Processamento de Linguagem Natural

Aula 04

Marcação Textual - Part of Speech Tagging

PUC Minas

Part-of-Speech Tagging

- O Processamento de Linguagem Natural (PLN) se preocupa, entre outras tarefas, em transformar linguagem humana em informação útil computacionalmente.
- Diversas técnicas emergem como formas de tratar a linguagem, sendo uma delas o **Part-of-Speech Tagging**.

Part-of-Speech Tagging

- O POS Tagging se trata de um processo de rotulação de elementos textuais - tipicamente palavras e pontuação - com o fim de evidenciar a estrutura gramatical de um determinado trecho de texto.

Part-of-Speech Tagging

- A tarefa de rotulação morfossintática (POS Tagging) se preocupa em classificar os tokens de uma sentença de acordo com suas classes morfológicas - substantivo, verbo, adjetivo, entre outros

Part-of-Speech Tagging

- Categorizar sinteticamente as palavras é útil pois revela muito sobre a própria palavra e seus vizinhos
  - Saber que uma palavra é um substantivo ou verbo pode nos alertar que ela é seguida de um adjetivo - no caso do substantivo - ou por um adjetivo, ou verbo.

Part-of-Speech Tagging

- Por isso, a marcação é um aspecto chave no tratamento de textos.
- É útil na marcação de entidades nomeadas (que veremos em breve) como pessoas e organizações durante o processo de extração da informação.

Part-of-Speech Tagging

- Em reconhecimento e síntese de fala, seu uso é útil para:
  - extração de termos,
  - desambiguação,
  - composição de novas frases
  - pesquisa lexicográfica.

Part-of-Speech Tagging

- Existe um grande número de aplicações possíveis de serem construídas usando POS Taggers e há uma diversa quantidade de trabalhos sendo desenvolvidos nesta área

Part-of-Speech Tagging

- Para português do Brasil ainda existe campo para avanços, seja em correção de corpora, seja em aplicação de técnicas utilizadas até então apenas para o inglês.

Part-of-Speech Tagging

- A tarefa parece ser intuitiva para grande parte das pessoas mas o processo de automatização não é trivial.
- Um dos maiores problemas presentes em todos os idiomas é a presença de ambiguidade entre palavras
  - Português: morro (substantivo) e morro (verbo)
  - Inglês: object (substantivo) e object (verbo)
  - Alemão: sein (verbo) e sein (pronome)

Métodos para o português

- A acurácia dos taggers adaptados para o português são por volta dos 97%
- Ainda são trabalhos acadêmicos
- Existem várias abordagens para resolver o problema de classificação das palavras, como:
  - Método Estocástico
  - Método baseado em regras
  - Etiquetador baseado em transformação

PUC Minas

Método Estocástico

- Simples
- Bom desempenho
- Possui implementação no Natural Language Toolkit (NLTK)

NLTK - DefaultTagger

- Anotador mais básico da biblioteca
- Etiqueta os tokens de acordo com o tipo de caractere utilizado.
  - Se o Token for número será classificado como numeral
- Precisão média de 20-30%
- Baixíssima performance se utilizado sozinho.

NLTK - UnigramTagger

- Também conhecido como Lookup Tagger
- Calcula a probabilidade de um token receber uma dada etiqueta com base nas frequências obtidas de um corpus de treinamento.

NLTK - UnigramTagger

- O cálculo é feito a partir de um modelo probabilístico com a fórmula:

$$P(t_i|w) = \frac{c(w, t_i)}{c(w, t_1) + ... + c(w, t_k)}$$

Sendo w a palavra a ser classificada e *t1, ..., tk* uma lista das tags (etiquetas) possíveis. c(*w, ti*) indica quantas vezes a correspondência da palavra *w* com a etiqueta *ti* apareceu no corpus de treinamento.

NLTK - UnigramTagger

- Se a palavra “morro”, em português, foi etiquetada 15 vezes como verbo e 65 como substantivo no corpus de treinamento, então:

$$P(\text{verbo} \mid \text{morro}) = 15/80 = 0{,}19$$
$$P(\text{substantivo} \mid \text{morro}) = 65/80 = 0,\ 81$$

- Nesta estratégia, a etiqueta a ser selecionada é aquela com maior probabilidade de ser a correta de acordo com o corpus de treinamento.

