

### III Simpósio Doutoral da Linguateca

Estado da arte:  
- Extração de Informação (geográfica)

Marcirio Silveira Chaves

### Contexto / Motivação

- ✓ Aproximação entre conhecimento expresso em textos e aquele que pode ser representado formalmente
- ✓ Domínio geográfico – transversal
- ✓ Expansão do conhecimento geográfico existente
  - Integração de informação

03-Out-2006

III Simpósio Doutoral da Linguateca

2

### Roteiro

- ✓ Extração de Informação (EI)
- ✓ Avaliação de EI
- ✓ Extração de informação geográfica
- ✓ Integração de conhecimento geográfico
- ✓ Trabalhos correlatos

03-Out-2006

III Simpósio Doutoral da Linguateca

3

### Extração de Informação

- ✓ qualquer processo que **seletivamente estrutura e combina** dados encontrados por esse processo, os quais estão explícita ou implicitamente declarados em um ou mais textos [Cowie and Wilks 2000].
- ✓ ... reconhecimento das **propriedades e relações** que são mencionadas em um ponto particular de um texto [DowdallEtal04].
- ✓ **processo** de derivar **dado quantificável desambiguado** a partir da LN para servir a alguma necessidade de **informação precisa e pré-especificada** [Cunningham05].

03-Out-2006

III Simpósio Doutoral da Linguateca

4

### Extração de Informação

1. Regras
  - ✓ Ex.: cidade [W]
  - ✓ Listas negras
2. Aprendizagem automática
  - ✓ *Bootstrapping*
3. Apoiada por listas de EMs
  - ✓ Usados em 1 e 2

03-Out-2006

III Simpósio Doutoral da Linguateca

5

### Extração de Informação

- ✓ Avaliações
  - MUCs
  - ACE - *Automatic Content Extraction*
  - HAREM

03-Out-2006

III Simpósio Doutoral da Linguateca

6

## MUC dividido em 5 tarefas em 1998

*O largo do Marquês de Pombal foi ponto de encontro dos adeptos após a vitória de Portugal na última terça-feira. Ele foi primeiro-ministro de Portugal e um dos principais personagens na reconstrução de Lisboa após o terremoto de 1755.*

- ✓ Reconhecimento de Entidades Mencionadas (REM)
  - REM:** reconhece que as EMs presentes são: o largo do Marquês de Pombal, Portugal, Lisboa e terremoto de 1755.
- ✓ Resolução de Co-referências (RCO)
  - Identifica quais entidades e referências (pronomes, por exemplo) se referem para a mesma coisa.
  - RCO:** reconhece que Ele se refere a Marquês de Pombal.

03-Out-2006

III Simpósio Doutoral da Linguatca

7

## MUC dividido em 5 tarefas em 1998

- ✓ Construção do Modelo de Elementos (CME)
  - Adiciona informação descritiva para os resultados gerados em REM (usando RCO).
  - CME:** reconhece que Marquês de Pombal é um largo (no modelo de elementos, Marquês de Pombal preencheria o campo largo, que geralmente são pontos de referência nas cidades).
- ✓ Construção do Modelo de Relacionamentos (CMR)
  - Encontra relacionamentos entre as EMs com informação descritiva.
  - CMR:** reconhece que Marquês de Pombal foi primeiro-ministro.
- ✓ Produção do Cenário do Modelo (PCM)
  - Analisa os resultados produzidos em CME e CMR e reconhece o cenário e as entidades mencionadas envolvidas.
  - PCM:** reconhece que Portugal venceu um jogo e como consequência os adeptos foram para junto ao largo do Marquês de Pombal.

03-Out-2006

III Simpósio Doutoral da Linguatca

8

## ACE - Automatic Content Extraction

MUC	ACE
Reconhecimento de Entidades Mencionadas (REM)	Entity Detection and Tracking
Resolução de Co-referências (RCO)	
Construção do Modelo de Elementos (CME)	Relation Detection and Tracking
Construção do Modelo de Relacionamentos (CMR)	
Produção do Cenário do Modelo (PCM)	Deteção e caracterização de evento

03-Out-2006

III Simpósio Doutoral da Linguatca

9

## ACE - Automatic Content Extraction

- ✓ Categorias
  - Pessoa, Organização, Entidade geo-política, Local (entidade geográfica com extensão física), facility (artefatos feitos por humanos)
- ✓ Tarefas mais complexas no ACE
  - Taxonomia das entidades é mais granular
  - Interpretação de metonímia (análise semântica dos textos)
  - Múltiplos domínios utilizados e fontes de informação
- ✓ Sw público
- ✓ Avaliação não é pública (restrita aos participantes)

03-Out-2006

III Simpósio Doutoral da Linguatca

10

## El geográfica

- ✓ Uso de almanaques/ontologias
- ✓ Padrões
  - [conceito geográfico] tais como ...,
- ✓ Regras
  - localizado em [Local]
  - situado em [Local]
  - rua [Local]
  - ...

