RICE SOCIAL SCIENCES
Department of Linguistics

named entities & annotation

LING 430 Computational Linguistics Fall 2026
lecture 10a
Oct 27 2026

Siyu Liang
siyu.liang@rice.edu

logistics

- **HW 5 due Thursday, Oct 29, 11:59 pm**

last time

- hidden Markov model & Viterbi
  - Markov chains
  - hidden Markov models
  - the Viterbi algorithm
  - HMM tagging

agenda

- named entities recognition
- evaluating NER
- annotation

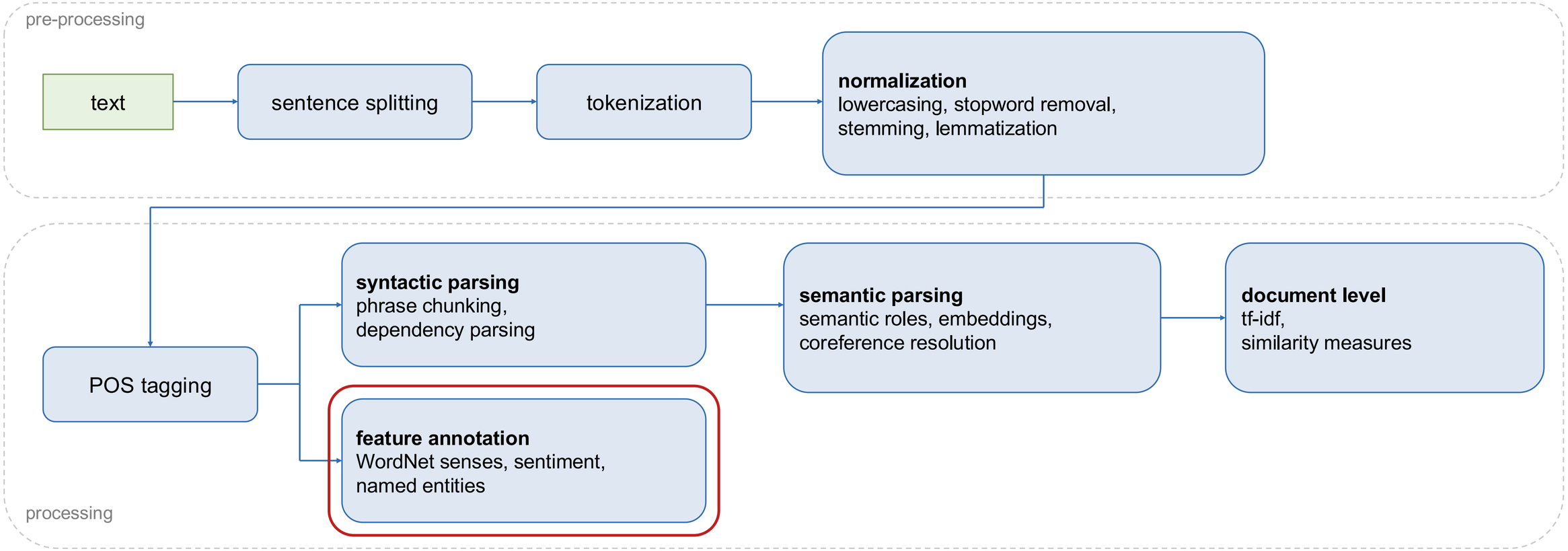
today's notebook

class10a.ipynb

- course website → week 10 → activity
  - **File → Save a copy in Drive**

the NLP pipeline

the NLP pipeline



the NLP pipeline

- plans
  - syntactic parsing (Thursday)
  - semantic parsing & document level analysis: embedding, tf-idf and similarity (next week)
- most current approaches are end-to-end (E2E) instead of a pipeline
  - neural networks, transformers, large language models (week 12)

named entities

named entities

- who is Washington?
  - *Washington was born into slavery on James Burroughs's farm.*
  - *Washington went up 2 games to 1 in the four-game series.*
  - *Blair arrived in Washington for his last state visit.*
  - *In June, Washington passed a primary seatbelt law.*
  - *Washington admitted its largest class in Seattle this fall.*











named entities

- a POS tagger would tag all cases of *Washington* as NNP
  - which is right
- but they refer to different kinds of things: a person, a team, a city, a state, a school
- a **named entity** is anything that can be referred to with a proper name
  - a person, a place, an organization
- **named entity recognition** (NER): find the spans of text that are proper names, and tag the type of each

