

Capítulo 7

Processamento de Línguas Naturais

João Luís Garcia Rosa

2004

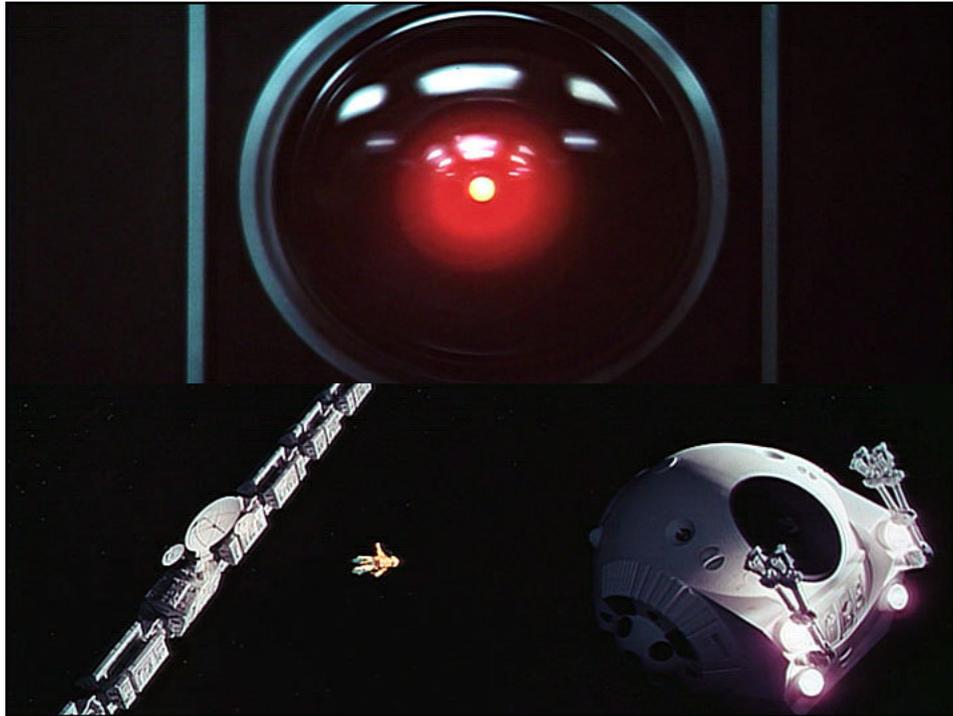
Parte dos slides foi baseado no material disponibilizado por Dr. Donald Loveland em <http://www.cs.duke.edu/courses/cps270/fall00/>



A STANLEY KUBRICK PRODUCTION

2001: a space odyssey

DIGITALLY RESTORED & RE-MASTERED SOUND • 70 MILLIMETER LIMITED RELEASE



Introdução

- David Bowman:
 - Abra as portas da nave, HAL.
- HAL:
 - Sinto muito, Dave, mas temo não ser capaz de abrí-las.

**COMPUTER
MALFUNCTION**

Stanley Kubrick e Arthur C. Clarke,
roteiro de 2001: Uma Odisséia no Espaço

HAL 9000

- AGENTE:
 - reconhecimento de fala
 - entendimento de língua natural (leitura de lábios)
 - geração de língua natural
 - síntese de fala
 - recuperação de informação
 - extração de informação
 - inferência.

Áreas do PLN

- 1. Fonética/Fonologia:
 - como as palavras são pronunciadas
- 2. Morfologia:
 - captura informações sobre a forma e comportamento das palavras no contexto
- 3. Sintaxe:
 - conhecimento necessário para ordenar e agrupar palavras

Áreas do PLN

- 4. Semântica Lexical:
 - conhecimento dos significados das palavras que compõem uma sentença
- 5. Semântica Composicional:
 - conhecimento de como esses componentes se combinam para formar significados maiores

Áreas do PLN

- 6. Pragmática:
 - uso apropriado da linguagem polida e indireta. O estudo de como a linguagem é usada para alcançar os objetivos
- 7. Discurso:
 - o estudo de unidades lingüísticas maiores que uma sentença.

