Processamento lexical, morfologia e morfossintaxe

SCC5908 Introdução ao Processamento de Língua Natural

Thiago A. S. Pardo

Preâmbulo

- Em processamento de texto, é comum
  - Substituir uma palavra por outra
  - Procurar por uma informação, como data, nome, etc.
  - Analisar determinadas palavras
  - Mais genericamente, procurar por padrões no texto
    - Padrões simples: palavras
    - Padrões mais complexos: expressões, segmentos maiores

2

Exemplo

- Busca por todos os valores monetários em um texto

Levantamento da consultoria Economática aponta que empresas brasileiras de capital aberto tiveram os maiores lucros em 2010, considerando somente o setor de construção civil das Américas.

No topo da lista está a PDG Realty, com um lucro líquido de US$ 473,9 milhões apurados em 2010, bem acima da segunda colocada, a americana Walter Industries, com ganhos de US$ 385,8 milhões para o mesmo período.

As companhias brasileiras ocupam as próximas cinco posições (da 3ª à 7ª) no ranking preparado pela Economática, a saber: MRV, Cyrela, Gafisa, Brookfield e Rossi Residencial. Numa lista de dez posições, sete são ocupados por companhias nacionais.

A consultoria também preparou uma lista baseada em volume de vendas, desta vez com presença mais expressiva das construtoras americanas, a exemplo do primeiro lugar -- a Pulte Homes-- com um volume de US$ 4,44 bilhões em imóveis comercializados, e do segundo lugar --a Horton-- com uma cifra de US$ 3,96 bilhões.

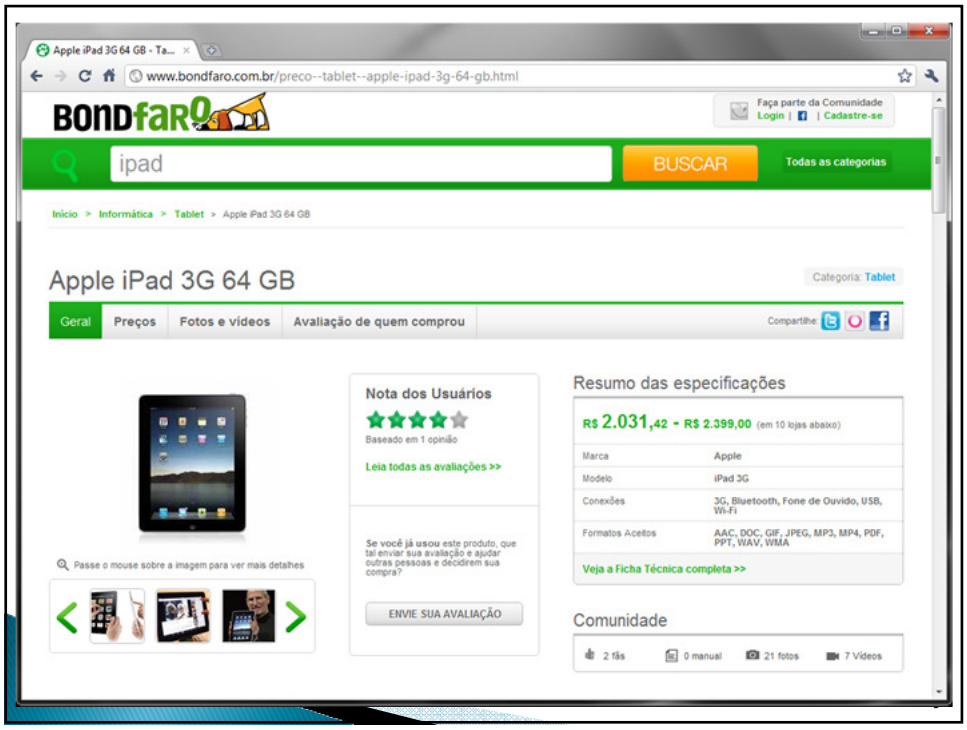
...