named entities

- there are four frequent types
- dates, times and prices are often added, though they are technically not entities
- a biology application would add proteins and genes
- a company can add lines of products

tag	type	example
PER	people, characters	*Turing is a giant of computer science.*
ORG	companies, sports teams	*The IPCC warned about the cyclone.*
LOC	regions, mountains, seas	*Mt. Sanitas is in Sunshine Canyon.*
GPE	countries, states, cities	*Palo Alto is raising the fees for parking.*

named entities

- named entities are ambiguous
  - *JFK*: a person, an airport, and any number of schools, bridges and streets
  - *Washington*: different named entity types

sentence	type
Washington was born into slavery on James Burroughs's farm.	PER
Washington went up 2 games to 1 in the four-game series.	ORG
Blair arrived in Washington for his last state visit.	GPE
In June, Washington passed a primary seatbelt law.	GPE
Washington admitted its largest class in Seattle this fall.	ORG

named entities

- the output of an NER tagger looks like the following
- *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.*

named entities

why find them?

- what would we do with the entities once we have them?
- sentiment analysis: what does the reviewer think of *this* airline?
- question answering: *who was the first President of the US?* wants an answer of type PER
  - so *the Washington Monument* and *Washington, D.C.* can be skipped
- linking the text to Wikipedia or another knowledge source
- extracting events and the relations between their participants: *X acquired Y*

BIO tagging

BIO tagging

- in part-of-speech tagging every word gets one tag
- an entity is a **span**: one or more words with a start and an end
  - *[PER Jane Villanueva] of [ORG United Airlines Holding] discussed the [LOC Chicago] route.*
- and most words in a text are in no span at all
- how do we turn spans into one label per word?

BIO tagging

one label per word

- a first try: label every word with its entity type, or O for “outside”
  - *Jane*/PER *Villanueva*/PER *of*/O *United*/ORG *Airlines*/ORG *Holding*/ORG …
- this is **IO tagging**, and it loses information
  - *Rice*/ORG *Houston*/ORG *game*/O: one team or two?
- two neighboring entities of the same type melt into one

BIO tagging

B, I and O

- Ramshaw and Marcus (1995): mark where each span begins
  - **B**-type: the word that begins a span
  - **I**-type: a word inside a span, after the first
  - **O**: outside any span
- one O tag, and a B and an I for each entity type
  - $2n$ + 1 tags, so 9 tags for our 4 types
- *Rice*/B-ORG *Houston*/B-ORG *game*/O: two teams

BIO tagging

- [PER Jane Villanueva] of [ORG United Airlines Holding] discussed the [LOC Chicago] route.

Jane	***Villanueva***	***of***	***United***	***Airlines***	***Holding***	***discussed***	***the***	***Chicago***	***route***	***.***
B-PER	I-PER	O	B-ORG	I-ORG	I-ORG	O	O	B-LOC	O	O

BIO tagging

IO, BIO and BIOES

- **BIOES** adds E for the end of a span and S for a span of a single word

word	IO	BIO	BIOES
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
Chicago	I-LOC	B-LOC	S-LOC

BIO tagging

sequence labeling

- the brackets and the BIO tags carry exactly the same information
- but now there is one label per word, so NER looks just like part-of-speech tagging
- so we have sequence labeling again!
- any sequence labeling algorithm will do: an HMM, a CRF, a neural network
- the tags have a grammar of their own: I-PER can follow B-PER or I-PER, never O
  - a transition probability of 0, and Viterbi never proposes a span without a beginning

evaluating NER

evaluating NER

- an NER system is scored by **precision**, **recall** and **F1,** similar to a part-of-speech tagger (and in other tasks)
- the **entity**, not the word, is the unit of response
  - an entity counts as correct only if its boundaries and its type both match the gold

evaluating NER

gold standard vs. system output

- gold: *[PER Jane Villanueva] of [ORG United Airlines Holding] discussed the [LOC Chicago] route.*
- system: *[PER Jane] Villanueva of [ORG United Airlines Holding] discussed the [LOC Chicago] route.*

gold entity	output entity	correct?
Jane Villanueva PER	*Jane* PER	✗
United Airlines Holding ORG	*United Airlines Holding* ORG	✓
Chicago LOC	*Chicago* LOC	✓

evaluating NER

precision, recall, F1

- the system proposed 3 entities, 2 of them correct

$$\text{precision} = \frac{\text{correct}}{\text{proposed}} = \frac{2}{3} = 0.67$$