PLN

- Teste de Turing
- ELIZA
- O Estado da Arte

- Processamento de Linguagem Natural
- (Psico)Lingüística Computacional

Níveis de entendimento

- **Fonético**
 - Como as palavras são relacionadas aos sons que as criaram?
 - /ka-za/ = casa
 - Importante para sistemas de reconhecimento de fala

Níveis de entendimento

- Fonético
- Morfológico
- Como as palavras são construídas a partir de componentes mais básicos
- Des-necessariamente
- Dá informação sobre a função das palavras

Níveis de entendimento

- Fonético
- Morfológico
- Sintático
- Como as palavras são combinadas para formar sentenças corretas? Qual é o papel das palavras?
- Melhor entendido – Bem estudado por linguagens formais

Níveis de entendimento

- Fonético
 - Morfológico
 - Sintático
 - Semântico
- O que as palavras significam? Como esses significados se combinam nas sentenças?

Níveis de entendimento

- Fonético
 - Morfológico
 - Sintático
 - Semântico
 - Pragmático
- Como as sentenças são usadas em situações diferentes? Como isto afeta a interpretação de uma sentença?

Níveis de entendimento

- Fonético
 - Morfológico
 - Sintático
 - Semântico
 - Pragmático
 - **Nível do discurso**
- Como o conteúdo da linguagem avizinhada de uma sentença afeta a interpretação da mesma?
 - Resolução de pronomes, referências temporais

Níveis do entendimento

- Fonético
 - Morfológico
 - Sintático
 - Semântico
 - Pragmático
 - Nível do discurso
 - **Conhecimento do Mundo**
- Conhecimento geral sobre o mundo necessário para a comunicação. Inclui conhecimento sobre metas e intenções de outros usuários.

Ambigüidade na linguagem

- A língua natural pode ser ambígua em vários níveis
 - manga
 - O homem viu o menino com o telescópio.
 - Todo garoto ama um cachorro.
 - Idéias verdes têm narizes grandes.
 - Você pode passar o sal?

Sintaxe

- A estrutura sintática de uma sentença indica a forma na qual as palavras da sentença estão relacionadas umas com as outras.
- A estrutura pode indicar os relacionamentos entre palavras e sintagmas e pode armazenar informação que pode ser usada mais tarde no processamento.

Exemplos

- O menino viu o gato
- O gato viu o menino
- A garota viu o homem na loja
 - A garota estava na loja?

Processamento sintático

- Metas principais
 - Determinar se uma seqüência de símbolos constitui uma sentença legal.
 - Atribuir uma estrutura de sintagma/constituente para sentenças legais para processamento posterior.

Classes de Palavras

- Substantivo, Verbo, Pronome, Preposição, Advérbio, etc.
- *Part-of-Speech (POS) tagging*:
 - rotular palavras de acordo com um *tagset*
 - baseado em regras;
 - estocástico (probabilístico);
 - *Transformation-Based Learning* - TBL.
 - Resolução de ambigüidade lexical

Gramáticas e Processamento Humano

- A forma como o cérebro processa linguagem
 - Psicolingüística
 - Constituintes

Gramáticas e técnicas de *parsing*

- Precisamos de uma **gramática** para realizar o *parsing*
 - Gramática = especificação formal de estruturas permitidas em uma linguagem
- Dada uma gramática, precisamos também de uma **técnica de *parsing***, ou um método de análise de sentença para determinar sua estrutura de acordo com a gramática

LN e GLC

- Gramática Livre de Contexto (GLC) é uma boa escolha
 - Poderosa o suficiente para descrever a maioria das estruturas da LN;
 - Restrita o suficiente para permitir *parsing* eficiente.
- Uma GLC tem regras com um único símbolo não-terminal no lado esquerdo.

Parsing como busca

- O *parsing* pode ser visto como um caso especial do problema de busca
- Quais são as similaridades?

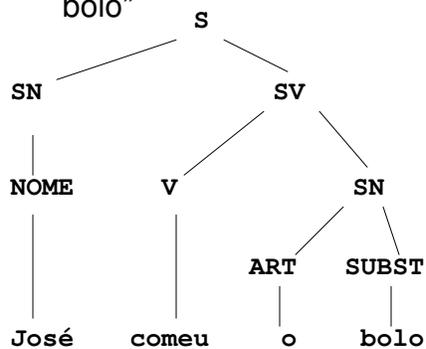
Parsing

- *Parsing*:
 - *top-down*
 - *bottom-up*
 - Resolução da ambigüidade estrutural
- *Parsing* Humano: probabilístico
 - preferências de subcategorização lexical
 - sentenças *garden path*: temporariamente ambíguas

Uma gramática (muito) simples

$S \rightarrow SN SV$
 $SV \rightarrow V SN$
 $SN \rightarrow NOME$
 $SN \rightarrow ART SUBST$
 $NOME \rightarrow José$
 $V \rightarrow comeu$
 $ART \rightarrow o$
 $SUBST \rightarrow bolo$

