# Stanford·CS124 | From Languages to Information (2021)

## CS124(2021)· 课程资料包 @ShowMeAI

**视频**
中英双语字幕

**课件**
一键打包下载

**笔记**
官方笔记翻译

**代码**
作业项目解析

视频 ·B 站 [ 扫码或点击链接 ]

*https://www.bilibili.com/video/BV1YA411w7ym*

课件 & 代码 · 博客 [ 扫码或点击链接 ]

*http://blog.showmeai.tech/cs124/*

文本处理
信息检索
编辑距离
神经嵌入
序列标注
PageRank
倒排索引
对话系统
推荐系统
社交网络分析
协同过滤

Awesome AI Courses Notes Cheatsheets是 **ShowMeAI** 资料库的分支系列，覆盖最具知名度的 **TOP20+** 门 AI 课程，旨在为读者和学习者提供一整套高品质中文学习笔记和速查表。

**点击**课程名称，跳转至课程**资料包**页面，**一键下载**课程全部资料！

| 机器学习 | 深度学习 | 自然语言处理 | 计算机视觉 |
|---|---|---|---|
| Stanford · CS229 | Stanford · CS230 | Stanford · CS224n | Stanford · CS231n |

### # Awesome AI Courses Notes Cheatsheets· 持续更新中

| 知识图谱 | 图机器学习 | 深度强化学习 | 自动驾驶 |
|---|---|---|---|
| Stanford · CS520 | Stanford · CS224W | UCBerkeley · CS285 | MIT · 6.S094 |

**微信公众号**

资料下载方式 2：扫码点击底部菜单栏

称为 **AI 内容创作者？** 回复 [ 添砖加瓦 ]

# Sequence Labeling for Part of Speech and Named Entities

## Part of Speech Tagging

# Parts of Speech

From the earliest linguistic traditions (Yaska and Panini 5th C. BCE, Aristotle 4th C. BCE), the idea that words can be classified into grammatical categories

- part of speech, word classes, POS, POS tags

8 parts of speech attributed to Dionysius Thrax of Alexandria (c. 1st C. BCE):

- noun, verb, pronoun, preposition, adverb, conjunction, participle, article

- These categories are relevant for NLP today.

# Two classes of words: Open vs. Closed

## Closed class words
- Relatively fixed membership
- Usually **function** words: short, frequent words with grammatical function
  - determiners: *a, an, the*
  - pronouns: *she, he, I*
  - prepositions: *on, under, over, near, by, …*

## Open class words
- Usually **content** words: Nouns, Verbs, Adjectives, Adverbs
  - Plus interjections: **oh, ouch, uh-huh, yes, hello**
- New nouns and verbs like *iPhone* or *to fax*

