

Processamento Estatístico da Linguagem Natural

Aula 2

Professora Bianca
(Sala 302 – Bloco E)
bianca@ic.uff.br

<http://www.ic.uff.br/~bianca/peln/>

Aula 2 - 02/09/2008

1

Camadas do Processamento Linguístico

- O que distingue o processamento linguístico de outros tipos de processamento é o uso de conhecimento sobre a linguagem.
- Esse conhecimento pode ser dividido em 6 camadas:
 1. Fonética
 2. Morfologia
 3. Sintaxe
 4. Semântica
 5. Pragmática
 6. Discurso

Aula 2 - 02/09/2008

2

Fonética

- É o estudo dos sons da língua e de como eles são realizados fisicamente.
 - Determina como as palavras são pronunciadas em termos das seqüências de sons.

dis-k&-'nekt

disconnect

"It is easy to recognize speech."

"It is easy to wreck a nice beach."

Aula 2 - 02/09/2008

3

Morfologia

- É o estudo de como as palavras são quebradas em partes menores que tem significado próprio.
 - É mais importante em línguas como turco e alemão.

disconnect

"not" "to attach"

TURCO:

uygarlastiramadiklarimizdanmissinizcasina

uygar las tir ama dik lar imiz dan mis siniz casina
(behaving) as if you are among those whom we could not civilize

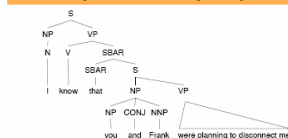
Aula 2 - 02/09/2008

4

Sintaxe

- É o estudo das relações estruturais entre palavras.
 - As palavras de uma língua devem ser ordenadas de uma certa maneira para que as sentenças sejam válidas.
 - Uma mudança na ordem pode alterar o sentido da frase.

I know that you and Frank were planning to disconnect me.



Not same structure:

You know me--Frank and I were planning to disconnect that.

Aula 2 - 02/09/2008

5

Semântica

- É o estudo do significado literal.
 - Semântica léxica: significado de cada palavra.
 - Semântica composicional: significado das palavras em conjunto.
 - Exemplo: "Western Europe", "by the end of the 18th century" vs. "by southern merchants"

I know that you and Frank were planning to disconnect me.

ACTION = disconnect
ACTOR = you and Frank
OBJECT = me

Aula 2 - 02/09/2008

6

Pragmática

- É o estudo de como a linguagem é utilizada para atingir objetivos.
 - O que você deve concluir sobre o fato de alguém ter dito alguma coisa.

I'm sorry Dave, I'm afraid I can't do that.

REQUEST: HAL, open the pod bay door.
STATEMENT: HAL, the pod bay door is open.
INFORMATION QUESTION: HAL, is the pod bay door open?

Aula 2 - 02/09/2008

7

Discurso

- É o estudo de unidades lingüísticas maiores do que uma única "frase".
 - Leva em conta a estrutura dos diálogos.
 - Uma pessoa fala de cada vez.
 - O significado de um termo depende do que já foi falado (exemplo: pronomes).

David Bowman:
Open the pod bay doors, Hal.
HAL:
I'm sorry, Dave, I'm afraid I can't do that.
David Bowman:
What are you talking about, Hal?
HAL:
I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.

Aula 2 - 02/09/2008

8

Ambigüidade

- Um fato surpreendente sobre essas camadas é que a maioria das tarefas de processamento lingüístico pode ser vista como o processo de resolver **ambigüidades** em um desses níveis.
- Dizemos que uma entrada é ambígua quando múltiplas estruturas lingüísticas podem ser construídas para ela.

Aula 2 - 02/09/2008

9

Exemplo: Ambigüidade

- I made her duck.
- (1.5) I cooked waterfowl for her.
(1.6) I cooked waterfowl belonging to her.
(1.7) I created the (plaster?) duck she owns.
(1.8) I caused her to quickly lower her head or body.
(1.9) I waved my magic wand and turned her into undifferentiated waterfowl.
- As palavras "her" e "duck" são morfologicamente e sintaticamente ambíguas.
 - A palavra "make" é semanticamente e sintaticamente ambígua.

Aula 2 - 02/09/2008

10

Resolvendo ambigüidades

- Ambigüidade sintática
 - Decidir se uma palavra é um verbo ou substantivo.
 - É resolvida com "**Part-of-Speech Tagging**" (POS tagging) = anotação sintática.
- Ambigüidade semântica
 - Decidir qual é o significado da palavra.
 - É resolvida com "**Word Sense Disambiguation**".
- Ambigüidade pragmática.
 - Determinar se uma frase é uma afirmação ou uma pergunta.
 - É resolvido com "**Speech Act Interpretation**".

