

spaCy

Extração de Informação

André Câmara (acan@cesar.school)

Agenda

1. Introdução
2. Abordagens
3. Demonstração
4. Exercício

Sequence Labeling

- A idéia é ir além do problema de classificação padrão
 - **Casos individuais são desconectados e independentes** (*i.i.d.: independently and identically distributed*)
- Utilizar informação dos padrões na **ordem em que são observados**
- Cada **token** possui um **"label"** a ele associado
 - Labels são **dependentes** dos seus vizinhos

Sequence Labelling

Problema

Dado uma **sequência** de *tokens*, inferir qual a sequência **mais provável** de labels para estes *tokens*

Exemplos:

Part of Speech Tagging

Slot Filling (chatbots)

Named Entity Recognition

Part Of Speech Tagging

- Nível mais baixo de análise sintática

John saw the saw and decided to take it to the table.

PN V Det N Con V Part V Pro Prep Det N

- Utilizado para análises subsequentes

Extração de informação

- Identificação de **dados relevantes** nos **textos**
- Encontrar informações **específicas**
 - **PESSOAS**, **ORGANIZAÇÕES**, etc.
- Transformar de um **texto** em um banco de dados
 - Extrai **informações relevantes** baseando-se no domínio de conhecimento do documento
 - Exemplo:

MAKE **MODEL** **YEAR** **MILEAGE** **PRICE**

For sale, 2002 Toyota Prius, 20,000 mi, \$15K or best offer.

Available starting July 30, 2006.

Texto Livre

4 de abril em Dallas – cedo na noite passada, um tornado varreu todo o noroeste da área de Dallas, causando extensos danos. Testemunhas confirmam que o ciclone passou sem advertência, aproximadamente às 7:15 da noite, e destruiu dois *motor-homes*. O posto Texaco, na Rua Principal, 102, Farmers Branch, TX, também foi severamente danificado, mas nenhuma morte foi informada. O valor total calculado dos danos é de US\$200.000.

Jornal



**Sistema de
Extração de
Informações**

Template

Evento: tornado
Data: 4/4/2000
Hora: 19:15
Local: Farmers Branch : "noroeste de Dallas" : TX : USA
Danos: "*motor-homes*" (2) : "Posto Texaco" (1)
Perdas Estimadas: US\$200.000
Mortes: nenhuma

Extração de Informação

October 14, 2002, 4:00 a.m. PT

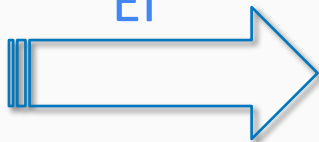
For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

EI



Associação

Segmentação

Clusterização

Classificação

Cluster A

Microsoft Corporation

CEO

Bill Gates

Microsoft

Gates

Microsoft

Cluster B

Bill Veghte

Microsoft

VP

Cluster C

Richard Stallman

founder

Free Software Foundation

Extração de Informação

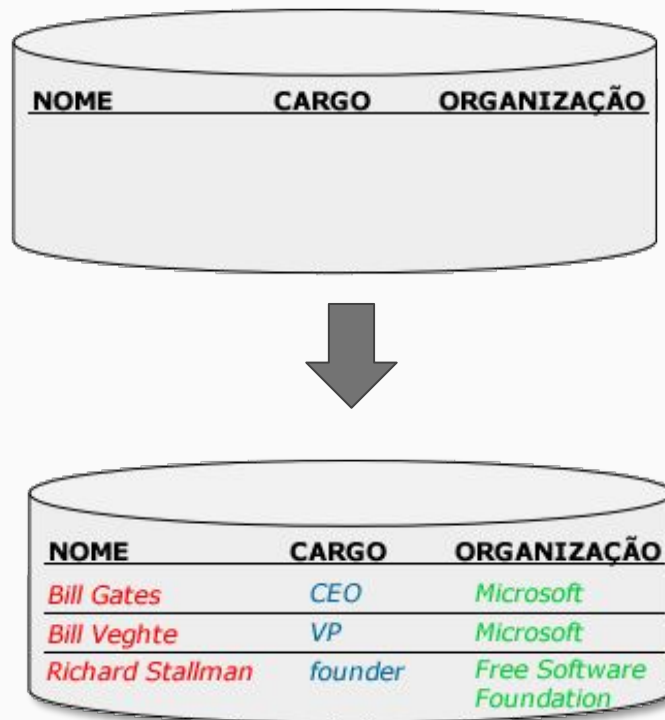
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



Principais abordagens

1. Rule-based models (REGEX, FSA, ...)
2. Token classification
3. Probabilistic Sequence models (HMM, MEMM, CRF)
4. Neural networks

Autômatos Finitos

Bons para textos estruturados.

Definidos manualmente ou aprendidos automaticamente.

Tipos:

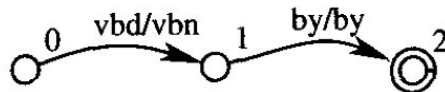
- Acceptors: com resposta **sim** ou **não**
- Recognizers: um ou mais estados finais (categorização)

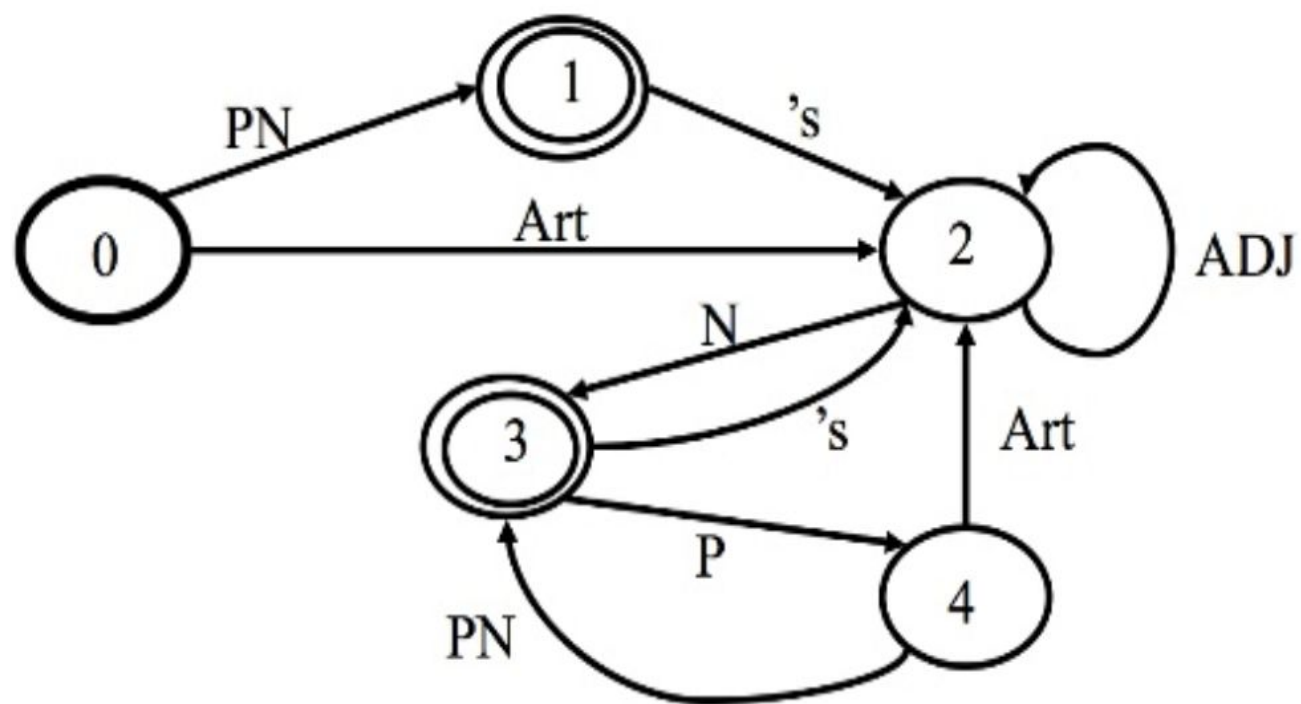
Chapman/np killed/vbn John/np Lennon/np

John/np Lennon/np was/bedz shot/vbd by/by Chapman/np

He/pps witnessed/vbd Lennon/np killed/vbn by/by Chapman/np

VB - Verb, base form
VBD - Verb, past tense
VBG - Verb, gerund or present participle
VBN - Verb, past participle
VBP - Verb, non-3rd person singular present
VBZ - Verb, 3rd person singular present



