Natural Language Processing

Week 6

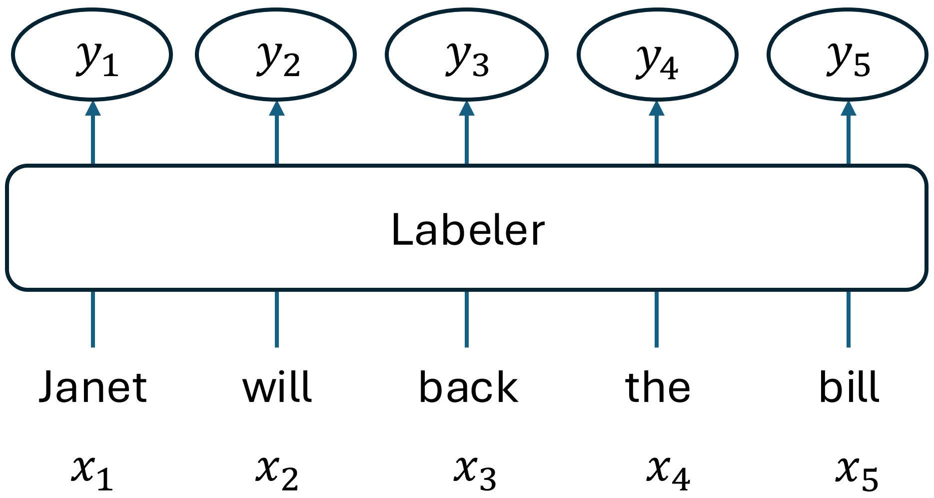
Agenda

- Project Introduction
- Token Classification
  - POS
  - NER
- Extractive Summarization

Token Classification

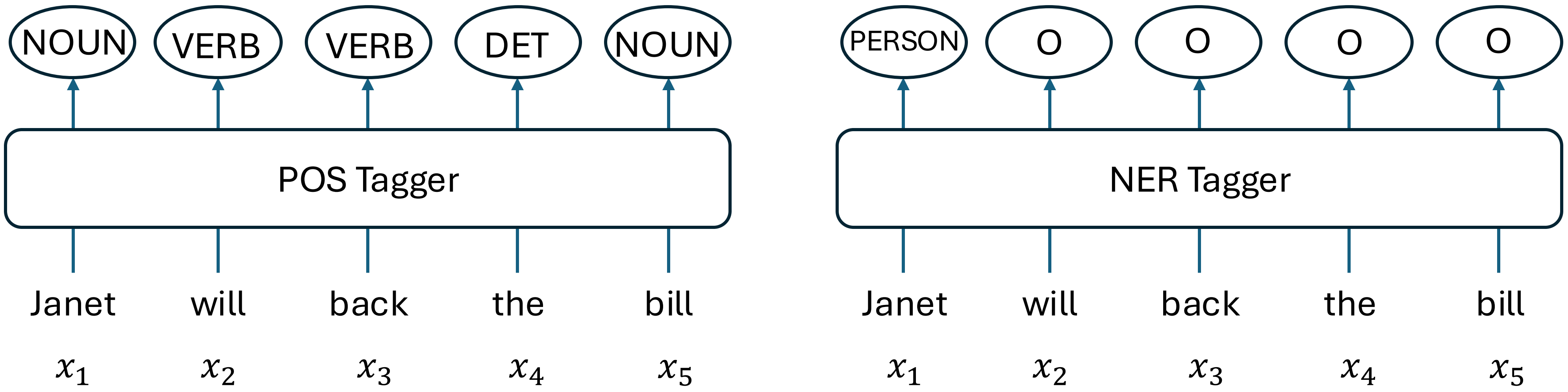
Token Classification

- Token classification is the task of assigning a categorical label for every element in sequence of data
  - Text data, biological sequences, time-series
- For NLP applications, the task is to **assign each word x_i in an input sequence a label y_i** (also called **sequence labeling**)



Token Classification

- **Parts-of-Speech (POS) tagging** and **Named Entity Recognition (NER)** are two of the most common forms of sequence modeling in NLP



Parts of Speech

Parts of Speech

- Parts of speech are formally defined based on their grammatical relationship with neighboring words or the morphological properties of their affixes

	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, 2026, 11:00, hundred*
	PART	Particle: a function word that must be associated with another word	*'s, not, (infinitive) to*
	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	*whether, because*
Other	**PUNCT**	Punctuation	; , ()
	SYM	Symbols like $ or emoji	$, %
	X	Other	asdf, qwfg

