



Tópico 1: O quê é NLP?

 Materiais

Leitura 1 - Med...

Leitura 2

Leitura 3.pdf

Vídeo

 Due date

19 juillet 2023


 Status

● Done

Introdução à área de Processamento de Linguagem Natural e suas aplicações.

Leitura inicial I: - O que é Processamento de Linguagem Natural?, Medium.

- mecanismos como o google tradutor são considerados por muitos como precursores do PLN
- são necessários pré-processamentos para abstrair e estruturar a língua para que a máquina a entenda de forma mais eficiente

 o pré-processamento torna os dados **menos esparsos** e **reduz o vocabulário** (muito conveniente para o processamento computacional)

Algumas tarefas muito utilizadas no pré-processamento textual

Normalização

- tokenização:
 - lexical: marca cada palavra como um token no texto ("oi, está frio." → ['oi', ',', 'está', 'frio', '.'])
 - sentencial: identifica e marca sentenças

Esta é a primeira sentença. Esta é a segunda. Esta é a terceira!

['Esta é a primeira sentença.', 'Esta é a segunda.', 'Esta é a terceira!']

- maiúsculas para minúsculas
- remoção de caracteres especiais, etc

Os processos seguintes à normalização atuam sobre essas unidades sentenciais e lexicais.

Remoção de stopwords

- palavras muito frequentes, como preposições, artigos, pronomes relativos ("a", "de", "o", "da", "que", "e", "do", etc)
- são removidas porque geralmente não são relevantes para a construção do modelo (se forem, devem ser mantidas)
- tem listas de stopwords na internet

Remoção de numerais

- removidos por não possuírem carga semântica
- remover também as unidades de medida que os acompanham

? mas tudo isso depende do contexto, né? os números podem ser bem relevantes também

Correção ortográfica

- existem spell checkers que tratam datasets e corrigem erros de digitação, abreviações e vocabulário informal
- esses erros são grandiosos porque geram novos tokens, aumentando a esparsidade dos dados
- tem um [artigo](#) que mostra uma implementação de corretor ortográfico em Python

Stemização e Lematização

- Stemmização: processo que consiste reduzir uma palavra ao seu **radical**
- Lematização: reduz a palavra ao seu lema (como no dicionário, é a forma no **masculino e singular**)

⚠ quando lidando com verbos, o lema é o **infinitivo**

- o uso desses dois processos faz com que, novamente, o vocabulário seja reduzido e ocorra abstração de significado

i Todas as etapas de pré-processamento vistas acima são de cunho morfossintático (atuam em cima de itens **lexicais, palavras**). Também existem processamentos de nível sintático e semântico.

Leitura inicial II: - What is Natural Language Processing

- natural language understanding (NLU): usar computadores para compreender linguagem humana
- natural language generation (NLG): usar computadores para **produzir** linguagem humana

(ler até o tópico 'Machine learning models for NLP' em NLP Technology Overview).

Leitura inicial III: - Abordagens clássicas de NLP, UFU.

(ler a seção 5.2, exceto o item '5.2.5. Análise Pragmática')

- separação do processamento de linguagem, tornando a análise mais gerenciável

Tokenização

- segmentação de palavras (marca o ponto onde uma palavra termina e outra começa)
- essas palavras são chamadas de tokens
- tokenização para languages não segmentadas
 - chinês e tailandês são exemplos de linguagens assim
 - palavras são escritas sem indicação de limite de palavras
 - por isso, a tokenização requer informação léxica e morfológica adicional

- tokenização para linguagens delimitadas por espaço
 - limites de palavras indicados por espaço em branco
 - a maior parte das ambiguidades vem de sinais de pontuação, que podem ter significados diferentes em uma mesma sentença (exemplo do Av., R\$ 200.000 e fim de sentença)
 - há casos em que as pontuações são tratadas como token separado, mas também há casos em que devem ser anexadas a outro token (como em abreviações, tipo a Av.)

Análise léxica

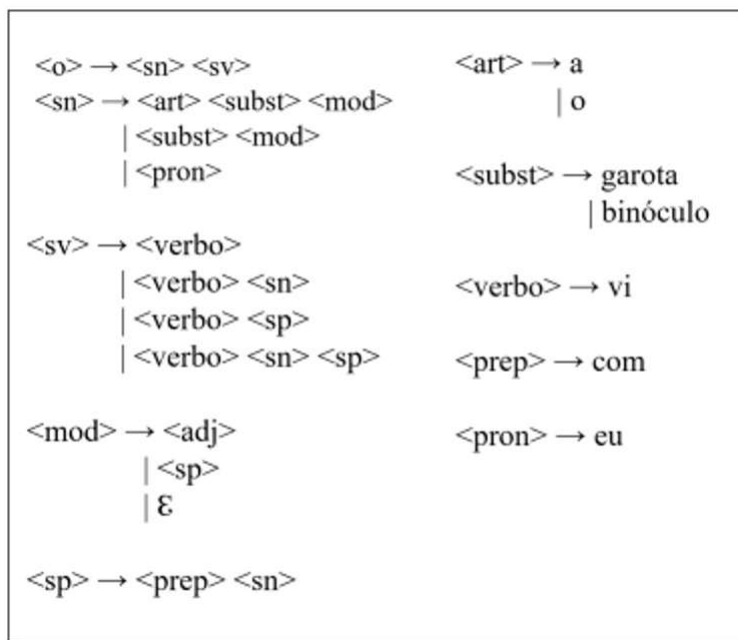
- depois da tokenização, faz-se a análise a nível de **palavras**
- a tarefa básica é relacionar variantes morfológicas das palavras aos seus lemmas (forma canônica, do jeito que vemos no dicionário)
- a análise léxica pode ser dividida em dois lados:
 - parsing side: mapeamento da palavra até o lemma
 - geração morfológica: mapeamento do lemma para a palavra

