Tagging (POS, NER)

CS 485, Fall 2023
Applications of Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs485_f23/

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

- HW2 - how's it going? Phase 1 tomorrow, Phase 2 next week (Fri 10/20)!
- Project proposals: due **Wed 10/25**
  https://people.cs.umass.edu/~brenocon/cs485_f23/project.html
- after that, HW3, syntax
- after that, Midterm: either 11/7 or 11/9. Will know soon. Practice questions will be available.

Topics overview

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

Open class (lexical) words

- Nouns
  - Proper: *IBM*, *Italy*
  - Common: *cat / cats*, *snow*
- Verbs
  - Main: *see*, *registered*
- Adjectives: *old older oldest*
- Adverbs: *slowly*
- Numbers: *122,312*, *one*
- … more

Closed class (functional)

- Verbs
  - Modals: *can*, *had*
- Determiners: *the some*
- Conjunctions: *and or*
- Pronouns: *he its*
- Prepositions: *to with*
- Particles: *off up*
- Interjections: *Ow Eh*
- … more

slide credit: Chris Manning

Why 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.

Congressional bills

(Top terms, ranked by relative log-odds z-scores)

Uni. Dem.	and, deleted, health, mental, domestic, inserting, grant, programs, prevention, violence, program, striking, education, forensic, standards, juvenile, grants, partner, science, research
Uni. Rep.	any, offense, property, imprisoned, whoever, person, more, alien, knowingly, officer, not, united, intent, commerce, communication, forfeiture, immigration, official, interstate, subchapter
NPs Dem.	
NPs Rep.	

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/

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***

POS Tagging: lexical ambiguity

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

		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%**)

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: most frequent tag. 92.7% accuracy
  - Simple baselines are very important to run!

- Is this actually that high?
  - I get 0.918 accuracy for token tagging
  - ...but, 0.186 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

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

BIO tagging

	Barack	Obama	Michelle	Obama	were ...
Goal: represent two spans	Barack Obama		Michelle Obama		
NAME vs O doesn't work	N	N	N	N	O
BIO	B-N	I-N	B-N	I-N	O

make cross-product of "B"egin and "I"nside against each class type:

O, B-PER, I-PER, B-LOC,I-LOC, ...

... then spans can easily be extracted from tagger output.

[slide made after lecture]

Features for tagging

- Word-based features
  - Word itself
  - Word shape ("Aa" "aa")
  - Contextual (word window) variants: versions of these at position t-1, t-2, t-3 … 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)

Intuition from Warren Weaver (1955):

“If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words…

But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word…

The practical question is : ``What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?”

[slide: SLP3]

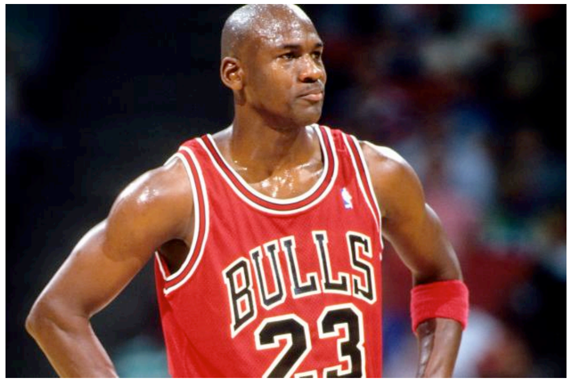
Gazetteers example

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]

Word sense disambiguation

- Task: Choose a word's sense in context
- Given KB and text:
  Want to tag spans in text with concept IDs
- Disambiguation problem
  - "I saw the bank" => bank#1 or bank#2?
  - "Michael Jordan was here" => ?



- Many terms for this: concept tagging, entity linking, "wikification", WSD

Word sense disambiguation

- Supervised setting: need ground-truth concept IDs for words in text
- Main approach: use *contextual information* to disambiguate.

Intuition from Warren Weaver (1955):

"If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words…

But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word…

The practical question is : ``What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?"

[slide: SLP3]

Two kinds of features in the vectors

- **Collocational** features and **bag-of-words** features
  - **Collocational**
    - Features about words at **specific** positions near target word
      - Often limited to just word identity and POS
  - **Bag-of-words**
    - Features about words that occur anywhere in the window (regardless of position)
      - Typically limited to frequency counts

[slide: SLP3]

Examples

- Example text (WSJ):

  An electric guitar and **bass** player stand off to one side not really part of the scene

- Assume a window of +/- 2 from the target

[slide: SLP3]

Examples

- Example text (WSJ)

  An electric $\boxed{\text{guitar}}\boxed{\text{and}}$ **bass** $\boxed{\text{player}}\boxed{\text{stand}}$ off to one side not really part of the scene,

- Assume a window of +/- 2 from the target

[slide: SLP3]

Collocational features

- Position-specific information about the words and collocations in window
- $\boxed{\text{guitar}}\boxed{\text{and}}$ bass $\boxed{\text{player}}\boxed{\text{stand}}$

$$[w_{i-2}, \mathrm{POS}_{i-2}, w_{i-1}, \mathrm{POS}_{i-1}, w_{i+1}, \mathrm{POS}_{i+1}, w_{i+2}, \mathrm{POS}_{i+2}, w_{i-2}^{i-1}, w_{i}^{i+1}]$$

```
[guitar, NN, and, CC, player, NN, stand, VB, and guitar, player stand]
```

- word 1,2,3 grams in window of ±3 is common

[slide: SLP3]

Bag-of-words features

- “an unordered set of words” – position ignored
- Counts of words occur within the window.
- First choose a vocabulary
- Then count how often each of those terms occurs in a given window
  - sometimes just a binary “indicator” 1 or 0

[slide: SLP3]

Word sense disambiguation

- Supervised setting: need ground-truth concept IDs for words in text
- Contextual features
  - Word immediately to left ... to right ...
  - Word within 10 word window (20 word window? entire document?)
- Features from matching a concept description, if your KB has one
  - *Michael Jeffrey Jordan (born February 17, 1963), also known by his initials, MJ,[1] is an American former professional basketball player. He is also a businessman, and principal owner and chairman of the Charlotte Hornets. Jordan played 15 seasons in the National Basketball Association (NBA) for theChicago Bulls and Washington Wizards.*
- Overall (prior) sense frequency
  - For WN, hard to beat Most Frequent Sense baseline (?!)
  - For major real-world named entities: consider "Obama", "Trump"
    - This task is also called "Entity Linking"