03-Out-2006

III Simpósio Doutoral da Linguatca

11

## Extração de informação geográfica

- ✓ Reconhecimento de EMs geográficas (locais)
- ✓ Identificação das propriedades dos conceitos geográficos
- ✓ Ex.:
  - O rio Douro é um rio que nasce na província de Sória, nos picos da Serra de Urbião, 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.
  - Rio: Douro
  - Comprimento do rio: 850 km
  - Foz: cidade do Porto
  - Província: Sória
  - Serra: Serra de Urbião
  - Altitude da serra: 2.080 metros

03-Out-2006

III Simpósio Doutoral da Linguatca

12

## El geográfica

### ✓ Relações espaciais [Delboni et al. 05]

- Fuzzy
  - Próximo, perto, abaixo de
- Topológica
  - Dentro de, no coração de, embaixo de
- Direcional
  - Em frente ao, ao lado de, atrás de
- Métrica
  - A ? Km de, a ? minutos de, a ? quarteirões de

03-Out-2006

III Simpósio Doutoral da Linguateca

13

## El geográfica

### ✓ Principais ambigüidades existentes

- Com ORG e PES

Ex.:

O **Porto** está muito a frente dos demais clubes.

A **Armênia** entrou para a escola esse ano.

- Reconhecimento em contexto

03-Out-2006

III Simpósio Doutoral da Linguateca

14

## El geográfica

### Integração de conhecimento geográfico

## Integração de conhecimento geográfico

- ✓ Conhecimento proveniente de fontes de autoridades administrativas
- ✓ Conhecimento adquirido dos textos (LN)
- ✓ Novo conhecimento
  - Nomes alternativos
    - Históricos, raros, ...
  - Extensão dos atributos geográficos
    - Comprimento, altitude, nascente, ...
- ✓ Problemas
  - Conceitos distintos, complementares

03-Out-2006

III Simpósio Doutoral da Linguateca

16

## Integração de conhecimento geográfico

### ✓ Problemas na integração

- Texto – Texto
  - Reconhecer que diferentes nomes se referem ao mesmo local
    - Rio, Rio de Janeiro, Cidade Maravilhosa
    - Grande Lisboa, região metropolitana de Lisboa, *Great Lisbon*
- Texto – Ontologia
  - Reconhecer que diferentes nomes se referem ao mesmo local
  - Em que local da ontologia integrar o novo conhecimento?

03-Out-2006

III Simpósio Doutoral da Linguateca

17

## Trabalhos correlatos

03-Out-2006

III Simpósio Doutoral da Linguateca

18

## Tarefas envolvendo ontologias

- ✓ Construção de ontologias: Dado um conjunto de documentos, produz-se uma ontologia com esses documentos como ocorrências.
- ✓ Extensão de ontologias: Dada uma ontologia parcial e um conjunto de ocorrências, estende-se a ontologia com novos conceitos respeitantes às ocorrências.
- ✓ População de ontologias: Dada uma estrutura conceitual com uma população parcial, desenvolve-se um modelo que possa assumir novas ocorrências para os conceitos.

03-Out-2006

III Simpósio Doutoral da Linguatca

19

## Tarefas envolvendo ontologias

- ✓ Tipos de entradas
  - Texto
  - Dicionário legível por máquina
  - Base de conhecimento
  - Dados semi-estruturados (esquemas XML)
  - Base de dados

03-Out-2006

III Simpósio Doutoral da Linguatca

20

## Aprendizado de ontologias – texto

- ✓ Extração baseada em padrões
  - $n$  nomes são detectados, então os  $n-1$  primeiros nomes são hiperônimos de  $n$ .
  - Hearst 1992
- ✓ Regras de associação
  - Algoritmo recebe como entrada exemplos de regras e percorre a hierarquia para detectar novas associações que serão validadas pelo engenheiro do conhecimento
  - Ex.: (evento;área).
- ✓ Agrupamento conceitual
  - Conceitos são agrupados de acordo com a distância semântica entre cada conceito para formar hierarquias
- ✓ Aprendizagem de conceitos
  - Dada uma taxonomia, ela é incrementalmente expandida com novos conceitos dos textos