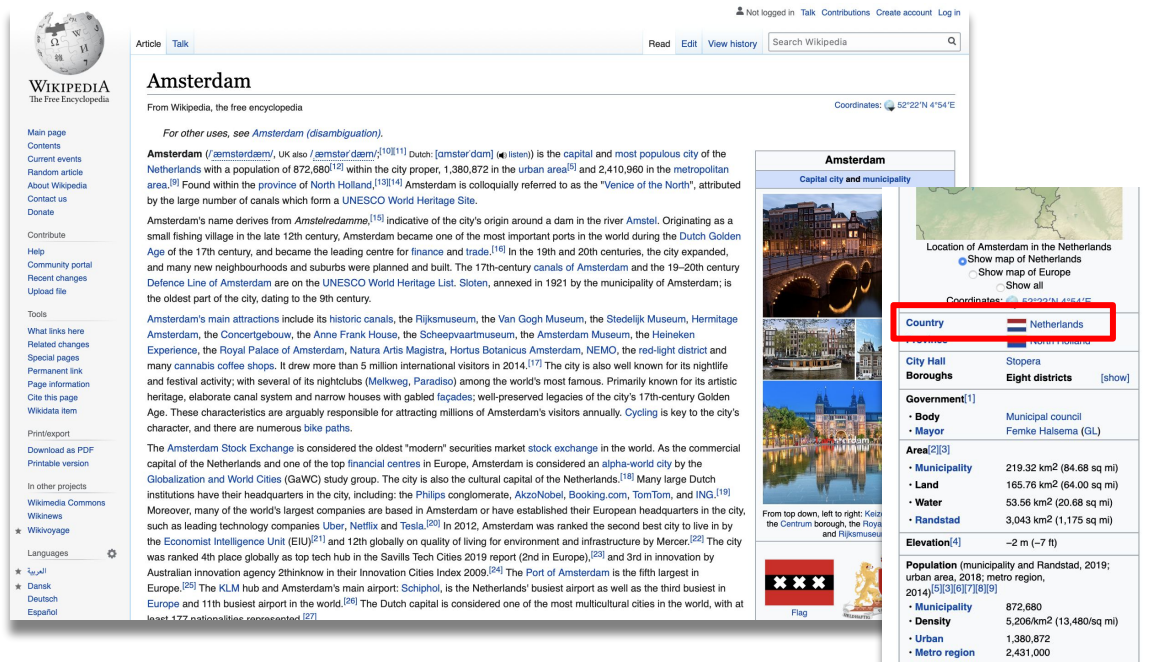


=> John's interesting book with a nice cover

REGEX

Muito útil para termos destacados por tags ou formatação

Ex.: HTML



The screenshot displays the Wikipedia page for Amsterdam. The title "Amsterdam" is prominently displayed at the top. Below it, the introductory text describes the city's location, population, and historical significance. A sidebar on the left contains various navigation links, including "Main page", "Contents", "Current events", "Random article", "About Wikipedia", "Contact us", "Donate", "Community portal", "Recent changes", "Upload file", "Tools", "What links here", "Related changes", "Special pages", "Permanent link", "Page information", "Cite this page", "Wikidata item", "Print/export", "Download as PDF", "Printable version", "In other projects", "Wikimedia Commons", "Wikispecies", "Wikivoyage", "Languages", and "Flag". The main content area includes a map of Amsterdam, a table of the city's districts, and a list of the city's government bodies. The sidebar on the right contains a list of the city's districts and a table of the city's government bodies.

Amsterdam

Capital city and municipality

Coordinates: 52°22′N 4°54′E﻿ / ﻿52.367°N 4.9°E﻿ / 52.367; 4.9

For other uses, see *Amsterdam* (disambiguation).

Amsterdam (/ˈæmstərdɑːm/, UK also /ˈæmstərdæm/^{[1][9][11]} Dutch: [ˈɑmstərˈdɑm] listen) is the capital and most populous city of the Netherlands with a population of 872,680^[12] within the city proper, 1,380,872 in the urban area^[6] and 2,410,960 in the metropolitan area.^[9] Found within the province of North Holland^{[13][14]} Amsterdam is colloquially referred to as the "Venice of the North", attributed by the large number of canals which form a UNESCO World Heritage Site.

Amsterdam's name derives from *Amstelredamme*,^[15] indicative of the city's origin around a dam in the river *Amstel*. Originating as a small fishing village in the late 12th century, Amsterdam became one of the most important ports in the world during the Dutch Golden Age of the 17th century, and became the leading centre for finance and trade.^[16] In the 19th and 20th centuries, the city expanded, and many new neighbourhoods and suburbs were planned and built. The 17th-century canals of Amsterdam and the 19–20th century Defence Line of Amsterdam are on the UNESCO World Heritage List. *Sloten*, annexed in 1921 by the municipality of Amsterdam; is the oldest part of the city, dating to the 9th century.

Amsterdam's main attractions include its historic canals, the Rijksmuseum, the Van Gogh Museum, the Stedelijk Museum, Hermitage Amsterdam, the Concertgebouw, the Anne Frank House, the Scheepvaartmuseum, the Amsterdam Museum, the Heineken Experience, the Royal Palace of Amsterdam, Natura Artis Magistra, Hortus Botanicus Amsterdam, NEMO, the red-light district and many cannabis coffee shops. It drew more than 5 million international visitors in 2014.^[17] The city is also well known for its nightlife and festival activity; with several of its nightclubs (Melkweg, Paradiso) among the world's most famous. Primarily known for its artistic heritage, elaborate canal system and narrow houses with gabled façades; well-preserved legacies of the city's 17th-century Golden Age. These characteristics are arguably responsible for attracting millions of Amsterdam's visitors annually. *Cycling* is key to the city's character, and there are numerous *bike paths*.

The Amsterdam Stock Exchange is considered the oldest "modern" securities market stock exchange in the world. As the commercial capital of the Netherlands and one of the top financial centres in Europe, Amsterdam is considered an *alpha-world city* by the Globalization and World Cities (GaWC) study group. The city is also the cultural capital of the Netherlands.^[18] Many large Dutch institutions have their headquarters in the city, including: the Philips conglomerate, AkzoNobel, Booking.com, TomTom, and ING.^[19] Moreover, many of the world's largest companies are based in Amsterdam or have established their European headquarters in the city, such as leading technology companies Uber, Netflix and Tesla.^[20] In 2012, Amsterdam was ranked the second best city to live in by the Economist Intelligence Unit (EIU)^[21] and 12th globally on quality of living for environment and infrastructure by Mercer.^[22] The city was ranked 4th place globally as top tech hub in the Savills Tech Cities 2019 report (2nd in Europe),^[23] and 3rd in innovation by Australian innovation agency 2thinknow in their Innovation Cities Index 2009.^[24] The Port of Amsterdam is the fifth largest in Europe.^[25] The KLM hub and Amsterdam's main airport: Schiphol, is the Netherlands' busiest airport as well as the third busiest in Europe and 11th busiest airport in the world.^[26] The Dutch capital is considered one of the most multicultural cities in the world, with at least 177 nationalities represented.^[27]

Country Netherlands

City Hall Stopera

Boroughs Eight districts [show]

Government^[1]

- Body** Municipal council
- Mayor** Femke Halsema (GL)

Area^{[2][3]}

- Municipality** 219.32 km² (84.68 sq mi)
- Land** 165.76 km² (64.00 sq mi)
- Water** 53.56 km² (20.68 sq mi)
- Randstad** 3,043 km² (1,175 sq mi)

Elevation^[4] −2 m (−7 ft)

Population (municipality and Randstad, 2019; urban area, 2018; metro region, 2014)^{[5][3][6][7][8][9]}

- Municipality** 872,680
- Density** 5,206/km² (13,480/sq mi)
- Urban** 1,380,872
- Metro region** 2,431,000