NLTK - UnigramTagger

- No caso do exemplo anterior, uma nova ocorrência de “morro” seria etiquetada como substantivo, pois (substantivo, morro)
- A performance do *UnigramTagger* muito da qualidade do corpus de treinamento fornecido

NLTK - N-Gram Tagging

- Quando realizar o processamento utilizando *unigrams* nós estamos utilizando somente um item do contexto, levamos em consideração somente um *token* isolado de um contexto maior

NLTK - N-Gram Tagging

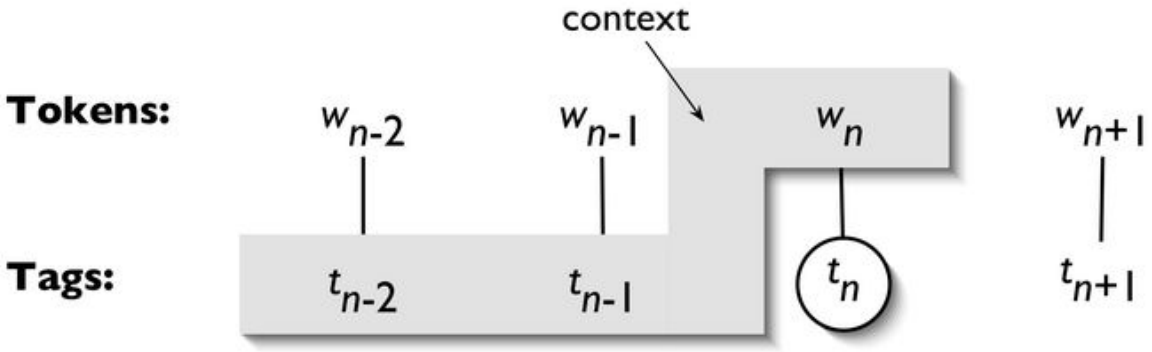
- No exemplo dado com a palavra “morro”, nos dois casos abaixo ela seria classificada como substantivo

Eu *morro* de frio durante o trabalho

O *morro* da minha casa é íngrime

NLTK - N-Gram Tagging

- O *N-Gram Tagger* é uma generalização do *Unigram* onde o contexto utilizado para a palavra atual é ela junto com as tags dos *n-1* tokens anteriores



- O rotulador *N-Gram* escolhe a classificação mais adequada para o contexto dado.

PUC Minas

NLTK - N-Gram Tagging

- Sua implementação é mais complexa, porém ela gera uma precisão maior por depender, também, do contexto onde as palavras se encontram e não apenas da palavra em si.

NLTK - N-Gram Tagging

- Quanto maior o número de tokens anteriores verificados pelo anotador automático, maior o tempo de execução da ferramenta e maior a necessidade de uma ampla variedade de sequências de etiquetas possíveis no corpus de treinamento (abrangência) pois os contextos presentes no texto a ser classificado podem não se apresentar no treino gerando, assim, erros de etiquetação.

Construindo um Tagger

- A construção de um part-of-speech tagger necessita de um *Corpus* de treinamento anotado.
- Os *Corpus* etiquetados podem usar diferentes convenções para a rotulação das palavras, os conjuntos de etiquetas usadas para essa tarefa em particular é conhecida como *tagset*.

Construindo um Tagger

- Mac-Morpho é um Corpus fechado, formado por artigos publicados no jornal Folha de São Paulo, em 1994, contendo mais de 1 milhão de palavras, anotadas pelo etiquetador de palavras (BICK 2000)

Construindo um Tagger

- Teve sua primeira versão criada em 2003, com revisões de melhoria da qualidade em 2013 e 2015.
- Formado por 1,1 milhões de palavras validadas manualmente com anotações morfossintáticas

http://nilc.icmc.usp.br/macmorpho/#ref3

Construindo um Tagger

Tabela 5.2. Etiquetas MacMorpho

CLASSE GRAMATICAL	ETIQUETA
ADJETIVO	ADJ
ADVÉRBIO CONECTIVO SUBORDINATIVO	ADV-KS
ADVÉRBIO RELATIVO SUBORDINATIVO	ADV-KS-REL
ARTIGO (def. ou indef.)	ART
CONJUNÇÃO COORDENATIVA	KC
CONJUNÇÃO SUBORDINATIVA	KS
INTERJEIÇÃO	IN
SUBSTANTIVO	N
SUBSTANTIVO PRÓPRIO	NPROP
NUMERAL	NUM
PARTICÍPIO	PCP
PALAVRA DENOTATIVA	PDEN
PREPOSIÇÃO	PREP
PRONOME ADJETIVO	PROADJ
PRONOME CONECTIVO SUBORDINATIVO	PRO-KS
PRONOME PESSOAL	PROPESS
PRONOME RELATIVO CONECTIVO SUBORDINATIVO	PRO-KS-REL
PRONOME SUBSTANTIVO	PROSUB
VERBO	V
VERBO AUXILIAR	VAUX
SIMBOLO DE MOEDA CORRENTE	CUR

Construindo um Tagger

- Como utilizar o NLTK

```
import nltk
#nltk.download('mac_morpho')

#Importa a biblioteca
from nltk.corpus import mac_morpho
#Carrega as sentença rotuladas do Corpus
sentencas_etiquetadas = mac_morpho.tagged_sents()
```

Links

- http://nilc.icmc.usp.br/macmorpho/
- http://www.nltk.org/howto/portuguese_en.html
- https://www.nltk.org/book/ch05.html