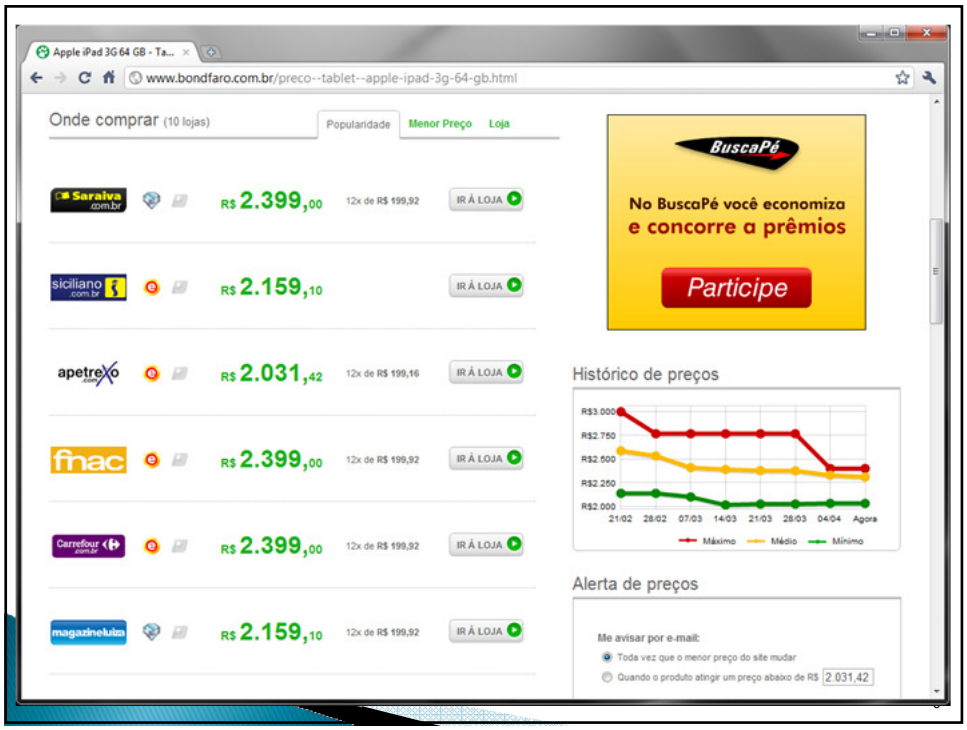
3

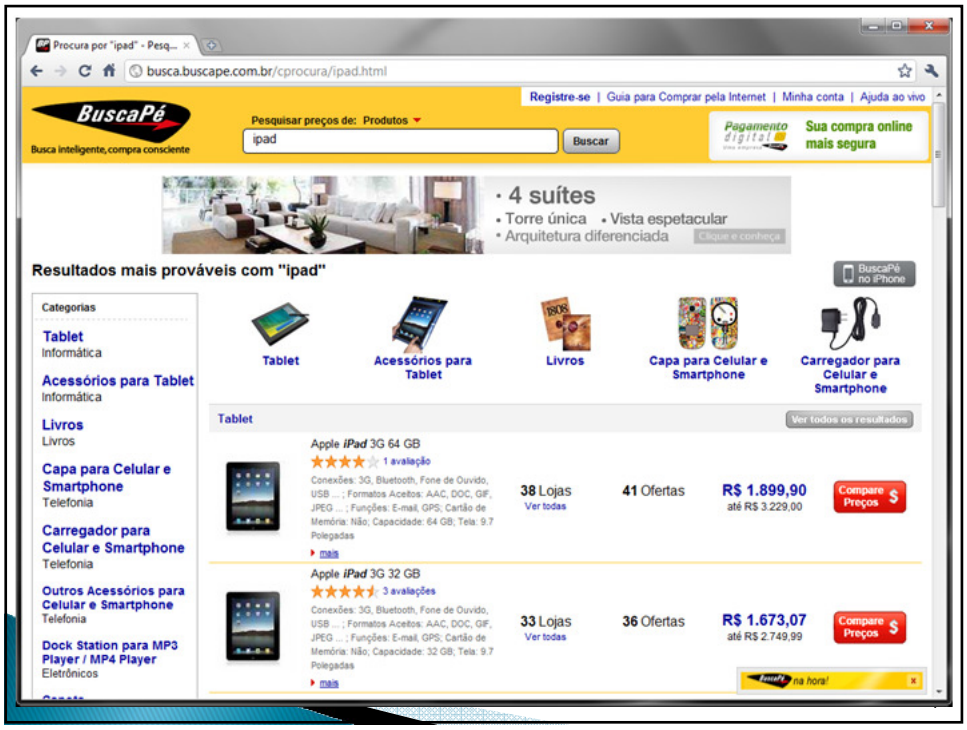
Processamento textual

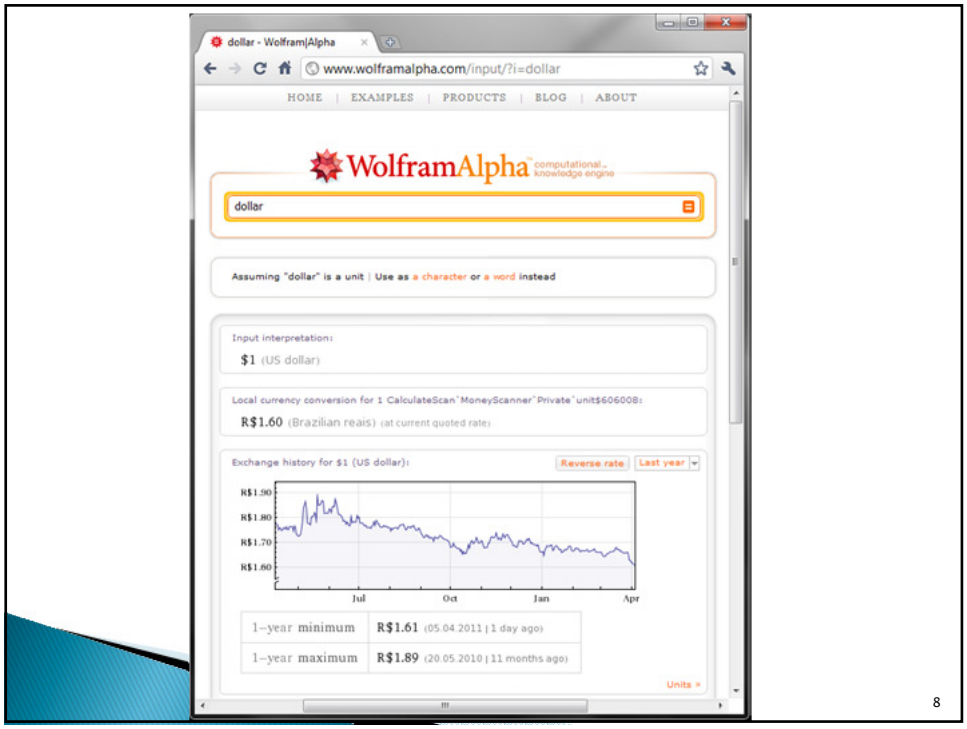
- Útil para
  - Tarefas particulares: buscar algo que leu
  - Tarefas científicas: sintomas e tratamentos de uma doença
  - Tarefas comerciais: sistemas on-line

4


Apple iPad 3G 64 GB - Ta...
www.bondfaro.com.br/preco--tablet--apple-ipad-3g-64-gb.html
BONDFARO
Faça parte da Comunidade
Login | | Cadastre-se
ipad
BUSCAR
Todas as categorias
Inicio > Informática > Tablet > Apple iPad 3G 64 GB
Apple iPad 3G 64 GB
Categoria: Tablet
Geral
Preços
Fotos e vídeos
Avaliação de quem comprou
Compartilhe:
Passe o mouse sobre a imagem para ver mais detalhes
Nota dos Usuários
Baseado em 1 opinião
Leia todas as avaliações >>
Se você já usou este produto, que tal enviar sua avaliação e ajudar outras pessoas a decidirem sua compra?
ENVIE SUA AVALIAÇÃO
Resumo das especificações
R$ 2.031,42 - R$ 2.399,00 (em 10 lojas abaixo)
Marca Apple
Modelo iPad 3G
Conexões 3G, Bluetooth, Fone de Ouvido, USB, Wi-Fi
Formatos Aceitos AAC, DOC, GIF, JPEG, MP3, MP4, PDF, PPT, WAV, WMA
Veja a Ficha Técnica completa >>
Comunidade
2 fãs
0 manual
21 fotos
7 Vídeos


Apple iPad 3G 64 GB - Ta...
www.bondfaro.com.br/preco--tablet--apple-ipad-3g-64-gb.html
Onde comprar (10 lojas)
Popularidade
Menor Preço
Loja
Saraiva .com.br R$ 2.399,00 12x de R$ 199,92 IR À LOJA
siciliano .com.br R$ 2.159,10 IR À LOJA
apetrexo .com.br R$ 2.031,42 12x de R$ 199,16 IR À LOJA
fnac R$ 2.399,00 12x de R$ 199,92 IR À LOJA
Carrefour .com.br R$ 2.399,00 12x de R$ 199,92 IR À LOJA
magazineluiza R$ 2.159,10 12x de R$ 199,92 IR À LOJA
BuscaPé
No BuscaPé você economiza e concorre a prêmios
Participe
Histórico de preços
R$3.000
R$2.750
R$2.500
R$2.250
R$2.000
21/02 28/02 07/03 14/03 21/03 28/03 04/04 Agora
Máximo
Médio
Mínimo
Alerta de preços
Me avisar por e-mail:
Toda vez que o menor preço do site mudar
Quando o produto atingir um preço abaixo de R$ 2.031,42


Procura por "ipad" - Pesq...
busca.buscape.com.br/cprocura/ipad.html
BuscaPé
Busca inteligente, compra consciente
Pesquisar preços de: Produtos
ipad
Buscar
Registre-se | Guia para Comprar pela Internet | Minha conta | Ajuda ao vivo
Pagamento digital
Sua compra online mais segura
· 4 suítes
• Torre única • Vista espetacular
• Arquitetura diferenciada
Clique e conheça
Resultados mais prováveis com "ipad"
BuscaPé no iPhone
Categorias
Tablet
Informática
Acessórios para Tablet
Informática
Livros
Livros
Capa para Celular e Smartphone
Telefonia
Carregador para Celular e Smartphone
Telefonia
Outros Acessórios para Celular e Smartphone
Telefonia
Dock Station para MP3 Player / MP4 Player
Eletrônicos
Tablet
Acessórios para Tablet
Livros
Capa para Celular e Smartphone
Carregador para Celular e Smartphone
Tablet
Ver todos os resultados
Apple iPad 3G 64 GB
1 avaliação
38 Lojas
Ver todas
41 Ofertas
R$ 1.899,90
até R$ 3.229,00
Compare Preços
Apple iPad 3G 32 GB
3 avaliações
33 Lojas
Ver todas
36 Ofertas
R$ 1.673,07
até R$ 2.749,99
Compare Preços
mais


dollar - Wolfram|Alpha
www.wolframalpha.com/input/?i=dollar
HOME | EXAMPLES | PRODUCTS | BLOG | ABOUT
WolframAlpha computational knowledge engine
dollar
Assuming "dollar" is a unit | Use as a character or a word instead
Input interpretation:
$1 (US dollar)
Local currency conversion for 1 CalculateScan`MoneyScanner`Private`unit$606008:
R$1.60 (Brazilian reais) (at current quoted rate)
Exchange history for $1 (US dollar):
Reverse rate
Last year
R$1.90
R$1.80
R$1.70
R$1.60
Jul
Oct
Jan
Apr
1-year minimum
R$1.61 (05.04.2011 | 1 day ago)
1-year maximum
R$1.89 (20.05.2010 | 11 months ago)
Units »

8

Exemplo

- Busca por todos os valores monetários em um texto

Levantamento da consultoria Economática aponta que empresas brasileiras de capital aberto tiveram os maiores lucros em 2010, considerando somente o setor de construção civil das Américas.

No topo da lista está a PDG Realty, com um lucro líquido de US$ 473,9 milhões apurados em 2010, bem acima da segunda colocada, a americana Walter Industries, com ganhos de US$ 385,8 milhões para o mesmo período.

As companhias brasileiras ocupam as próximas cinco posições (da 3ª à 7ª) no ranking preparado pela Economática, a saber: MRV, Cyrela, Gafisa, Brookfield e Rossi Residencial. Numa lista de dez posições, sete são ocupados por companhias nacionais.

A consultoria também preparou uma lista baseada em volume de vendas, desta vez com presença mais expressiva das construtoras americanas, a exemplo do primeiro lugar -- a Pulte Homes-- com um volume de US$ 4,44 bilhões em imóveis comercializados, e do segundo lugar --a Horton-- com uma cifra de US$ 3,96 bilhões.

...

Como fariam?

9

Expressões Regulares (ER)

- Notação tradicional para caracterizar segmentos textuais de todo tipo
  - Especificam sequências de símbolos a serem buscados/caracterizados
  - Vários sistemas de busca de expressões regulares
    - grep, no Linux/UNIX
    - Lex/flex
      - Há variações de sistema para sistema, mas são muito parecidas

10

ER: notação

- Exemplos
  - Casamento direto: **preço**
  - Letra maiúscula ou minúscula no início: **[Pp]reço**
    - [ ] indicam disjunção, ou seja, um único elemento do conjunto
  - Identificação de um único dígito do texto: **[0123456789]**
  - Identificação de uma letra em um intervalo de letras: **[a-z]**
  - Qualquer caractere diferente de a: **^a**

11

ER: notação

- Exemplos
  - Singular ou plural: **preços?**
  - 1 ou mais ocorrências (+) de algum elemento: **Aa+i+!**
    - Aai!, Aaaaaaiiiii!
  - 0 ou maiz ocorrências (*) de algum elemento: **Aa*i*!**
    - Aaaaiii!, Aaaiiiii!, Ai!, Aaaa!
  - Caractere curinga (.): **beg.n**
    - begin, began, begun
  - Alternativa (|): **preço|os** ou **(gato)|(cão)**
    - O que acontece se tivermos **gato|cão** sem parênteses?

12

Exercícios

- Como identificar nomes próprios?
- E e-mails?

