Introdução ao Processamento de Línguas Naturais

SCC5908 Introdução ao Processamento de Língua Natural

Thiago A. S. Pardo

1

Recapitulando...

- Abordagens superficiais vs. profundas
- Simbolismo vs. estatística
- Racionalismo (gerativismo) vs. empirismo (estruturalismo, distribucionalismo)
  - Dominância atual do empirismo, trabalhos com base em córpus e em evidência linguística
  - Análises e modelos estatísticos, frequências de fenômenos textuais

2

Abordagens: PLN

- Exemplo: livros de Tom Sawyer (de Mark Twain)

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Tokens = 71.370
Types = 8.018 (poucas para um texto tão grande)
→ para crianças

Taxa type/token = 0,11 (11%)

Em geral, quanto maior o córpus, menor a taxa

3

Abordagens: PLN

- Distribuição de palavras
  - Lei de **Zipf**
    - *George Kingsley Zipf*
    - Baseada em trabalho de Estoup (1916)
    - Proveniente do "Princípio do Mínimo Esforço", publicado no livro *Human Behavior and the Principle of Least Effort* (1949)

4

Abordagens: PLN

- Distribuição de palavras
  - Lei de Zipf
    - Contam-se quantas vezes cada palavra ocorre em um córpus grande, montando-se um ranque em função da frequência delas
      - Há uma relação entre a frequência e a posição da palavra no ranque
        - **Frequência x posição no ranque = constante k**
        - Palavra na posição 50 deve ocorrer 3 vezes mais do que palavra na posição 150

5

Abordagens: PLN

- Exemplo: livros de Tom Sawyer
  - Há distorções, comuns na lei de Zipf

Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	410	20	8400
be	294	30	8820
there	222	40	8880
one	172	50	8600
about	158	60	9480
more	138	70	9660
never	124	80	9920
Oh	116	90	10440
two	104	100	10400

Word	Freq. (f)	Rank (r)	$f \cdot r$
turned	51	200	10200
you'll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000
brushed	4	2000	8000
sins	2	3000	6000
Could	2	4000	8000
Applausive	1	8000	8000

6

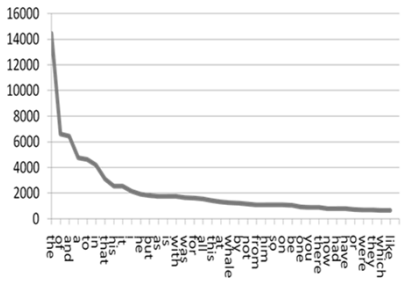
Abordagens: PLN

- Distribuição de palavras
  - Lei de Zipf
    - Poucas palavras muito frequentes
    - Número significativo de palavras de frequência média
    - Muitas palavras de frequência baixa
      - É possível plotar um gráfico

7

Abordagens: PLN

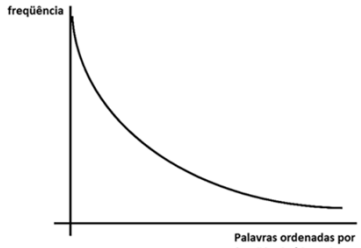
- Exemplo: parte inicial da curva de Zipf para Moby Dick



8

Abordagens: PLN

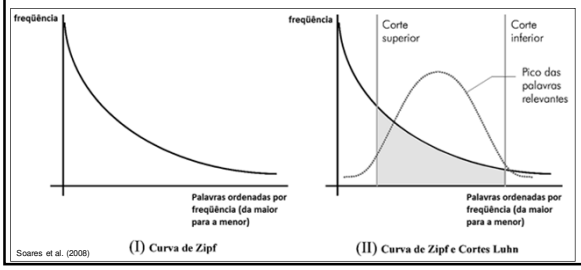
- Curva de Zipf



9

Abordagens: PLN

- Distribuição de palavras
  - Curva de Zipf e corte de Luhn (1958)
    - Busca por termos importantes