Speech and Language Processing, Jurafsky, 2023, p.163, Fig. 8-1

Parts of Speech

- **Open class**: category of parts of speech that is open to being expanded / updated / changed as language evolves
  - Nouns: bitcoin, iPhone, etc
  - We’ve updated the Merriam-Webster.com Dictionary with 690 New Words | Merriam-Webster
- **Closed class:** category of parts of speech that is generally fixed / doesn’t evolve
  - Prepositions: of, about, around

Parts of Speech (Penn Treebank)

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coord. conj.	*and, but, or*	NNP	proper noun, sing.	*IBM*	TO	infinitive to	*to*
CD	cardinal number	*one, two*	NNPS	proper noun, plu.	*Carolinas*	UH	interjection	*ah, oops*
DT	determiner	*a, the*	NNS	noun, plural	*llamas*	VB	verb base	*eat*
EX	existential 'there'	*there*	PDT	predeterminer	*all, both*	VBD	verb past tense	*ate*
FW	foreign word	*mea culpa*	POS	possessive ending	*'s*	VBG	verb gerund	*eating*
IN	preposition/ subordin-conj	*of, in, by*	PRP	personal pronoun	*I, you, he*	VBN	verb past partici- ple	*eaten*
JJ	adjective	*yellow*	PRP$	possess. pronoun	*your*	VBP	verb non-3sg-pr	*eat*
JJR	comparative adj	*bigger*	RB	adverb	*quickly*	VBZ	verb 3sg pres	*eats*
JJS	superlative adj	*wildest*	RBR	comparative adv	*faster*	WDT	wh-determ.	*which, that*
LS	list item marker	*1, 2, One*	RBS	superlatv. adv	*fastest*	WP	wh-pronoun	*what, who*
MD	modal	*can, should*	RP	particle	*up, off*	WP$	wh-possess.	*whose*
NN	sing or mass noun	*llama*	SYM	symbol	*+, %, &*	WRB	wh-adverb	*how, where*

Speech and Language Processing, Jurafsky, 2023, p.165, Fig. 8-2

Evolution of POS Tagging

- **Rule-Based Tagging (60s-80s):** manual tagging and rule-based systems utilizing dictionaries and manually engineering grammatical rules
- **Probabilistic Methods (80s-90s):** Hidden Markov Models (HMMs) tagged sequences based on word occurrence probabilities, leading to better ambiguity resolution than rule-based systems
- **Machine Learning Methods (2000s):** Conditional Random Fields (CRFs) and other ML algorithms utilized feature context and interactions
- **Deep Learning Methods (2010s):** Bidirectional RNNs and LSTMs enhanced handling of long-range dependencies, minimizing manual feature engineering.
- **Transformers (current):** Efficiency, multi-language improvements, etc.

Speech and Language Processing, Jurafsky, 2023, p.168

Task Performance

- Accuracy of part-of-speech tags is very high
- 97% accuracy across 15 languages from Universal Dependency treebank
- Accuracies for other treebanks are 97% no matter the algorithm (Hidden Markov Model, Conditional Random Field, BERT, etc)

Parts of Speech: How hard is the task?

- Most word types (unique words) are unambiguous (85-86%)
- Ambiguous tokens (instances of words), however, appear very often
- Most-frequent-tag baseline gives 92% accuracy (STOA is 97%)

		WSJ	Brown
Types:			
Unambiguous	(1 tag)	44,432 (**86%**)	45,799 (**85%**)
Ambiguous	(2+ tags)	7,025 (**14%**)	8,050 (**15%**)
Tokens:			
Unambiguous	(1 tag)	577,421 (**45%**)	384,349 (**33%**)
Ambiguous	(2+ tags)	711,780 (**55%**)	786,646 (**67%**)

earnings growth took a **back/JJ** seat

a small building in the **back/NN**

a clear majority of senators **back/VBP** the bill

Dave began to **back/VB** toward the door

enable the country to buy **back/RP** debt

I was twenty-one **back/RB** then

Speech and Language Processing, Jurafsky, 2023, p.166, Fig. 8.3

Parts of Speech: Applications

- POS tagging could be used for:
    - NER Models (creating features, used for post-processing)
    - Models for sentence segmentation (features)
    - Sentiment analysis (features)
    - Improving speech recognition
- Currently, neural networks and modern transformer architecture **has reduced the need to explicitly engineer features for NLP applications**
- However, there may still be places where POS is needed (environments that can't support transformers models)

