Introdução ao Processamento de Línguas Naturais

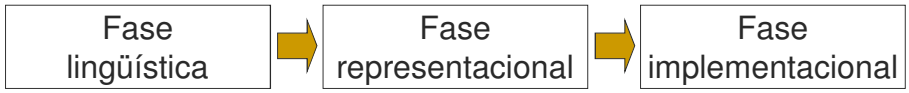
SCC5869 Tópicos em Processamento de Língua Natural

Thiago A. S. Pardo

1

PLN

- Trabalho em PLN



2

PLN

- Trabalho em PLN

Fase lingüística	Fase representacional	Fase implementacional
Sintaxe de sentenças da língua portuguesa	Regras sintáticas e formalismos sintáticos	Desenvolvimento do analisador sintático
Resumos de artigos de jornais	Formalização das regras para resumir	Desenvolvimento do sumarizador automático
Tradução espanhol-português	Regras de tradução, dicionários bilíngues	Desenvolvimento do tradutor automático

3

PLN

- Trabalho em PLN

Fase lingüística	Fase representacional	Fase implementacional
Teorias e dados lingüísticos	Formalismos/esquemas de representação	Métodos e modelos computacionais

- Aspectos da língua que são possíveis capturar e automatizar
  - Maioria das teorias lingüísticas são sofisticadas demais para o PLN... alguns recursos também (exemplo?)

4

PLN

- Interação delicada entre informatas e lingüistas
  - Como na maioria das áreas interdisciplinares
    - Informata: sujeito, predicado, relações semânticas e lógico-conceptuais, vozes do texto, Chomsky?
    - Lingüista: scripts, GUI, usabilidade, autômato, Turing?

5

PLN

- Interação delicada entre informatas e lingüistas
  - Preconceitos, “pérolas” e “sabedorias milenares”
    - Informata só quer implementar... não tem fundamentação e não entendem com o que estão lidando
    - Informata não fala direito
    - PLN é bobagem: a língua nunca será automatizável
    - Para que testar “numericamente” o que já é consenso entre lingüistas?
    - Lingüistas não conseguem formalizar, não percebem que muitos detalhes não interessam

6

PLN

- Interação delicada entre informatas e lingüistas
  - Preconceitos, “pérolas” e “sabedorias milenares”
    - Lingüistas sempre conhecem exceções
    - O bonde está andando... com ou sem lingüistas, vai ser feito
    - O lingüista sempre terá o seu lugar
    - *Every time I fire a linguist, the system performance goes up* (Fred Jelinek, 1980s)

7

PLN & IA

- Classificações... nem sempre triviais

Critérios	**Paradigmas**
Uso de conhecimento lingüístico	Superficial, profundo e híbrido
Representação do conhecimento	Simbólico, não-simbólico e híbrido
Obtenção do conhecimento	Manual, automática e híbrida

8

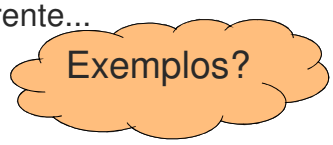
Tendências do PLN

- No início, métodos superficiais e simbólicos
- Métodos profundos e simbólicos
- Atualmente, métodos estatísticos
- Em direção ao hibridismo
  - E o ciclo recomeça, mas diferente...

9

Tendências do PLN

- No início, métodos superficiais e simbólicos
- Métodos profundos e simbólicos
- Atualmente, métodos estatísticos
- Em direção ao hibridismo
  - E o ciclo recomeça, mas diferente...