Exercícios

- Como identificar nomes próprios?
  - [A-Z][a-z]+
- E e-mails?
  - [a-z0-9_]+@[a-z\.]+
    - Cuidado: alguns caracteres são especiais e, para serem usados em seu sentido original, precisam de \ ou “”
      - Exemplos: . $ -

Exercício

- Expressão regular para reconhecer os valores monetários?

Levantamento da consultoria Economática aponta que empresas brasileiras de capital aberto tiveram os maiores lucros em 2010, considerando somente o setor de construção civil das Américas.

No topo da lista está a PDG Realty, com um lucro líquido de US$ 473,9 milhões apurados em 2010, bem acima da segunda colocada, a americana Walter Industries, com ganhos de US$ 385,8 milhões para o mesmo período.

As companhias brasileiras ocupam as próximas cinco posições (da 3ª à 7ª) no ranking preparado pela Economática, a saber: MRV, Cyrela, Gafisa, Brookfield e Rossi Residencial. Numa lista de dez posições, sete são ocupados por companhias nacionais.

A consultoria também preparou uma lista baseada em volume de vendas, desta vez com presença mais expressiva das construtoras americanas, a exemplo do primeiro lugar -- a Pulte Homes-- com um volume de US$ 4,44 bilhões em imóveis comercializados, e do segundo lugar --a Horton-- com uma cifra de US$ 3,96 bilhões.

...

15

Exercício

- Expressão regular para reconhecer os valores monetários?

Levantamento da consultoria Economática aponta que empresas brasileiras de capital aberto tiveram os maiores lucros em 2010, considerando somente o setor de construção civil das Américas.

No topo da lista está a PDG Realty, com um lucro líquido de US$ 473,9 milhões apurados em 2010, bem acima da segunda colocada, a americana Walter Industries, com ganhos de US$ 385,8 milhões para o mesmo período.

As companhias brasileiras ocupam as próximas cinco posições (da 3ª à 7ª) no ranking preparado pela Economática, a saber: MRV, Cyrela, Gafisa, Brookfield e Rossi Residencial. Numa lista de dez posições, sete são ocupados por companhias nacionais.

A consultoria também preparou uma lista baseada em volume de vendas, desta vez com presença mais expressiva das construtoras americanas, a exemplo do primeiro lugar -- a Pulte Homes-- com um volume de US$ 4,44 bilhões em imóveis comercializados, e do segundo lugar --a Horton-- com uma cifra de US$ 3,96 bilhões.

...

US\$ [0-9]+,[0-9]+ [mb]ilhões

16

Exercício

Sentença:

O *homem* viu a mulher de binóculos na montanha.

Análise automática:

O_DET homem_N viu_V a_DET mulher_N de_PRP binóculos_N em_PRP a_DET montanha_N .

Expressão regular para os substantivos e os verbos?

17

Exercício

Sentença:

O *homem* viu a mulher de binóculos na montanha.

Análise automática:

O_DET homem_N viu_V a_DET mulher_N de_PRP binóculos_N em_PRP a_DET montanha_N .

Expressão regular para os substantivos e os verbos?

[A-Za-z][a-z]*_N|V

18

Exercício

Sentença:

O *homem* viu a mulher de binóculos na montanha.

Análise automática:

O_DET homem_N viu_V a_DET mulher_N de_PRP binóculos_N em_PRP a_DET montanha_N .

Expressão para substantivos seguidos de verbos?

19

Exercício

Sentença:

O *homem* viu a mulher de binóculos na montanha.

Análise automática:

O_DET homem_N viu_V a_DET mulher_N de_PRP binóculos_N em_PRP a_DET montanha_N .

Expressão para substantivos seguidos de verbos?

[A-Za-z][a-z]*_N [a-z]+_V

20

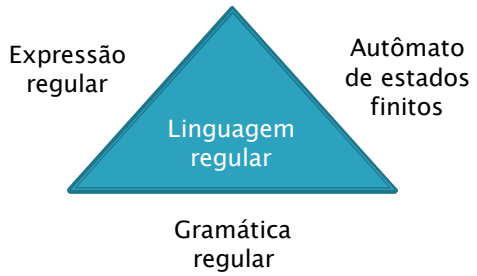
Autômatos

- Expressões regulares implementadas como autômatos de estados finitos
  - Autômato: modelo matemático eficaz e elegante para lidar com expressões regulares
- Autômatos utilizados para revisão ortográfica, síntese e reconhecimento de fala, extração de informação, tradução automática, análise morfológica, análise morfossintática, etc.

21

Autômatos

- Poder representacional equivalente



22

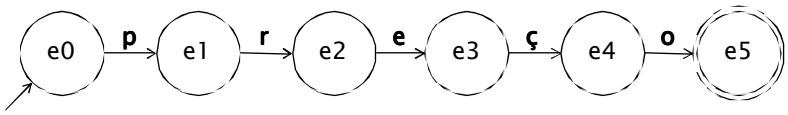
Autômatos

- **Componentes**
  - Estados que modelam o "sistema"
    - Pontos da análise sendo realizada, por exemplo
  - Símbolos de entrada
    - Letras das palavras, números, símbolos, etc.
  - Estados inicial e final
    - Início e fim do processo
  - Transições entre estados

23

Exemplo

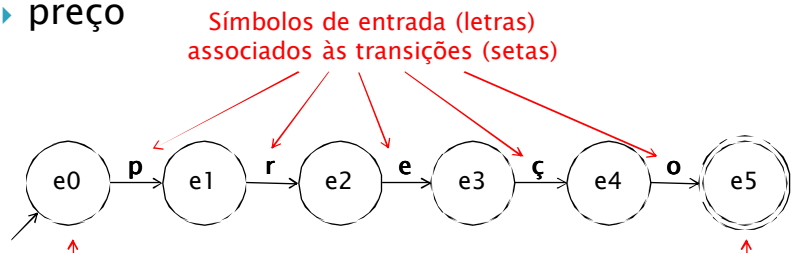
- preço



24

Exemplo

- preço



Estado inicial (indicado pela seta, em geral)

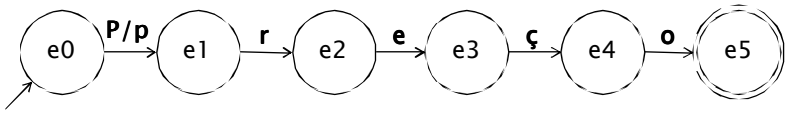
Começa-se no e0: a cada transição, percorre-se uma letra da palavra de entrada; se atingiu estado final, palavra reconhecida

Estado final (indicado pelo contorno duplo, em geral)

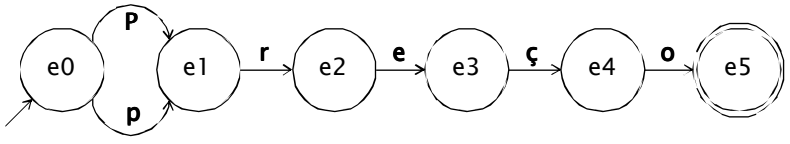
25

Exemplo

- [Pp]reço



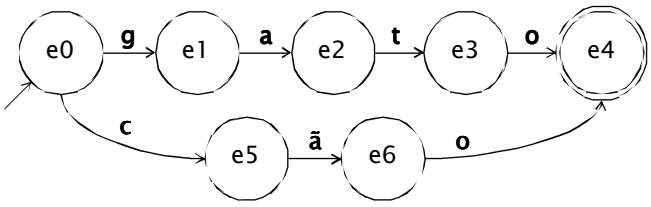
ou



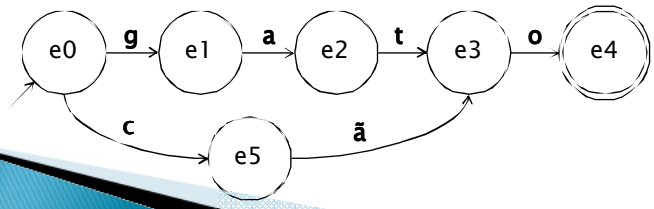
26

Exemplo

- (gato)|(cão)



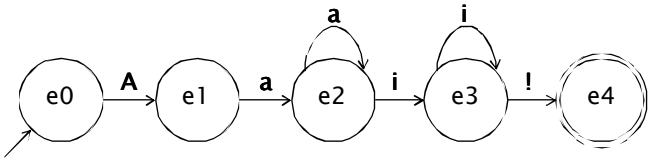
ou



27

Exemplo

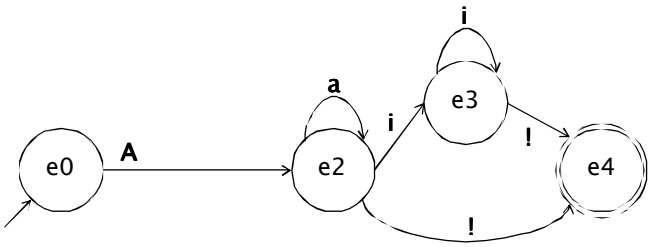
- Aa+i+!