**Open class ("content") words**

**Nouns**

Proper

*Janet*
*Italy*

Common

*cat, cats*
*mango*

**Verbs**

Main

*eat*
*went*

Auxiliary

*can*
*had*

Adjectives *old green tasty*

Adverbs *slowly yesterday*

Interjections *Ow hello*

Numbers

*122,312*
*one*

*... more*

**Closed class ("function")**

Determiners *the some*

Conjunctions *and or*

Pronouns *they its*

Prepositions *to with*

Particles *off up*

*... more*

# Part-of-Speech Tagging
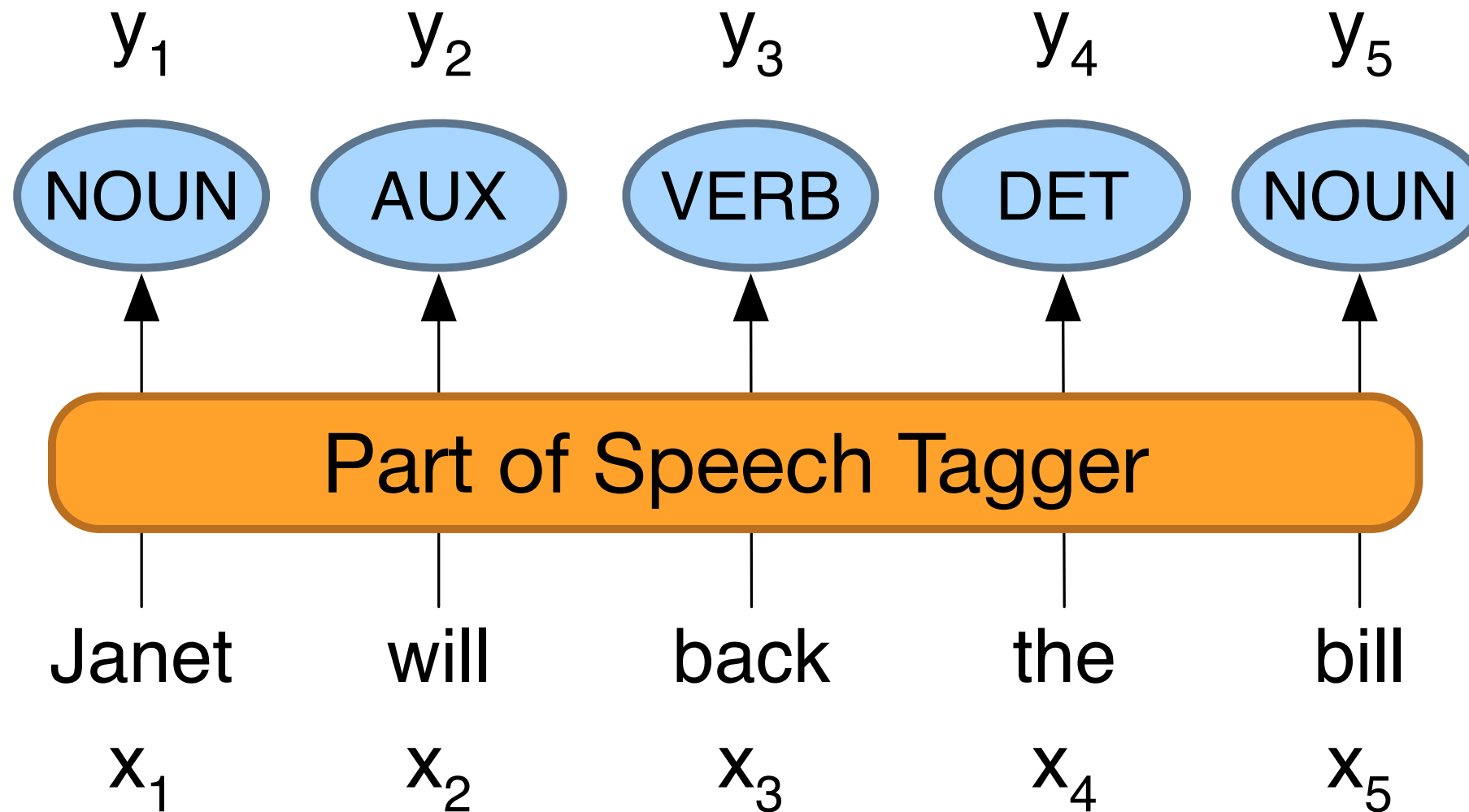
Assigning a part-of-speech to each word in a text.

Words often have more than one POS.

**book**:

- VERB: (**Book** *that flight*)
- NOUN: (*Hand me that* **book**).

# Part-of-Speech Tagging

Map from sequence $x_1,...,x_n$ of words to $y_1,...,y_n$ of POS tags

# "Universal Dependencies" Tagset

Nivre et al. 2016

| | Tag | Description | Example |
|---|---|---|---|
| **Open Class** | **ADJ** | Adjective: noun modifiers describing properties | *red, young, awesome* |
| | **ADV** | Adverb: verb modifiers of time, place, manner | *very, slowly, home, yesterday* |
| | **NOUN** | words for persons, places, things, etc. | *algorithm, cat, mango, beauty* |
| | **VERB** | words for actions and processes | *draw, provide, go* |
| | **PROPN** | Proper noun: name of a person, organization, place, etc.. | *Regina, IBM, Colorado* |
| | **INTJ** | Interjection: exclamation, greeting, yes/no response, etc. | *oh, um, yes, hello* |
| **Closed Class Words** | **ADP** | Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation | *in, on, by under* |
| | **AUX** | Auxiliary: helping verb marking tense, aspect, mood, etc., | *can, may, should, are* |
| | **CCONJ** | Coordinating Conjunction: joins two phrases/clauses | *and, or, but* |
| | **DET** | Determiner: marks noun phrase properties | *a, an, the, this* |
| | **NUM** | Numeral | *one, two, first, second* |
| | **PART** | Particle: a preposition-like form used together with a verb | *up, down, on, off, in, out, at, by* |
| | **PRON** | Pronoun: a shorthand for referring to an entity or event | *she, who, I, others* |
| | **SCONJ** | Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement | *that, which* |
| **Other** | **PUNCT** | Punctuation | ; , () |
| | **SYM** | Symbols like $ or emoji | $, % |
| | **X** | Other | asdf, qwfg |