Named Entity Recognition

Named Entity Recognition (NER)

- Named entities are words / phrases in text that refer to proper nouns
- NER is an NLP task in which the goal is to find **spans of text** that constitute proper names and **assign the correct entity type** to the span
- Most common are **people**, **places**, **organizations**, or **geo-political** entities
- Does not have to be an entity per se (could be time, or any other short span of text you deem as an entity (such as cost of a contractual agreement)

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

Speech and Language Processing, Jurafsky, 2023, p.167

NER Labels

- For NER, every word in a sequence is labeled
- **BIO/IOB2** – label token that begins a span with **B**, tokens that occur inside are labeled with **I**, tokens outside the spans are with **O**
- **IO** – loses indicator for the beginning of a span
- **BIOES/IOBES** – adds indicator for end of a span, and a single-token span

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

Speech and Language Processing, Jurafsky, 2023, p.168

Evaluation of NER

- NER is a highly **imbalanced, multi-class classification task**
- Precision, Recall, F-score as standard evaluation metrics
- NER has two dimensions of correctness: **type** and **span**
- A popular library call **seqeval** penalizes on both dimensions
- Other libraries exist that may not be as strict: nervaluate · PyPI

Evaluation of NER

tokens	true	underspan	overspan	type
Steve	B-PER	B-PER	B-PER	B-PER
Jobs	I-PER	O	I-PER	I-PER
and	O	O	O	O
Tim	B-PER	B-PER	B-PER	B-PER
Cook	I-PER	O	I-PER	I-PER
both	O	O	O	O
led	O	O	O	O
Apple	B-ORG	B-ORG	B-ORG	B-LOC
,	O	O	I-ORG	O
which	O	O	I-ORG	O
is	O	O	O	O
based	O	O	O	O
in	O	O	O	O
Cupertino	B-LOC	B-LOC	B-LOC	B-LOC
.	O	O	O	O

UNDER

Accuracy: 0.87

	precision	recall	f1-score	support
LOC	1.00	1.00	1.00	1
ORG	1.00	1.00	1.00	1
PER	0.00	0.00	0.00	2
micro avg	0.50	0.50	0.50	4
macro avg	0.67	0.67	0.67	4
weighted avg	0.50	0.50	0.50	4

OVER

Accuracy: 0.87

	precision	recall	f1-score	support
LOC	1.00	1.00	1.00	1
ORG	0.00	0.00	0.00	1
PER	1.00	1.00	1.00	2
micro avg	0.75	0.75	0.75	4
macro avg	0.67	0.67	0.67	4
weighted avg	0.75	0.75	0.75	4

TYPE

Accuracy: 0.93

	precision	recall	f1-score	support
LOC	0.50	1.00	0.67	1
ORG	0.00	0.00	0.00	1
PER	1.00	1.00	1.00	2
micro avg	0.75	0.75	0.75	4
macro avg	0.50	0.67	0.56	4
weighted avg	0.62	0.75	0.67	4

Span-Flexible Evaluation of NER

tokens	true	underspan	overspan	type
Steve	B-PER	B-PER	B-PER	B-PER
Jobs	I-PER	O	I-PER	I-PER
and	O	O	O	O
Tim	B-PER	B-PER	B-PER	B-PER
Cook	I-PER	O	I-PER	I-PER
both	O	O	O	O
led	O	O	O	O
Apple	B-ORG	B-ORG	B-ORG	B-LOC
,	O	O	I-ORG	O
which	O	O	I-ORG	O
is	O	O	O	O
based	O	O	O	O
in	O	O	O	O
Cupertino	B-LOC	B-LOC	B-LOC	B-LOC
.	O	O	O	O

UNDER

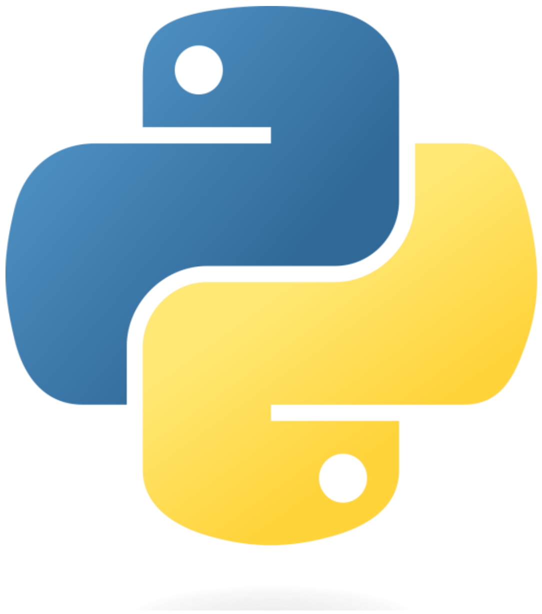
	precision	recall	f1
LOC	1.0	1.0	1.0
ORG	1.0	1.0	1.0
PER	1.0	1.0	1.0

