Introdução ao Processamento de Línguas Naturais

SCC5908 Introdução ao Processamento de Língua Natural

Thiago A. S. Pardo

1

Recapitulando: níveis de PLN

- Morfologia: construção de palavras
  - Morfemas, raiz, afixos, etc.
- Morfossintaxe: classes gramaticais
- Sintaxe: função e estruturação das partes das sentenças
  - Sintagmas, sujeito, predicado, etc.

2

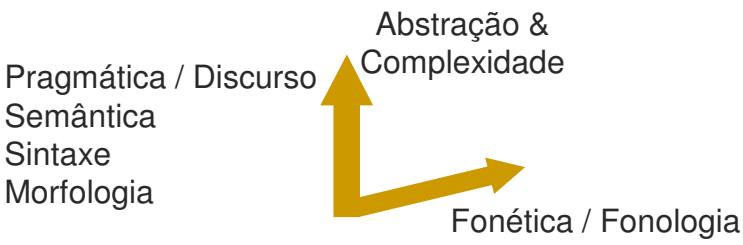
Recapitulando: níveis de PLN

- Semântica: significado de palavras, expressões, orações, sentenças, textos
  - Conceitos, traços semânticos, papéis semânticos/temáticos, classes semânticas de entidades, entidades nomeadas e mencionadas, relações lexicais, ontologias, relações semânticas entre partes do texto
- Discurso: relacionamento entre partes do texto, está além da sentença
  - Correferência (anáfora), estruturação textual, intenções
- Pragmática: língua em uso, contexto
  - Estilo, fatores pragmáticos (força, hierarquia, educação, familiaridade, etc.), modelagem de usuário (produtor e receptor)

3

PLN

- Vários níveis de conhecimento
  - Tradicionalmente distinguidos em PLN, apesar dos limites entre eles serem nebulosos na maioria dos casos



4

PLN

- Considerações para uso por um computador
  - Os níveis de conhecimento precisam ser representados (formalizados) e manipulados automaticamente
  - Interação entre os níveis
    - Morfologia e sintaxe
    - Sintaxe e semântica
    - Semântica e discurso

5

PLN

- Considerações para uso por um computador
  - Os níveis de conhecimento precisam ser representados (formalizados) e manipulados automaticamente
  - Interação entre níveis mais distantes
    - Morfologia e semântica (goleiro e porteiro vs. padeiro)
    - Morfologia e pragmática (são carlense vs. são carlino, laranjada e limonada vs. cajuada)
    - Sintaxe e discurso (subordinadas)

6

PLN e humanos

- Processamento sequencial vs. paralelo
  - Arquiteturas em *pipeline* vs. integradas

7

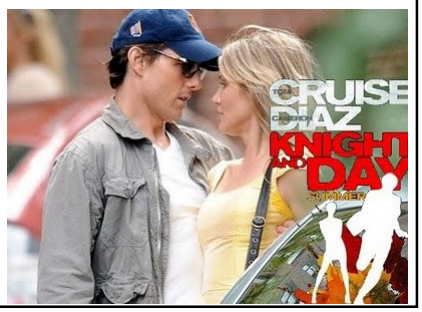
PLN e humanos

- Humanos lidam naturalmente com
  - Ambiguidade
  - Irregularidade
  - Vagueza
  - Dinamicidade
  - Variabilidade
- ... máquinas não!

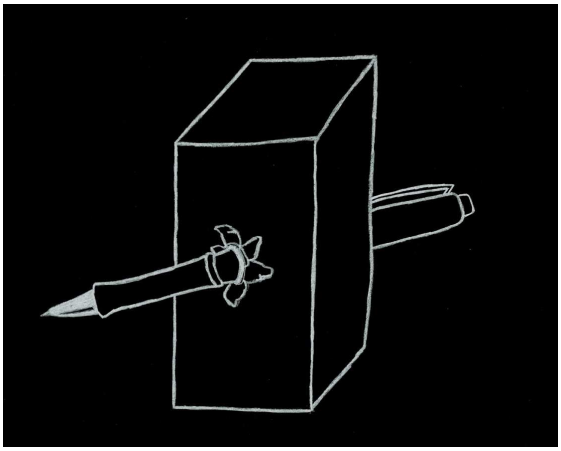
8

Exemplos de dificuldades

- O homem viu a mulher na montanha de binóculos
- Filme *Knight and Day*
- Você sabe as horas?
- O coelho foi servido
- O homem foi servido
- A caneta está na caixa
- A caixa está na caneta



