Etiquetagem Morfossintática

Análise de Dependências

Prof. Marcos Lopes

Departamento de Linguística – USP

Partes do Discurso (POS)

Etiquetas Universais

POS-tagging

Gramática de Dependências

Definição

- As *Partes do Discurso* ou *Partes da Oração* (POS ou PoS: *Part Of Speech*) são classes atribuídas às palavras segundo critérios *morfossintáticos* ou *semânticos* (conceituais).

Definição

- As *Partes do Discurso* ou *Partes da Oração* (POS ou PoS: *Part Of Speech*) são classes atribuídas às palavras segundo critérios *morfossintáticos* ou *semânticos* (conceituais).
- Vamos nos concentrar sobre os morfossintáticos. São de dois tipos:

Definição

- As *Partes do Discurso* ou *Partes da Oração* (POS ou PoS: *Part Of Speech*) são classes atribuídas às palavras segundo critérios *morfossintáticos* ou *semânticos* (conceituais).
- Vamos nos concentrar sobre os morfossintáticos. São de dois tipos:

Sintático A interrelação das palavras nas frases (por ex., um verbo exige um substantivo para formar uma frase);

Definição

- As *Partes do Discurso* ou *Partes da Oração* (POS ou PoS: *Part Of Speech*) são classes atribuídas às palavras segundo critérios *morfossintáticos* ou *semânticos* (conceituais).
- Vamos nos concentrar sobre os morfossintáticos. São de dois tipos:

Sintático A interrelação das palavras nas frases (por ex., um verbo exige um substantivo para formar uma frase);

Morfológico As flexões, isto é, modificações na forma da palavra para indicar gênero, número, grau e pessoa (em português).

Classes abertas e fechadas

Em qualquer língua, há duas classes de palavras divididas em função da possibilidade ou não de alteração dos elementos nessas classes.

Classes abertas e fechadas

Em qualquer língua, há duas classes de palavras divididas em função da possibilidade ou não de alteração dos elementos nessas classes.

Fechada Praticamente não surgem novos elementos nem são eliminados os que já existem.
Na verdade, existem mudanças, mas costumam levar muito tempo a ocorrer, de forma que os falantes não têm consciência delas.
Fazem parte da classe fechada as palavras “funcionais” (ou “gramaticais”) da língua: artigos, preposições, pronomes...
Há poucos *types* e muitos *tokens* na classe fechada.

Classes abertas e fechadas

Em qualquer língua, há duas classes de palavras divididas em função da possibilidade ou não de alteração dos elementos nessas classes.

Fechada Praticamente não surgem novos elementos nem são eliminados os que já existem.
Na verdade, existem mudanças, mas costumam levar muito tempo a ocorrer, de forma que os falantes não têm consciência delas.
Fazem parte da classe fechada as palavras "funcionais" (ou "gramaticais") da língua: artigos, preposições, pronomes...
Há poucos *types* e muitos *tokens* na classe fechada.

Aberta Novos elementos surgem e desaparecem constantemente.
Fazem parte da classe aberta as palavras "conceituais": substantivos, verbos, adjetivos, advérbios.
Seu número é grande, mas a relação type/token se inverte por relação às palavras fechadas: muitos *types*, pouco *tokens*.
$\approxeq$ 70% das palavras dos dicionários são substantivos.

Por que se importar com as Partes do Discurso?

- Objetividade: Elas representam a classificação mais antiga de expressões linguísticas de que se tem notícia no mundo ocidental. Testadas muitas vezes e, em geral, aprovadas.

Por que se importar com as Partes do Discurso?

- Objetividade: Elas representam a classificação mais antiga de expressões linguísticas de que se tem notícia no mundo ocidental. Testadas muitas vezes e, em geral, aprovadas.
- Úteis nos modelos sequenciais: Saber que uma palavra é um artigo aumenta as chances de a seguinte ser um substantivo. A recíproca é verdadeira.

Por que se importar com as Partes do Discurso?

