Tagging (POS, NER)

CS 485, Fall 2024
Applications of Natural Language Processing

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

- Announcement: Project proposals are due by the end of next week: **Wed, 10/16** (the week after HW2 is due)

Upcoming NLP topics

- From bags-of-words to ordered structure....

Part of speech tags

- Syntax = how words compose to form larger meaning-bearing units
- POS = syntactic categories for words
  - You could substitute words within a class and have a syntactically valid sentence.
  - Give information how words can combine.

  - I saw the dog
  - I saw the cat
  - I saw the {table, sky, dream, school, anger, ...}

Schoolhouse Rock: Conjunction Junction
https://www.youtube.com/watch?v=ODGA7ssL-6g&index=1&list=PL6795522EAD6CE2F7

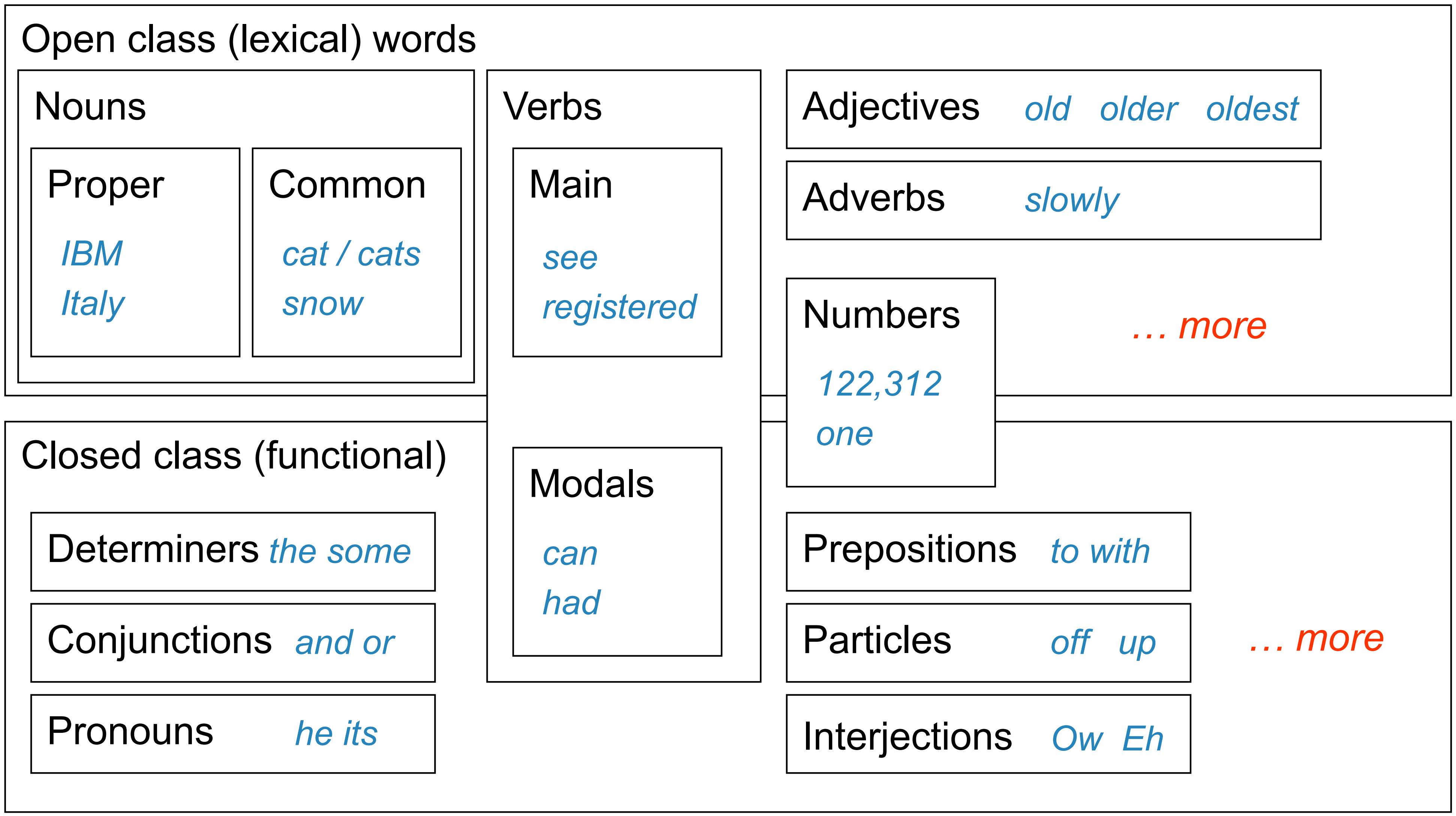
Demo

- https://corenlp.run/

Part of speech tagging

- I saw the fire today
- Fire!

Open vs closed classes



slide credit: Chris Manning

Do we want POS?

- Useful for many syntactic and other NLP tasks.
    - Phrase identification ("chunking")
    - Named entity recognition (proper nouns are often names)
    - Syntactic/semantic dependency parsing
    - Sentiment
- Either as features or heuristic filtering
- Esp. useful when not much training data
- Limitations
    - Coarse approximation of grammatical features
    - Sometimes cases are hard and ambiguous

POS patterns: simple noun phrases

POS patterns: simple noun phrases

- Quick and dirty noun phrase identification (Justeson and Katz 1995, Handler et al. 2016)
  - BaseNP = (Adj | Noun)* Noun
  - PP = Prep Det* BaseNP
  - NP = BaseNP PP*

Grammatical structure: Candidate strings are those multi-word noun phrases that are specified by the regular expression $((A \mid N)^+ \mid ((A \mid N)^*(NP)^?)(A \mid N)^*)N$,

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

Table 5.2 Part of speech tag patterns for collocation filtering. These patterns were used by Justeson and Katz to identify likely collocations among frequently occurring word sequences.