Location of Amsterdam in the Netherlands

☒ Show map of Netherlands
☐ Show map of Europe
☐ Show all

Coordinates: 52°22′N 4°54′E﻿ / ﻿

Country	 Netherlands
Province	 North Holland
City Hall	Stopera
Boroughs	Eight districts [show]
Government ^[1]	
 • Body	Municipal council
 • Mayor	Femke Halsema (GL)
Area ^{[2][3]}	
 • Municipality	219.32 km ² (84.68 sq mi)
 • Land	165.76 km ² (64.00 sq mi)
 • Water	53.56 km ² (20.68 sq mi)
 • Randstad	3,043 km ² (1,175 sq mi)
Elevation ^[4]	−2 m (−7 ft)
Population (municipality and Randstad, 2019; urban area, 2018; metro region, 2014) ^{[5][3][6][7][8][9]}	
 • Municipality	872,680
 • Density	5,206/km ² (13,480/sq mi)
 • Urban	1,380,872
 • Metro region	2,431,000

```
<div class="shortdescription nomobile noexcerpt noprint searchaux" style="display:none">Capital city and municipality in North Holland, Netherlands</div>
<table class="infobox geography vcard" style="width:22em; width:23em">
  <tbody>
    <tr>...</tr>
    <tr>...</tr>
    <tr class="mergedtoprow">...</tr>
    <tr class="mergedtoprow">...</tr>
    <tr class="mergedtoprow">...</tr>
    <tr class="mergedrow">...</tr>
    <tr class="mergedtoprow">...</tr>
    <tr class="mergedtoprow">...</tr>
    <tr class="mergedbottomrow">...</tr>
    <tr class="mergedtoprow">
      <th scope="row">
        <a href="/wiki/Country" title="Country">Country</a>
      </th>
      <td>
        <span class="flagicon">...</span>
        <a href="/wiki/Netherlands" title="Netherlands">Netherlands</a> = $0
      </td>
    </tr>
    <tr class="mergedrow">...</tr>
    <tr class="mergedtoprow">...</tr>
    <tr class="mergedrow">...</tr>
```

```
>>> import re
>>> my_reg_exp = '<td><a [^>]+>(.*?)</a></td>\'
>>> line='<td><a href="/wiki/Netherlands" title="Netherlands">Netherlands</a></td>\'
>>> print re.findall(my_reg_exp,line)
['Netherlands']
>>>
>>> line='<td><a href="/wiki/Spain" title="Spain">Spain</a></td>\'
>>> print re.findall(my_reg_exp,line)
['Spain']
>>>
```

REGEX

Outros termos possuem uma estrutura muito peculiar

- Números de telefone
- CPF
- E-mails
- URLs
- etc.

Também é utilizado para capturar relações entre entidades

[PER], [POSITION] of [ORG]

[ORG] (named, appointed,...) [PER] Prep [POSITION]

- o *Nokia has appointed Rajeev Suri as President*

[ORG] headquarters in [LOC]

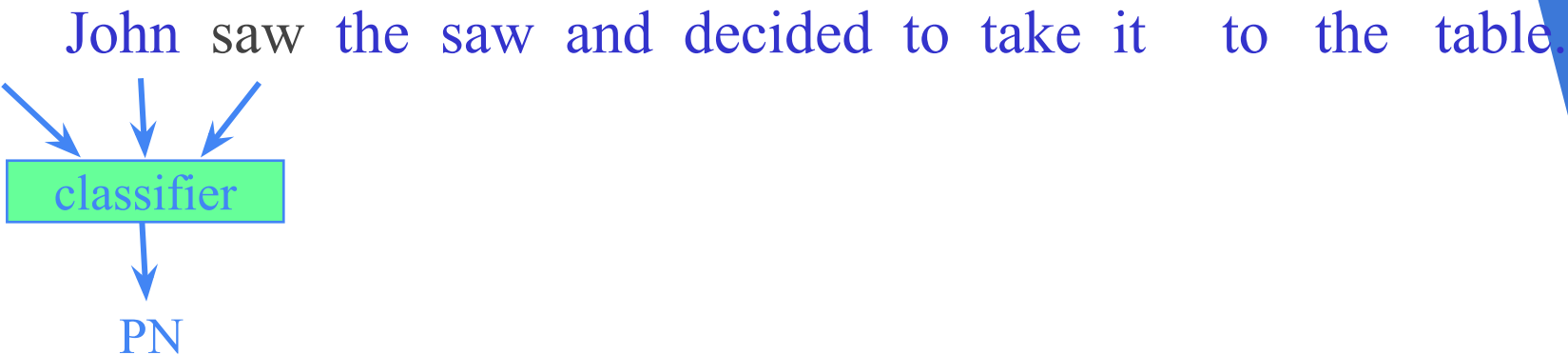
- o *NATO headquarters in Brussels*

[ORG][LOC] (division, branch, headquarters...)

- o *KFOR Kosovo headquarters*

Sequence Labeling como um problema de Classificação

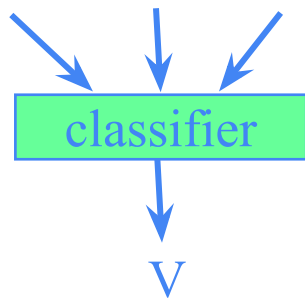
- Classificar cada token **independentemente**, mas utilizar como atributos as informações sobre os tokens na **vizinhança** (*sliding window*).



Sequence Labeling como um problema de Classificação

- Classificar cada token **independentemente**, mas utilizar como atributos as informações sobre os tokens na **vizinhança** (*sliding window*).

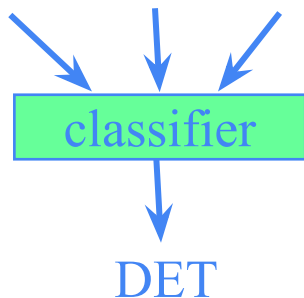
John saw the saw and decided to take it to the table.



Sequence Labeling como um problema de Classificação

- Classificar cada token **independentemente**, mas utilizar como atributos as informações sobre os tokens na **vizinhança** (*sliding window*).

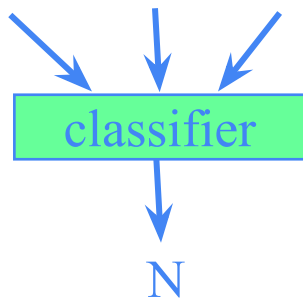
John saw the saw and decided to take it to the table.



Sequence Labeling como um problema de Classificação

- Classificar cada token **independentemente**, mas utilizar como atributos as informações sobre os tokens na **vizinhança** (*sliding window*).

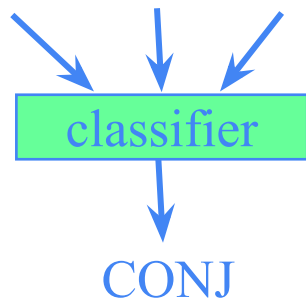
John saw the saw and decided to take it to the table.



Sequence Labeling como um problema de Classificação

- Classificar cada token **independentemente**, mas utilizar como atributos as informações sobre os tokens na **vizinhança** (*sliding window*).

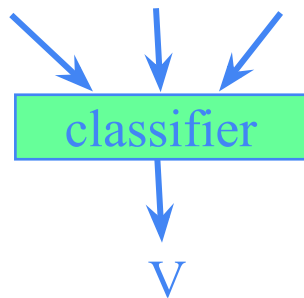
John saw the saw and decided to take it to the table.



Sequence Labeling como um problema de Classificação

- Classificar cada token **independentemente**, mas utilizar como atributos as informações sobre os tokens na **vizinhança** (*sliding window*).

