Parts of speech

CS 585, Fall 2015
Introduction to Natural Language Processing
http://people.cs.umass.edu/~brenocon/inlp2015/

Brendan O'Connor

Where we are going

- Bags vs. Sequences
- Classification vs. Sequence Tagging

What's a part-of-speech (POS)?

- 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, ...}

POS is an old idea

- Dionysius Thrax of Alexandria (100 BCE):
  8 parts of speech
- Common in grammar classes today:
  noun, verb, adjective, preposition, conjunction, pronoun, interjection
- Many other more fine-grained possibilities

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

Open vs closed classes

Open class (lexical) words

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

… more

Closed class (functional)

- Determiners *the some*
- Conjunctions *and or*
- Pronouns *he its*
- Prepositions *to with*
- Particles *off up*
- Interjections *Ow Eh*

… more

slide credit: Chris Manning

Many tagging standards

- Penn Treebank (45 tags) ... *the most common one*
- Brown corpus (85 tags)
- Coarse tagsets
  - Petrov et al. "Universal" tagset (12 tags)
    - http://code.google.com/p/universal-pos-tags/
    - Motivation: cross-linguistic regularities
      - e.g. adposition: pre- and postpositions
    - For English, collapsing of PTB tags
  - Gimpel et al. tagset for Twitter (25 tags)
    - Motivation: easier for humans to annotate
    - We collapsed PTB, added new things that were necessary for Twitter

Why do we want POS?

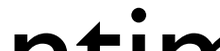
- Useful for many syntactic and other NLP tasks.
  - Phrase identification ("chunking")
  - Named entity recognition
  - Full parsing
  - Sentiment

POS patterns: sentiment

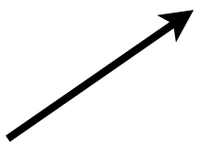
- Turney (2002): identify bigram phrases, from unlabeled corpus, useful for sentiment analysis.

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



(plus sentiment labels stuff)



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 sentiment labels stuff)

POS patterns: simple noun phrases

- Quick and dirty noun phrase identification

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.

How to get POS tags?

- Classification or sequence labeling problem

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%**)

Most words types are unambiguous ...

- Ambiguous wordtypes tend to be very 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: 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 very 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!

POS Tagging: baseline

- Baseline: most frequent tag. 92.7% accuracy
  - Simple baselines are very important to run!
- Why so high?
  - Many ambiguous words have a skewed distribution of tags
  - Credit for easy things like punctuation, “the”, “a”, etc.

POS Tagging: baseline

- Baseline: most frequent tag. 92.7% accuracy
  - Simple baselines are very important to run!
- Why so high?
  - Many ambiguous words have a skewed distribution of tags
  - Credit for easy things like punctuation, “the”, “a”, etc.
- Is this actually that high?
  - I get 0.918 accuracy for token tagging
  - ...but, 0.186 whole-sentence accuracy (!)

POS tagging can be hard for humans, too

- Mrs/NNP Shaefer/NNP never/RB got/VBD **around/RP** to/TO joining/VBG
- All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB **around/IN** the/DT corner/NN
- Chateau/NNP Petrus/NNP costs/VBZ **around/RB** $/$ 250/CD

Need careful guidelines (and do annotators always follow them?)
PTB POS guidelines, Santorini (1990)

4 Confusing parts of speech

This section discusses parts of speech that are easily confused and gives guidelines on how to tag such cases.

CD or JJ

Number-number combinations should be tagged as adjectives (JJ) if they have the same distribution as adjectives.

EXAMPLES: a 50–3/JJ victory (cf. a handy/JJ victory)

Hyphenated fractions *one-half*, *three-fourths*, *seven-eighths*, *one-and-a-half*, *seven-and-three-eighths* should be tagged as adjectives (JJ) when they are prenominal modifiers, but as adverbs (RB) if they could be replaced by *double* or *twice*.

EXAMPLES:	one-half/JJ cup;	cf. a full/JJ cup
	one-half/RB the amount;	cf. twice/RB the amount; double/RB the amount

Some other lexical ambiguities

- Prepositions versus verb particles
  - turn into/P a monster
  - take out/T the trash
  - check it out/T, what's going on/T, shout out/T

Test:
turn slowly into a monster
*take slowly out the trash

Careful annotator guidelines are necessary to define what to do in many cases.

- http://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports
- http://www.ark.cs.cmu.edu/TweetNLP/annot_guidelines.pdf

Some other lexical ambiguities

- Prepositions versus verb particles
  - turn into/P a monster
  - take out/T the trash
  - check it out/T, what's going on/T, shout out/T

Test:
turn slowly into a monster
*take slowly out the trash

- this,that -- pronouns versus determiners
  - i just orgasmed over this/O
  - this/D wind is serious

Careful annotator guidelines are necessary to define what to do in many cases.

- http://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports
- http://www.ark.cs.cmu.edu/TweetNLP/annot_guidelines.pdf

How to build a POS tagger?

- Key sources of information:
  - 1. The word itself
  - 2. Word-internal characters
  - 3. POS tags of surrounding words: syntactic context