- Objetividade: Elas representam a classificação mais antiga de expressões linguísticas de que se tem notícia no mundo ocidental. Testadas muitas vezes e, em geral, aprovadas.
- Úteis nos modelos sequenciais: Saber que uma palavra é um artigo aumenta as chances de a seguinte ser um substantivo. A recíproca é verdadeira.
- Úteis nos modelos BoW: A classe da palavra indica o tipo de informação que ela representa (uma ação, uma entidade...).

Por que se importar com as Partes do Discurso?

- Objetividade: Elas representam a classificação mais antiga de expressões linguísticas de que se tem notícia no mundo ocidental. Testadas muitas vezes e, em geral, aprovadas.
- Úteis nos modelos sequenciais: Saber que uma palavra é um artigo aumenta as chances de a seguinte ser um substantivo. A recíproca é verdadeira.
- Úteis nos modelos BoW: A classe da palavra indica o tipo de informação que ela representa (uma ação, uma entidade...).
- Fácil de aplicar e com bons resultados: Existem muitas bibliotecas prontas e gratuitas para a etiquetagem morfossintática, todas com acurácia acima de 90%.

Partes do Discurso (POS)

Etiquetas Universais

POS-tagging

Gramática de Dependências

Etiquetas de Classes Abertas

Abrev.	Classe	Exemplos
ADJ	Adjetivo	alto, caríssima
ADV	Advérbio	geralmente, talvez
INTJ	Interjeição	Oi!, Ufa!
NOUN	Substantivo	carro, escola
PROPN	Nome próprio	João
VERB	Verbo	comprar, saísse

Fonte: `https://universaldependencies.org/docs/u/pos/`

Etiquetas de Classes Fechadas

Abrev.	Classe	Exemplos
ADP	Adposição	preposições (*de, para*) e posposições (ing.: *ago*, *Mary's*)
AUX	Verbo Auxiliar	tem sido, vai comprar
CONJ	Conjunção	e, ou
DET	Determinante	o, uns, nenhum
NUM	Numeral	50, dez, IV
PART	Partícula	Ing.: *not, up, off*
PRON	Pronome	eu, lhe, quem
SCONJ	Conjunção subordinativa	que, se

Outras classes

Abrev.	Classe	Exemplos
PUNCT	Pontuação	. , ! ()
SYM	Símbolo	+, $
X	Outros	(não classificado)

Partes do Discurso (POS)

Etiquetas Universais

POS-tagging

Gramática de Dependências

Desambiguação

- Muitas palavras, por si mesmas (fora da oração), são ambíguas quanto à sua classe morfossintática. Exemplos:

Desambiguação

- Muitas palavras, por si mesmas (fora da oração), são ambíguas quanto à sua classe morfossintática. Exemplos:
  - "vela" é um substantivo ou um verbo?

Desambiguação

- Muitas palavras, por si mesmas (fora da oração), são ambíguas quanto à sua classe morfossintática. Exemplos:
  - "vela" é um substantivo ou um verbo?
  - "francês" é um substantivo ou um adjetivo?

Desambiguação

- Muitas palavras, por si mesmas (fora da oração), são ambíguas quanto à sua classe morfossintática. Exemplos:
  - "vela" é um substantivo ou um verbo?
  - "francês" é um substantivo ou um adjetivo?
- A tarefa de etiquetagem morfossintática é, assim, uma das tarefas de *desambiguação*.

Desambiguação

- Muitas palavras, por si mesmas (fora da oração), são ambíguas quanto à sua classe morfossintática. Exemplos:
  - "vela" é um substantivo ou um verbo?
  - "francês" é um substantivo ou um adjetivo?
- A tarefa de etiquetagem morfossintática é, assim, uma das tarefas de *desambiguação*.
- A maioria das palavras não é morfossintaticamente ambígua, mas as que são ambíguas formam a maioria dos tokens em qualquer documento – ou seja, são as palavras mais comuns.

