Processamento Profundo

Parte 3/3

trabalho conjunto com

Francisco Costa

Divisão de trabalho

- Processamento superficial
  - Usa janela limitada de contexto
  - Vai resolvendo ambiguidades, restringindo alternativas
  - Ferramentas específicas para fenómenos específicos
  - Autómatos finitos como modelo de computação subjacente

- Profundo
  - Autómatos de pilha como modelo de computação
  - Maior complexidade
    - mas cobre dependências de longa-distância
    - **De quem** é que [a Ana gosta **Ø**]?
    - **De quem** é que [a Ana disse que [a Sara gosta **Ø**]]?
    - **De quem** é que [a Ana disse que [... [que a Joana gosta **Ø**]...]]?

LX-GRAM

- Gramática para processamento linguístico profundo
  - Enquadramento tecnológico: Delph-in
  - Enquadramento linguístico: HPSG
  - Formalismo de representação semântica: MRS
    - Minimal Recursion Semantics (Copestake et al., 2001)
  - Sistema de desenvolvimento: LKB
    - Linguistic Knowledge Builder (Copestake, 2002)
  - Gramática nuclear: MATRIX
    - (Bender et al, 2002)
  - Cobertura em desenvolvimento
    - Basic phrase structure
    - Basic declarative sentences
    - Full NP coverage

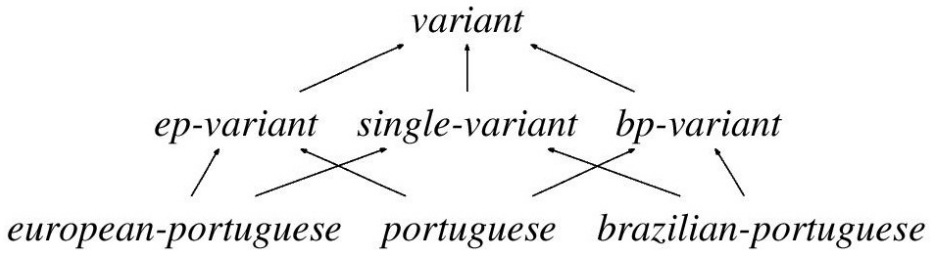
Processamento de variantes

- Vantagens
  - Escrita de código
    - evita proliferação de gramáticas e versões
  - Análise
    - reduz sobregeração espúria de parses
  - Síntese
    - geração consistente numa variante

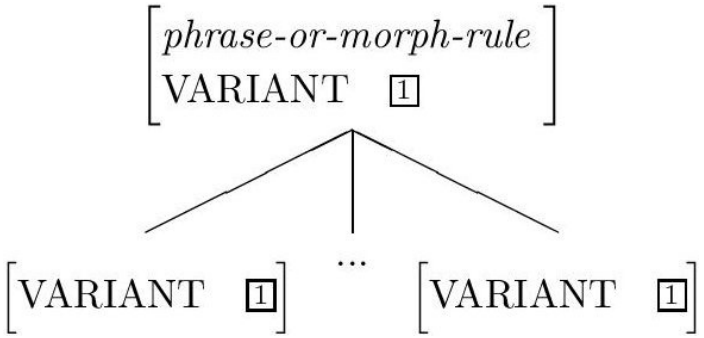
- Desafios
  - **1 - Configuração**
    - como restringir gramática a apenas uma variante?
  - **2 - Ajuste**
    - como detectar a variante do input?

❶ Configuração: traços

- Estender declaração de traços
  - tipo *sign* também com traço `[VARIANT variant]`
- Hierarquia do tipo *variant*



- Cálculo
  - Percolação por todos os signos da árvores de parse

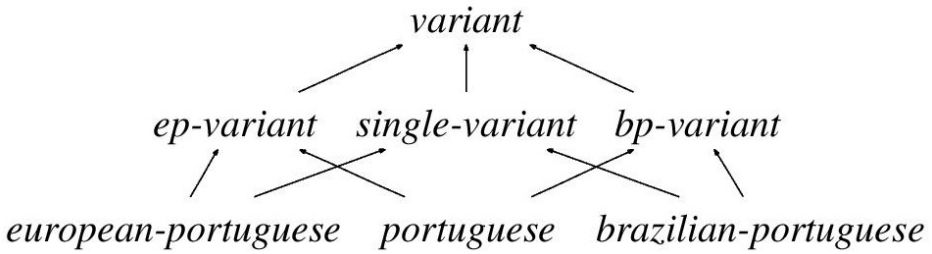


Configuração: fixar valor

- Casos de base
  - `ep-variant, bp-variant, variant`
  - Construções específicas de cada variante
  - Entradas lexicais específicas, ex:

$$\begin{bmatrix} \text{ORTH} & \langle \text{"idéia"} \rangle \\ \text{VARIANT} & \textit{bp-variant} \end{bmatrix} \qquad \begin{bmatrix} \text{ORTH} & \langle \text{"ideia"} \rangle \\ \text{VARIANT} & \textit{ep-variant} \end{bmatrix}$$

