Um analisador semântico para português

Luís Sarmento

Resumo

- Breve contextualização
- Análise Semântica
- Motivação para o doutoramento
- A Arquitectura
  - Módulo de Análise
  - Base de conhecimento
  - Módulo de EI
- Avaliação
- Trabalho Presente e Futuro

Breve contextualização (1)

- MIAC (Outubro de 2000 a Maio de 2004)
  - Agentes Emocionais
  - Arquitectura de Agentes + Simulador
  - Trabalho conceptual e complexo
  - Bom "background" conceptual
- Linguateca (Outubro de 2002 a Outubro de 2005)
  - processamento computacional da lingua portuguesa
  - desenvolvimento de ferramentas para linguístas
  - ambiente "hostil", só recentemente mais apoio técnico
  - Bom "background" prático
- FEUP (Outubro de 2003 a Agosto de 2005)
  - Cadeiras de AIAD, AED, TW
  - Muita aprendizagem (IA e CC)

Breve contextualização (2)

- A partir de Outubro 2005
  - ProDEI – FEUP
  - Orientação:
    - Prof. Eugénio Oliveira
    - Prof. Diana Santos
  - Bolsa da FCT
  - Tema:
    - Analisador Semântico Robusto e de Cobertura Larga para o Português

Análise Semântica

- A análise semântica é uma capacidade fundamental para vários sistemas de PLN
  - sistemas de sumarização
  - resposta automática a perguntas
  - tradução automática
  - recuperação de informação
- Mas tem muitas dificuldades de concretização:
  - Cobertura (diferentes domínio, géneros)
  - Robustez (erros, variações ortográficas, velocidade)

Análise Semântica

- Desenvolvimento de analisadores especializados:
  - extracção de terminologia e variantes semânticas
    - Morin & Jacquemin, 2004
  - a desambiguação de sentidos
    - Purandare & Pedersen, 2004
  - a identificação de relações e a criação de tesauros ou léxicos semânticos
    - Grefenstette, 1994; Richardson et al., 1998; Caraballo, 1999
  - o reconhecimento de entidades mencionadas
    - Mikheev et al., 1999
  - a determinação de papéis semânticos
    - Carreras & Màrquez, 2004
- Operam normalmente em condições controladas relativamente a género e domínio.

Análise Semântica

- Existem alguns sistemas de grande cobertura para o inglês:
  - Shi & Mihalcea, 2004;
  - Fillmore & Collin, 2001;
  - Forbus et al, 2005;
- Estes sistemas são geralmente baseados em recursos de conhecimento abrangentes e complexos
  - WordNet
  - FrameNet
  - Cyc
- para o português não existem recursos semelhantes publicamente disponíveis:
  - WordNet.pt? WordNet.br?
  - construção é difícil e dispendiosa
  - são projectos de muitos anos

Motivação

- Desenvolver um analisador semântico de larga cobertura para o português
- Capaz de executar várias tarefas semânticas:
  - identificar e classificar terminologia
  - identificar e classificar entidades mencionadas
  - identificar conceitos e suas relações semânticas existentes em texto
  - determinar papéis semânticos
- Capaz processar robustamente:
  - grandes quantidades de texto
  - diferentes géneros e domínios
- O analisador seguirá uma filosofia "data-intensive"
  - funcionamento será suportado numa ampla base de conhecimento
  - Base criada usando técnicas de extracção de informação "recentes"