Superficial vs. profundo

- Superficial
  - Mais fácil aplicação e desenvolvimento, mais robusto
  - Resultados piores
- Profundo
  - De mais difícil modelagem e aquisição
  - Resultados melhores
- Híbrido: como fazer?
- Métodos profundos "explicam" a língua, mas alguns métodos superficiais são muito bons
  - Por exemplo, sumarização de notícias jornalísticas
- "Métodos cada vez mais sofisticados para fazer a mesma coisa"
  - Dilema da sumarização automática

11

Simbolismo vs. estatística

- Regras são muito "rígidas" para a fluidez e flexibilidade da língua
  - Por exemplo, regras gramaticais para boa formação de sentenças
- Padrões mais freqüentes de organização da língua podem ser aprendidos (estatisticamente)
- Mas alguns tipos de regras são muito bons
  - Regras de formação de sintagmas nominais

12

Abordagens conflitantes

- Simbolismo/profundidade e a **validação de teorias**
  - Explicitação do conhecimento
- Grande **utilidade** da estatística
  - O conhecimento está lá... "codificado" (controverso)
    - Dilemas da TA estatística
      - Funciona melhor que outras abordagens, codifica conhecimento, conhecimento pode estar errado (quem se importa?)

Racionalismo

- 1960-1985: racionalismo entre lingüistas, informatas, etc.
  - Racionalismo: crença de que parte significativa do conhecimento humano não vem dos sentidos, mas é herdada geneticamente
- Noam Chomsky
  - Linguagem inata
    - Argumento: muito pouco estímulo para um aprendizado muito eficiente de algo complexo
      - Como é possível aprender tanto a partir de tão pouco evidência lingüística?
- IA: sistemas com muito conhecimento manualmente fornecido e com mecanismos de inferência

Empirismo

- 1920-1960: empirismo
  - Mente não vem com princípios e procedimentos pré-determinados
  - Mas vem com operações gerais de associação, reconhecimento de padrões e generalizações
    - Importância do estímulo sensorial para o aprendizado da língua
- Ressurgimento na atualidade
  - Aprendizado da estrutura da linguagem com modelos de língua parametrizáveis

Empirismo

- Não temos como observar uma grande quantidade de uso da língua em seu contexto no mundo
- Alternativa: textos
  - *Corpus* e *corpora*
    - Ou córpus, simplesmente
- Estruturalismo americano, representado por Zellig Harris
  - Distribucionalismo
- Firth (1957): *You shall know a word by the company it keeps*
- Como é possível aprender tão pouco a partir de tanta evidência lingüística?
  - Questão importante para a área de Aprendizado de Máquina

Racionalismo vs. empirismo

- Racionalismo
  - Lingüística a la Chomsky (*gerativismo*)
    - Descrição da módulo lingüístico da mente humana, sendo córpus somente evidência indireta, suplementado pela intuição humana
      - "Regras" e "princípios" que regem/geram a linguagem
- Empirismo
  - Descrição da língua em uso, representada em córpus

Racionalismo vs. empirismo

- Distinção importante de Chomsky (1965)
  - Competência lingüística: conhecimento da língua pelo falante
    - Foco do racionalismo/gerativismo
      - Argumentam que é possível isolar esse componente para estudo e formalização
  - Desempenho lingüístico: afetado por vários fatores, como memória disponível, distrações do ambiente, etc.
    - Foco do empirismo

Racionalismo vs. empirismo

- Lingüística a la Chomsky
  - Princípios categóricos
    - Sentenças satisfazem ou não
- Empirismo/estatística
  - Usual e "não usual"
    - Preferências, padrões mais comuns, convenções

Gerativismo

- Argumentos **contra "binariedade"** das sentenças (van Riemsdijk e Williams, 1986)
  - *John I believe Sally said Bill believed Sue saw.*
  - *John wants very much for himself to win.*
  - *Those are the books you should read before it becomes difficult to talk about.*
  - *Who did Jo think said John saw him?*
  - *That a serious discussion could arise here of this topic was quite unexpected.*
- Difícil dizer que são gramaticais, mas são! Elas não são "usuais"
- Gerativistas dizem que é "problema de desempenho"
  - Desacreditados, muitas vezes, pois outros fenômenos além da gramaticalidade também não são categóricos
    - Exemplos?