- the gold has 3 entities, 2 of them found

$$\text{recall} = \frac{\text{correct}}{\text{gold}} = \frac{2}{3} = 0.67$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} = \frac{2 \times 0.67 \times 0.67}{0.67 + 0.67} = 0.67$$

gold entity	output entity	correct?
Jane Villanueva PER	*Jane* PER	✗
United Airlines Holding ORG	*United Airlines Holding* ORG	✓
Chicago LOC	*Chicago* LOC	✓

evaluating NER

- segmentation makes NER harder
- word by word, the system got 10 of 11 words right: 91%
- entity by entity, one wrong boundary costs two errors
  - a false positive: *Jane* is not an entity in the gold
  - a false negative: *Jane Villanueva* was never found

evaluating NER

NER with nltk

- `ne_chunk`, an NER classifier trained on newswire, takes a part-of-speech tagged sentence and returns a tree

```
import nltk
nltk.download("maxent_ne_chunker_tab")
nltk.download("words")
from nltk import word_tokenize, pos_tag, ne_chunk
s = "Blair flew to Washington."
print(ne_chunk(pos_tag(word_tokenize(s))))
```

→ (S (PERSON Blair/NNP) flew/VBD to/TO (GPE Washington/NNP) ./.)

evaluating NER

NER with nltk

```
s = "i love berlin."
print(ne_chunk(pos_tag(word_tokenize(s))))
```

→ (S i/NN love/VBP berlin/NN ./.)

```
s = "Rice beat Houston on Saturday."
print(ne_chunk(pos_tag(word_tokenize(s))))
```

→ (S
→ (PERSON Rice/NNP)
→ beat/VBD
→ (PERSON Houston/NNP)
→ on/IN
→ Saturday/NNP
→ ./.)

activity 1

part 1: annotate by hand

- in pairs, bracket every named entity and give it a type: PER, ORG, LOC or GPE
  - *Siri thinks Fondren is a person, not a library.*
  - *My roommate from Austin says Texas will beat Rice again this year.*
  - *I missed the talk at the Ken Kennedy Institute because the bus was late.*
  - *We went for drinks at Texas Tailgate after the Dynamo game.*
  - *Beyoncé grew up in the Third Ward of Houston.*
- then compare, entity by entity: same boundaries, same type?

activity 1

part 2: run the tagger

- in the notebook, run `ne_chunk` on the same four sentences
- then try with your own sentence
- does the tagger disagree with you? on the boundary, or on the type?

annotation

annotation

guidelines

- every disagreement in evaluation could be a rule somebody has to write down
  - “the determiner is not part of the span”
  - “a sports team named after a city is ORG”
- **annotation guidelines** define the task and document the markup conventions
- they grow as difficult cases turn up, and annotators are trained on them
- a gold standard is the text, plus the guidelines, plus the people who applied them

annotation

the life cycle of a corpus

- where do the tagged corpora we train on come from?
- raw data is collected, cleaned up, documented, and stored in a systematic structure
- layers of annotation are added, some needing specialist knowledge of the language
- quality control finds inconsistencies in the annotation
- a large corpus takes years, and tens or hundreds of person-years of work

annotation

quality control

- how do we know the annotation is any good?
- have a portion of the material annotated independently by two people, say 10% of the corpus
- report the **inter-annotator agreement**: an upper bound on any system trained on the corpus
- the number only means something for its task
  - 90% agreement is terrible for part-of-speech tagging, but exceptional for semantic role labeling
- cases of uncertainty or disagreement go to an expert for **adjudication**

annotation

inter-annotator agreement

- raw agreement counts the cases where two people guessing at random would agree anyway
- **Cohen's kappa** (Cohen, 1960) corrects for the agreement we expect by chance

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

- P_o: the observed agreement; P_e: the agreement expected by chance
- P_e comes from each annotator's own label frequencies, so a label that both use often counts for little
- κ = 1 is perfect agreement; κ = 0 is what chance alone would give

annotation

kappa

- two annotators A and B annotate 20 words as “entity” or O
  - they agree on 17 of them

$$P_o = (1 + 16)/20 = 0.85$$
$$P_e = 2/20 \cdot 3/20 + 18/20 \cdot 17/20 = 0.015 + 0.765 = 0.78$$
$$\kappa = (0.85 - 0.78)/(1 - 0.78) = 0.07/0.22 = 0.32$$

- since almost every word is O, most of that agreement is chance

	B: entity	**B: O**
A: entity	1	1
A: O	2	16

next time

- Thursday, Oct 29
  - grammars and parsing
- **HW 5 due Thursday, 11:59 pm**