SINTAXE – PARTE 2

SCC5908 Tópicos em Processamento de Língua Natural

Thiago A. S. Pardo

GLCP: problemas

- 2 principais limitações
  - Suposições fracas de independência
  - Falta de informação lexical

2

GLCP: problemas

- 2 principais limitações
  - Suposições fracas de independência
    - A probabilidade de uma regra independe de onde ela é usada
      - SN→art subst [0.28]
      - SN→pronome [0.25]
    - Sabe-se que isso não é verdade
      - Pronomes são muito mais prováveis de acontecerem como sujeito → recuperam o tópico ou a informação antiga
      - Sintagmas nominais não pronominais são mais prováveis como objeto → introduzem informação nova

3

GLCP: problemas

- 2 principais limitações
  - Suposições fracas de independência
    - Estudo para o inglês (Francis et al., 1999)

	Pronome	Não pronome
Sujeito	91%	9%
Objeto	34%	66%

    - Para representar tal fenômeno, faz-se necessário ter a informação do pai do elemento sendo expandido

4

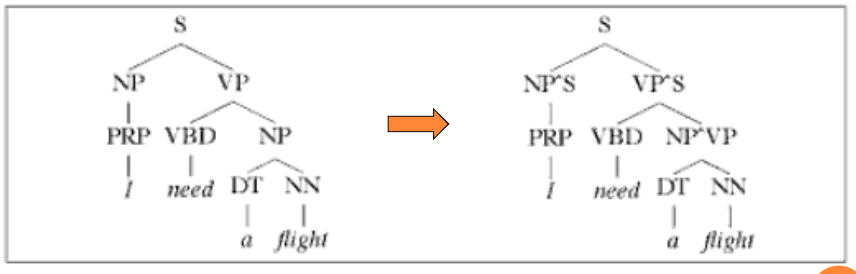
GLCP: problemas

- 2 principais limitações
  - Suposições fracas de independência
    - Solução possível: dividir as regras
      - $SN_{SUJEITO}$→pronome [0.91]
      - SN_{OBJETO}→pronome [0.34]
    - Forma de implementação: anexar a cada símbolo o símbolo de seu nó pai → nó_filho^nó_pai

5

GLCP: problemas

- 2 principais limitações
  - Suposições fracas de independência

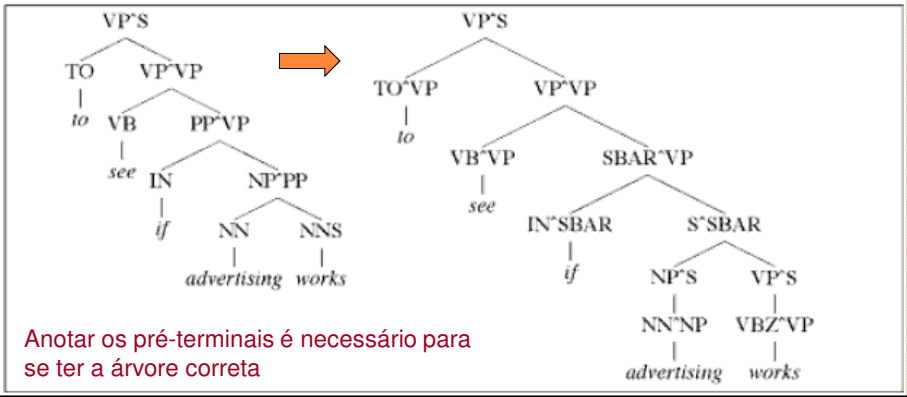


6

Sem anotar os pré-terminais (etiquetas morfossintáticas)

GLCP: problemas

- 2 principais limitações
  - Suposições fracas de independência



Anotar os pré-terminais é necessário para se ter a árvore correta

GLCP: problemas

- 2 principais limitações
  - Suposições fracas de independência
    - Anotar os pré-terminais permite representar mais fenômenos
      - Por exemplo, SVs são comuns com o **advérbio^SV** *não* e SNs são comuns com os **advérbios^SN** *apenas* e *somente*

8

GLCP: problemas

- 2 principais limitações
  - Suposições fracas de independência
    - Problemas dessa abordagem?