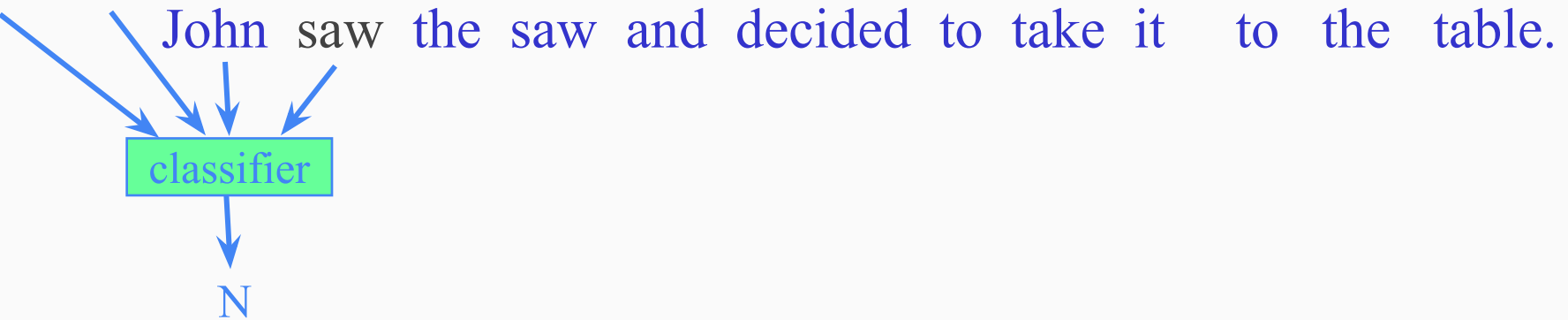
John saw the saw and decided to take it to the table.



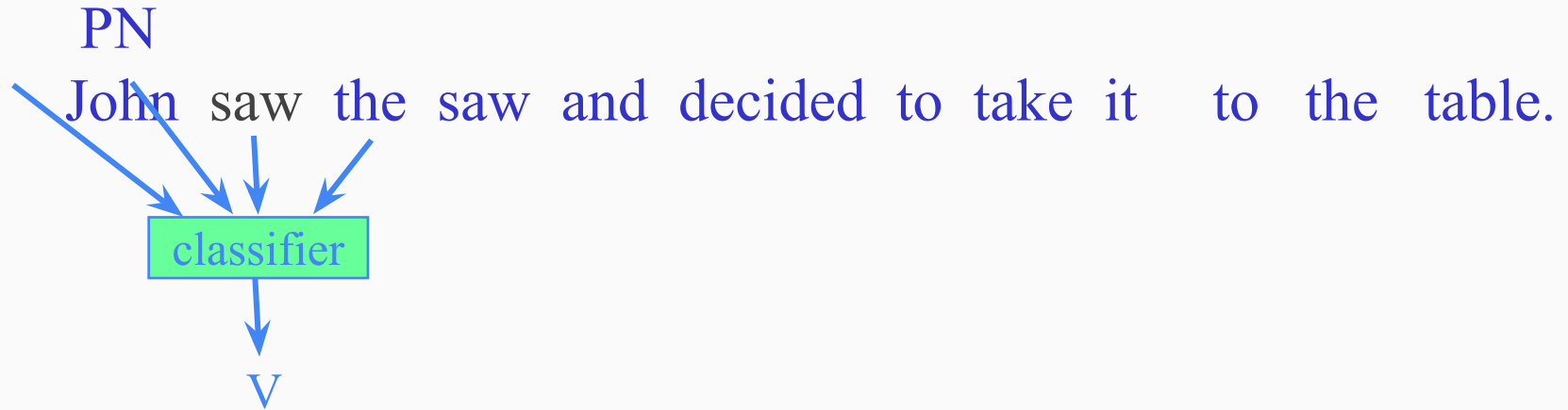
Sequence Labeling como um problema de Classificação

- Utilizar **outputs** (vizinhança) como **inputs**
- Seguir nas duas direções para ter melhor contexto

