Projeto de mestrado

Uma aplicação da Teoria da Previsão à análise semântica de linguagens naturais

Aluno: Iúri Chaer

Orientador: Prof. Dr. Ricardo Luis de Azevedo da Rocha

Laboratório de Tecnologia Adaptativa

Escola Politécnica da USP

Organização da apresentação

1. Motivação
2. Objetivos
3. Fundamentação
    1. Abordagens à análise semântica de LN
    2. A Teoria da Previsão de Solomonoff
    3. Gramática gerativa
4. Resultados pretendidos
5. Plano de trabalho
6. Plano de publicações

Motivação

- 1997: A quantidade de informação escrita no mundo dobra a cada 5 anos [1].
- 2005: A Internet conta com 1 bilhão de usuários [2].

- Como aproveitar toda essa informação tão acessível para tantos?
- Problema da conversão de informação em conhecimento.

Objetivos

- Principal
  - Desenvolver um analisador semântico – simplificado, mas fiel à idéia de geração de conhecimento a partir de descrições em Linguagem Natural.
- Secundário
  - Estudar
    - Cognição humana
    - Mecanismos de indução
    - Modelo gerativo de lingüística
    - Formalismos adequados aos requisitos de adaptatividade e capacidade indutiva do problema
  - Integrar contribuições dessas diferentes linhas de pesquisa

Fundamentação
Abordagens à análise semântica de LN

Correlação estatística

- Abordagem comum na computação – correlação estatística ≈ correlação semântica (Latent Semantic Analysis [3])
- Relativamente fácil de implementar (dados os componentes semânticos, basta fazer análises em *corpora* anotados)
- Resultados excelentes na maioria dos casos
- Mas LNs seguem a Power Law
- Como contabilizar a estrutura sintática?

Fundamentação
Abordagens à análise semântica de LN

Web Semântica e o conceito difundido de ontologias

Exemplo extraído de [4]:

```
<rdfs:Class rdf:ID="WINE">
  <rdfs:subClassOf rdf:resource="#POTABLE-
LIQUID"/>
  <rdfs:subClassOf>
     <daml:Restriction>
        <daml:onProperty rdf:resource="#MAKER"/>
        <daml:minCardinality>
           1
        </daml:minCardinality>
     </daml:Restriction>
  </rdfs:subClassOf>
.........
</rdfs:Class>
```

Fundamentação
Abordagens à análise semântica de LN

Web Semântica e o conceito difundido de ontologias

- Hierarquia entre conceitos
- Conhecimento descrito como atributos e relacionamentos
- Possibilita consultas complicadas, mas exige definições prévias dos conceitos e só permite inferências superficiais – nada de extrapolações (há estudos sobre o uso de quantificadores fuzzy associados, mas incipientes e adequados artificialmente ao modelo original).
- Problema de Platão!

Fundamentação
Abordagens à análise semântica de LN

Cyc

- Como a idéia da Web semântica, mas não colaborativo
- O autor (Lenat [5]) discute a aplicação do sistema como uma base de bom-senso, algo aparentemente essencial para decisão de diversas ambigüidades
  - *João viu [[a montanha] no avião].*
- Mas crianças, com uma conhecimento de mundo relativamente pequeno, entendem frases assim!
- Capacidade de extrapolação limitada à lógica de primeira ordem; exige um tamanho gigantesco para extrapolações não-superficiais

Fundamentação
Abordagens à análise semântica de LN

Lingüística

- Estabelecimento da Forma Lógica (LF)
  - Quem você acha que viu João?
  - Para que pessoa *x* você acha que *x* viu João?
- Remove ambigüidades, estabelece relações lógicas diretas
- Não está preocupada com o conhecimento contido em cada termo, mas nos papéis e relações entre eles

Fundamentação
Abordagens à análise semântica de LN

- Mas...
  - E o Problema de Platão?
  - A proposta de ontologias mostrada é inadequada para diversas aplicações:
    - Conhecimento restrito a atributos e relações?
    - Onde começam as hierarquias?