9

GLCP: problemas

- 2 principais limitações
  - Suposições fracas de independência
    - Problemas dessa abordagem?
      - Aumento do tamanho da gramática
      - Dados mais esparsos

→ há procedimentos automáticos para se achar o nível ótimo de anotação

10

GLCP: problemas

- 2 principais limitações
  - Falta de informação lexical
    - Informação lexical é determinante para se decidir onde ligar sintagmas preposicionais

Workers dumped sacks into a bin.

MAIS PROVÁVEL: *dumped* e *into* têm mais afinidade do que *sacks* e *into*

VS.

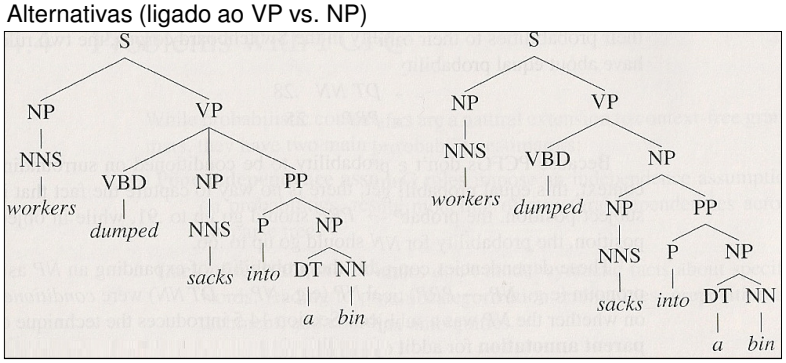
Workers dumped sacks into a bin.

11

GLCP: problemas

- 2 principais limitações
  - Falta de informação lexical

Alternativas (ligado ao VP vs. NP)

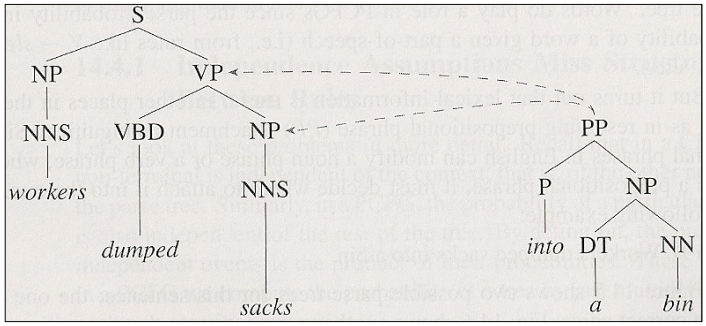


12

GLCP: problemas

- 2 principais limitações
  - Falta de informação lexical

Alternativas (mesmas regras, ligações diferentes)



13

GLCP: problemas

- 2 principais limitações
  - Falta de informação lexical
    - Informação lexical é determinante para se decidir onde ligar sintagmas preposicionais

Fishermen caught tons of herring.

VS.

Fishermen caught tons of herring.

MAIS PROVÁVEL: *tons* e *of* têm mais afinidade do que *caught* e *of*

14

GLCP: problemas

- 2 principais limitações
  - Falta de informação lexical
    - Informação lexical é determinante para resolver coordenações

dogs in houses and cats

- *[dogs in houses] and [cats]*

- *dogs in [houses and cats]*

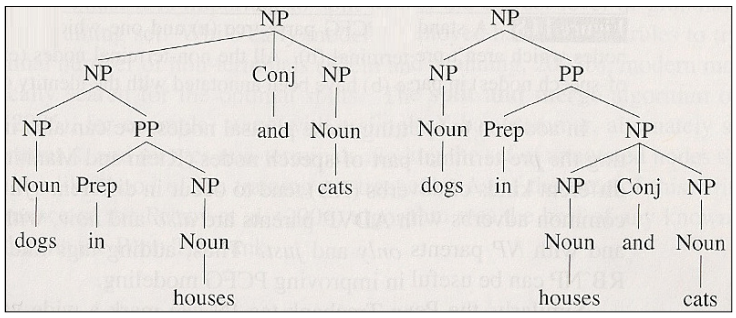
MAIS PROVÁVEL: *dogs* e *cats* são mais afins... e *dogs* não cabem dentro de *cats*

15

GLCP: problemas

- 2 principais limitações
  - Falta de informação lexical

Alternativas



16

GLCP: problemas

- 2 principais limitações
  - Falta de informação lexical
    - É necessário estender as GLCPs para lidar com dependências lexicais