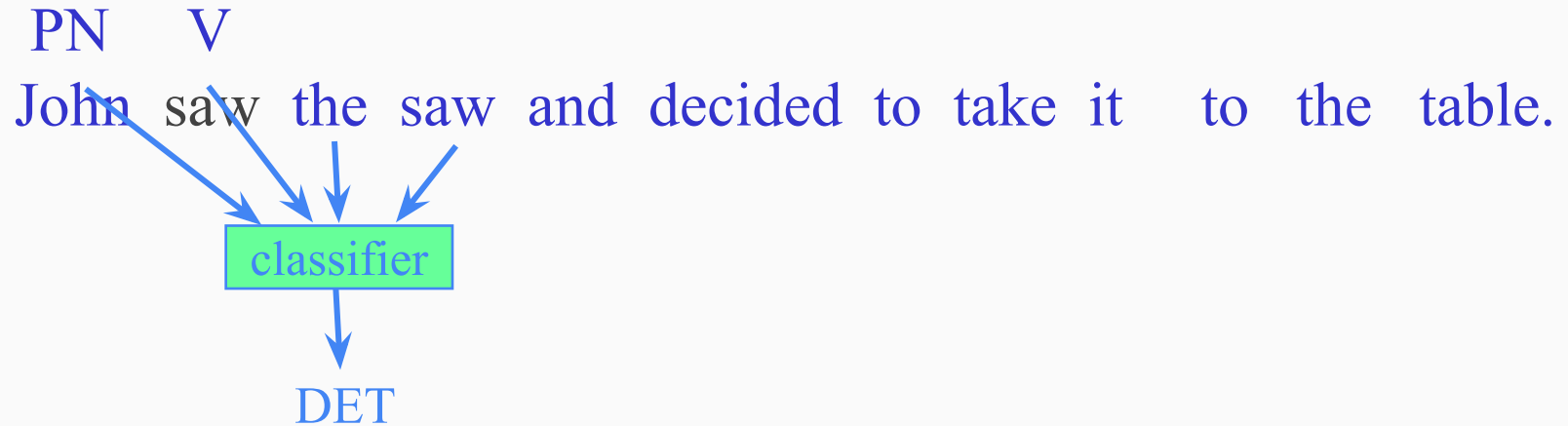
Forward Classification



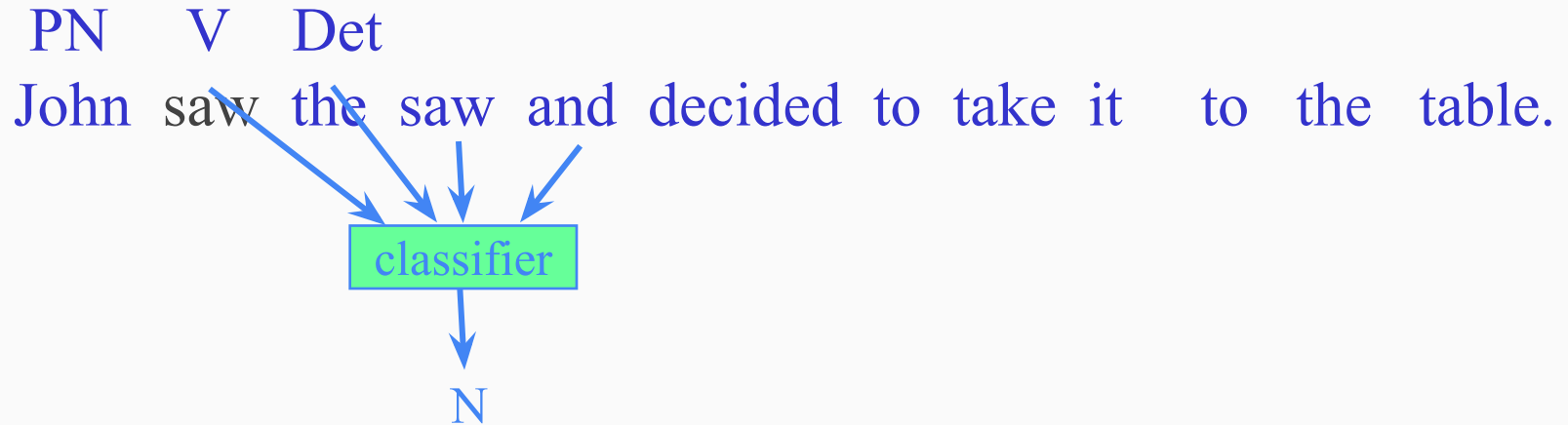
Forward Classification



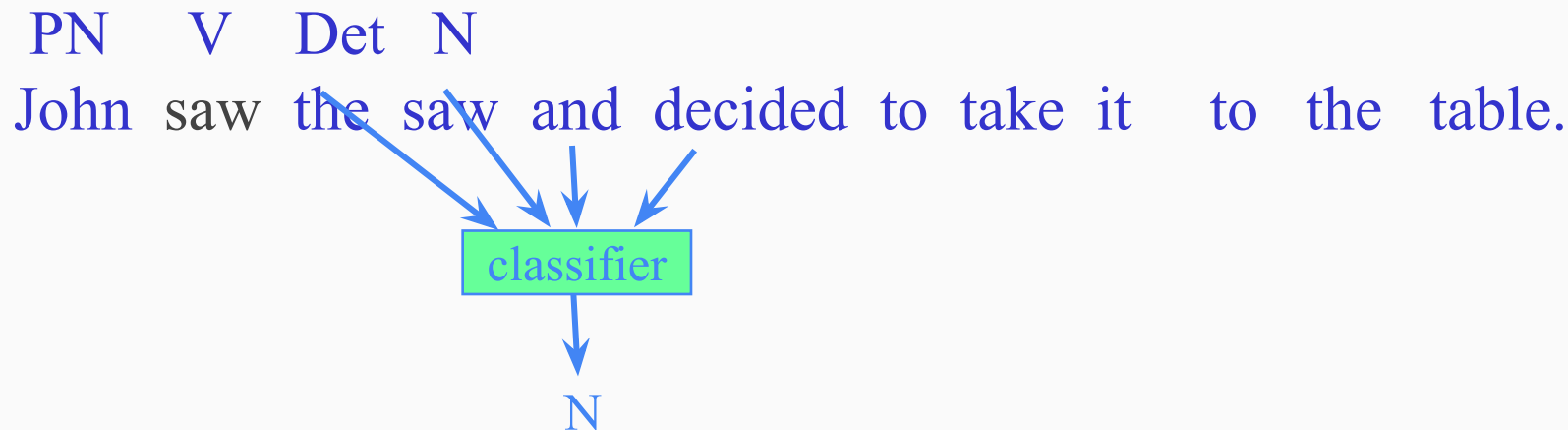
Forward Classification



Forward Classification

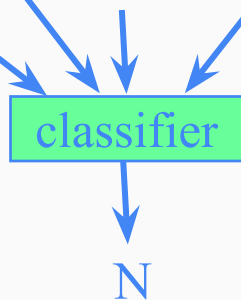


Forward Classification



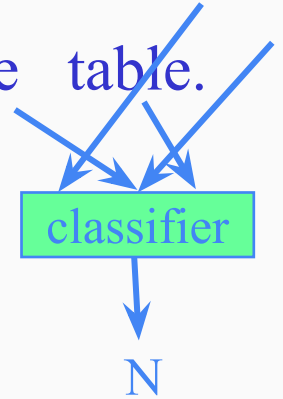
Forward Classification

PN V Det N Conj V Part V Pro Prep Det N
John saw the saw and decided to take it to the table.



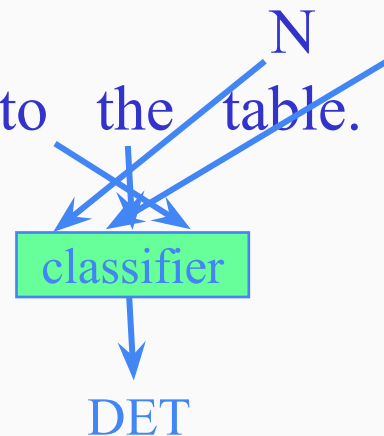
Backward Classification

John saw the saw and decided to take it to the table.

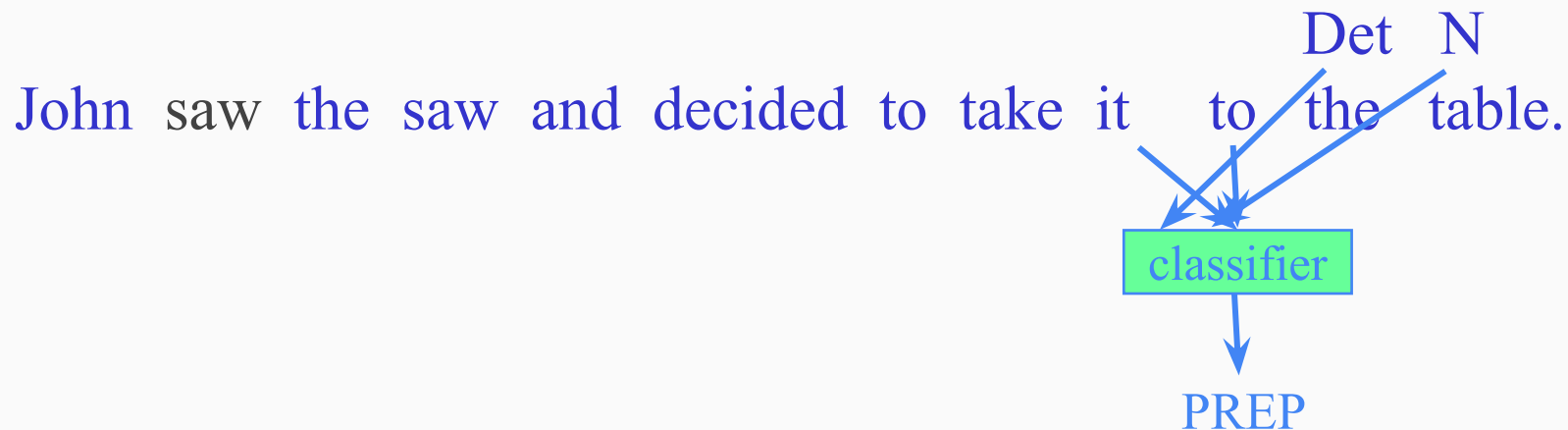


Backward Classification

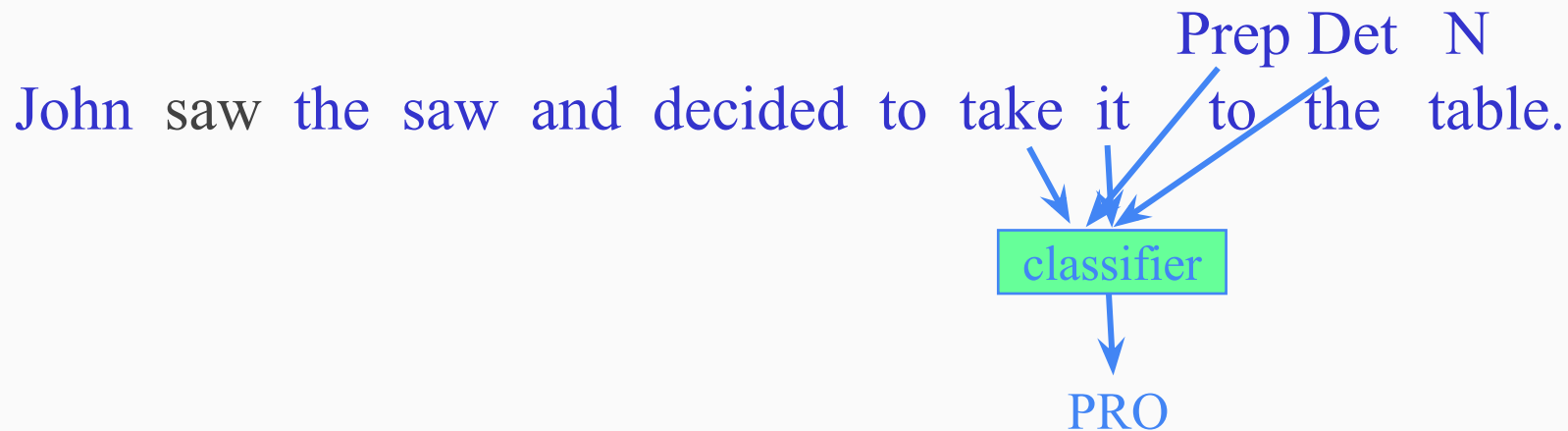
John saw the saw and decided to take it to the table.