 o stem é o **radical** da palavra

- com o exemplo do texto: entregar é o lemma e entreg é o stem (a partir do qual se pode criar novas palavras)

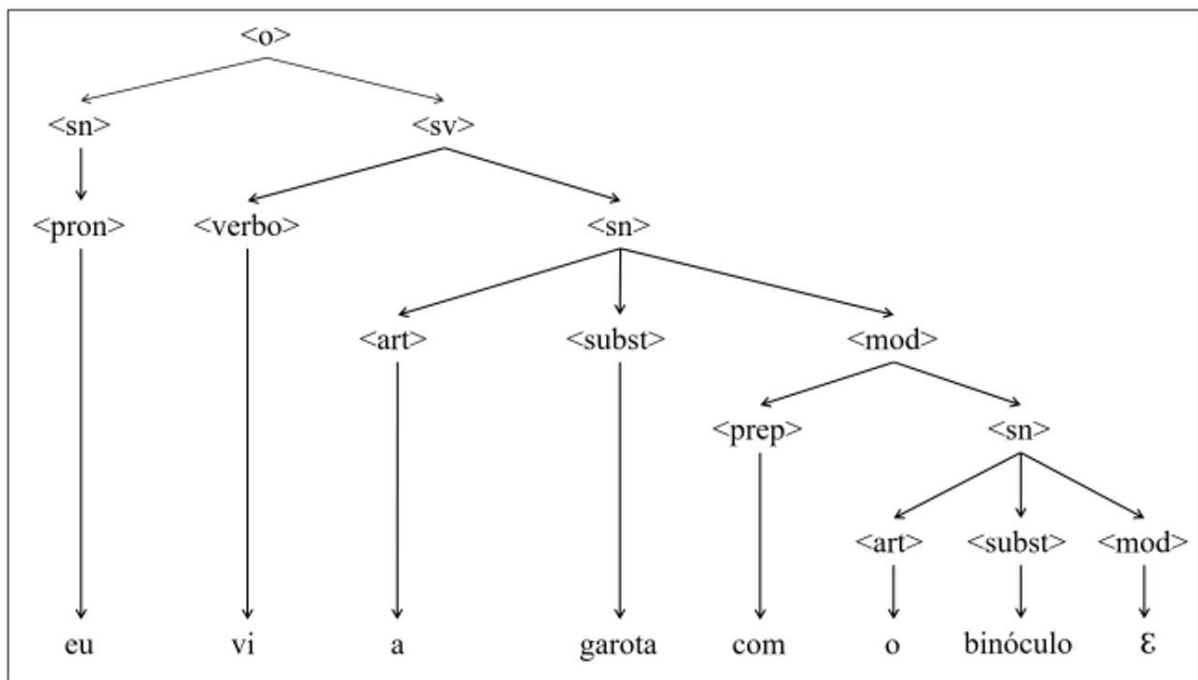
Análise sintática

- a unidade básica de análise de significado é a **frase**
- a análise sintática pode ser representada por gramáticas e árvores sintáticas



Página 6 - exemplo de uma gramática

- a gramática acima é composta por dois lados:
 - no esquerdo, símbolos pré-terminais, que se decompõem em uma
 - produção, ao lado direito, que são os termos léxicos



Página 6 - árvore sintática da frase "eu vi a garota como binóculo"

- cada nó interno na árvore representa a aplicação de uma regra da gramática

Análise semântica

- refere-se à análise do **significado** das palavras, expressões, sentenças inteiras e enunciados **no contexto**

- ou seja, traduz as expressões originais em um tipo de metalinguagem
- a evidência primária para a semântica vem das **interpretações do orador nativo** no uso das expressões no contexto (são detectáveis usando técnicas de linguísticas em Córpus diversos)
- resolução de ambiguidade: para uma máquina, um enunciado humano pode ter **diversas interpretações**
 - ambiguidade léxica: palavras com mais de um significado (manga) ou polissemia (diferentes sentidos para uma mesma palavra, diferenças bem