28

Exemplo

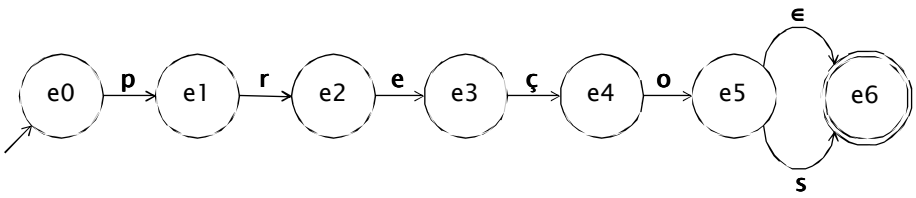
- Aa*i*!



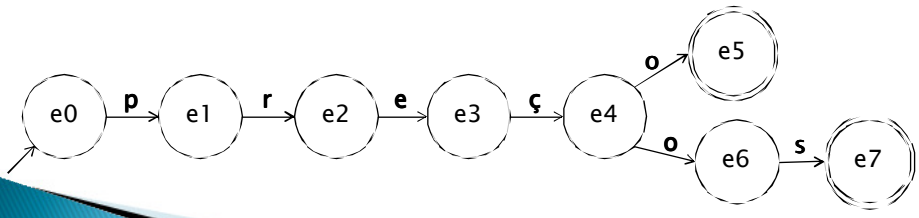
29

Exemplo

- preços?



ou



30

Exercício em duplas

- Criar autômato para reconhecer valores monetários
  - US\$ [0-9]+,[0-9]+ [mb]ilhões

31

Autômatos

- Variações
  - Transdutores
    - Além de reconhecerem a entrada, geram saída
      - Usados em análise morfológica
  - Modelos de Markov
  - Redes de transição

32

Análise morfológica

- *Parsing* morfológico
  - Analisar uma palavra e identificar seus componentes
    - Morfemas
    - Possibilidades
      - meninos → lema (menino), masculino (o), plural (+s), subst
      - meninos → radical (menin), masculino (+o), plural (+s), subst
      - meninas → lema (menino), feminino (–o +a), plural (s), subst

33

Análise morfológica

- Relevância da tarefa
  - Reconhecer palavras e suas variações
    - Revisão ortográfica, busca na web, sumarização, extração de informação
    - Stemming, lematização
  - ... e também produzir a forma adequada das palavras, derivar palavras novas, lidar com neologismos
    - Geração textual, tradução automática
      - "Máquina morfológica"
  - Caracterização léxica da língua, no geral

34

Terminologia básica

- Morfemas: unidade mínima de significado
  - Raiz/radical
    - Alguns diferenciam esses termos, outros não
  - Afixos

- Afixos
  - Prefixo: desamor, infeliz
  - Sufixo: lealdade, facilmente, quebrado, comia
  - Infixo: rabiscar
    - Raro, alguns dizem que não existe para o português
  - Circunfixo: anoitecer, descampado

35

Terminologia básica

- Morfe
  - Realização de um morfema
    - Morfema é abstrato, enquanto morfe é concreto
    - Exemplo: morfema de negação pode ser expresso pelos morfes in (de infeliz) ou i (de imutável)

- Alomorfes
  - Morfes que expressam um mesmo morfema
    - In e i para negação
    - Ante, pré e pró para anterioridade

36

Terminologia básica

- Processos principais de formação de palavras
  - Flexional: variações de uma mesma palavra
    - Flexão nominal: número, gênero
    - Flexão verbal: modo-tempo, número-pessoa
      - Adição de morfemas gramaticais
  - Derivacional: palavras novas
    - Podem mudar classe e sentido
      - "modelo" → "modelagem"
        - Adição de morfemas lexicais

37

Análise morfológica

- Para construir um parser morfológico, são necessários
  - Léxico
    - Radicais e afixos e suas possíveis classificações (substantivos, verbos, etc.)
  - Conhecimento de morfotática
    - Como os morfemas se ordenam para que as palavras se formem
      - Exemplo: em português, o morfema de plural aparece após e substantivo, e não antes
      - "Sintaxe da morfologia"
  - Regras ortográficas
    - Modelam mudanças que ocorrem nas palavras quando morfemas se combinam
      - Exemplo: casa+PL=casaS, mas flor+PL=florES

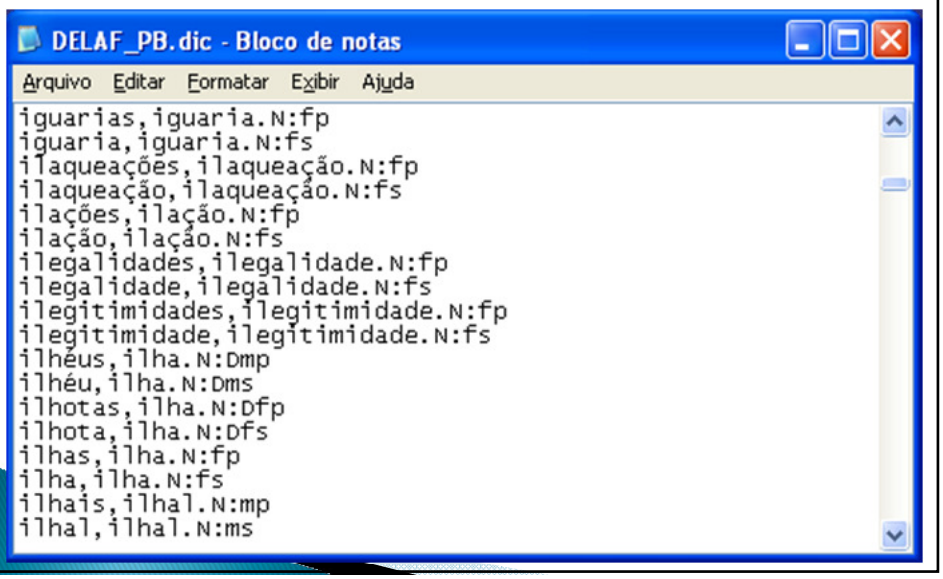
38

Análise morfológica

- Alternativa 1
  - Listagem de palavras
    - Exaustiva: léxico de formas analisadas (também chamadas flexionadas ou plenas)
      - Palavras com todas as suas variações
        - Pouca economia, redundância, compactação de arquivos

39

Exemplo do UNITEX-PB



Análise morfológica

- Alternativa 1
  - Listagem de palavras
    - Econômica: léxico de raízes (ou de morfemas)
      - Listagem de raízes + regras de formação das palavras (morfotática e regras ortográficas)
        - Mais economia, processo mais caro

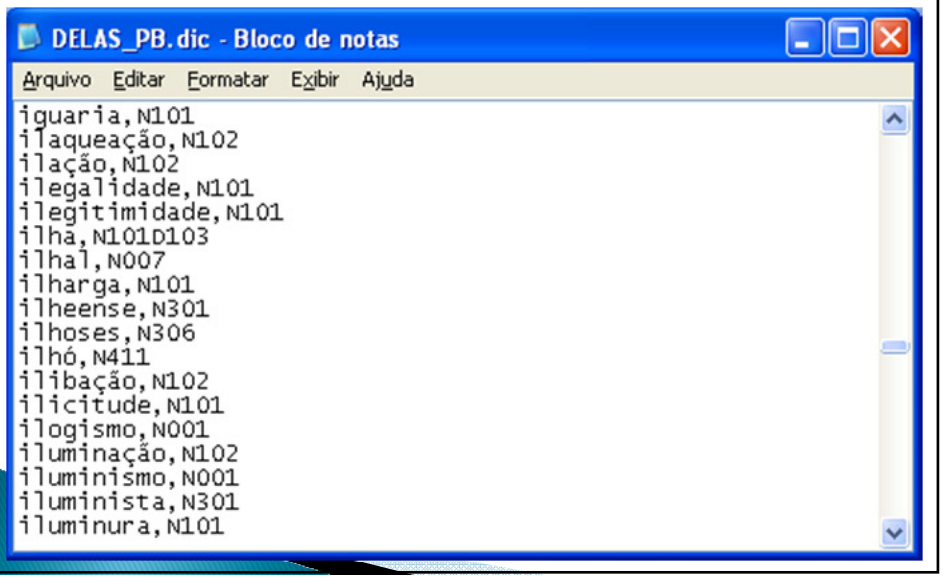
41

Análise morfológica

- Alternativa 1
  - Listagem de palavras
    - Meio termo
      - Léxico de lemas (ou formas canônicas) associados as suas variações
      - Palavras irregulares em formas plenas no léxico + léxico de raízes para palavras regulares
      - Etc.