OVER

	precision	recall	f1
LOC	1.0	1.0	1.0
ORG	1.0	1.0	1.0
PER	1.0	1.0	1.0

TYPE

	precision	recall	f1
LOC	1.0	1.0	1.0
ORG	0.0	0.0	0.0
PER	1.0	1.0	1.0



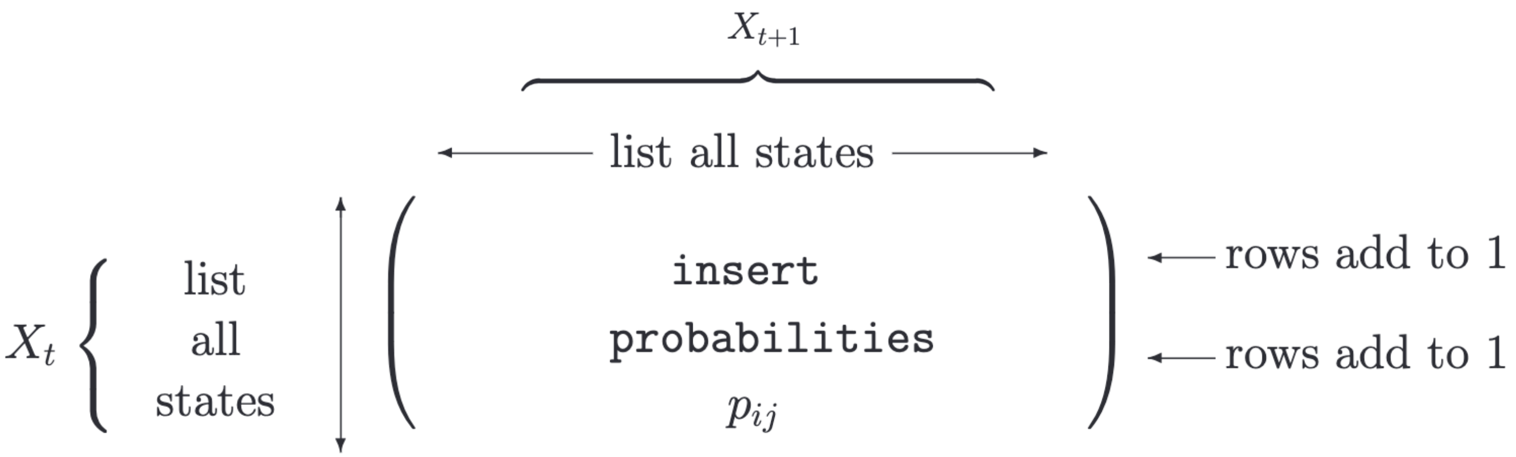
Extractive Summarization

Summarization

- There are two types of summarization:
  - **Extractive:** Method wherein key sentences are **extracted** without modification from the original document.
  - **Abstractive:** Method wherein new sentences are **generated** to represent a summary of the original document.
- Extractive Pro:
  - Faithful to the original text
- Extractive Con:
  - Can lead summaries that are not fluent

PageRank

- The text rank algorithm is based of the **PageRank** algorithm
- Given a matrix of transition probabilities for web pages (web pages as nodes, links between them as edges)



- Use PageRank algorithm to compute stationary distribution (long-term probabilities)

$$r_{new} = (\alpha P + (1-\alpha)E) \times r_{old}$$

- Higher values in PageRank indicate a webpage is “more important”

TextRank

- Instead of a transition matrix of webpages, **TextRank relies on similarities between sentences** (sentences as nodes, similarities as edges)
- Sentences are represented as vectors
  - Embeddings
  - BOW
- The pairwise similarity of those vectors is calculated
  - Cosine Similarity
  - Jaccard
- The core convergence algorithm is still applied
- **A sentence is considered important if it is similar to many other sentences**