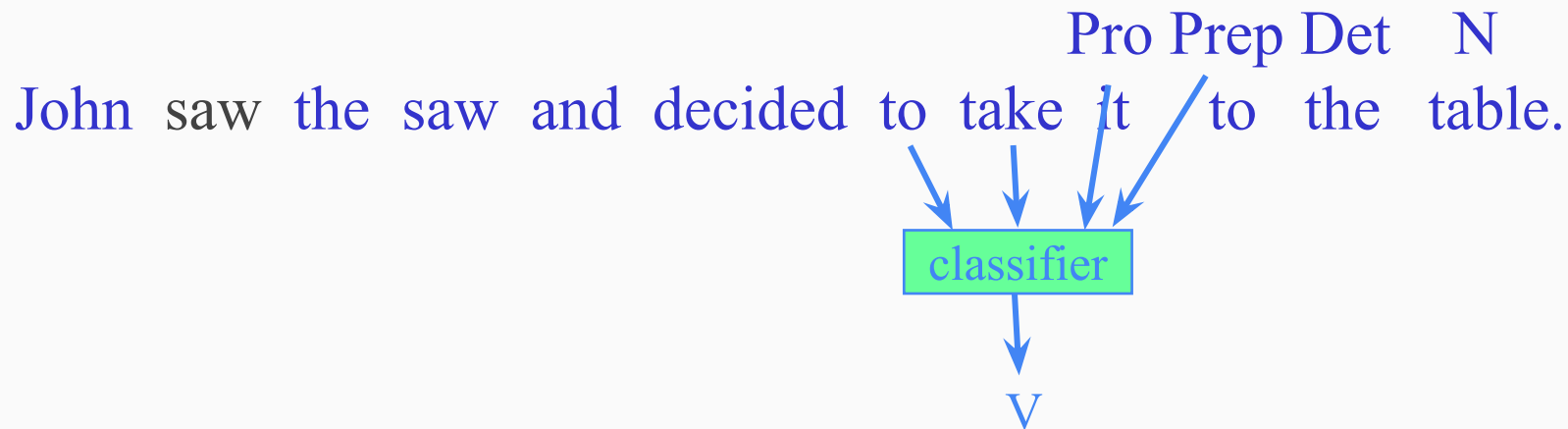
Backward Classification



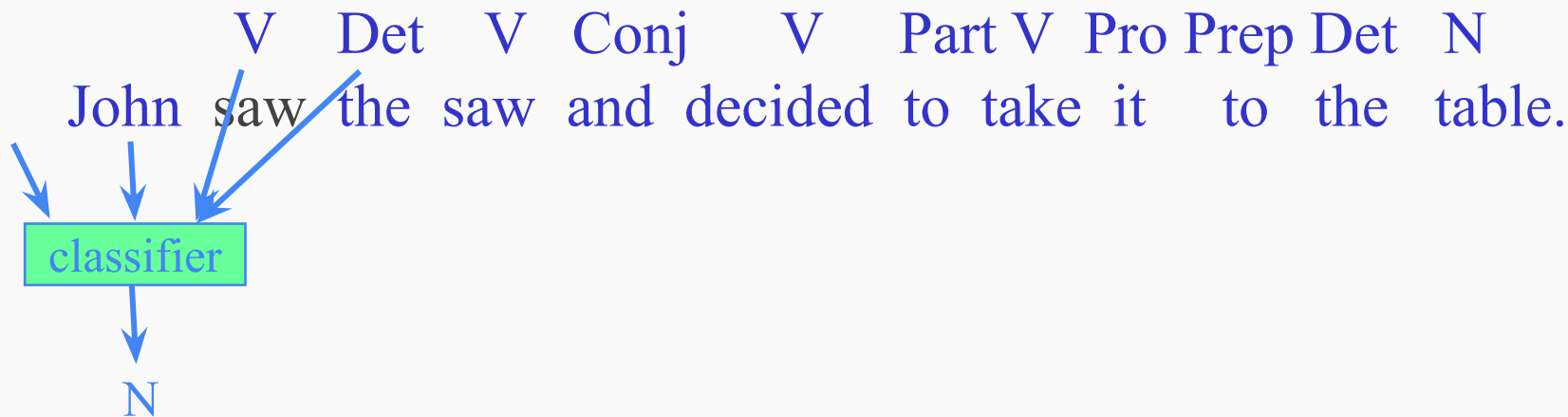
Backward Classification



Backward Classification



Backward Classification



Forward+Backward Classification

FORWARD	PN	V	Det	N	Conj	V	Part	V	Pro	Prep	Det	N
	John	saw	the	saw	and	decided	to	take	it	to	the	table.
BACKWARD		V	Det	V	Conj	V	Part	V	Pro	Prep	Det	N
	John	saw	the	saw	and	decided	to	take	it	to	the	table.

Features for Token Classification

- **Token Features**: características do próprio token.
- **Local Features**: características do contexto local (proximidades) do token atual.
- **Global Features**: características sobre a ocorrência do token em outros pontos do documento.
- **Gazetteer Features**: características extraídas das ocorrências do token em outras bases de dados

Sequence Labelling como classificação

Limitações

- Integração de dados/labels não é tão simples
- Dificuldade de propagar incertezas e determinar, de forma conjunta, qual a classe mais provável de cada token

Modelos probabilísticos

- Textos livres e semi-estruturados.
- Verifica a ocorrência de padrões em sequência no texto de entrada.
- Assume-se que a probabilidade de se visitar um estado depende do estado que foi visitado anteriormente.
- Maximiza a probabilidade de acerto para o conjunto todo de padrões.
- HMM, CRFs