Desambiguação

- Muitas palavras, por si mesmas (fora da oração), são ambíguas quanto à sua classe morfossintática. Exemplos:
  - "vela" é um substantivo ou um verbo?
  - "francês" é um substantivo ou um adjetivo?
- A tarefa de etiquetagem morfossintática é, assim, uma das tarefas de *desambiguação*.
- A maioria das palavras não é morfossintaticamente ambígua, mas as que são ambíguas formam a maioria dos tokens em qualquer documento – ou seja, são as palavras mais comuns.
- Por isso, os *parsers* (analisadores) morfossintáticos costumam incluir um viés na classificação: diante da incerteza, apostar na classe mais comum.

Mac-Morpho

O NLTK oferece um corpus anotado morfossintaticamente por humanos (padrão ouro): é o Mac-Morpho, com textos recolhidos de notícias da Folha de S. Paulo.

Mac-Morpho

O NLTK oferece um corpus anotado morfossintaticamente por humanos (padrão ouro): é o Mac-Morpho, com textos recolhidos de notícias da Folha de S. Paulo.

O corpus tem 1.170.095 palavras etiquetadas. Também é possível segmentá-los por sentenças. São 51.397 sentenças.

Mac-Morpho

O NLTK oferece um corpus anotado morfossintaticamente por humanos (padrão ouro): é o Mac-Morpho, com textos recolhidos de notícias da Folha de S. Paulo.

O corpus tem 1.170.095 palavras etiquetadas. Também é possível segmentá-los por sentenças. São 51.397 sentenças.

Por fim, existe a possibilidade de se usar o corpus como conjunto textual não etiquetado.

Mac-Morpho (cont.)

```
import nltk

# Lista das palavras etiquetadas do corpus
palavras_etiquetadas = nltk.corpus.mac_morpho.tagged_words()

# Lista das palavras do corpus sem etiquetas
palavras = nltk.corpus.mac_morpho.words()

# Lista de listas com as sentenças etiquetadas
sents_tags = nltk.corpus.mac_morpho.tagged_sents()

# Lista de listas com as sentenças não etiquetadas
sents = nltk.corpus.mac_morpho.sents()
```

O POS-tagger do spaCy

O etiquetador morfossintático do spaCy é certamente o mais usado para a língua portuguesa hoje em dia.

O POS-tagger do spaCy

O etiquetador morfossintático do spaCy é certamente o mais usado para a língua portuguesa hoje em dia.

Nele, a parte do discurso (POS) é um dos atributos dos *tokens* de um documento.

O POS-tagger do spaCy

O etiquetador morfossintático do spaCy é certamente o mais usado para a língua portuguesa hoje em dia.

Nele, a parte do discurso (POS) é um dos atributos dos *tokens* de um documento.

```
import spacy
nlp = spacy.load('pt_core_news_sm')

doc = nlp('O rato roeu a roupa do rei de Roma')
for token in doc:
        print(token.text, token.pos_)
```

Partes do Discurso (POS)

Etiquetas Universais

POS-tagging

Gramática de Dependências

Definição

- As Gramáticas de Dependências são geradas por funções binárias de associações entre palavras.

Definição

- As Gramáticas de Dependências são geradas por funções binárias de associações entre palavras.
- Nessa gramática, as palavras são de dois tipos:

Definição

- As Gramáticas de Dependências são geradas por funções binárias de associações entre palavras.
- Nessa gramática, as palavras são de dois tipos:
  - Núcleos (*heads*)

Definição

- As Gramáticas de Dependências são geradas por funções binárias de associações entre palavras.
- Nessa gramática, as palavras são de dois tipos:
  - Núcleos (*heads*)
  - Dependentes

Definição

- As Gramáticas de Dependências são geradas por funções binárias de associações entre palavras.
- Nessa gramática, as palavras são de dois tipos:
  - Núcleos (*heads*)
  - Dependentes
- Um tipo especial de núcleo é a *raiz* (*root*), que é o núcleo da sentença inteira.