03-Out-2006

III Simpósio Doutoral da Linguatca

21

## População de ontologias a partir de menções [B. Magnini et al. 06]

- ✓ Uma entidade pode ter várias menções  
Ex.:
  - Entidade: Vila Nova de Gaia
  - Menções: Gaia, Vila Nova de Gaia, V. N. de Gaia
- ✓ Sub-tarefa de população de ontologias
- ✓ Coleção de documentos anotados manualmente
- ✓ Agrupados em categorias (notícia, cultura, local, ...)
- ✓ Usa modelo com uma lista de atributos que devem ser encontrados nos textos

03-Out-2006

III Simpósio Doutoral da Linguatca

22

## População de ontologias a partir de menções [B. Magnini et al. 06]

- ✓ Experimento realizado com atributos da categoria Pessoa
  - 49% das menções distintas foram incluídas nos atributos
- ✓ Com Locais
  - Nome
  - Tipo
  - Comprimento
  - Altitude
  - ...

03-Out-2006

III Simpósio Doutoral da Linguatca

23

## Snowball

- ✓ Baseline
  - Encontra co-ocorrências de nomes de organizações e locais na mesma linha de um doc..
- ✓ Adaptação do método DIPRE (*Dual Iterative Pattern Expansion*)
  - Extrai relações estruturadas de docs. HTML
  - Recebe conjunto de tuplas [organização,localização] definidas manualmente para treinamento
  - Utiliza o REM - MITRE Corporation's Alembic Workbench
  - Tenta identificar padrões de ocorrências baseada nas etiquetas colocadas pelo REM.
- ✓ Associa um grau de confiança aos padrões encontrados

03-Out-2006

III Simpósio Doutoral da Linguatca

24

## Snowball

- ✓ Avaliação manual dos resultados
  - 100 pares ORG-LOC escolhidos aleatoriamente
  - Erros divididos em 3 categorias:
    - Local etiquetado erroneamente
    - Org. etiquetada erroneamente
    - Relacionamento errado, ex.: "FCUL localizada no Porto".
- ✓ Limitações
  - Não integra o conhecimento extraído com o conhecimento existente
  - Não extrai propriedades de conceitos

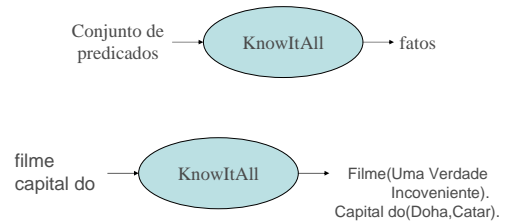
03-Out-2006

III Simpósio Doutoral da Linguatca

25

## KnowItAll

- ✓ Atribui uma probabilidade a cada fato extraído



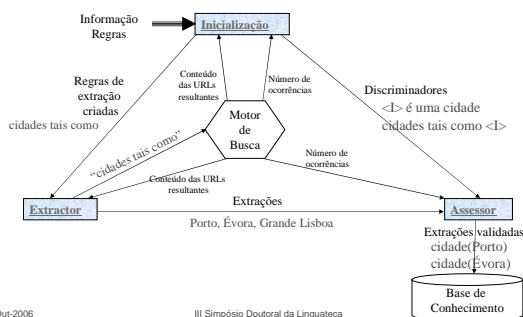
03-Out-2006

III Simpósio Doutoral da Linguatca

26

KnowItAll

## Principais componentes



03-Out-2006

III Simpósio Doutoral da Linguatca

27

## Modelo com aprendizado



03-Out-2006

III Simpósio Doutoral da Linguatca

28

## Métodos utilizados para melhorar a abrangência

- ✓ Aprendizado de padrões
- ✓ Extração de subclasses
- ✓ Extração de listas

03-Out-2006

III Simpósio Doutoral da Linguatca

29

## Aprendizado de padrões

- ✓ Recebe um conjunto de instâncias geradas a partir do *Extractor*
- ✓ Consulta a web
- ✓ "melhores padrões são selecionados" = **Padrões aprendidos**