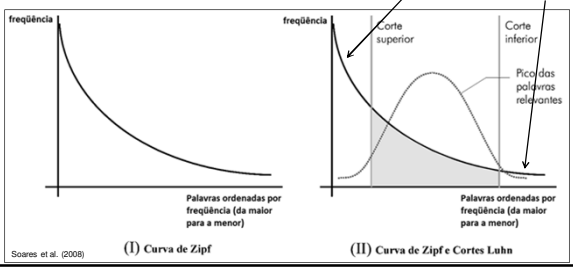
Soares et al. (2008)

Abordagens: PLN

- Distribuição de palavras
  - Curva de Zipf e corte de Luhn (1958)
    - Busca por termos importantes

preposições, conjunções, etc.

termos raros



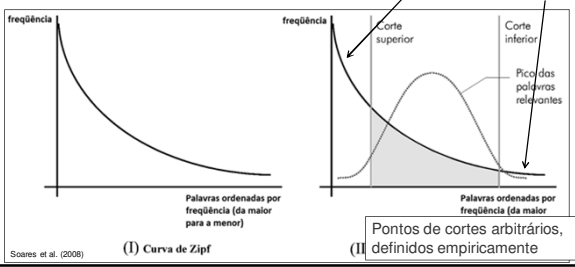
Soares et al. (2008)

Abordagens: PLN

- Distribuição de palavras
  - Curva de Zipf e corte de Luhn (1958)
    - Busca por termos importantes

preposições, conjunções, etc.

termos raros



Pontos de cortes arbitrários, definidos empiricamente

Soares et al. (2008)

Abordagens: PLN

- Distribuição de palavras
  - Outra lei de Zipf
    - O número de significados de uma palavra é correlacionado com sua frequência
      - Palavra com 10.000 ocorrências → 2.1 significados
      - Palavra com 5.000 ocorrências → 3 significados
      - Palavra com 2.000 ocorrências → 4.6 significados

13

Abordagens: PLN

- Distribuição de palavras
  - Ainda outras leis de Zipf
    - Uma palavra de conteúdo tende a ocorrer próxima a outra ocorrência sua
    - A frequência de uma palavra é inversamente proporcional ao seu tamanho
    - Quanto maior a frequência de uma palavra, mais "permutações" há (em seus componentes morfológicos)

14

Abordagens: PLN

- Leis de Zipf
  - Exageradamente valorizadas
    - Não deveriam ser "leis", mas "observações" aproximadas
  - Até alguns eventos aleatórios obedecem essas leis
    - Forma de gerar os dados, de construir a curva

15

PLN

- Classificação
  - Recursos
  - Ferramentas
  - Aplicações

Recursos

- Córpus
  - Anotação: humana e/ou automática
    - XML, XCES, TEI, etc.
  - Paralelo, comparável, alinhado, etc.
- Dicionários monolíngues e bilíngues
  - *Machine readable* vs. *machine tractable*
- Léxicos
  - Vários paradigmas

17

Ferramentas

- Segmentadores textuais: palavras (*tokenizador*), sentenças, parágrafos, tópicos
- Stemmers, lematizadores, nominalizadores
- Etiquetadores morfossintáticos (*taggers*)
- Analisadores sintáticos *shallow* (*chunkers*) e *deep* (*parsers*)
- Analisadores semânticos e discursivos
- Alinhadores textuais: lexicais, sentenciais, etc.
- Concordanceadores, *word counting*, etc.

18

Aplicações

- Tradutores automáticos
- Revisores ortográficos e gramaticais
- Ferramentas de auxílio à escrita
- Sumarizadores automáticos
- Simplificadores textuais

19

Recursos, ferramentas e aplicações

- Atenção
  - Classificação difusa, às vezes
  - Dependente do uso
    - Sumarizador como passo intermediário para recuperação da informação → ferramenta
    - Dicionário eletrônico para consulta → aplicação