42

Exemplo do UNITEX-PB



Análise morfológica

- Alternativa 1
  - Listagem de palavras
    - Problemas para lidar com
      - Novas palavras e variações: novos verbos (denominais, inclusive), nomes próprios, etc.
      - Línguas morfologicamente complexas
        - Turco, por exemplo

44

Turco – exemplo

(Jurafsky e Martin, 2008)

uygarlaştıramadıklarımızdanmışsınızcasına

uygar +laş +tır +ama +dık +lar +ımız +dan +mış +sınız +casına
civilized +BEC +CAUS +NABL +PART +PL +P1PL +ABL +PAST +2PL +AsIf

"(behaving) as if you are among those whom we could not civilize"

+BEC	"become"
+CAUS	the causative verb marker ('cause to X')
+NABL	"not able"
+PART	past participle form
+P1PL	1st person pl possessive agreement
+2PL	2nd person pl
+ABL	ablative (from/among) case marker
+AsIf	derivationally forms an adverb from a finite verb

45

Análise morfológica

- Alternativa 2
  - Codificação em forma de autômatos: maior eficiência computacional
    - De forma complementar com o léxico
      - Formas básicas/raízes no léxico e regras de formação de palavras (morfotática e regras ortográficas) mapeadas em autômatos
    - De forma isolada
      - Todo o léxico da língua mapeado em autômatos

46

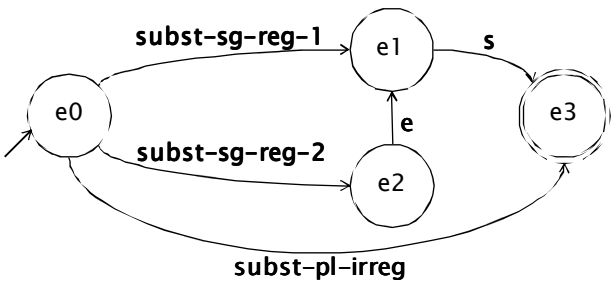
Análise morfológica

- Mapear palavras em seus componentes
  - gatos → gato + SUBST + MASC + PL
  - canto → canto + SUBST + MASC + SG
  - canto → cantar + V + 1P + SG + Pind
    - A rigor, a tarefa de desambiguar "canto" (SUBST ou V) está além da análise morfológica
      - Morfossintaxe
    - No momento, listam-se todas as possibilidades

47

Exemplo simples

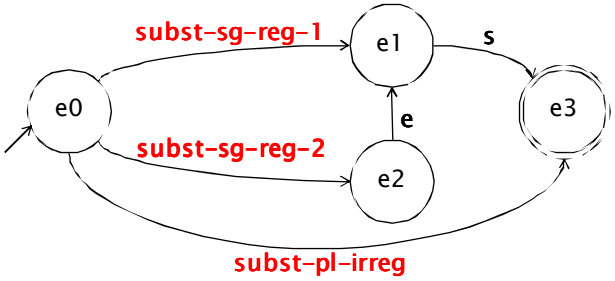
- Reconhecimento/geração de alguns substantivos no plural
  - Léxico de lemas + autômato



subst-sg-reg-1	subst-sg-reg-2	subst-pl-irreg
casa	flor	lápis
porta	lar	córpus
...	...	...

Exemplo simples

- Reconhecimento/geração de alguns substantivos no plural
  - Léxico de lemas + autômato



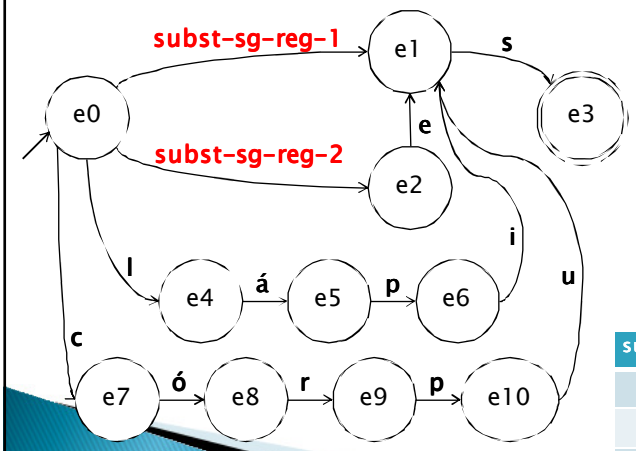
Podem ser substituídos pelos autômatos correspondentes!

Como?

subst-sg-reg-1	subst-sg-reg-2	subst-pl-irreg
casa	flor	lápis
porta	lar	córpus
...	...	...

Exemplo simples

- Reconhecimento/geração de alguns substantivos no plural
  - Léxico de lemas + autômato



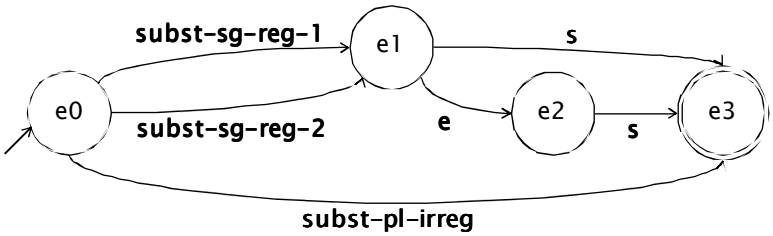
Podem ser substituídos pelos autômatos correspondentes!

Continuem!

subst-sg-reg-1	subst-sg-reg-2
casa	flor
porta	lar
...	...

Exemplo simples

- Cuidado com ***overgeneration***!
  - O que acontece no caso abaixo?

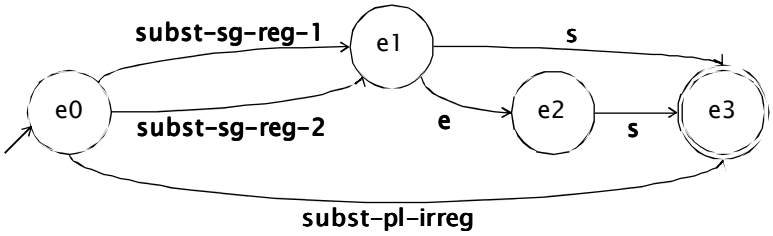


subst-sg-reg-1	subst-sg-reg-2	subst-pl-irreg
casa	flor	lápis
porta	lar	córpus
...	...	...

Exemplo simples

- Cuidado com ***overgeneration***!
  - O que acontece no caso abaixo?

casas
casaes
flors
flores
...



subst-sg-reg-1	subst-sg-reg-2	subst-pl-irreg
casa	flor	lápis
porta	lar	córpus
...	...	...

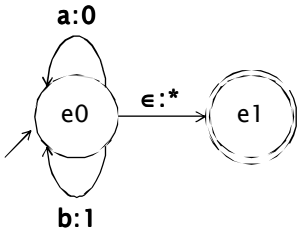
Análise morfológica

- Para nossa tarefa, precisamos de mais poder
  - Além de se reconhecer/gerar as palavras, é necessário identificar os componentes
    - gatos → gato + SUBST + MASC + PL
    - canto → canto + SUBST + MASC + SG
    - canto → cantar + V + 1P + SG + Pind
  - Transdutores
    - Reconhecem a entrada e, em paralelo, geram saída

53

Transdutores

- Lendo a_s e b_s e gerando 0_s e 1_s, respectivamente, terminando com *

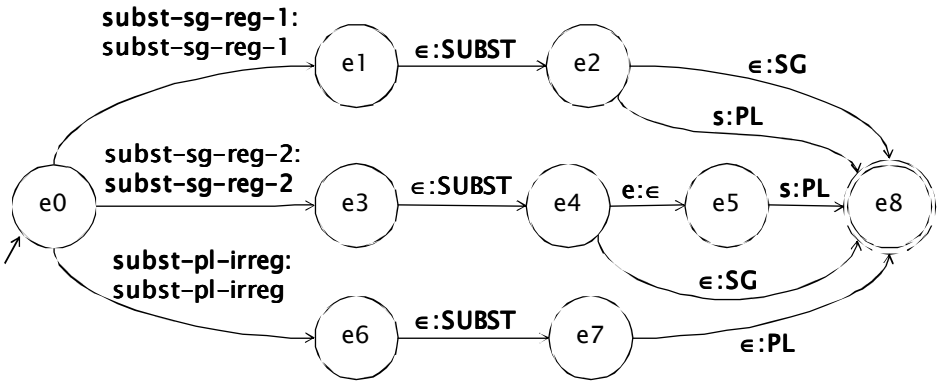


Análise de abba

54

Transdutores: exemplo

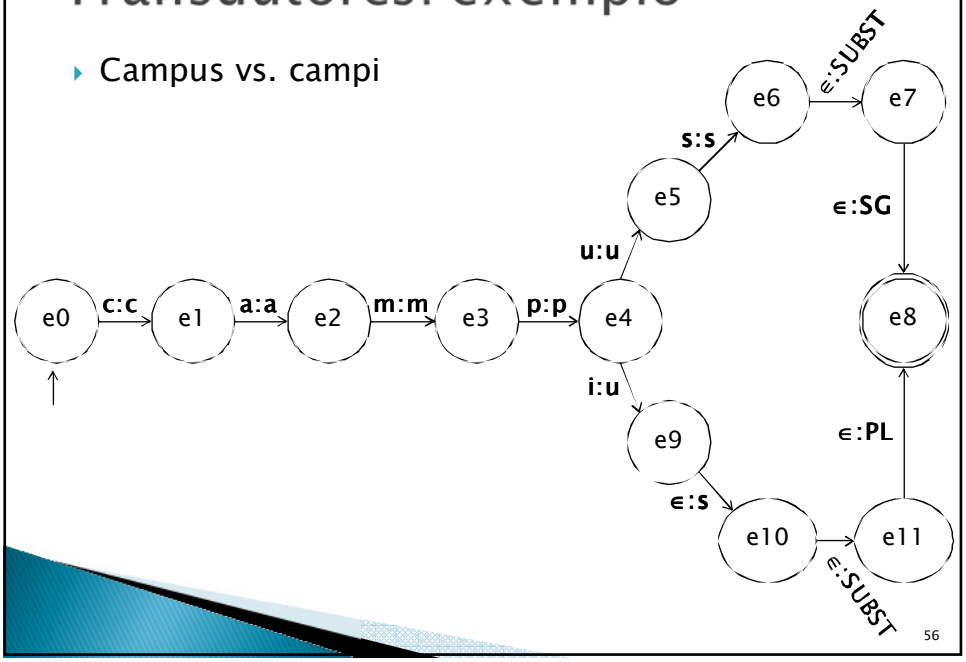
- Releitura do autômato de substantivos



subst-sg-reg-1	subst-sg-reg-2	subst-pl-irreg
casa	flor	lápis
...	...	...