Definição

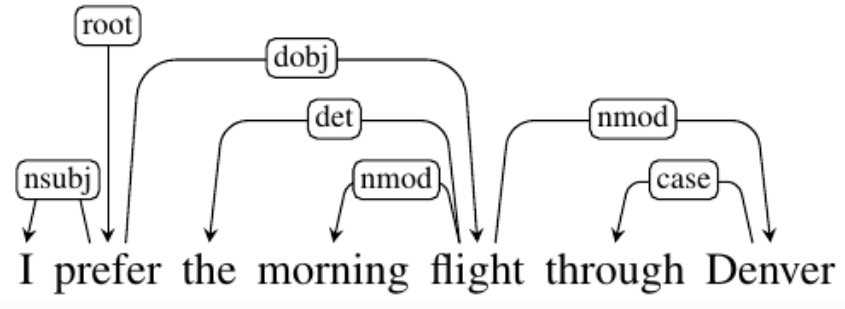
- As Gramáticas de Dependências são geradas por funções binárias de associações entre palavras.
- Nessa gramática, as palavras são de dois tipos:
  - Núcleos (*heads*)
  - Dependentes
- Um tipo especial de núcleo é a *raiz* (*root*), que é o núcleo da sentença inteira.
- As Gramáticas de Dependência são dos modelos de linguagem formalmente mais simples.

Definição

- As Gramáticas de Dependências são geradas por funções binárias de associações entre palavras.
- Nessa gramática, as palavras são de dois tipos:
  - Núcleos (*heads*)
  - Dependentes
- Um tipo especial de núcleo é a *raiz* (*root*), que é o núcleo da sentença inteira.
- As Gramáticas de Dependência são dos modelos de linguagem formalmente mais simples.