20

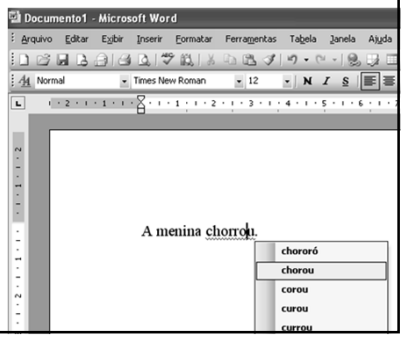
PLN e áreas correlatas

- Limites entre PLN e outras áreas: como percebem isso?
  - Recuperação de informação
  - Extração de informação
  - Inteligência artificial
  - Banco de dados
  - Interação humano-computador
  - Tradução automática
  - Tradução
  - Mineração de textos
  - Linguística de córpus

21

Exemplos

- Revisão ortográfica
  - Tokenizador
  - Léxico
  - Regras para ordenar sugestões

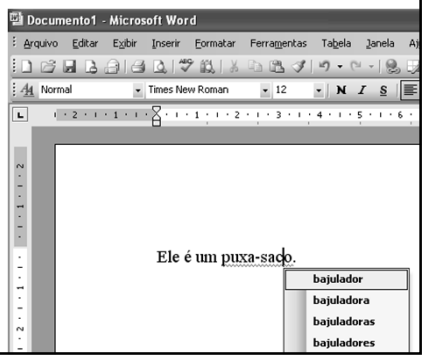


Exemplos

- Revisão gramatical
  - Tokenizador
  - Segmentador sentencial
  - Etiquetador morfossintático
  - Analisador sintático
  - Léxico
  - Regras gramaticais

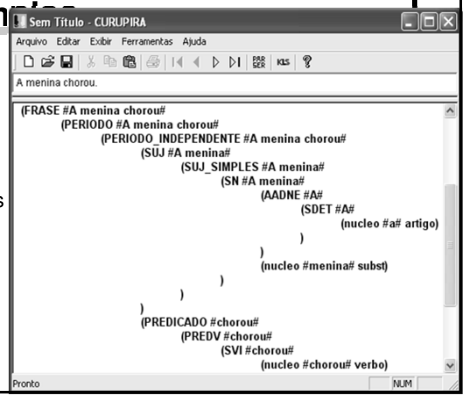
Exemplos

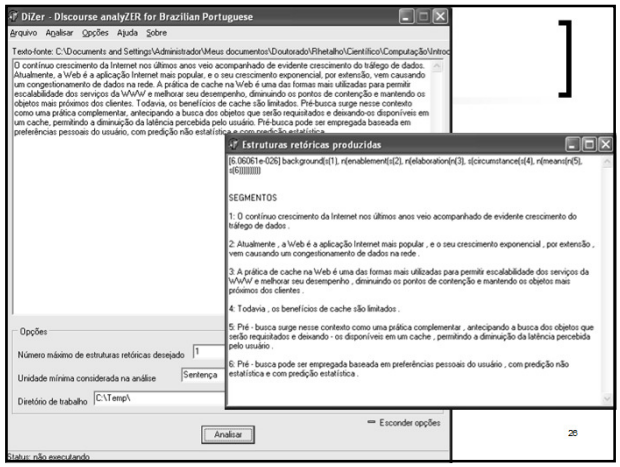
- Revisão estilística
  - Tokenizador
  - Regras estilísticas
  - ...



Exemplos

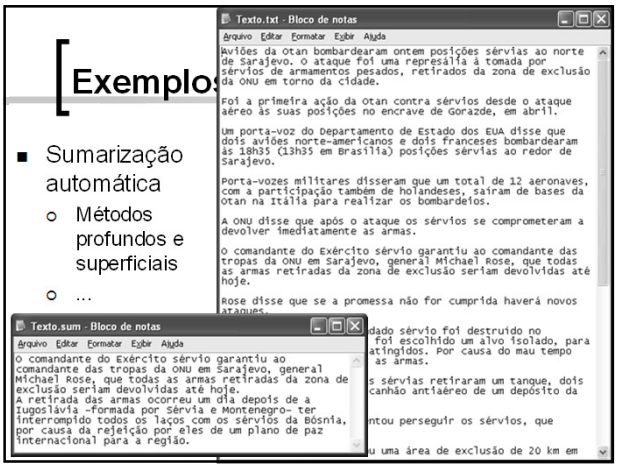
- Análise sintática
  - Léxico
  - Regras sintáticas
  - ...