Fundamentação

A Teoria da Previsão de Solomonoff

- Teorema de Bayes:

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

  - Estabelece uma distribuição de probabilidade para hipóteses: dado um gerador de hipóteses e P(H), podemos usá-lo para indução

- Indução resolveria o Problema de Platão
  - Não é necessário saber tudo, extrapola-se a partir dos padrões do que se conhece

Fundamentação

A Teoria da Previsão de Solomonoff

- Teoria da Informação de Shannon
  - Define quantidade de informação em uma cadeia C dada uma linguagem = 1/log(1/P(C))
- Complexidade algorítmica de Kolmogorov
  - Tese de Church: todos os computadores universais são equivalentes
  - Quantidade de informação do menor algoritmo que produz a cadeia
  - Medida universal

Fundamentação

A Teoria da Previsão de Solomonoff

- Probabilidade algorítmica = probabilidade de uma hipótese
  - O conjunto de algoritmos que geram uma cadeia é o universo de hipóteses para aquele evento
  - Probabilidade inversamente proporcional ao comprimento do algoritmo
- A distribuição de probabilidades das hipóteses é dominada por aquela com a menor complexidade de Kolmogorov

Fundamentação

A Teoria da Previsão de Solomonoff

Solomonoff, além de ser um dos proponentes da Complexidade Algorítmica, integrou essa idéia [6] ao teorema de Bayes e métodos de busca de soluções (LSearch, PPM, Alg. Genéticos) para formalizar um processo de indução

Fundamentação
A Teoria da Previsão de Solomonoff

Efeito colateral:

- Descrições algorítmicas – uma abordagem interessante para descrição do conhecimento
  - Tese de Church: Máquina de Turing = conceito de computabilidade
- Complexidade de implementação muito maior que atributos+relacionamentos, mas cobre deficiências
- Como começa o treinamento? Quanto tempo para ter algo útil?

Fundamentação
Gramática Gerativa

- Proposta por Noam Chomsky [7] tendo em mente o Problema de Platão
  - Toda criança aprende a sua primeira língua no mesmo período da vida **mesmo que ninguém tente ensiná-la**
- Gramática Universal: conjunto de princípios que regem todas as Línguas Naturais
  - Cada língua tem parâmetros diferentes

Fundamentação
Gramática Gerativa

- Princípios simples que regem os papeis e relacionamentos entre termos em LN
- Ainda em desenvolvimento, mas já é utilizável e contribui para a extração de formas lógicas (LF)

- Juntando isso a alguns artifícios estatísticos, temos um patamar inicial para o mecanismo de indução

Resultados pretendidos

- Propor integração de todos os avanços apresentados

- Implementar um analisador semântico
  - Simplificado, restrito, longe de comparável ao que é feito atualmente; mas capaz de aprender e descrever informações a partir do sistema indutivo proposto por Solomonoff

Plano de trabalho

Obs.: Tarefas não seqüenciais

- Levantamento bibliográfico
  - Teoria da Predição – 8 meses
  - Processamento de LN – 4 meses
  - Gramática gerativa – 6 meses
- Proposição e verificação de artifícios para processamento de LN – 4 meses
- Desenv. do protótipo e dissertação – 8 meses

Plano de trabalho

Estado atual (trabalho iniciado em 02/2008)

- Levantamento bibliográfico adiantado – no momento, concentrado nas teorias de lingüística
- Proposta de método de definição de domínios semânticos em textos adequada para uso em mecanismo indutivo (implementação e verificação dos resultados pendentes)
- Texto da dissertação em andamento

Plano de publicações

- Artigo em revisão no Congresso de Computação da Grande Dourados

Em estudo (prazos até o final do ano):

- Congreso de Inteligencia Computacional Aplicada – CICA '09
- IEEE Congress on Evolutionary Computation – CEC '09

E outros não listados – pretende-se uma publicação em revista (sem prazo de aceitação).

Bibliografia

1. Reuters Magazine. "Information overload causes stress". Reuters Magazine, disponível em: http://library. humboldt. edu/ccm/ngertips/ioverloadstats.html, 1997.
2. The World Factbook. "Rank Order - Internet users". CIA disponível em: https://www.cia.gov/library/publications/the-world-factbook/rankorder/2153rank.html, 2008.
3. Landauer, T.K. e Dumais, S.T. "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge". Psychological Review (New York), 1997.
4. McGuiness, D.L. e Chang C. "Wine Agent 1.0 – How does it work?". Disponível em: http://www-ksl.stanford.edu/projects/wine/explanation.html , 2008.

Bibliografia

5. Lenat, D., Parkash, M., Shepard, M., "CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks", The AI Magazine, Spring 1986.
6. Solomonoff, R.J. "Machine Learning – Past and Future", 2002.
7. Chomsky, N. "Knowledge of Language: Its Nature, Origin, and Use", Praeger/Greenwood, 1986.