Gerativismo

- Exemplos no inglês
  - Mudanças históricas de significado e classe gramatical das palavras
    - Evidências de mudanças *graduais*
      - *While*
        - Antigamente, somente "tempo": *to take a while*
        - Atualmente, principalmente usada como introdução a orações subordinadas: *while you were out...*

21

Gerativismo

- Exemplos no inglês
  - Mudanças recentes de significado e classe gramatical das palavras
    - Evidências de mudanças *graduais*
      - *Kind of* e *sort of*
        - Nome + preposição (no sentido de "tipo"): *What sort of animal made these tracks?*
        - Modificadores (no sentido de *somewhat* ou *slightly*): *We are kind of hungry. He sort of understood what was going on.*

22

Gerativismo

- Exemplos no inglês
  - *Near*: adjetivo ou preposição?
    - Adjetivo: *We will review that decision in the near future.*
      - Evidências: entre determinante e nome, pode formar um advérbio pela adição de *-ly*
    - Preposição: *He lives near the station.*
      - Evidências: componente principal da frase locativa que complementa o verbo *live* (papel clássico de preposições), pode ser modificado por *right*
    - Adjetivo e preposição: *We live nearer the water than you thought.*
      - Evidências: forma comparativa (*-er*) é marca registrada de adjetivos, age como preposição ao ser o componente principal da frase locativa

23

Abordagens: Lingüística

- Estruturalismo e gerativismo saem um pouco de foco
- Atualmente, na Lingüística, tendência/paradigma "pragmático"
  - Falante é o sujeito da ação, funcionalidade de língua
  - Gramática de uso
  - Língua como código de comunicação e de **interação**
- E no Brasil? Problema para PLN?

Abordagens: PLN

- Tendência: empirismo
  - Córpus para estudo e formalização de fenômenos, verificação e validação de hipóteses, evidências lingüísticas
  - Freqüência e leis de **distribuição de palavras/n-gramas**
    - Eduard Hovy (ACL 2010): "Não contrato chomskyanos!"
- Exceções
  - Modelos simplistas vs. sofisticados
    - Modelos simplistas → má impressão original da área
- Atenção aos "erros"

25

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 = 11%

Em geral, quanto maior o córpus, menor a taxa

26

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)

27

Abordagens: PLN

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

28

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

29

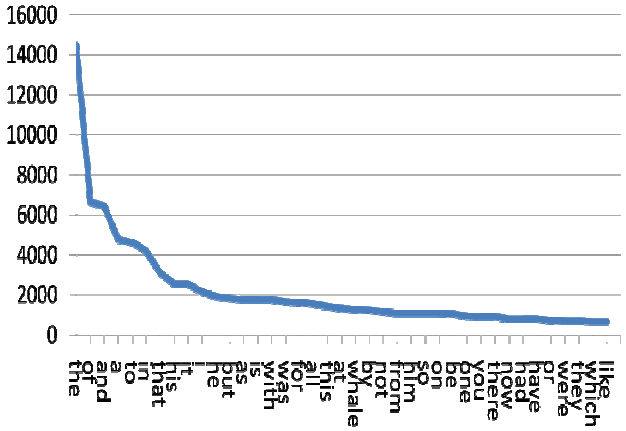
Abordagens: PLN

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

30

Abordagens: PLN

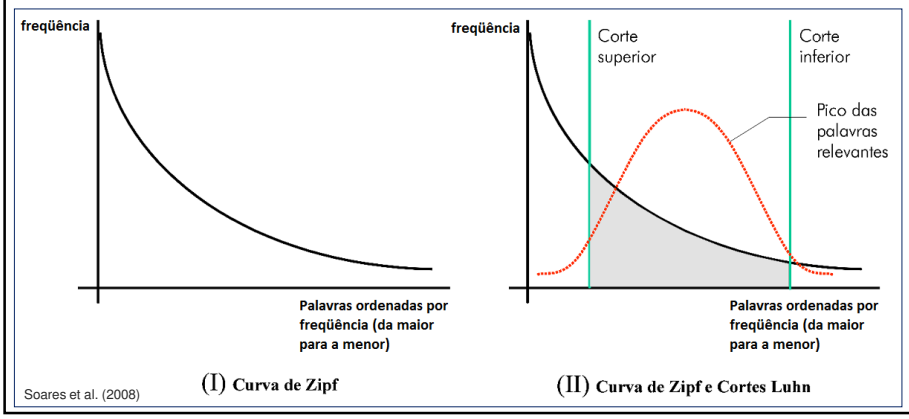
- Exemplo: parte inicial da curva de Zipf para Moby Dick