- Uma árvore de *parsing* para "José comeu o bolo"



Parsing top-down simples

$S \rightarrow SN SV$
 $SV \rightarrow V SN$
 $SN \rightarrow NOME$
 $SN \rightarrow ART SUBST$
 $NOME \rightarrow José$
 $V \rightarrow comeu$
 $ART \rightarrow o$
 $SUBST \rightarrow bolo$

S
 $SN SV$
 $NOME SV$
 $José SV$
 $José V SN$
 $José comeu SN$
 $José comeu ART SUBST$
 $José comeu o SUBST$
 $José comeu o bolo$

Parsing bottom-up simples

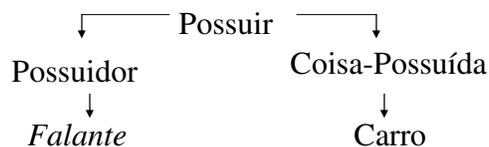
S -> SN SV	NOME comeu o bolo
SV -> V SN	NOME V o bolo
SN -> NOME	NOME V ART bolo
SN -> ART SUBST	NOME V ART SUBST
NOME -> José	SN V ART SUBST
V -> comeu	SN V SN
ART -> o	SN SV
SUBST -> bolo	S

Linguagens de representação de significado

– CPPO:

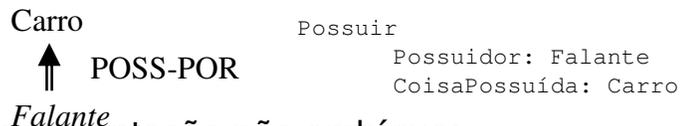
- $\exists x,y \text{Possuir}(x) \wedge \text{Possuidor}(\textit{Falante},x) \wedge \text{CoisaPossuída}(y,x) \wedge \text{Carro}(y)$

– Rede Semântica:



Linguagens de representação de significado

– Diagrama de dependência conceitual



– Representação não ambígua:

- ambigüidade
- vagueza
- forma canônica

Semântica Lexical: lexemas

- Lexema: uma entrada individual no léxico.
- Léxico: lista finita de lexemas.
- Relações entre lexemas:
 - Homonímia: relação entre palavras que têm a mesma forma com significados não relacionados: Ex. *banco*.
 - Homofonia
 - Homografia

Lexemas

- Polissemia: único lexema com múltiplos significados relacionados: cravo.
- Sinonímia: lexemas diferentes com o mesmo significado: trabalho e labuta.
- Hiperonímia: grupo de lexemas onde um denota uma subclasse de outro(s):
 - hiperônimo: mais geral: mamífero;
 - hipônimo: mais específico: baleia.

Estrutura Interna das Palavras

- A Estrutura Interna das Palavras
 - Papéis temáticos:
 - O menino bateu na bola com o taco
 - [agente, paciente, instrumento]
 - Restrições de seleção
 - O menino comeu a gelatina
 - *A gelatina comeu o menino.

Criatividade e o Léxico

- Metáfora:
 - Aquele político é uma *raposa* velha.
 - O cachorro urinou no *pé* da mesa.
- Metonímia:
 - Para relaxar, costumava beber um *porto*.
 - Quando jovem, lia tanto *Camões* quanto *Machado*.

O Estado da Arte

- Programa de computador canadense recebe dados sobre o tempo diariamente e gera relatórios metereológicos em inglês e francês (Chandioux, 1976);
- *Babel Fish*: sistema tradutor trabalha com um milhão de pedidos de tradução por dia, a partir do site de busca AltaVista;
- Cambridge, Massachusetts: sistema com reconhecimento de voz para informações sobre restaurantes (Zue *et al.*, 1991);

O Estado da Arte

- Um programa lê centenas de ensaios digitados por alunos e dá nota a eles de uma forma similar aos humanos (Landauer *et al.*, 1997);
- Um tutor de leitura ajuda crianças leitoras de estórias, usando um reconhecedor de fala para intervir quando o leitor pede ajuda ou comete erros (Mostow and Aist, 1999);

O Estado da Arte

- Um computador equipado com um sistema de visão assiste a uma partida de futebol e fornece um relatório do jogo em língua natural (Wahlster, 1989).
- Um computador prevê próximas palavras ou expande fala telegráfica para ajudar pessoas com problemas de fala e de comunicação (Newell *et al.*, 1998; McCoy *et al.*, 1998).