Texto de Entrada

R. C. Schank and R. P. Abelson,
1977, Scripts, Plans, Goals and
Understanding, Lawrence Erlbaum
Associates, New Jersey

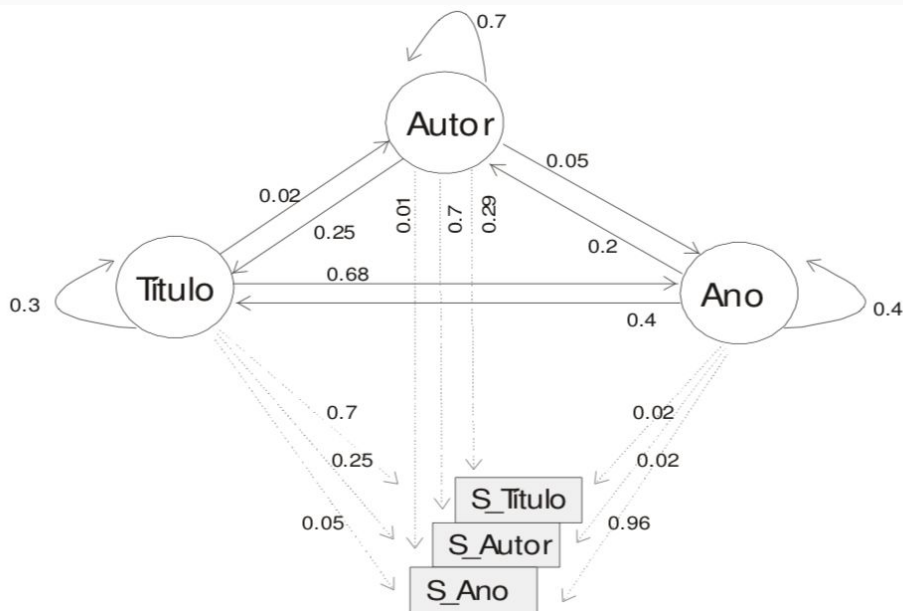
Sistema de EI

Formulário de Saída

Autor: R. C. Schank and R. P. Abelson
Ano: 1977

Título: Scripts, Plans, Goals and
Understanding

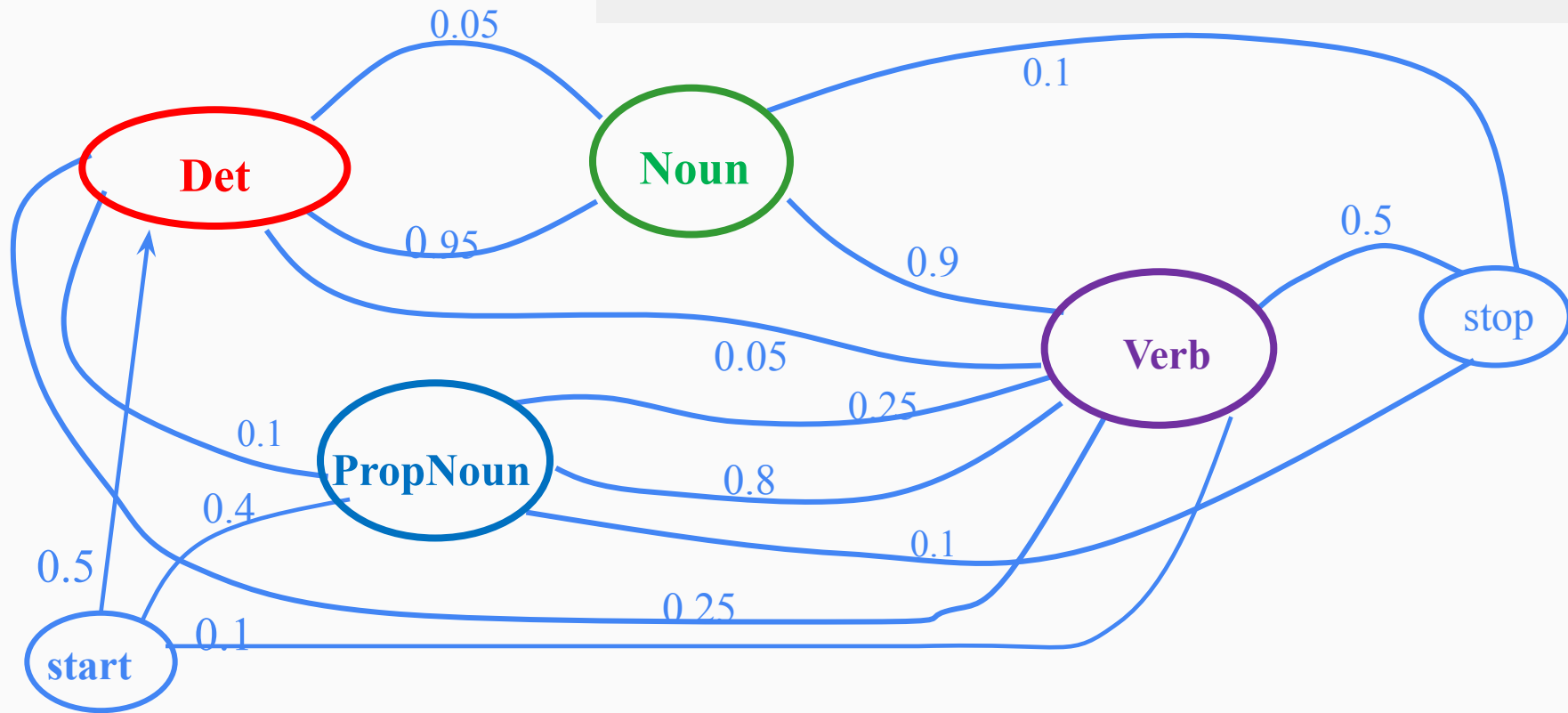
Páginas:
Editora: Lawrence Erlbaum Associates
Local: New Jersey



HMM para POS

$P(\text{PropNoun Verb Det Noun}) =$

$$0.4 * 0.8 * 0.25 * 0.95 * 0.1 = 0.0076$$



Hidden Markov Model

- Modelo gerador probabilístico para **sequências**.
- Supõe um conjunto subjacente de estados ocultos (não observados) nos quais o modelo pode estar (por exemplo, classes gramaticais).
- Assume transições probabilísticas entre estados ao longo do tempo (por exemplo, transição de um POS para outro POS conforme a sequência é gerada).
- Supõe uma geração probabilística de tokens de estados (por exemplo, palavras geradas para cada POS).

Conditional Random Fields

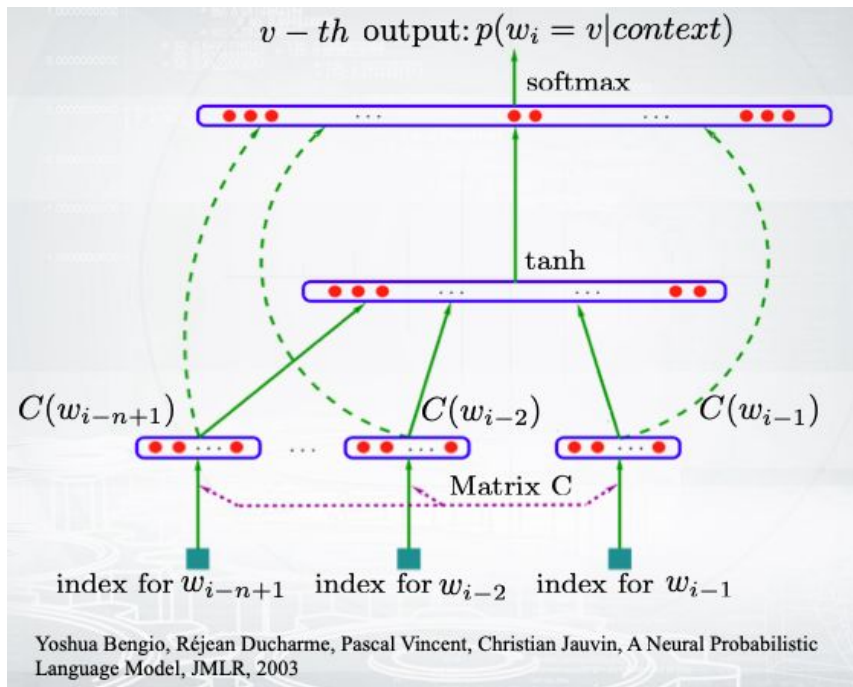
- **Conditional Random Fields** (CRFs) são modelos discriminativos especificamente projetados e treinados para rotulagem de sequência.
- Em geral, têm uma precisão superior em várias tarefas de rotulagem de sequência.
 - Noun phrase chunking
 - Named entity recognition
 - Semantic role labeling
- Modelos são mais complexos do que HMM, e mais custosos para treinar

DEMO

https://github.com/ufrpe-ensino/workshop-extracao-informacao/blob/main/notebooks/01_ExtracaoInformacao_CRF.ipynb

Neural Networks

Considerado estado da arte em sequence labelling



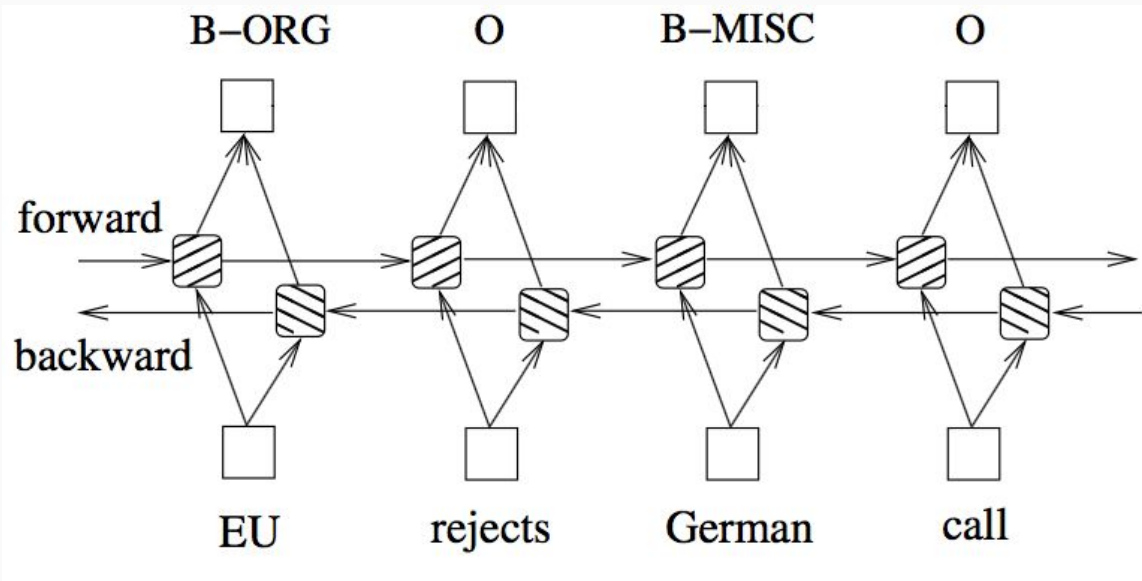
LSTM

LSTM - Long Short-Term Memory - é um tipo de Rede Neural Recorrente, com uma estrutura computacional mais complexa, que tem tido sucesso na resolução de tarefas sequenciais [Tai, 2015];

- Uma LSTM permite manter, alterar ou descartar informações anteriores para relacionar com uma informação atual;