Objectivo: Encontrar padrões de alta qualidade

03-Out-2006

III Simpósio Doutoral da Linguatca

30

## Aprendizado de padrões

- ✓ Classe <cidade> e conjunto de instâncias (Lisboa, Algarve, Vila Nova de Gaia)
- ✓ Procura as instâncias na web e guarda contexto
  - (w w w <|> w w w)
- ✓ Identifica padrões candidatos  
Ex.:  
**localizado em Lisboa**  
**está situado no Algarve a 40 Km de**  
**perto de Vila Nova de Gaia no norte do país**
- ✓ Avalia os padrões (Precisão estimada)

03-Out-2006

III Simpósio Doutoral da Linguatca

31

## Aprendizado de padrões

- Padrões aprendidos atuando como Extratores
  - **localizado em** <cidade>
  - **situado no** <cidade>
  - **perto de** <cidade>
- Consulta a web.
- Qualquer nome próprio ocorrendo após **os padrões aprendidos** torna-se um candidato a cidade.

03-Out-2006

III Simpósio Doutoral da Linguatca

32

## Aprendizado de padrões

- Padrões aprendidos atuando como Discriminadores
  - <|> é uma cidade
    - Limiar aprendido: 0.000016
  - cidades tais como <|>
    - Limiar aprendido: 0.000044
  - <|> e outros municípios
    - Limiar aprendido: 0.000032
  - ...
  - Output
    - Fatos
      - vila (Sintra)
      - concelho (Silves)

03-Out-2006

III Simpósio Doutoral da Linguatca

33

## KnowItAll

- ✓ Limitações
  - Não integra o conhecimento extraído com o conhecimento existente
  - Não extrai propriedades de conceitos
  - Não relaciona os fatos extraídos

03-Out-2006

III Simpósio Doutoral da Linguatca

34

## Alfonseca e Manandhar

- ✓ Enriquecimento de ontologias de modo não-supervisionado com informação dependente de domínio
- ✓ Desambiguação do sentido das palavras
  - Submete um synset da WordNet para um motor de busca
  - Calcula a frequência das palavras retornadas
  - Submete essas palavras a duas funções que pesam a proximidade de cada uma com um determinado termo
- ✓ Experimento
  - Documentos (versão eletrônica do livro O senhor dos Anéis, 1968) – 478.000 palavras
  - Ontologia utilizada: 7 conceitos em 3 níveis hierárquicos
  - Sem avaliação

03-Out-2006

III Simpósio Doutoral da Linguatca

35

## Uryupina

- ✓ Objetivo
  - Aquisição de almanaques a partir de páginas web
- ✓ Uso de padrões
  - "X island", "island of X"
  - Selecionados manualmente
- ✓ Almanaque utilizado inicialmente é codificado manualmente (1.260 locais)
- ✓ Usa treinamento
  - Ripper (*machine learner*)

03-Out-2006

III Simpósio Doutoral da Linguatca

36

## Uryupina

- ✓ **Bootstrapping**
  - Nomes de Locais do almanaque são pesquisados no AltaVista
  - Para cada nome, extrai o contexto *2E Local 2D* das 100 primeiras páginas retornadas
  - Seleciona os melhores padrões aprendidos para captar novos nomes
- ✓ **Avaliação**
  - 86,5% de precisão média para nomes de cidades, países, ilhas, montanhas e regiões.
- ✓ **Resultado**
  - Listas de nomes de locais
- ✓ **Limitações**
  - Não há hierarquia nos nomes aprendidos nos documentos.
  - Não integra os nomes aprendidos ao almanaque codificado inicialmente

03-Out-2006

III Simpósio Doutoral da Linguatca

37

## Comparação: trabalhos correlatos

	PAD	Onto	EEM	ICA	PT
Snowball	✓	X	✓	X	X
KnowItAll	✓	X	✓	X	X
Alfonseca e Manandhar	X	✓ (WordNet)	✓	?	X
Uryupina	✓	✓ (almanaque)	✓	X	X

- ✓ **Legenda:**
  - PAD = padrões
  - Onto = ontologias
  - EEM = Extração de EMs
  - ICA = Integra o conhecimento adquirido
  - PT = processa textos em português

03-Out-2006

III Simpósio Doutoral da Linguatca

38

## Considerações Finais

- ✓ EI
- ✓ EI geográfica
- ✓ EI -> Integração de informação
- ✓ Trabalhos relacionados e suas limitações

03-Out-2006

III Simpósio Doutoral da Linguatca

39