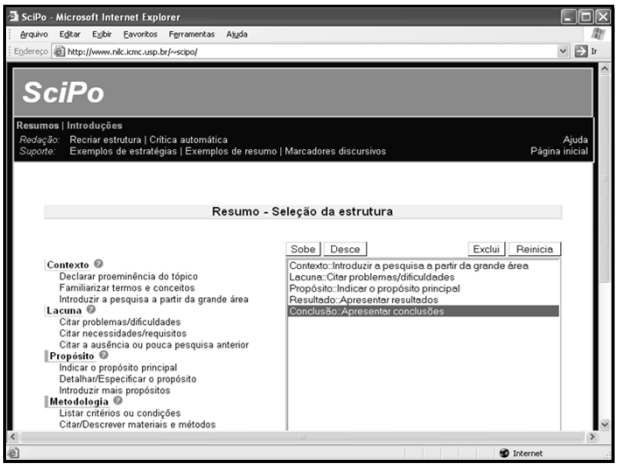
Exemplos

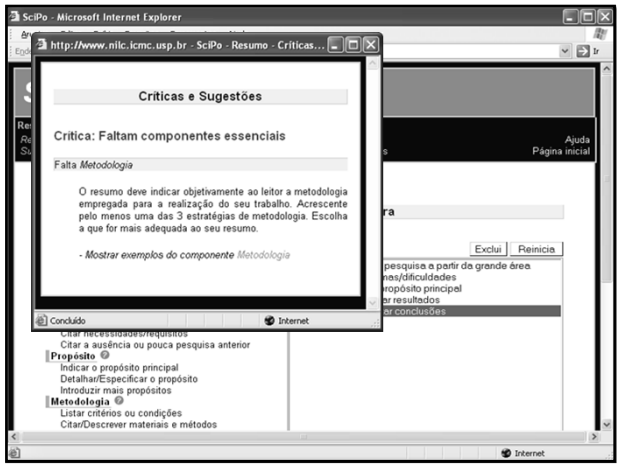
- Sumarização automática
  - Métodos profundos e superficiais
  - ...



Exemplos

- Auxílio à escrita de textos científicos
  - Regras de estruturação textual
  - Exemplos da estruturas de outros textos
  - Crítica de cada parte do texto

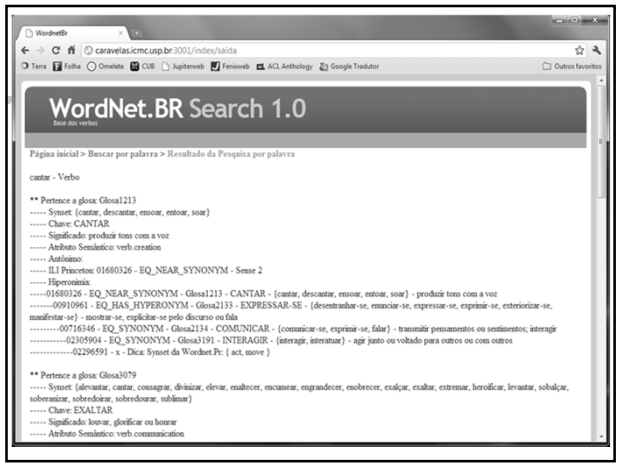




Exemplos

- WordNet
  - Base de dados lexicais e conceituais
  - Relações entre palavras
    - Sinonímia
    - Antonímia
    - Acarretamento
    - Etc.
  - Relações ontológicas