Aula 2 - 02/09/2008

11

Modelos e Algoritmos

- Os vários tipos de conhecimento sobre linguagens podem ser capturados através de **modelos** formais.
 - Máquinas de estado e sistemas de regras formais
 - Máquinas de estado finito, gramáticas regulares e gramáticas livres de contexto
 - Lógica de primeira ordem
 - Modelos probabilísticos
 - Modelos de Markov (HMMs)
 - Espaços vetoriais

Aula 2 - 02/09/2008

12

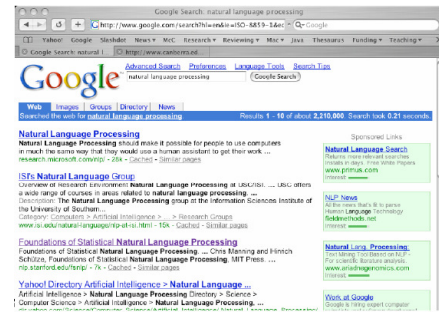
Modelos e Algoritmos

- O processamento de linguagem com quaisquer desses modelos requer o uso de **algoritmos**.
 - Busca num espaço de estados
 - Busca em profundidade, A*
 - Exemplo: busca num espaço de árvores pela representação sintática de uma frase.
 - Aprendizado automático
 - Classificação: baseado nos atributos que descrevem um objeto, o classificador atribui uma classe ao objeto.
 - Previsão sequencial: atribui uma sequência de objetos a uma sequência de classes.

Aula 2 - 02/09/2008

13

Exemplos de Aplicações de PLN

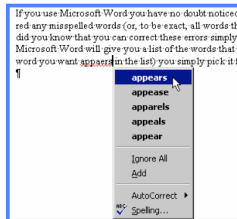


Aula 2 - 02/09/2008

14

Exemplos de Aplicações de PLN

- Correção ortográfica e gramatical



Aula 2 - 02/09/2008

15

Exemplos de Aplicações de PLN

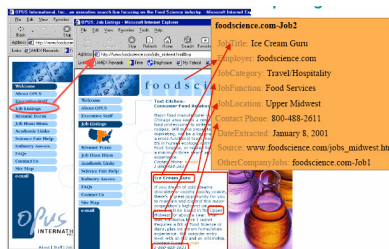


Aula 2 - 02/09/2008

16

Exemplos de Aplicações de PLN

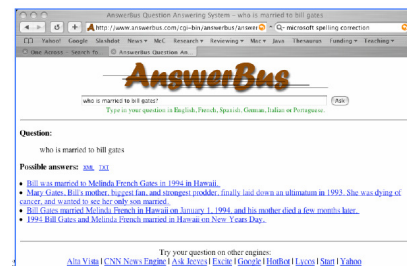
- Extração de informação



Aula 2 - 02/09/2008

17

Exemplos de Aplicações de PLN

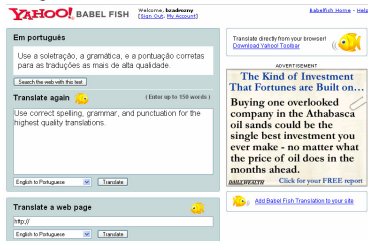


Aula 2 - 02/09/2008

18

Exemplos de Aplicações de PLN

- Tradução automática

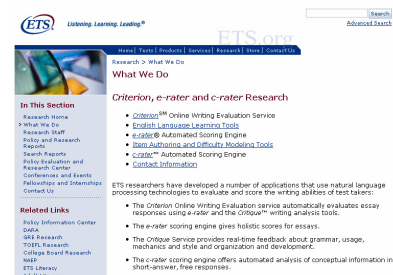


Aula 2 - 02/09/2008

19

Exemplos de Aplicações de PLN

- Correção automática de redações.

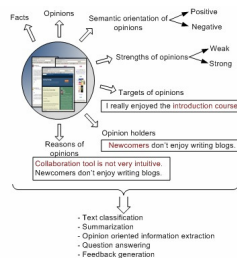


Aula 2 - 02/09/2008

20

Exemplos de Aplicações de PLN

- Análise de "sentimento"
 - Medição automática de opiniões e preferências de usuários a partir de posts em blogs e fóruns de discussão.



Aula 2 - 02/09/2008

21

Breve histórico

- Idéias fundamentais: décadas de 1940 e 1950.
 - Modelos probabilísticos e teoria da informação.
 - Modelo de neurônio de McCulloch-Pitts.
 - Teoria da linguagem de Chomsky - gramáticas livres de contexto.
- As duas frentes: 1957-1970.
 - Processamento de linguagem se dividiu em dois paradigmas: simbólico e estatístico.
 - Simbólico
 - Primeiros sistemas de "parsing" (análise sintática).
 - Busca de padrões usada em sistemas de pergunta-resposta.
 - Estatístico
 - Primeiros sistemas de OCR.
 - Primeiro "corpus" em língua inglesa: o corpus de Brown com 500 textos e um milhão de palavras.

Aula 2 - 02/09/2008

22

Breve Histórico

- Quatro paradigmas: 1970-1983
 - Paradigma estocástico
 - Desenvolvimento de sistemas de reconhecimento e síntese de voz.
 - Paradigma lógico
 - Paradigma de entendimento de linguagem natural
 - Paradigma de modelagem de discurso
 - Resolução automática de referências.
 - BDI (Belief-Desire-Intention)
- Empiricismo e volta aos modelos de estado finito: 1983-1993
- Unificação: 1994-1999
- A ascensão do aprendizado automático: 2000-2008

Aula 2 - 02/09/2008

23