- Interruptor
  - Valor de VARIANT na condição root da gramática
  - `brazilian-portuguese, european-portuguese, single-variant`



Dados para a experiência

- Corpora
  - Corpus americano: CETENFolha (32Mtokens)
  - Corpus europeu: CETEMPúblico (204Mtokens)
    - Texto jornalístico
    - Minimamente anotado (parágrafo e frases)
- Selecção
  - Reter 2x90K linhas/frases parsáveis pela gramática (<9 tokens)
  - Retirar aleatoriamente 2x1800 frases
  - Classificar manualmente
- Metodologia de classificação
  - Pedir a falante de PE que seleccione frases do corpus americano que lhe pareçam apenas PB
  - Pedir a falante de PB que seleccione frases do corpus europeu que lhe pareçam apenas PE

A - Pré-ajuste

- Classificador
  - Detector de linguagem Bayesiano
    - (Dunning, 1994)
  - Probabilidade da sequência ocorrer numa variante ou ambas
    - distinção tripla
  - Produto de probabilidades de caracteres
    - prob condicionada por ngrams
- Dados
  - 280 frases pb + 280 frases pe + 280 frases pc
  - Divisão 50/50 entre treino e teste
- Avaliação
  - **59%** (correcção)
  - para bigramas, com sequência=1 frase

B - Auto-ajuste: itens marcados (1/2)

- Português comum: 91%
  - 81% coincidência estrita
  - 10% coincidência lata
    - Meras diferenças ortográficas: acção/ação, aguentar/agüentar,...
- Itens e construções marcadas
  - Entradas lexicais múltiplas: 5,59%
    - `time/equipa`
  - Diferenças de acepção lexical: 0,44%
    - `policial`: agente/romance

Auto-ajuste: construções marcadas (2/2)

- Diferenças sintácticas: 3,97%
  - Co-ocorrência artigos def e possessivos: 1,22%
    - (O) meu pai cuida de tudo
  - Diferentes quadros de subcategorização: 0,98%
    - Estar GER / a INF: 0,54%
  - Colocação dos clíticos 0,64%
    - O Pedro a viu(-a).
  - Bare NP singulares 0,54%
    - (Um) Médico também é ser humano.
  - Todo + Art 0,09%
    - Toda (a) operação do gênero foi proibida.
  - Contrações prep + art: 0,09%
    - Eles estão em uma (numa) creche.
  - Interrogativas sem inversão 0,09%
    - O que (é que) ele viu?
  - Negação frásica pós-verbal 0,05%
    - Isso não existe (não), bonitinha.

B - Auto-ajuste: experiência

- Valor base
  - Para cada variante, contar # itens lexicais no input
  - Seleccionar variante com maior contagem
  - 53% (Medida-F)
- Items marcados
  - Léxico: 800 entradas
  - Gramática: cobre apenas 20% das diferenças
    - Colocação dos clíticos
    - Coocorrência dos artigos e possessivos
- Avaliação
  - 57% (Medida-F)

Discussão

- Comparação Pré- vs. Auto-ajuste
  - Desempenho similar: 59% vs 57%

- A - Pré-ajuste
  - Espaço para melhorar: apenas 15 Kbytes de treino
  - Alargamento a maior número de variantes
    - Tecnicamente fácil (hier. de tipos porém cresce exponencialmente...)
    - Como se vai degradar desempenho?

- B - Auto-ajuste
  - Espaço para melhorar: 1ª exp. cobriu apenas 20% das diferenças
  - Menos sensível aos dados de treino e domínio textual
  - Maior promessa para lidar com diálogo entre falantes de diferentes dialectos, com com falas curtas

Trabalho em curso

- Projecto SemanticShare
  - UL + PUCRS, Fev 2008
  - Anotação de corpora paralelos
    - Constituência sintáctica
    - Função gramatica
    - Papel semântico
    - Forma lógica (LQG)
  - Desenvolvimento da LX-Gram
    - ...
    - Melhorar capacidade de ajuste à variante do input

Obrigado.