32

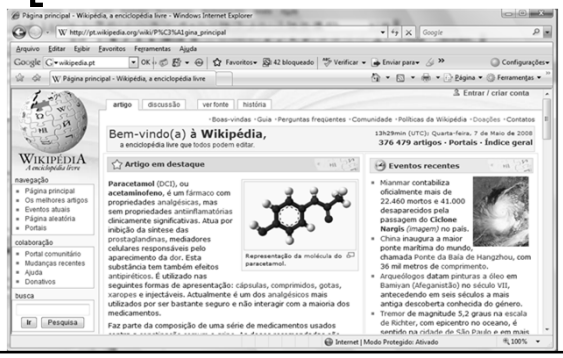


PLN

- Conhecimento linguístico é a base para muitos sistemas que manipulam língua natural
  - Extração de conhecimento de córpus
    - Regras gramaticais, sintáticas e discursivas
    - Estrutura textual
    - Regras de tradução
    - Critérios para resumir

35

Conhecimento de mundo



Senso comum

38

PLN no Brasil

- Poucos grupos de pesquisa no país
  - São Carlos
  - Porto Alegre
  - Rio de Janeiro
  - Outros?

39

Recentemente

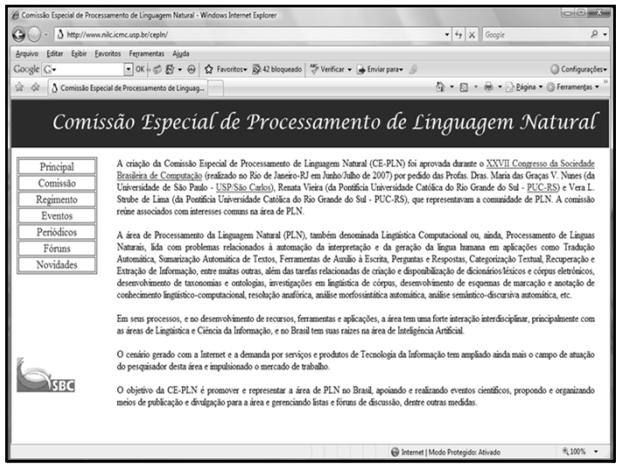
- A área de PLN tem crescido no Brasil
  - Tecnologia da Informação
  - Google
  - Comissão especial da SBC
  - Eventos científicos próprios melhores e maiores a cada ano
    - Além dos eventos típicos de IA
  - Iniciativas internacionais importantes

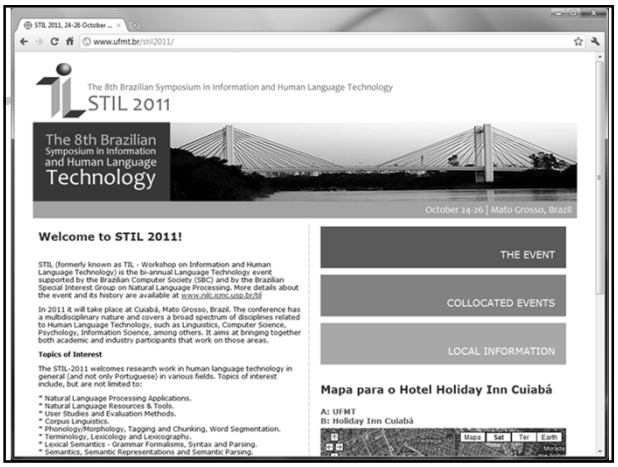
40

Comissão Especial de PLN

- Responsável pela condução da área e representação nacional
- www.sbc.org.br/ce-pln
  - Aproximadamente 200 membros na lista de discussão
  - Não precisa ser membro da SBC

41





Outras iniciativas

- ACL (aclweb.org)
  - *ACL anthology*, listas de discussão, wiki
  - *Registry of Latin American Researchers in Natural Language Processing and Computational Linguistics*
- Linguateca (www.linguateca.pt)
  - Oficialmente finalizado
- forum-lp
- Eventos correlatos
  - Escola Brasileira de Linguística Computacional
  - Encontro de Linguística de Córpus
  - Workshop de Descrição do Português
    - Junto ao STIL
- Toolkits
  - GATE, NLTK, Giza++ e Moses, AntMover, etc.

46