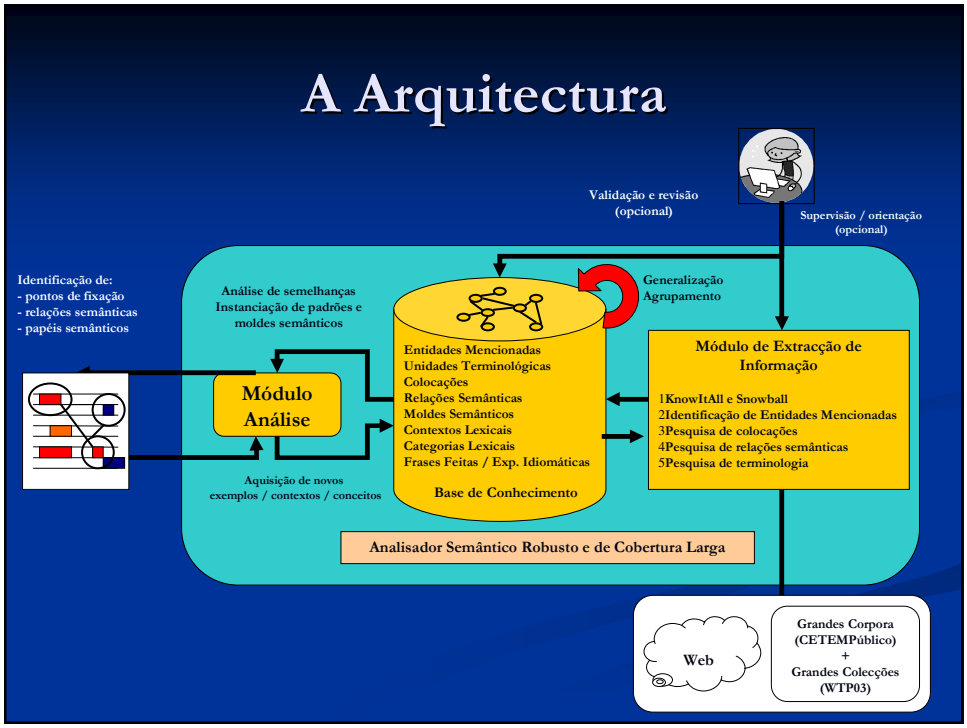
Motivação

- Têm surgido na EI técnicas simples mas eficientes:
  - KnowItAll e Snowball
  - exploram a redundância da web
  - extraem conhecimento variado
    1. factóides: "[Luís Sarmento] {nasceu no} [Porto]"
    2. entidades mencionadas cujas categorias não são conhecidas à partida: "[Luís Sarmento] {é} [engenheiro]"
    3. conceitos e suas relações: [engenheiro] {é um tipo de} [profissão]
- Sistemas apresentam características evolutivas
  - aprendizagem "levemente supervisionada"
  - aprendem novas formas de recolher mais conhecimento
- Parecem ser uma boa forma de compensar a inexistência de recursos de conhecimento para a análise semântica!

Motivação

- Disponibilização de cada vez mais informação na web pt / br
- Torna-se viável a construção automática de grandes bases de conhecimento para o português
  - aplicando os referidos métodos de EI
  - métodos de pesquisa de colocações (Smadja, 1993 e Yarowsky, 1993),
  - de extracção terminologia (Sarmento, 2005)
  - ou outros que explorem a redundância de informação (Brill, 2003).
- Conclusão:
  - É possível considerar aproximações do estilo "data-intensive" para desenvolvimento de analisadores semânticos para o português
    - incluir no analisador capacidades próprias de EI

A Arquitectura



Módulo de análise

- Responsável pela análise semântica propriamente dita.
- Identificará e clasificará no texto pontos de fixação semânticos:
  - terminologia
  - entidades mencionadas
  - colocações
  - frases feitas
- Identificará também
  - relações entre conceitos
  - papéis semânticos dos agentes no contexto.
- Consultará activamente a base de conhecimento:
  1. encontrar informação já conhecida acerca dos pontos de fixação
  2. detectar semelhanças / analogias com pontos de fixação já conhecidos
  3. encontrar padrões lexicais e moldes semânticos
- Usará tecnicas "rápidas" de pesquisa de padrões
- Enriquecer a base de conhecimento com os produtos da análise

Base de conhecimento

- Armazenará todo a informação a ser utilizada pelo módulo de análise:
  - conhecimento do mundo
    - entidades, terminologia e suas relações, classes e categorias semânticas, moldes semântico possíveis de relacionamento
  - conhecimento acerca das formas e contextos em que poderão vir a ser encontrados em texto livre
    - expressões idiomáticas, colocações, regras de forma, padrões lexicais
- Esta base deverá ser capaz de:
  - generalizar o conhecimento através de técnicas de agrupamento
    - para poder gerar novas classes e categorias semânticas que permitam gerar regras de análise mais compactas
  - iniciar novos processos de extracção de informação para colmatar possiveis faltas de cobertura no seu conhecimento.

O módulo de extracção

- Combinará:
  - várias técnicas de extracção de informação:
    - inspiradas no KnowItAll e SnowBall
  - técnicas robustas de:
    - extracção de terminologia
    - extracção de reconhecimento de entidades mencionadas
    - detecção de relações semânticas
- Serão também integradas neste módulo métodos de:
  - descoberta automática de colocações
  - frases feitas
  - expressões idiomáticas.
- Detecção de novos padrões e contextos lexicais:
  - para uso no processo de descoberta de conhecimento e de análise.

Avaliação - Outro do objectivo

- Estabelecer um conjunto:
  - de métodos de avaliação
  - recursos a estes associados para português
  - Exemplo o trabalho do ACE
    - Doddington et al., 2004
- A avaliação do sistema desenvolvido
- Teste em actividade do sistema:
  - acoplamento do analisador a um sistema de RaP
  - participação na pista QA@CLEF em português
    - (http://clef-qa.itc.it/2005/)

Ponto de Partida

- Experiência em análise semântica:
  - Corpógrafo (texto de domínio específico):
  - extracção terminológica
  - extracção de definições e de relacões semanticas
  - SIEMÊS: sistema de reconhecimento de entidades mencionadas
- Experiência com grandes colecções de texto:
  - técnicas de manipulação e análise de texto eficientes
  - emprego de SGBD para armazenamento / indexação

Muito para fazer...

- Módulos para:
  - identificação de colocações, frases feitas e expressões idiomáticas
  - identificação de relações e papéis semânticos
- Modelização de uma ontologia:
  - armaz. conhecimento necessário à análise
- Implementação de métodos:
  - agrupamento
  - Extracção de informação (tipo KnowItAll)
- Desenvolvimento do protótipo final
- Preparação da metodologia, métricas e recursos de avaliação
- Testes de validação e participação na pista QA@CLEF

Principais contribuições

- O que se espera serem as principais contribuições:
  1. O analisador semântico para o português
  2. A base de conhecimento (muitos usos)
  3. A metodologia híbrida:
     - Extracção de Informação de grandes colecções
     - Análise de textos concretos