# Sample "Tagged" English sentences

There/PRO were/VERB 70/NUM children/NOUN there/ADV ./PUNC

Preliminary/ADJ findings/NOUN were/AUX reported/VERB in/ADP today/NOUN 's/PART New/PROPN England/PROPN Journal/PROPN of/ADP Medicine/PROPN

# Why Part of Speech Tagging?

- Can be useful for other NLP tasks
  - Parsing: POS tagging can improve syntactic parsing
  - MT: reordering of adjectives and nouns (say from Spanish to English)
  - Sentiment or affective tasks: may want to distinguish adjectives or other POS
  - Text-to-speech (how do we pronounce "lead" or "object"?)
- Or linguistic or language-analytic computational tasks
  - Need to control for POS when studying linguistic change like creation of new words, or meaning shift
  - Or control for POS in measuring meaning similarity or difference

# How difficult is POS tagging in English?

Roughly 15% of word types are ambiguous

- Hence 85% of word types are unambiguous
- *Janet* is always PROPN, *hesitantly* is always ADV

But those 15% tend to be very common.

So ~60% of word tokens are ambiguous

E.g., *back*

earnings growth took a back/ADJ seat
a small building in the back/NOUN
a clear majority of senators back/VERB the bill
enable the country to buy back/PART debt
I was twenty-one back/ADV then

# POS tagging performance in English

How many tags are correct? (Tag accuracy)

- About 97%
  - Hasn't changed in the last 10+ years
  - HMMs, CRFs, BERT perform similarly .
  - Human accuracy about the same

But baseline is 92%!

- Baseline is performance of stupidest possible method
  - "Most frequent class baseline" is an important baseline for many tasks
    - Tag every word with its most frequent tag
    - (and tag unknown words as nouns)
- Partly easy because
  - Many words are unambiguous

# Sources of information for POS tagging

Janet **will** back the **bill**
  AUX/NOUN/VERB?        NOUN/VERB?

Prior probabilities of word/tag
- "will" is usually an AUX

Identity of neighboring words
- "the" means the next word is probably not a verb

Morphology and wordshape:
- Prefixes          unable:          un- $\rightarrow$ ADJ
- Suffixes          importantly:     -ly $\rightarrow$ ADJ
- Capitalization    Janet:           CAP $\rightarrow$ PROPN

# Standard algorithms for POS tagging

Supervised Machine Learning Algorithms:

- Hidden Markov Models

- Conditional Random Fields (CRF)/ Maximum Entropy Markov Models (MEMM)

- Neural sequence models (RNNs or Transformers)

- Large Language Models (like BERT), finetuned

All required a hand-labeled training set, all about equal performance (97% on English)

All make use of information sources we discussed

- Via human created features: HMMs and CRFs
- Via representation learning:  Neural LMs

# Sequence Labeling for Part of Speech and Named Entities

## Part of Speech Tagging

# Sequence Labeling for Part of Speech and Named Entities

## Named Entity Recognition (NER)

# Named Entities

- **Named entity**, in its core usage, means anything that can be referred to with a proper name. Most common 4 tags:
  - PER (Person): "Marie Curie"
  - LOC (Location): "New York City"
  - ORG (Organization): "Stanford University"
  - GPE (Geo-Political Entity): "Boulder, Colorado"
- Often multi-word phrases
- But the term is also extended to things that aren't entities:
  - dates, times, prices

# Named Entity tagging

The task of named entity recognition (NER):

- find spans of text that constitute proper names

- tag the type of the entity.

# NER output

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

# Why NER?

Sentiment analysis: consumer's sentiment toward a particular company or person?