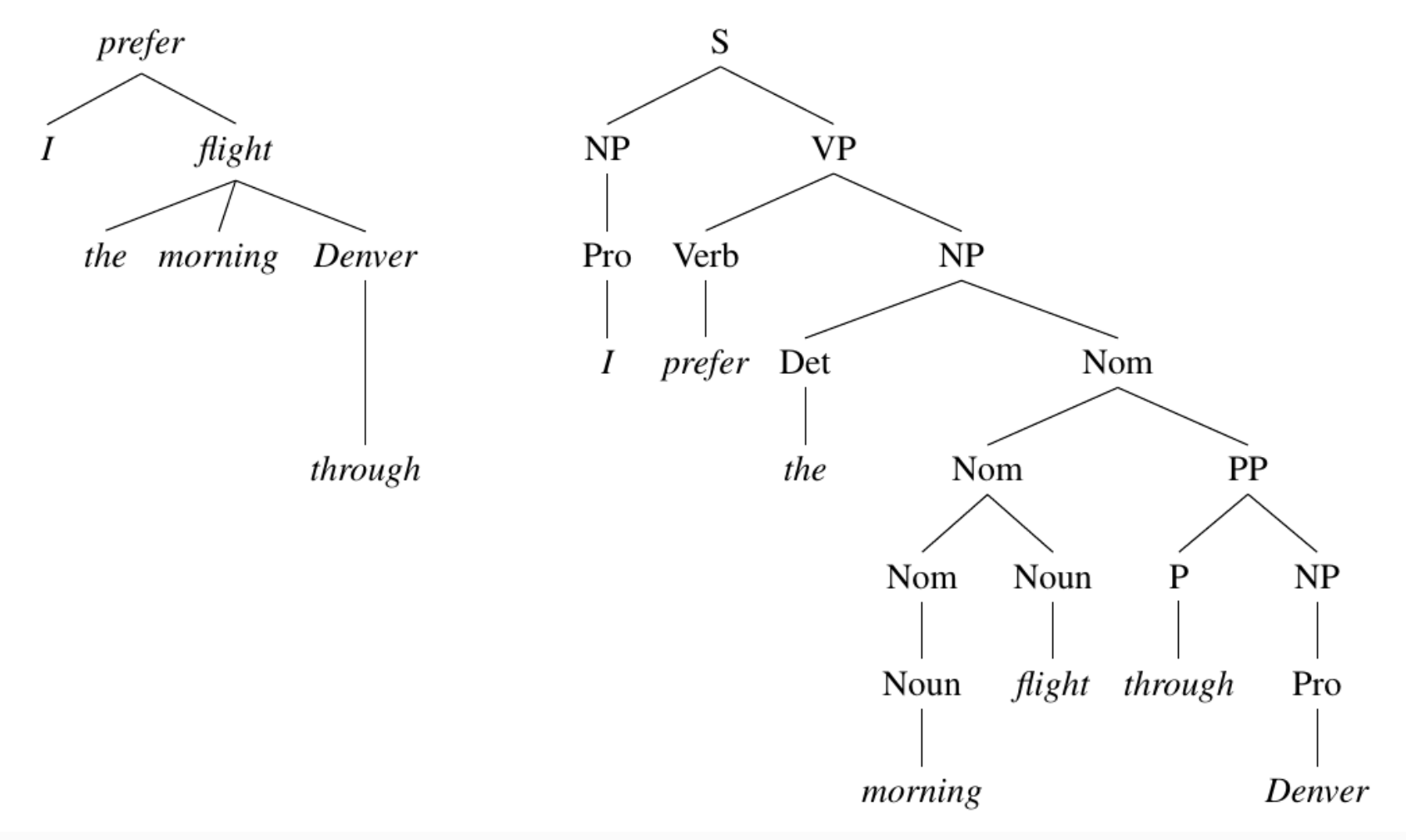
Definição

- As Gramáticas de Dependências são geradas por funções binárias de associações entre palavras.
- Nessa gramática, as palavras são de dois tipos:
  - Núcleos (*heads*)
  - Dependentes
- Um tipo especial de núcleo é a *raiz* (*root*), que é o núcleo da sentença inteira.
- As Gramáticas de Dependência são dos modelos de linguagem formalmente mais simples.



Fonte: Jurafsky & Martin (2020).

Gramática de Dependências *versus* Gramática de Constituintes



Fonte: Jurafsky & Martin (2020).

Funções Gramaticais (parte do conjunto UD)

Relações ligadas ao verbo	
ROOT	Núcleo da oração
NSUBJ	Sujeito nominal
DOBJ	Objeto direto
IOBJ	Objeto indireto
ADVMOD	Modificador adverbial
Relações nominais	
NMOD	Modificador nominal
AMOD	Modificador adjetival
NUMMOD	Modificador numérico
DET	Determinante
CASE	Casos: preposições, possessivos etc
CONJ	Relação de Conjunção
CC	Coordenador de conjunção

Tabela completa em `http://universaldependencies.org/u/dep/index.html`

Link para artigo com a proposta das Dependências Universais (UD).

Análise de Dependências com o spaCy

O módulo spaCy fornece um analisador de dependências para os modelos de língua portuguesa.

Análise de Dependências com o spaCy

O módulo spaCy fornece um analisador de dependências para os modelos de língua portuguesa.

Os resultados da análise são razoáveis. Os erros tendem a aumentar conforme a distância entre o núcleo e seus dependentes.

Análise de Dependências com o spaCy

O módulo spaCy fornece um analisador de dependências para os modelos de língua portuguesa.

Os resultados da análise são razoáveis. Os erros tendem a aumentar conforme a distância entre o núcleo e seus dependentes.

Um recurso útil para ajudar a lembrar o significado das etiquetas é o método explain().

Análise de Dependências com o spaCy

O módulo spaCy fornece um analisador de dependências para os modelos de língua portuguesa.

Os resultados da análise são razoáveis. Os erros tendem a aumentar conforme a distância entre o núcleo e seus dependentes.

Um recurso útil para ajudar a lembrar o significado das etiquetas é o método explain().

Merece destaque também o submódulo displacy, que permite gerar gráficos das relações de dependência.

Análise de Dependências com o spaCy (cont.)

```
for token in doc:
        print(token.text, token.dep_)

spacy.explain('advmod')

from spacy import displacy
displacy.render(doc, style="dep", jupyter=True)
```