17

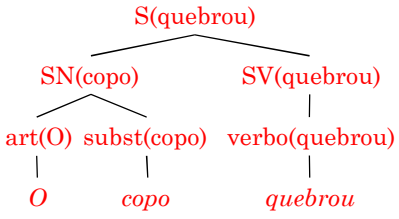
GLCP Lexicalizadas

- Modelos mais utilizados hoje
  - Parsers de Collins (1999) e Charniak (1997)
- Vantagens
  - Alternativa para a divisão de regras
  - Considera dependência lexical
    - Em vez de se alterarem as regras, altera-se o modelo probabilístico

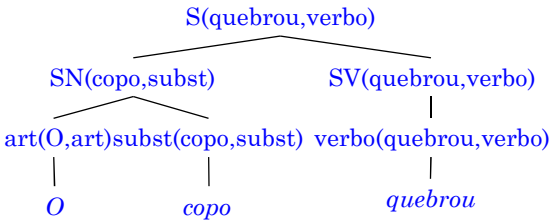
18

GLCP LEXICALIZADAS

- Extensão de modelos anteriores
  - Além da *head*, a *tag*



S(quebrou) → SN(copo), SV(quebrou)
SN(copo) → art(O), subst(copo)
...



S(quebrou,verbo) → SN(copo,subst), SV(quebrou,verbo)
SN(copo,subst) → art(O,art), subst(copo,subst)
...

19

GLCP LEXICALIZADAS

- Dois tipos de regras
  - Regras <u>lexicais</u>
    - subst(copo,subst)→copo
      - Atenção: probabilidade 1, pois não há outra opção (o terminal está explícito)
  - Regras <u>internas</u>
    - S(quebrou,verbo) → SN(copo,subst), SV(quebrou,verbo)
      - Probabilidades precisam ser estimadas

20

GLCP Lexicalizadas

- Estimativas de probabilidades
  - Regras internas
    - S(quebrou,verbo) → SN(copo,subst), SV(quebrou,verbo)

$$P(regra) = \frac{\text{Número}(\text{S}(\text{quebrou}, \text{verbo}) \rightarrow \text{SN}(\text{copo},\text{subst}), \text{SV}(\text{quebrou}, \text{verbo}))}{\text{Número}(\text{S}(\text{quebrou}, \text{verbo}))}$$

21

GLCP Lexicalizadas

- Estimativas de probabilidades
  - Regras internas
    - S(quebrou,verbo) → SN(copo,subst), SV(quebrou,verbo)

$$P(regra) = \frac{\text{Número}(\text{S}(\text{quebrou}, \text{verbo}) \rightarrow \text{SN}(\text{copo},\text{subst}), \text{SV}(\text{quebrou}, \text{verbo}))}{\text{Número}(\text{S}(\text{quebrou}, \text{verbo}))}$$

    - Qual o problema?

22

GLCP Lexicalizadas

- Estimativas de probabilidades
  - Regras internas
    - S(quebrou,verbo) → SN(copo,subst), SV(quebrou,verbo)

$$P(regra) = \frac{\text{Número}(\text{S}(\text{quebrou}, \text{verbo}) \rightarrow \text{SN}(\text{copo},\text{subst}), \text{SV}(\text{quebrou}, \text{verbo}))}{\text{Número}(\text{S}(\text{quebrou}, \text{verbo}))}$$

    - Qual o problema?
      - Regras muito mais específicas
      - Dados mais esparsos ainda
        - Maioria das probabilidades será zero!

23

    - Solução: ?

GLCP Lexicalizadas

- Estimativas de probabilidades
  - Regras internas
    - S(quebrou,verbo) → SN(copo,subst), SV(quebrou,verbo)

$$P(regra) = \frac{\text{Número}(\text{S}(\text{quebrou}, \text{verbo}) \rightarrow \text{SN}(\text{copo},\text{subst}), \text{SV}(\text{quebrou}, \text{verbo}))}{\text{Número}(\text{S}(\text{quebrou}, \text{verbo}))}$$

    - Qual o problema?
      - Regras muito mais específicas
      - Dados mais esparsos ainda
        - Maioria das probabilidades será zero!

24

    - Solução: mais suposições de independência!