Transdutores: exemplo

- Campus vs. campi



56

Transdutores: exemplo

- Menino, menina, meninos, meninas: exercício em duplas!
  - Reconhecer número, gênero, raiz e etiqueta morfossintática

57

Transdutores

- E casos como o de "canto"?
  - Como identificar que "canto" pode ser um verbo ou um substantivo, gerando-se os atributos correspondentes para cada caso?
    - canto → canto + SUBST + MASC + SG
    - canto → cantar + V + 1P + SG + Pind

58

Transdutores

- E casos como o de "canto"?
  - Como identificar que "canto" pode ser um verbo ou um substantivo, gerando-se os atributos correspondentes para cada caso?
    - canto → canto + SUBST + MASC + SG
    - canto → cantar + V + 1P + SG + Pind
  - **A palavra seria reconhecida por mais de um transdutor!**
    - Análise morfossintática para desambiguar

59

Origens da Morfossintaxe

- Dionísio Trácio, 100 AC
  - Esboço da gramática do grego
  - Cunhou o vocabulário atual
    - Sintaxe, ditongo, clítico, etc.
    - 8 etiquetas morfossintáticas: substantivo, verbo, pronome, preposição, advérbio, conjunção, particípio, artigo
      - Vocabulário usado até hoje!
- Morfossintaxe
  - Morfologia: tipos de afixos possíveis variam com a classe
  - Sintaxe: palavras com comportamentos/funções similares em seus contextos são de uma mesma classe
  - Algo mais?

60

Origens da Morfossintaxe

- Dionísio Trácio, 100 AC
  - Esboço da gramática do grego
  - Cunhou o vocabulário atual
    - Sintaxe, ditongo, clítico, etc.
    - 8 etiquetas morfossintáticas: substantivo, verbo, pronome, preposição, advérbio, conjunção, particípio, artigo
      - Vocabulário usado até hoje!
- Morfossintaxe
  - Morfologia: tipos de afixos possíveis variam com a classe
  - Sintaxe: palavras com comportamentos/funções similares em seus contextos são de uma mesma classe
  - Semântica: substantivos têm uma preferência por objetos, lugares e coisas, adjetivos por propriedades, etc.
  - Pragmática

61

Conjuntos de etiquetas

- Variam muito
  - Penn Treebank (Marcus et al., 1993): 45
  - Brown Corpus (Francis, 1979): 87
  - CLAWS 7 (Garside et al. 1997): 146
  - Palavras (Bick, 2000): 14
  - Mac-Morpho/Lácio-Web (Aluísio et al., 2003): 31

62

Exemplo: Penn Treebank

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP$	Possessive pronoun

RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Exemplo: Mac-Morpho

Tag	Definition
ADJ	open-class noun modifier
ADV-KS-REL	relative subordinating Adverb
ADV-KS	Non-relative subordinating Adverb
ADV	Non-subordinating adverb
ART	Article
KC	coordinating conjunction
KS	coordinating conjunction
IN	interjection
N	open-class noun phrase nucleus
NPROP	proper noun
NUM	numeral as a noun modifier
PCP	past participle or adjective
PDEN	emphasis/focus
PREP	preposition
PROPESS	personal pronoun
PRO-KS-REL	relative subordinating pronoun
PRO-KS	Non-relative subordinating pronoun
PROSUB	non-subordinating pronoun as a noun phrase nucleus
PROADJ	Non-subordinating pronoun as a modifier
VAUX	Auxiliary verb
V	Non-auxiliary verb
CUR	Currency symbol

Compl. Tag	Definition
|EST	foreign
|AP	apposition
|+	contraction/ enclitic
|!	mesoclitic
| [beginning, |... middle part, |] and end of discontinuous compound (further discussed in Section 3)	
|TEL	phone number
|DAT	date
|HOR	time
|DAD	formatted data not falling into above categories

64

Terminologia

- Palavras de classes fechadas, palavras funcionais
  - Conjunto de palavras varia pouco
    - Preposições, conjunções, artigos
- Palavras de classes abertas
  - O conjunto varia bastante, surgindo novas palavras
    - Substantivos, verbos
      - Conjuntos de palavras de classes abertas e fechadas não são iguais para todas as línguas
      - Nem todas as classes existem para todas as línguas ou são distinguidas das mesmas formas

65

Terminologia

- Substantivos/nomes
  - Comuns, próprios
  - Contáveis (abelha, casa), incontáveis (ar, açúcar)
- Verbos
  - Principais, auxiliares
- Advérbios
  - Tempo, local, modo, direção, etc.
- Conjunções
  - Coordenativas e subordinativas
- Pronomes
  - Pessoais, possessivos, interrogativos, relativos, etc.

66

Etiquetas morfossintáticas

- Nem sempre a distinção é simples
  - Advérbios vs. preposições
    - *Near*, *around*
  - Adjetivos vs. particípios
    - Eles estão casados.
  - Advérbios: tudo que não cabe nas outras classes

67

Etiquetação morfossintática

- Tagging, ou parsing morfossintático
  - Associação de etiquetas às palavras de uma sentença
    - Faz-se necessário, portanto, tokenização e segmentação sentencial
  - Tarefa de desambiguação: dentre as etiquetas (tags) possíveis previstas (pelo léxico, por exemplo), determinar a mais apropriada
    - Contexto desambigua!

68

Tagging

- Útil para um infinidade de tarefas de PLN
  - *Stemming*, lematização
  - Tradução, sumarização, auxílio à escrita
  - Identificação de autoria, extração de informação
  - Pesquisas lingüísticas variadas: neologismos, comportamento de palavras, etc.
  - Etc.

69

Tagging

- 2 principais abordagens
  - Regras
    - Por exemplo, uma palavra antecedida por um artigo é um substantivo
  - Probabilidades
    - Classe mais provável de uma palavra em função das palavras vizinhas, com aprendizado a partir de córpus
- Hibridismo também é possível
  - Por exemplo, aprendizado de regras a partir de córpus

70

Tagging: regras

- Primeiras abordagens (década de 60)
  - 2 passos tradicionais
    - Léxico fornece possíveis classes para cada palavra
    - Regras criadas manualmente são utilizadas para desambiguar
- Mais recentemente
  - Dicionários maiores e muito mais regras!

71

Exemplo: EngCG tagger (Voutilainen, 1999)

→ Análise morfológica da sentença (tag correta em negrito)

Pavlov had shown that salivation.

Pavlov:	**PAVLOV N NOM SG PROPER**
had:	**HAVE V PAST VFIN SVO**
	HAVE PCP2 SVO
shown:	**SHOW PCP2 SVOO SVO SV**
that:	ADV
	PRON DEM SG
	DET CENTRAL DEM SG
	CS
salivation:	**N NOM SG**
.:	**PUNC DOT**

72

Exemplo: EngCG tagger (Voutilainen, 1999)

→ Aplicação de regras para determinar as melhores tags

Exemplo de regra

WORD: that
IF
next word is adj, adverb, or quantifier AND
after this word there is the sentence boundary AND
the previous word is not a verb that allows adjs as complements
THEN eliminate non-adv tags
ELSE eliminate adv tag

73

Exemplo: ReGra (Martins et al., 1998)

Exemplo de entrada (com erros) para o revisor gramatical

OS	Definite article (the): masculine, plural
	Personal pronoun (them): masculine, plural
MENINO	Noun (boy): masculine, singular
PREFERE	Verb (to prefer): 3rd person, singular, present tense, indicative, transitive
BRINCAR	Verb (to play): infinitive
DO	Contraction: preposition (of) + definite article (the): masculine, singular
QUE	Relative Pronoun (which)
	Adverb (what)
	Conjunction (than)
ESTUDAR	Verb (to study): infinitive

Regras de desambiguação utilizadas

OS	Definite article (the): the following word is a masculine noun
QUE	Conjunction (than): the previous word is a contraction (preposition + article)

74

Tagging: regras

- Zellig Harris (1962) e o primeiro tagger (provavelmente)
  - 14 regras de desambiguação
- UNITEX-PB (Muniz, 2004)
  - 80 regras de desambiguação no formalismo ELAG

ELAG
16 regras para Adjetivos
30 regras para Advérbios
22 regras para Artigos
12 regras para Substantivos

75

Tagging: regras

- ELAG (*Elimination of Lexical Ambiguities by Grammars*) (Laporte e Monceaux, 1998)

> As seguintes premissas são seguidas pelo formalismo ELAG: análises corretas não devem ser removidas; os resultados de análise sintática não podem ser explicitamente utilizados, uma vez que eles não estão disponíveis quando a resolução de ambigüidade lexical é aplicada ao texto; a análise lingüística que desejamos aplicar à sentença deve ser levada em consideração, o que implica que o criador das gramáticas de resolução de ambigüidade lexical tem visões particulares sobre o resultado desejado da análise sintática.