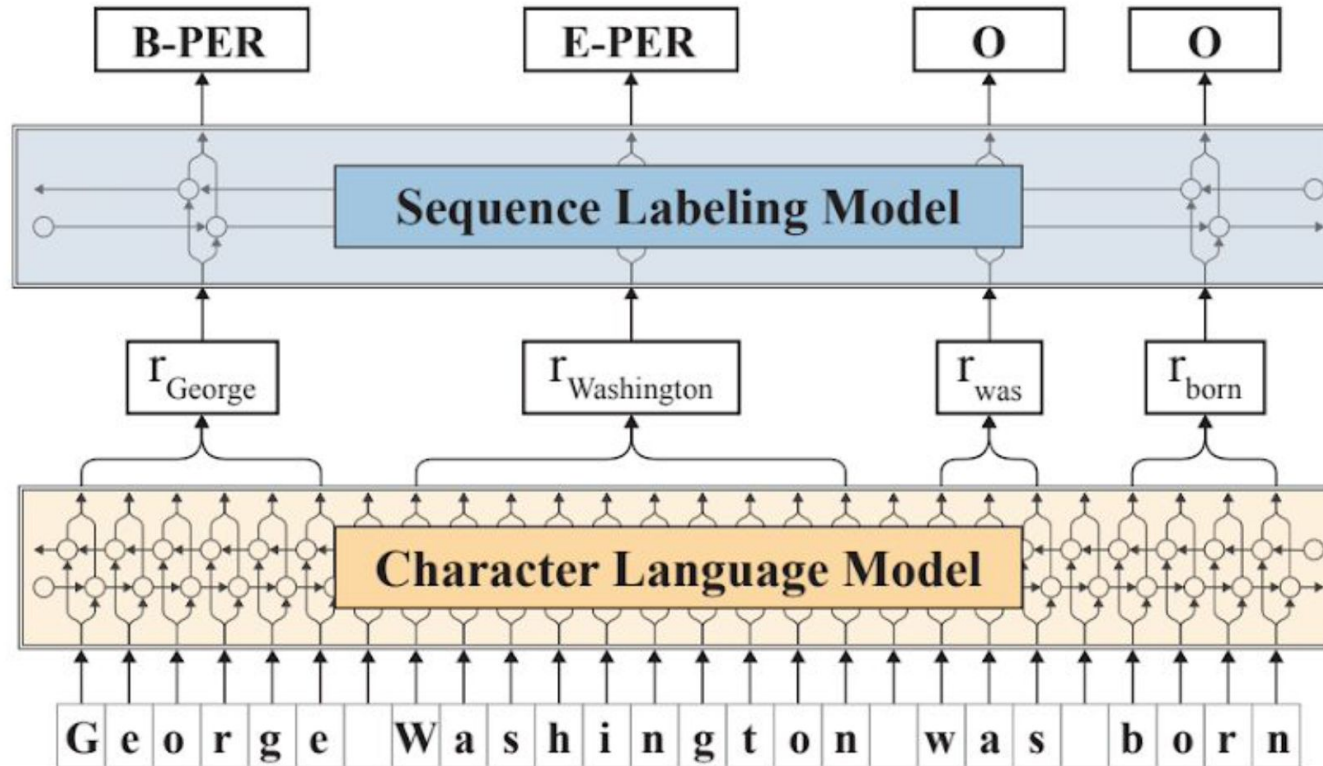
Há uma variação das redes LSTM, que são as Bidirectional LSTM (Bi-LSTM);

- As redes Bi-LSTM consistem de duas LSTM que funcionam em paralelo;



<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

FLAIR



Frameworks para NER

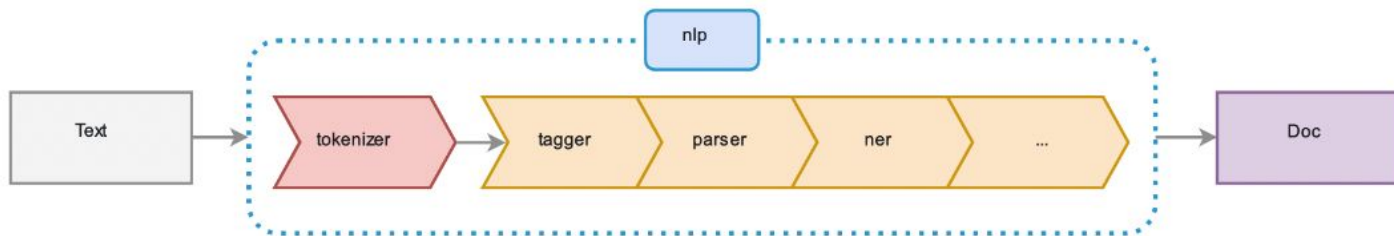
spaCy

flair

Spacy

*Slides adaptados do curso
Advanced NLP with Spacy*

Spacy



Edit the code & try spaCy

spaCy v3.4 · Python 3 · via Binder

```
# pip install -U spacy
# python -m spacy download en_core_web_sm
import spacy

# Load English tokenizer, tagger, parser and NER
nlp = spacy.load("en_core_web_sm")

# Process whole documents
text = ("When Sebastian Thrun started working on self-driving cars at "
        "Google in 2007, few people outside of the company took him "
        "seriously. "I can tell you very senior CEOs of major American "
        "car companies would shake my hand and turn away because I wasn't "
        "worth talking to," said Thrun, in an interview with Recode earlier "
        "this week.")
doc = nlp(text)

# Analyze syntax
print("Noun phrases:", [chunk.text for chunk in doc.noun_chunks])
print("Verbs:", [token.lemma_ for token in doc if token.pos_ == "VERB"])

# Find named entities, phrases and concepts
for entity in doc.ents:
    print(entity.text, entity.label_)
```

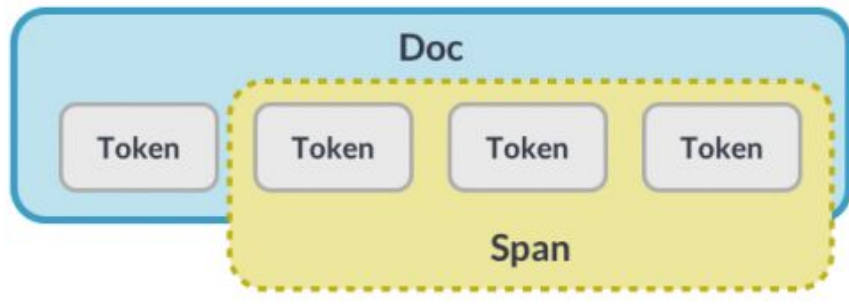
RUN

Features

- ✓ Support for **66+ languages**
- ✓ **76 trained pipelines** for 23 languages
- ✓ Multi-task learning with pretrained **transformers** like BERT
- ✓ Pretrained **word vectors**
- ✓ State-of-the-art speed
- ✓ Production-ready **training system**
- ✓ Linguistically-motivated **tokenization**
- ✓ Components for **named entity** recognition, part-of-speech tagging, dependency parsing, sentence segmentation, **text classification**, lemmatization, morphological analysis, entity linking and more
- ✓ Easily extensible with **custom components** and attributes
- ✓ Support for custom models in **PyTorch**, **TensorFlow** and other frameworks
- ✓ Built in **visualizers** for syntax and NER
- ✓ Easy **model packaging**, deployment and workflow management
- ✓ Robust, rigorously evaluated accuracy

Open Source
Foco em
desempenho e uso
industrial

Doc object



```
# Created by processing a string of text with the nlp object
doc = nlp("Hello world!")
# Iterate over tokens in a Doc
for token in doc:
    print(token.text)
```

```
Hello
world
!
```

Atributos Léxicos

```
doc = nlp("It costs $5.")
print('Index:  ', [token.i for token in doc])
print('Text:    ', [token.text for token in doc])
print('is_alpha:', [token.is_alpha for token in doc])
print('is_punct:', [token.is_punct for token in doc])
print('like_num:', [token.like_num for token in doc])
```

```
Index:  [0, 1, 2, 3, 4]
Text:    ['It', 'costs', '$', '5', '.']
is_alpha: [True, True, False, False, False]
is_punct: [False, False, False, False, True]
like_num: [False, False, False, True, False]
```

DEMOS!