POS patterns: sentiment

- Turney (2002): identify bigram phrases, from unlabeled corpus, useful for sentiment analysis.

Table 1. Patterns of tags for extracting two-word phrases from reviews.

	First Word	Second Word	Third Word (Not Extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Table 2. An example of the processing of a review that the author has classified as *recommended*.[6]

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
online experience	JJ NN	2.253
low fees	JJ NNS	0.333
local branch	JJ NN	0.421
small part	JJ NN	0.053
online service	JJ NN	2.780
printable version	JJ NN	-0.705
direct deposit	JJ NN	1.288
well other	RB JJ	0.237
inconveniently located	RB VBN	-1.541
other bank	JJ NN	-0.850
true service	JJ NN	-0.732

(plus co-occurrence information)

POS Taggers

- How do you predict POS tags?
- Off-the-shelf models widely available, at least for mainstream varieties of major world languages
  - e.g. Spacy, Stanza, CoreNLP, etc.
- Typically use logistic regression-like models
  - Each token instance is a classification problem
  - Labeled datasets: e.g. https://universaldependencies.org/

POS Tagging: lexical ambiguity

Can we just use a tag dictionary (one tag per word type)?

Types:		WSJ	Brown
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%**)

Most words types are unambiguous ...

But not so for *tokens*!

- Ambiguous wordtypes tend to be the common ones.
  - I know **that** he is honest = IN (relativizer)
  - Yes, **that** play was nice = DT (determiner)
  - You can't go **that** far = RB (adverb)

POS Tagging: baseline

- Baseline: each word type's most frequent tag. >90% accuracy!
  - Simple baselines are very important to run!

- Though...
  - I get 91.8% accuracy for token tagging
  - ...but, 18.6% whole-sentence accuracy (!)

- Next: many other NLP tasks can be cast as tagging
  - Named entities
  - Word sense disambiguation

Named entity recognition

SOCCER - [PER BLINKER] BAN LIFTED .
[LOC LONDON] 1996-12-06 [MISC Dutch] forward [PER Reggie Blinker] had his indefinite suspension lifted by [ORG FIFA] on Friday and was set to make his [ORG Sheffield Wednesday] comeback against [ORG Liverpool] on Saturday . [PER Blinker] missed his club's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was playing for [ORG Feyenoord].

Figure 1: Example illustrating challenges in NER.

- Goal: for a fixed entity type inventory (e.g. PERSON, LOCATION, ORGANIZATION), identify all *spans* from a document
  - Name structure typically defined as flat (is this good?)

[Ratinov and Roth 2009]

BIO tagging

[$_{\text{PER}}$ **Jane Villanueva**] of [$_{\text{ORG}}$ **United**] , a unit of [$_{\text{ORG}}$ **United Airlines Holding**] , said the fare applies to the [$_{\text{LOC}}$ **Chicago**] route.

- Can we map identify phrases (spans) identification to token-level tagging?

BIO tagging

[$_{\text{PER}}$ **Jane Villanueva**] of [$_{\text{ORG}}$ **United**] , a unit of [$_{\text{ORG}}$ **United Airlines Holding**] , said the fare applies to the [$_{\text{LOC}}$ **Chicago**] route.

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

Figure 17.7 NER as a sequence model,

"IO" tagging: issues?

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	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

Figure 17.7 NER as a sequence model, showing IO, BIO, and BIOES taggings.

BIO is a lossless representation of flat spans

Easy to extract spans from tagger output

Useful features for a tagger

- Key sources of information:
  - 1. The word itself
  - 2. Word-internal characters
  - 3. Nearby words in a *context window*
    - **Context window features are used for ALL tagging tasks!**
    - Necessary to deal with ***lexical ambiguity***

Features for tagging

- Current word features
  - Word itself
  - Word shape ("Aa" "aa"), affixes ("-ing")
- Contextual word features: versions of these at nearby positions (*e.g.*: t-3, t-2, t-1, t, t+1, t+2, t+3)

- External lexical knowledge
  - Gazetteer features: Does word/phrase occur in a list of known names?
  - Other hand-built lexicons

- Neural network embedding representations (later in course)

Gazetteers example

1)**People**: *people, births, deaths*. Extracts 494,699 Wikipedia titles and 382,336 redirect links. 2)**Organizations**: *cooperatives, federations, teams, clubs, departments, organizations, organisations, banks, legislatures, record labels, constructors, manufacturers, ministries, ministers, military units, military formations, universities, radio stations, newspapers, broadcasters, political parties, television networks, companies, businesses, agencies*. Extracts 124,403 titles and 130,588 redirects. 3)**Locations**: *airports, districts, regions, countries, areas, lakes, seas, oceans, towns, villages, parks, bays, bases, cities, landmarks, rivers, valleys, deserts, locations, places, neighborhoods*. Extracts 211,872 titles and 194,049 redirects. 4)**Named Objects**: *aircraft, spacecraft, tanks, rifles, weapons, ships, firearms, automobiles, computers, boats*. Extracts 28,739 titles and 31,389 redirects. 5)**Art Work**: *novels, books, paintings, operas, plays*. Extracts 39,800 titles and 34037 redirects. 6)**Films**: *films, telenovelas, shows, musicals*. Extracts 50,454 titles and 49,252 redirects. 7)**Songs**: *songs, singles, albums*. Extracts 109,645 titles and 67,473 redirects. 8)**Events**: *playoffs, championships, races, competitions, battles*. Extracts 20,176 titles and 15,182 redirects.

[Ratinov and Roth 2009]