Muniz (2004)

76

Tagging: probabilidades

- Abordagem antiga, desde a década de 60
- Modelo de Markov Oculto (HMM – *Hidden Markov Model*), um dos mais utilizados
  - Um tipo de inferência bayesiana
  - Tarefa de classificação: dadas algumas observações, quais as classes mais prováveis
    - Tagging: dada uma sequência de palavras, qual a sequência de tags mais provável

77

Tagging: probabilidades

- Tagging: dada uma sequência de palavras, qual a sequência de tags mais provável
  - Queremos a sequência de tags SeqTags que maximize a probabilidade P(SeqTags|SeqPalavras)

$$\hat{SeqTags} = \underset{SeqTags}{\operatorname{argmax}}\ P(SeqTags \mid SeqPalavras)$$

  - Exemplo: *O menino prefere brincar do que estudar*
    - SeqTags 1: art subst verbo verbo contração pro verbo
    - SeqTags 2: art subst verbo verbo contração adv verbo
    - SeqTags 3: art subst verbo verbo contração conj verbo
    - SeqTags 4: pro subst verbo verbo contração pro verbo
    - SeqTags 5: pro subst verbo verbo contração adv verbo
    - SeqTags 6: pro subst verbo verbo contração conj verbo
      - Qual a melhor sequência de tags, ou seja, qual destas sequências maximiza a probabilidade P(SeqTags|SeqPalavras)?

78

Tagging: probabilidades

- Qual a maior probabilidade?

$$\hat{SeqTags} = \underset{SeqTags}{\operatorname{argmax}}\ P(SeqTags \mid SeqPalavras)$$

  - P(art subst verbo verbo contração pro verbo | O menino prefere brincar do que estudar)
  - P(art subst verbo verbo contração adv verbo | O menino prefere brincar do que estudar)
  - P(art subst verbo verbo contração conj verbo | O menino prefere brincar do que estudar)
  - P(pro subst verbo verbo contração pro verbo | O menino prefere brincar do que estudar)
  - P(pro subst verbo verbo contração adv verbo | O menino prefere brincar do que estudar)
  - P(pro subst verbo verbo contração conj verbo | O menino prefere brincar do que estudar)

79

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(SeqTags \mid SeqPalavras)$$

80

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(SeqTags \mid SeqPalavras)$$

$$P(SeqTags \mid SeqPalavras) = \frac{P(SeqPalavras \mid SeqTags) \times P(SeqTags)}{P(SeqPalavras)}$$

81

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(SeqTags \mid SeqPalavras)$$

$$P(SeqTags \mid SeqPalavras) = \frac{P(SeqPalavras \mid SeqTags) \times P(SeqTags)}{P(SeqPalavras)} \longrightarrow \text{constante}$$

$$P(SeqTags \mid SeqPalavras) = P(SeqPalavras \mid SeqTags) \times P(SeqTags)$$

82

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(SeqTags \mid SeqPalavras)$$

$$P(SeqTags \mid SeqPalavras) = \frac{P(SeqPalavras \mid SeqTags) \times P(SeqTags)}{P(SeqPalavras)} \longrightarrow \text{constante}$$

$$P(SeqTags \mid SeqPalavras) = P(SeqPalavras \mid SeqTags) \times P(SeqTags)$$

- Simplificações para facilitar o cálculo
  - Cada palavra depende apenas de sua tag
  - Uma tag depende apenas da tag anterior na sentença

83

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(SeqTags \mid SeqPalavras)$$

$$P(SeqTags \mid SeqPalavras) = \frac{P(SeqPalavras \mid SeqTags) \times P(SeqTags)}{P(SeqPalavras)} \longrightarrow \text{constante}$$

$$P(SeqTags \mid SeqPalavras) = P(SeqPalavras \mid SeqTags) \times P(SeqTags)$$

- Simplificações para facilitar o cálculo
  - Cada palavra depende apenas de sua tag
  - Uma tag depende apenas da tag anterior na sentença

$$P(SeqTags \mid SeqPalavras) = \prod_{i=1}^{\text{número de palavras}} P(palavra_i \mid tag_i) \times P(tag_i \mid tag_{i-1})$$

84

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(\text{SeqTags} \mid \text{SeqPalavras})$$

$$P(\text{SeqTags} \mid \text{SeqPalavras}) = \frac{P(\text{SeqPalavras} \mid \text{SeqTags}) \times P(\text{SeqTags})}{P(\text{SeqPalavras})} \longrightarrow \text{constante}$$

$$P(\text{SeqTags} \mid \text{SeqPalavras}) = P(\text{SeqPalavras} \mid \text{SeqTags}) \times P(\text{SeqTags})$$

- Simplificações para facilitar o cálculo
  - Cada palavra depende apenas de sua tag
  - Uma tag depende apenas da tag anterior na sentença

$$P(\text{SeqTags} \mid \text{SeqPalavras}) = \prod_{i=1}^{\text{número de palavras}} P(\text{palavra}_i \mid \text{tag}_i) \times P(\text{tag}_i \mid \text{tag}_{i-1})$$

Como calcular essas 2 probabilidades?

85

Tagging: probabilidades

- Exemplo
  - Supondo que se usa o Brown Corpus

$$P(\text{SeqTags} \mid \text{SeqPalavras}) = \prod_{i=1}^{\text{número de palavras}} P(\text{palavra}_i \mid \text{tag}_i) \times P(\text{tag}_i \mid \text{tag}_{i-1})$$

$P(\text{palavra}_i \mid \text{tag}_i)$ = P("is"|VBZ)

tag=VBZ → 21.627 ocorrências no córpus
palavra="is" com tag=VBZ → 10.073 ocorrências no córpus

P("is"|VBZ) = número de vezes de "is" com VBZ / número de VBZ
P("is"|VBZ) = 10.073/21.627 = 0.47 ou 47%

86

Tagging: probabilidades

- Exemplo
  - Supondo que se usa o Brown Corpus

$$P(SeqTags \mid SeqPalavras) = \prod_{i=1}^{\text{número de palavras}} P(palavra_i \mid tag_i) \times P(tag_i \mid tag_{i-1})$$

$P(tag_i|tag_{i-1}) = P(NN|DT)$

tag_{i-1}=DT → 116.454 ocorrências no córpus
tag_i=NN com tag_{i-1}=DT → 56.509 ocorrências no córpus

P(NN|DT) = número de vezes de NN precedido por DT/ número de DT
P(NN|DT) = 56.509/116.454 = 0.49 ou 49%

87

Tagging: probabilidades

- Exemplo
  - P(art subst verbo verbo contração pro verbo | O menino prefere brincar do que estudar)

$$P(SeqTags \mid SeqPalavras) = \prod_{i=1}^{\text{número de palavras}} P(palavra_i \mid tag_i) \times P(tag_i \mid tag_{i-1})$$

  - = P(O|art) x P(art|<início da sentença>) x
    P(menino|subst) x P(subst|art) x
    P(prefere|verbo) x P(verbo|subst) x
    P(brincar|verbo) x P(verbo|verbo) x
    P(do|contração) x P(contração|verbo) x
    P(que|pro) x P(pro|contração) x
    P(estudar|verbo) x P(verbo|pro)

- Faz-se isso para todas as possíveis sequências de tags (com probabilidades aprendidas de córpus)
  - A sequência com maior probabilidade é escolhida

88

Tagging: probabilidades

- Considerando a probabilidade abaixo
  - P(art subst verbo verbo contração pro verbo | O menino prefere brincar do que estudar) =
    P(O|art) x P(art|<início da sentença>) x
    P(menino|subst) x P(subst|art) x
    P(prefere|verbo) x P(verbo|subst) x
    P(brincar|verbo) x P(verbo|verbo) x
    P(do|contração) x P(contração|verbo) x
    P(que|pro) x P(pro|contração) x
    P(estudar|verbo) x P(verbo|pro)
- Por que ela é provavelmente menor do que a abaixo – que é a interpretação correta?
  - P(art subst verbo verbo contração conj verbo | O menino prefere brincar do que estudar) =
    P(O|art) x P(art|<início da sentença>) x
    P(menino|subst) x P(subst|art) x
    P(prefere|verbo) x P(verbo|subst) x
    P(brincar|verbo) x P(verbo|verbo) x
    P(do|contração) x P(contração|verbo) x
    P(que|conj) x P(conj|contração) x
    P(estudar|verbo) x P(verbo|conj)