GLCP Lexicalizadas

- Estimativas de probabilidades
  - Modelo 1 do parser de Collins
    - Lado Direito da Regra (LDR): uma *head* + símbolos que precedem a *head* + símbolos que seguem a *head*
    - LER → E_N E_{N-1} ... E_1 *head* D_1 ... D_{M-1} D_M
  - Cálculo das probabilidades
    - Dado o lado esquerda da regra, computa-se a probabilidade se gerar a *head*
    - A partir da *head* e do lado esquerdo, gera-se cada um dos símbolos que precedem e seguem a *head*, individualmente
      - Deve-se controlar quando parar de gerar símbolos à esquerda e à direita da *head*

25

GLCP Lexicalizadas

- Estimativas de probabilidades
  - Exemplo
    - S(quebrou,verbo) → SN(copo,subst), SV(quebrou,verbo)
      - P(regra) =
        P_{HEAD}(SV(quebrou,verbo) | S(quebrou,verbo)) *
        P_{ESQ}(SN(copo,subst) | S,SV(quebrou,verbo))
    - Mais simples de se calcular, com menos dados esparsos

26

GLCP Lexicalizadas

- Estimativas de probabilidades
  - Variações dos modelos de Collins
    - Distância entre elementos
    - Subcategorização de verbos, identificando argumentos e adjuntos
    - Somente a *tag* em vez da *head* e da *tag*
    - Palavras "curinga"
    - Etc.

27

GLCP Lexicalizadas

- Collins (2003)
  - Extensão do CKY, incluindo as probabilidades e as lexicalizações

28

Re-ranqueamento de análises

- Modelos gerativos como os anteriores são muito bons
  - Relativamente fácil calcular probabilidades
  - Bons resultados
- Mas é difícil incorporar conhecimento externo
  - Por exemplo
    - Árvores sintáticas tendem a "pender para a direita"
    - Constituintes mais longos acontecem no fim da árvore
    - Certos falantes/escritores têm preferências por estruturas sintáticas particulares → questões de estilo de escrita

29

Re-ranqueamento de análises

- Possível solução
  - Re-ranqueamento discriminativo
    - Produz-se um ranque com as N melhores (mais prováveis) árvores sintáticas
      - Chamada *N-best list*
    - Novo ranqueamento com base em um conjunto de atributos relevantes
      - Por exemplo, probabilidade, regras aplicadas, número de ocorrências de cada constituinte, bigramas de não terminais adjacentes na árvore, etc.
    - Escolhe-se a melhor árvore

30

Re-ranqueamento de análises

- Possível solução
  - Re-ranqueamento discriminativo
    - Atenção: a qualidade do método depende diretamente da qualidade da *N-best list*
      - Se a análise correta não estiver na lista ou estiver muito mal ranqueada, o método será provavelmente ruim

31

Processamento humano & probabilidade

- Experimentos com humanos: **probabilidades na mente!**
  - Estruturas e palavras mais previsíveis (prováveis) são lidas mais rapidamente por humanos
    - Como se mede isso?

32

PROCESSAMENTO HUMANO & PROBABILIDADE

- Experimentos com humanos: **probabilidades na mente!**
  - Estruturas e palavras mais previsíveis (prováveis) são lidas mais rapidamente por humanos
    - Medidas empíricas: por exemplo, entropia vs. rastreamento do movimento dos olhos

33

PROCESSAMENTO HUMANO & PROBABILIDADE

- Experimentos com humanos: **probabilidades na mente!**
  - Humanos desambiguam análises, preferindo análises mais prováveis
    - Sentenças ***garden-path***: temporariamente ambíguas
      - *The students forgot the solution was in the back of the book.*
      - *The horse raced past the barn fell.*
      - *The complex houses married and single students and their families.*

34

PROCESSAMENTO HUMANO & PROBABILIDADE

- Experimentos com humanos: **probabilidades na mente!**
  - Humanos desambiguam análises, preferindo análises mais prováveis
    - Sentenças ***garden-path***: temporariamente ambíguas
      - *Por mais que Jorge continuasse lendo as histórias aborreciam as crianças da creche.*
      - *Maria beijou João e o irmão dele arregalou os olhos de espanto.*

35

PROCESSAMENTO HUMANO & PROBABILIDADE

- Experimentos com humanos: **probabilidades na mente!**
  - Humanos desambiguam análises, preferindo análises mais prováveis
    - Sentenças ***garden-path***: há casos mais complexos!
      - *Um navio brasileiro entrava na baía um navio japonês.*

36