Question Answering: answer questions about an entity?

Information Extraction: Extracting facts about entities from text.

# Why NER is hard

1) Segmentation
   - In POS tagging, no segmentation problem since each word gets one tag.
   - In NER we have to find and segment the entities!

2) Type ambiguity

[PER Washington] was born into slavery on the farm of James Burroughs.
[ORG Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [LOC Washington] for what may well be his last state visit.
In June, [GPE Washington] passed a primary seatbelt law.

# BIO Tagging

How can we turn this structured problem into a sequence problem like POS tagging, with one label per word?

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago ] route.

# BIO Tagging

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago ] route.

| Words | BIO Label |
|---|---|
| Jane | B-PER |
| Villanueva | I-PER |
| of | O |
| United | B-ORG |
| Airlines | I-ORG |
| Holding | I-ORG |
| discussed | O |
| the | O |
| Chicago | B-LOC |
| route | O |
| . | O |

Now we have one tag per token!!!

# BIO Tagging

B: token that *begins* a span

I: tokens *inside* a span

O: tokens outside of any span

# of tags (where n is #entity types):

1 O tag,

*n* B tags,

*n* I tags

 total of *2n+1*

| Words | BIO Label |
|-------|-----------|
| Jane | B-PER |
| Villanueva | I-PER |
| of | O |
| United | B-ORG |
| Airlines | I-ORG |
| Holding | I-ORG |
| discussed | O |
| the | O |
| Chicago | B-LOC |
| route | O |
| . | O |

# BIO Tagging variants: IO and BIOES

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago ] route.

| Words | IO Label | BIO Label | BIOES Label |
|---|---|---|---|
| Jane | I-PER | B-PER | B-PER |
| Villanueva | I-PER | I-PER | E-PER |
| of | O | O | O |
| United | I-ORG | B-ORG | B-ORG |
| Airlines | I-ORG | I-ORG | I-ORG |
| Holding | I-ORG | I-ORG | E-ORG |
| discussed | O | O | O |
| the | O | O | O |
| Chicago | I-LOC | B-LOC | S-LOC |
| route | O | O | O |
| . | O | O | O |

# Standard algorithms for NER

Supervised Machine Learning given a human-labeled training set of text annotated with tags

- Hidden Markov Models

- Conditional Random Fields (CRF)/ Maximum Entropy Markov Models (MEMM)

- Neural sequence models (RNNs or Transformers)

- Large Language Models (like BERT), finetuned

# Sequence Labeling for Part of Speech and Named Entities

## Named Entity Recognition (NER)

# Stanford·CS124 | From Languages to Information (2021)

## CS124(2021)· 课程资料包 @ShowMeAI

Awesome AI Courses Notes Cheatsheets是 **ShowMeAI** 资料库的分支系列，覆盖最具知名度的 **TOP20+** 门 AI 课程，旨在为读者和学习者提供一整套高品质中文学习笔记和速查表。

**点击**课程名称，跳转至课程**资料包**页面，**一键下载**课程全部资料！

**视频**
中英双语字幕

**课件**
一键打包下载

**笔记**
官方笔记翻译

**代码**
作业项目解析

| 机器学习 | 深度学习 | 自然语言处理 | 计算机视觉 |
|---|---|---|---|
| Stanford · CS229 | Stanford · CS230 | Stanford · CS224n | Stanford · CS231n |

**# Awesome AI Courses Notes Cheatsheets· 持续更新中**

| 知识图谱 | 图机器学习 | 深度强化学习 | 自动驾驶 |
|---|---|---|---|
| Stanford · CS520 | Stanford · CS224W | UCBerkeley · CS285 | MIT · 6.S094 |

视频 ·B 站 [ 扫码或点击链接 ]

https://www.bilibili.com/video/BV1YA411w7ym

课件 & 代码 · 博客 [ 扫码或点击链接 ]

http://blog.showmeai.tech/cs124/

文本处理

编辑距离 神经嵌入 PageRank 倒排索引

信息检索 序列标注 对话系统

推荐系统 社交网络分析 协同过滤

微信公众号

资料下载方式 2：扫码点击底部菜单栏

称为 **AI 内容创作者？** 回复 [ 添砖加瓦 ]