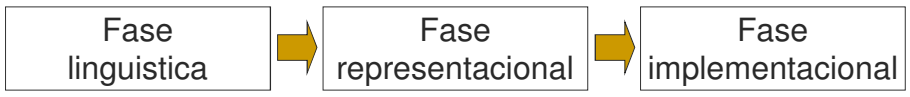
Exemplos de dificuldades



10

PLN

- Trabalho em PLN



11

PLN

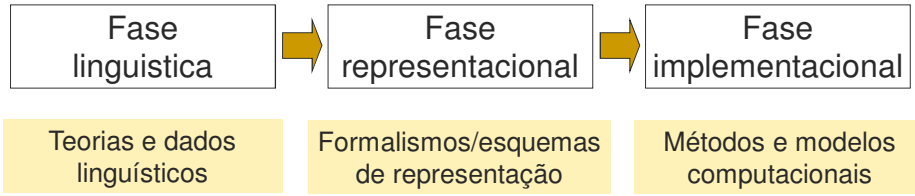
- Trabalho em PLN

Fase linguistica	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

12

PLN

- Trabalho em PLN



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

13

PLN

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

14

PLN

- Interação delicada entre informatas e linguistas
  - Preconceitos, “pérolas” e “sabedorias milenares”
    - Informata só quer implementar... não tem fundamentação e não entende com o que está lidando
    - Informata não fala direito e não sabe escrever
    - PLN é bobagem: a língua nunca será automatizada
    - Para que testar “numericamente” o que já é consenso entre linguistas?
    - O linguista sempre terá o seu lugar

15

PLN

- Interação delicada entre informatas e linguistas
  - Preconceitos, “pérolas” e “sabedorias milenares”
    - Linguistas não conseguem formalizar, não percebem que muitos detalhes não interessam
    - Linguistas sempre conhecem exceções
    - O bonde está andando... com ou sem linguistas, vai ser feito
    - *Every time I fire a linguist, the system performance goes up* (Fred Jelinek, 1980s)

16

PLN & IA

- Classificações... nem sempre triviais

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

17

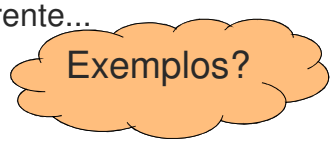
Tendências do PLN

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

18

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

20

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 frequentes de organização da língua podem ser aprendidos (estatisticamente)
- Mas alguns tipos de regras são muito bons
  - Regras de formação de sintagmas nominais

21

Abordagens conflitantes

- Simbolismo/profundidade e a **validação de teorias e modelos**
  - 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?)

Abordagens conflitantes

- 1994: ACL workshop "*The Balancing Act - Combining Symbolic and Statistical Approaches to Language*"
  - *A renaissance of interest in corpus-based statistical methods has rekindled old controversies – rationalist vs. empiricist philosophies, theory-driven vs. data-driven methodologies, symbolic vs. statistical techniques. The aim of this workshop was to set aside a priori biases and explore the balancing act that must take place when symbolic and statistical approaches are brought together*
  - *Research of this kind requires that the researcher make choices: What knowledge will be represented symbolically and how will it be obtained? What assumptions underlie the statistical model? What is the researcher gaining by combining approaches?*

    (Klavans e Resnik, 1994)

23

Abordagens conflitantes

- 1994: ACL workshop "*The Balancing Act - Combining Symbolic and Statistical Approaches to Language*"
  - *Throughout the book the advantages of statistical approaches are reiterated – the ability to produce robust systems with wide coverage and graceful degradation that can handle ambiguity in a tractable manner. These characteristics cannot be achieved by a symbolic system alone. The rationale for using symbolic information is given less emphasis: to ensure the statistics are collected over linguistically motivated data instances (for example to handle the multitude of situations where the relevant context is not local), to help diminish the inevitable effect of sparse data, and to ensure that any output is meaningful.*

    (McCarthy, 1998)

24

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)

25

História do PLN

- Direcionada por correntes filosófico-linguisticas
  - Às vezes complementares
  - Às vezes rivais até a morte

26

Racionalismo

- 1960-1985: racionalismo entre linguistas, 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 linguística?
- 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 linguística?
  - Questão importante para a área de Aprendizado de Máquina

Racionalismo vs. empirismo

- Racionalismo
  - Linguística a la Chomsky (*gerativismo*)
    - Descrição do módulo linguístico 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 linguistica: conhecimento da língua pelo falante
    - Foco do racionalismo/gerativismo
      - Argumentam que é possível isolar esse componente para estudo e formalização
  - Desempenho linguístico: afetado por vários fatores, como memória disponível, distrações do ambiente, etc.
    - Foco do empirismo

Racionalismo vs. empirismo

- Linguística 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...*

34

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.*

35

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

36

Abordagens: Linguística

- Estruturalismo e gerativismo saem um pouco de foco
- Atualmente, na Linguística, 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?