89

Tagging: probabilidades

- Considerando a probabilidade abaixo
  - P(art subst verbo verbo contração pro verbo | O menino prefere brincar do que estudar) =
    P(O|art) x P(art|<início da sentença>) x
    P(menino|subst) x P(subst|art) x
    P(prefere|verbo) x P(verbo|subst) x
    P(brincar|verbo) x P(verbo|verbo) x
    P(do|contração) x P(contração|verbo) x
    P(que|pro) x P(pro|contração) x
    P(estudar|verbo) x P(verbo|pro)
- Por que ela é provavelmente menor do que a abaixo – que é a interpretação correta?
  - P(art subst verbo verbo contração conj verbo | O menino prefere brincar do que estudar) =
    P(O|art) x P(art|<início da sentença>) x
    P(menino|subst) x P(subst|art) x
    P(prefere|verbo) x P(verbo|subst) x
    P(brincar|verbo) x P(verbo|verbo) x
    P(do|contração) x P(contração|verbo) x
    P(que|conj) x P(conj|contração) x
    P(estudar|verbo) x P(verbo|conj)

Esses termos devem ser mais prováveis do que os correspondentes na intepretação errada

90

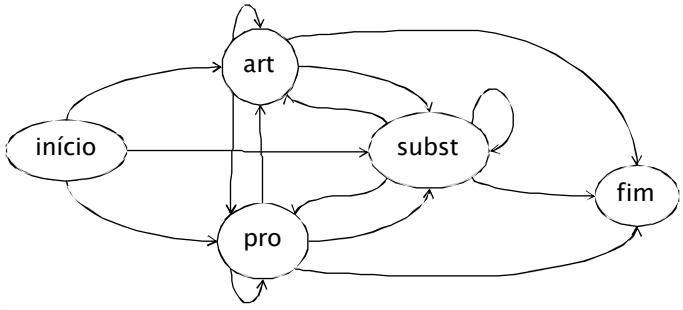
Tagging: probabilidades

- Modelo de Markov oculto
  - Modelam-se eventos observados (palavras) e eventos não observados, ou seja, ocultos (tags)
  - Como dito anteriormente, tipo especial de autômato
    - Probabilidades nos arcos (transições): $P(tag_i|tag_{i-1})$
    - Probabilidades nos nós: $P(palavra_i|tag_i)$

91

Tagging: probabilidades

- Modelo de Markov oculto
  - Exemplo hipotético

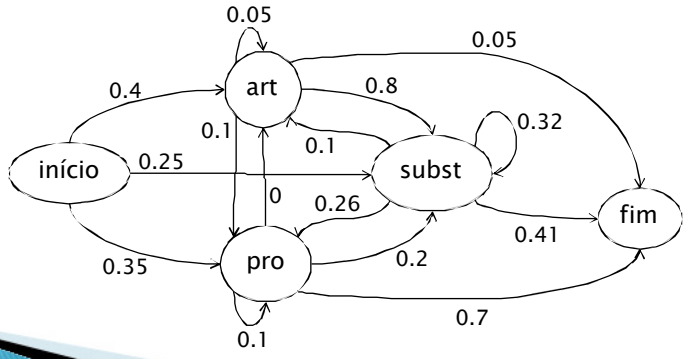


92

Tagging: probabilidades

- Modelo de Markov oculto
  - Exemplo hipotético

A soma das probabilidades dos arcos que saem de cada nó devem somar 1



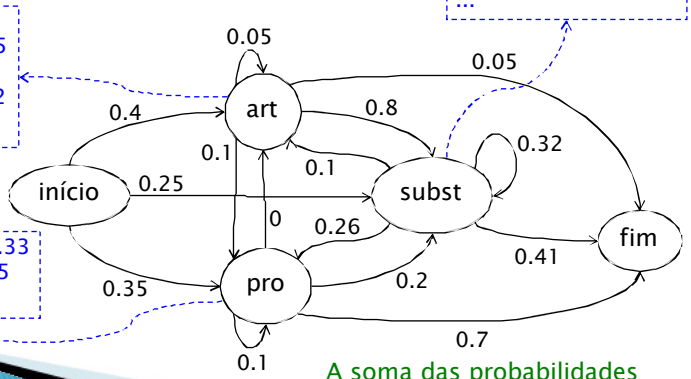
93

Tagging: probabilidades

- Modelo de Markov oculto
  - Exemplo hipotético

P(menino|subst)=0.1
P(casa|subst)=0.04
...

P(o|art)=0.2
P(os|art)=0.25
P(a|art)=0.2
P(as|art)=0.12
...



P(que|pro)=0.33
P(ele|pro)=0.5
...

A soma das probabilidades associadas a cada nó devem somar 1

Tagging: probabilidades

- Taggers atuais usam normalmente 2 tags anteriores como contexto
  - $P(tag_i|tag_{i-1},tag_{i-2})$
- É preciso otimizar a busca por sequências de tags mais prováveis, senão pode haver explosão combinatória
  - **Programação dinâmica** (estudaremos logo)
- É preciso lidar com probabilidades muito baixas ou próximas de zero
  - Por exemplo, situações em que $P(tag_i|tag_{i-1}, tag_{i-2})\approx 0$
  - Solução usual: combinar probabilidades ponderadas
    - $P(tag_i|tag_{i-1}, tag_{i-2}) = p_1*P(tag_i|tag_{i-1},tag_{i-2})+p_2*P(tag_i|tag_{i-1})+p_3*P(tag_i)$, com $p_1+p_2+p_3=1$

95

Tagging: Transformation-Based

- *Transformation-Based tagging*
  - Aplicação de TBL (*Transformation-Based Learning*), da linha de aprendizado de máquina
  - Regras são aprendidas e aprimoradas automaticamente
    - Várias iterações
    - A cada iteração, o processo melhora
    - Ao estabilizar, fim do processo de aprendizado

96

Tagging: Transformation-Based

- *Transformation-Based tagging*
  - Processo básico com base em um córpus anotado manualmente
    - Inicialmente, anota-se um córpus automaticamente, assumindo-se que a tag de uma palavra é a sua tag mais frequente (segundo um córpus/léxico)
    - Verificam-se os erros cometidos (comparando-se com a anotação humana correspondente) e, dentre todas as possibilidades de correção, monta-se uma regra de correção com maior precisão
    - Aplica-se essa regra nova em todo o córpus
    - Verificam-se novamente os erros cometidos e monta-se uma segunda regra de correção com maior precisão
    - E assim por diante, até não se obter mais melhora de performance

97

Tagging: Transformation-Based

- Exemplo

1. Inicialmente, anotação com base em frequência

   ... is/VBZ expected/VBN to/TO race/NN tomorrow/NN
   ... the/DT race/NN for/IN outer/JJ space/NN
   Book/VB the/DT flight/VB to/TO...

2. Verificando-se os erros, aprende-se uma nova regra

   Troque NN para VB quando a tag anterior é TO

3. Corrige-se a etiquetação anterior

   ... is/VBZ expected/VBN to/TO race/VB tomorrow/NN
   ... the/DT race/NN for/IN outer/JJ space/NN
   Book/VB the/DT flight/VB to/TO...

98

Tagging: Transformation-Based

- Exemplo

4. Aprende-se uma nova regra com base nos erros existentes

   Troque VB para NN quando a tag anterior é DT e a posterior é TO

5. Corrige-se a etiquetação anterior

   ... is/VBZ expected/VBN to/TO race/VB tomorrow/NN
   ... the/DT race/NN for/IN outer/JJ space/NN
   Book/VB the/DT flight/NN to/TO...

6. E assim por diante

99

Tagging: Transformation-Based

- Exemplo
  - **Resultado** do processo

Regra 1: etiquete as palavras com suas tags mais frequentes
Regra 2: troque NN para VB quando a tag anterior é TO
Regra 3: troque VB para NN quando a tag anterior é DT e a posterior é TO
...

100

Tagging: Transformation-Based

- *Transformation-Based tagging*
  - Ao final do processo, há um conjunto de regras ordenadas que devem ser aplicadas sequencialmente para etiquetar um novo texto
  - Como há muitas regras possíveis de serem aprendidas, costuma-se limitar as estruturas aceitas de regras
    - Troque a tag da palavra corrente de A para B se a palavra anterior tem a tag X
    - Troque a tag da palavra corrente de A para B se a palavra anterior tem a tag X e a posterior tem a tag Y
    - Etc.

Caso contrário, o que acontece?

101

Humanos e máquinas

- Palavras/morfologia e léxico mental
  - Nem a listagem exaustiva, nem todas as regras de flexão/derivação
  - Há indícios de que humanos armazenam em seu léxico mental os lemas das palavras e também algumas formas plenas
    - Stanners et al. (1979): são mantidas separadamente as palavras *happy* e *happiness*, mas somente o verbo *pour*, sem suas flexões
- Morfossintaxe
  - Experimentos mostram que humanos discordam em 3-4% das tags
    - Melhores taggers com 97%, normalmente com problemas justamente onde os humanos discordam
  - Voutilainen (1995) mostra que humanos atingem 100% se se permite que discutam as tags com problemas

102