exercises in front!

Tagging: Classification in Context

CS 490A, Fall 2021

https://people.cs.umass.edu/~brenocon/cs490a_f21/

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

- Housekeeping: where we're at in the semester
- How was HW2? How are projects going?

In-text tagging

- Previous: Text Classification
  - Input:
  - Output:
- Next: Token Tagging (Classif.)
  - Input:
  - Output:

- Let's move to classifying **within the text**!
  - Tasks you can do yourself, with the right heuristics or logistic regression features (or other NLP models)
  - Do it with a pretrained, off-the-shelf system as part of a larger system, especially for syntactic/semantic linguistic analyses

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.

[Ratinov and Roth 2009]

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, ...}

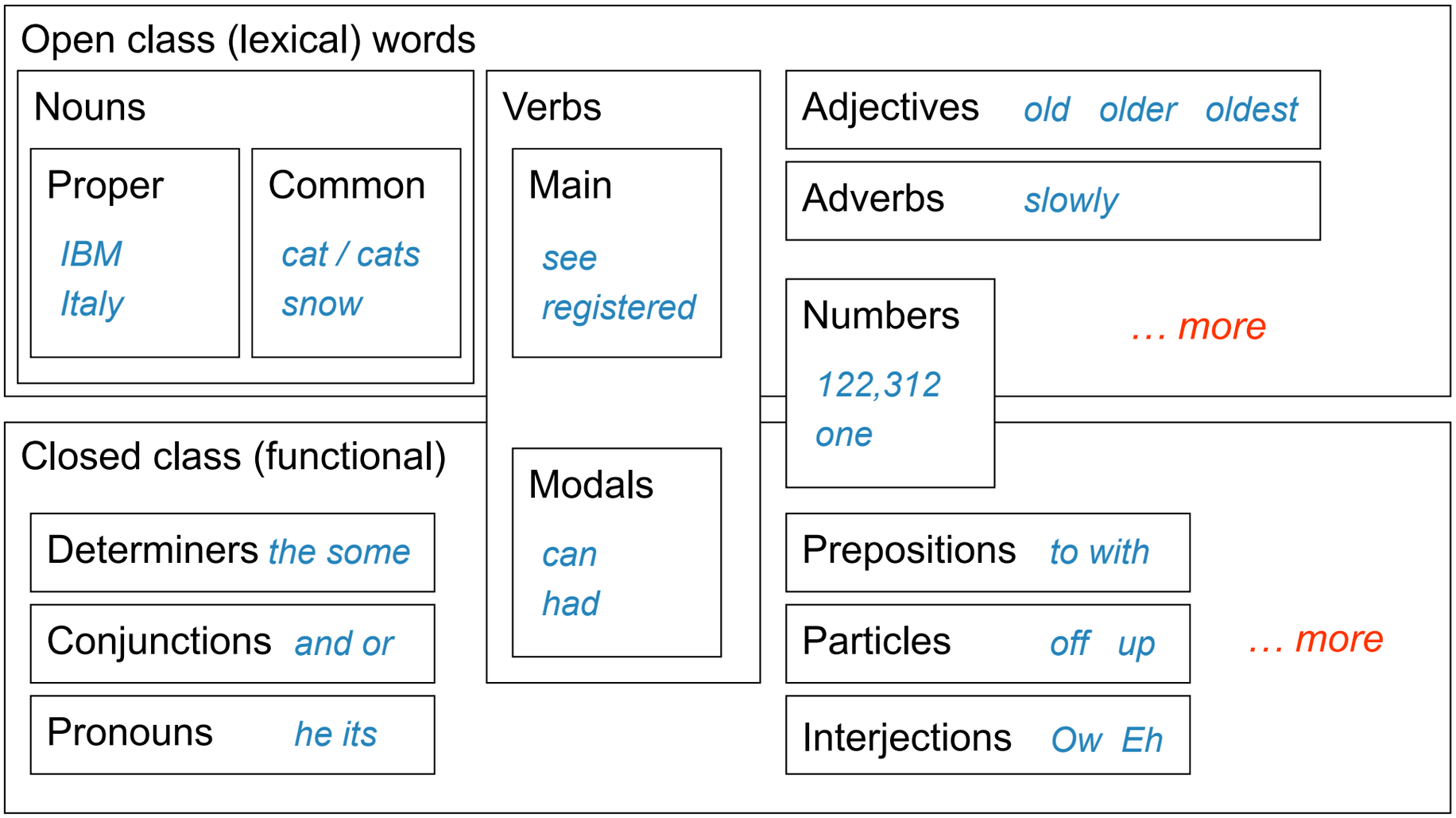
Schoolhouse Rock: Conjunction Junction
https://www.youtube.com/watch?v=ODGA7ssL-6g&index=1&list=PL6795522EAD6CE2F7

Part of speech tagging

- I saw the fire today

- Fire!

Open vs closed classes



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

- 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
    - (And possibly joint classification - we'll discuss next time)
  - Labeled datasets: e.g. https://universaldependencies.org/

- stopped here 10/12