A história mais a fundo

- Avanços da área no tempo
  - 1940-56: fundação da área
  - 1957-70: dois campos
  - 1970-83: quatro paradigmas
  - 1983-93: empirismo
  - 1994-99: fortalecimento da área
  - 2000-atual: aprendizado de máquina

38

A história mais a fundo

- Décadas de 40 e 50: fundação da área
  - Autômatos e máquinas de estados finitos de Turing: base para primeiro neurônio artificial (McCulloch e Pitts, 1943), expressões regulares (Kleene, 1951, 1956), gramáticas de Chomsky (1956)
  - Modelos probabilísticos: processamento de fala, *noisy-channel*

39

A história mais a fundo

- Dois campos
  - Fim dos anos 50 – início de 60: paradigmas simbólico e estatístico
    - Simbolismo
      - Primeira frente: a partir do trabalho de Chomsky e de informatas em *parsing* e programação dinâmica
        - Um dos primeiros parsers: Zellig Harris (1962)
      - Segunda frente: IA e as pesquisas em lógica e raciocínio
        - Provadores de teoremas

40

A história mais a fundo

- Dois campos
  - Fim dos anos 50 – início de 60: paradigmas simbólico e estatístico
    - Estatística
      - Departamentos de Estatística e Engenharia Elétrica
        - Aplicação de métodos bayesianos: OCR (Bledsoe e Browning, 1959), atribuição de autoria (Mosteller e Wallace, 1964)

41

A história mais a fundo

- Anos 60
  - Primeiros modelos psicológicos do processamento humano da linguagem
    - Gramáticas transformacionais
  - Primeiros córpus on-line
    - *Brown Corpus of American English* (Kucera e Francis, 1967)
    - Dicionário de dialeto chinês DOC (*Dictionary on Computer*) (Wang, 1967)

42

A história mais a fundo

- 1970-83: definição clara de quatro paradigmas
  - Paradigma estocástico: processamento de fala, modelos de Markov, *noisy-channel* e decodificação de texto
  - Paradigma lógico: *Definite Clause Grammars* (DCG) (Pereira e Warren, 1980), gramáticas funcionais (LFG, por exemplo) (Kay, 1979; Bresnan e Kaplan, 1982), unificação de estruturas de atributos

43

A história mais a fundo

- 1970-83: definição clara de quatro paradigmas
  - Interpretação de língua natural: mundo dos blocos de Winograd (1972) (manipulação de blocos a partir de comandos textuais), sistemas de interpretação textual com base em scripts, planos e objetivos (Schank e Abelson, 1977), uso de redes semânticas (Quillian, 1968) e papéis de Fillmore (1968)
  - Modelagem discursiva: estrutura do discurso e foco (Grosz, 1977), resolução anafórica (Hobbs, 1978), modelo BDI (*Belief-Desire-Intention*) para modelagem de atos de fala (Perrault e Allen, 1979)

44

A história mais a fundo

- 1983-1993
  - A valorização de modelos mais antigos
    - Modelos de estados finitos
      - Fonologia e morfologia (Kaplan e Kay, 1981)
      - Sintaxe (Church, 1980)
    - Empirismo
      - IBM e probabilidades para processamento de fala e de texto: *tagging*, *parsing*, *attachments*
      - Foco na avaliação: métricas, conjuntos de dados, comparações
  - Muito trabalho em geração textual

45

A história mais a fundo

- 1994-1999: maior reconhecimento e fortalecimento da área
  - Popularização científica de modelos probabilísticos e baseados em dados
  - Exploração comercial de processamento de fala, revisão gramatical e ortográfica
  - Web e recuperação e extração de informação

46

A história mais a fundo

- 2000-atual: a aceleração do empirismo e o aprendizado de máquina
  - 3 fatores
    - Grande quantidade de dados anotados
      - Tratamento de problemas mais complexos, grandes competições (NIST, por exemplo)
    - Interação com comunidade de aprendizado de máquina
    - Disponibilidade de computadores poderosos

47

A história mais a fundo

- 2000-atual: a aceleração do empirismo e o aprendizado de máquina
  - Recentemente
    - Aprendizado estatístico não supervisionado
      - Progresso significativo em tradução automática e outras áreas
      - Alto custo e dificuldade para produção de córpus anotados

48

Abordagens: PLN

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

49

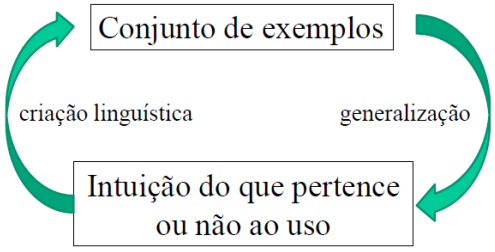
Abordagens: PLN

- Eric Laporte (2012) - *linguista*
  - As diferenças já não são tão evidentes
    - "Todo gerativista uso o Google escondido"
    - "Todo distribucionalista usa seu conhecimento e intuição"

50

Abordagens: PLN

- Eric Laporte (2012) - *linguista*
  - Dualidade córpus/introspecção



51