31

Abordagens: PLN

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



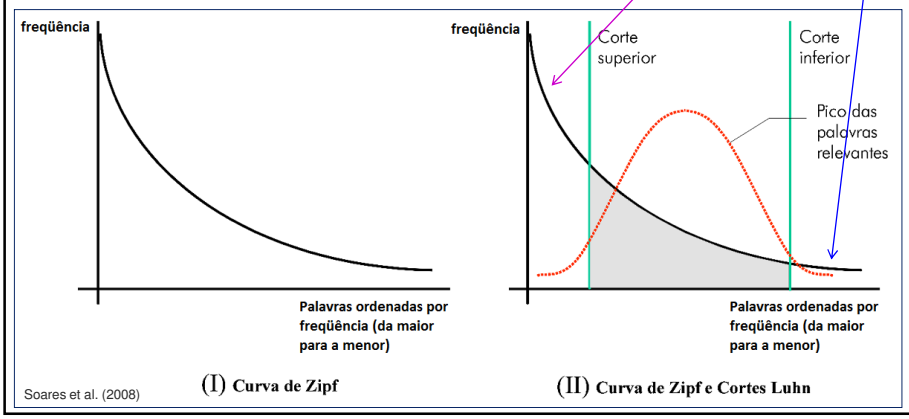
Soares et al. (2008)

Abordagens: PLN

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

preposições, conjunções, etc.

termos raros



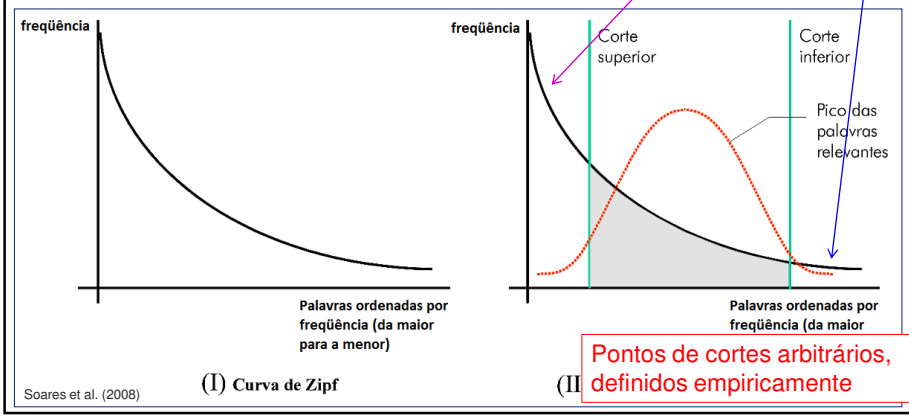
Soares et al. (2008)

Abordagens: PLN

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

preposições, conjunções, etc.

termos raros



Soares et al. (2008)

Pontos de cortes arbitrários, definidos empiricamente

Abordagens: PLN

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

35

Abordagens: PLN

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

36

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

37

Abordagens: PLN

- *The key to automatically processing human languages lies in the appropriate combination of symbolic [rationalist] and non-symbolic [empiricist] techniques*

  (Robert Dale, 2000)

38

PLN

- Classificação
  - Recursos
  - Ferramentas
  - Aplicações

Recursos

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

40

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.

41

Aplicações

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

42

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

43