto_web"/>
</gn:Geo_Feature>
```

03-Out-2006 III Simpósio Doutoral da Linguateca 20

```
<gn:Geo_Feature rdf:ID="GEO_400">
  <gn:geo_name xml:lang="pt">Paiva</gn:geo_name>
  <gn:geo_type_id rdf:resource="#RIO"/>
  <gn:info_source_id rdf:resource="#texto_web"/>
</gn:Geo_Feature>

<gn:Geo_Feature rdf:ID="GEO_401">
  <gn:geo_name xml:lang="pt">Sousa</gn:geo_name>
  <gn:geo_type_id rdf:resource="#RIO"/>
  <gn:info_source_id rdf:resource="#texto_web"/>
</gn:Geo_Feature>

<gn:Geo_Feature rdf:ID="GEO_402">
  <gn:geo_name xml:lang="pt">Tua</gn:geo_name>
  <gn:geo_type_id rdf:resource="#RIO"/>
  <gn:info_source_id rdf:resource="#texto_web"/>
</gn:Geo_Feature>

<gn:Geo_Feature rdf:ID="GEO_758">
  <gn:name>
    <rdf:Bag>
      <rdf:li xml:lang="pt">Serra de Urbião</rdf:li>
      <rdf:li xml:lang="es">Sierra de Urbión</rdf:li>
    </rdf:Bag>
  </gn:name>
  <gn:geo_type_id rdf:resource="#SERRA"/>
  <gn:altitude unit="m">2080</gn:altitude>
  <gn:info_source_id rdf:resource="#texto_web"/>
</gn:Geo_Feature>
```

03-Out-2006 III Simpósio Doutoral da Linguateca 21

Avaliação da metodologia proposta

- DE: Verificar quais expressões são mais produtivas para ocorrências geográficas em textos. Avaliação baseada na quantidade de átomos retornados e quantos desses são realmente geográficos.
- EF: Quantos fatos o módulo consegue extrair a partir dos extratos marcados com o sistema de REM?
- IF: Quantos fatos gerados pelo módulo EF são realmente integrados na Geo-Net-PT01?

03-Out-2006 III Simpósio Doutoral da Linguateca 22

Avaliação da metodologia proposta

- Discussão!!
- Quão úteis são as ontologias produzidas?
 - Aplicações utilizando-as
- Geração de ontologias existentes com conteúdo proveniente de texto?

03-Out-2006 III Simpósio Doutoral da Linguateca 23

Aplicações para ontologias geográficas

- Motores de busca geograficamente conscientes:
 - A ontologia geográfica fornece suporte às tarefas de desambiguação de consultas e expansão das mesmas.
- Sistemas de REM
- Sistemas de pergunta e resposta:
 - Quantos concelhos existem no distrito de Faro? O distrito de Faro contém 16 concelhos.
 - Quantos distritos têm todos os seus concelhos dentro de uma província portuguesa? Nove distritos.

03-Out-2006 III Simpósio Doutoral da Linguateca 24

Considerações Finais

- Panorâmica do desenvolvimento da tese
- Preocupação com o tratamento do conhecimento geográfico em textos tem sido uma tendência
- Mais importante do que extrair informação é integrar conhecimento

Trabalho futuro

- Desenvolver os módulos do SEI-Geo
- Buscar alternativas para avaliação da metodologia proposta
 - Comparação com outros trabalhos
- Ter uma ontologia geográfica **útil** no final do processo