Part of Speech Tagging

COMP90042

Natural Language Processing

Lecture 5

Semester 1 2021 Week 3

Jey Han Lau

THE UNIVERSITY OF MELBOURNE

What is Part of Speech (POS)?

- AKA word classes, morphological classes, syntactic categories
- Nouns, verbs, adjective, etc
- POS tells us quite a bit about a word and its neighbours:
  - nouns are often preceded by determiners
  - verbs preceded by nouns
  - *content* as a **noun** pronounced as *CONtent*
  - *content* as a **adjective** pronounced as *conTENT*

Information Extraction

- Given this:
  - “Brasilia, the Brazilian capital, was founded in 1960.”
- Obtain this:
  - capital(Brazil, Brasilia)
  - founded(Brasilia, 1960)
- Many steps involved but first need to know **nouns** (Brasilia, capital), **adjectives** (Brazilian), **verbs** (founded) and **numbers** (1960).

Outline

- Parts of speech
- Tagsets
- Automatic Tagging

POS Open Classes

Open vs **closed** classes: how readily do POS categories take on new words? Just a few open classes:

- Nouns
  - Proper (*Australia*) versus common (*wombat*)
  - Mass (*rice*) versus count (*bowls*)
- Verbs
  - Rich inflection (*go/goes/going/gone/went*)
  - Auxiliary verbs (*be*, *have,* and *do* in English)
  - Transitivity (*wait* versus *hit* versus *give*)
    — number of arguments

POS Open Classes

- Adjectives
  - Gradable (*happy*) versus non-gradable (*computational*)
- Adverbs
  - Manner (*slowly*)
  - Locative (*here*)
  - Degree (*really*)
  - Temporal (*today*)

POS Closed Classes (English)

- Prepositions (*in*, *on*, *with*, *for*, *of*, *over*,…)
  - ***on*** *the table*
- Particles
  - brushed himself ***off***
- Determiners
  - Articles (*a*, *an*, *the*)
  - Demonstratives (*this*, *that*, *these*, *those*)
  - Quantifiers (*each*, *every, some, two,…*)
- Pronouns
  - Personal (*I*, *me*, *she,…*)
  - Possessive (*my*, *our,…*)
  - Interrogative or *Wh* (*who*, *what*, …)

POS Closed Classes (English)

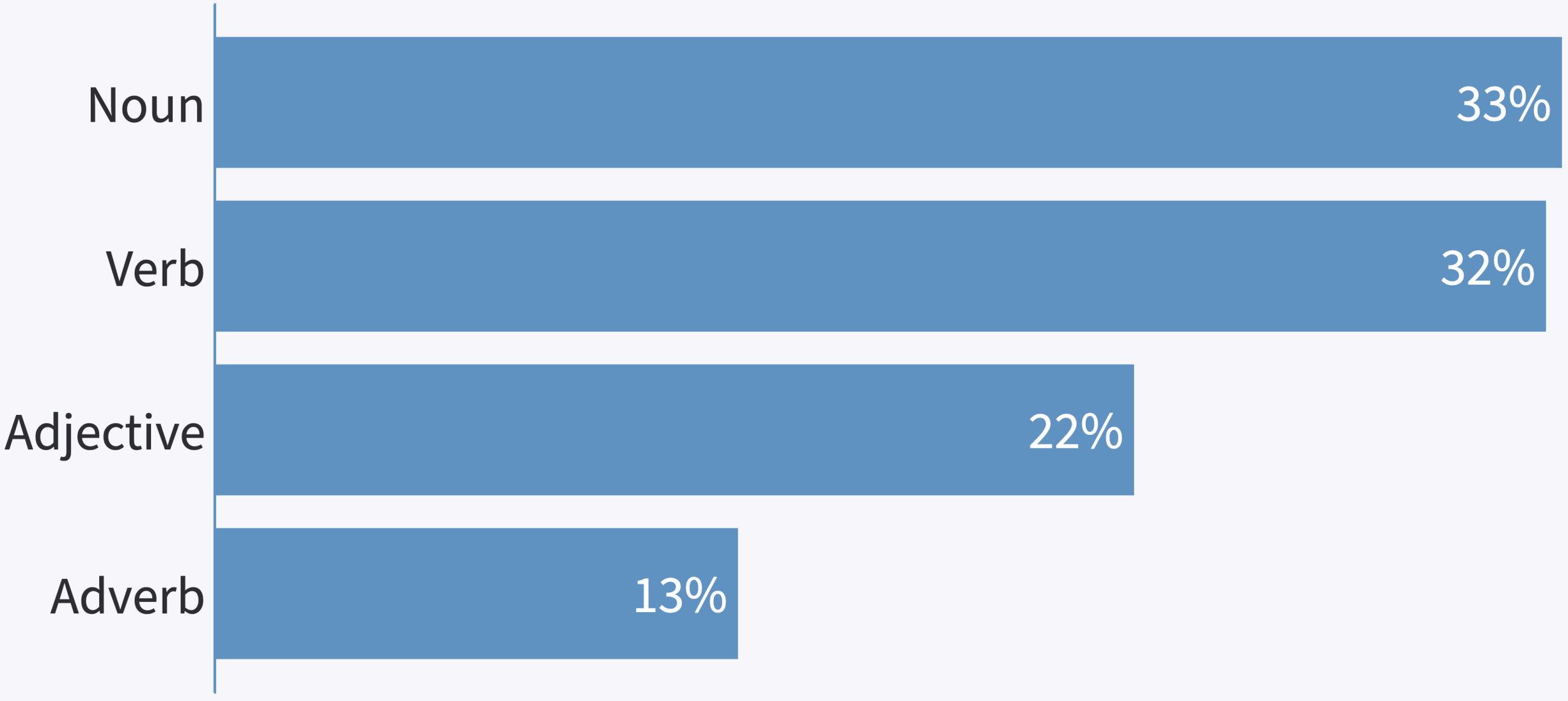
- Conjunctions
  - Coordinating (*and*, *or*, *but*)
  - Subordinating (*if*, *although, that, …*)
- Modal verbs
  - Ability (*can, could)*
  - Permission (*can, may*)
  - Possibility (*may, might, could, will*)
  - Necessity *(must*)
- And some more…
  - negatives, politeness markers, etc

Is POS universal? What open classes are seen in all languages?

- Noun
- Verb
- Adjective
- Adverb

PollEv.com/jeyhanlau569

Is POS universal? What open classes are seen in all languages?



Ambiguity

- Many word types belong to multiple classes
- POS depends on context
- Compare:
  - *Time flies like an arrow*
  - *Fruit flies like a banana*

Time	flies	like	an	arrow
noun	verb	preposition	determiner	noun

Fruit	flies	like	a	banana
noun	noun	verb	determiner	noun

POS Ambiguity in News Headlines

- British Left Waffles on Falkland Islands
  - [British Left] [Waffles] [on] [Falkland Islands]
- Juvenile Court to Try Shooting Defendant
  - [Juvenile Court] [to] [Try] [Shooting Defendant]
- Teachers Strike Idle Kids
  - [Teachers Strike] [Idle Kids]
- Eye Drops Off Shelf
  - [Eye Drops] [Off Shelf]

Tagsets

Tagsets

- A compact representation of POS information
  - Usually ≤ 4 capitalized characters (e.g. NN = noun)
  - Often includes inflectional distinctions
- Major English tagsets
  - Brown (87 tags)
  - Penn Treebank (45 tags)
  - CLAWS/BNC (61 tags)
  - “Universal” (12 tags)
- At least one tagset for all major languages

Major Penn Treebank Tags

NN	noun	VB	verb
JJ	adjective	RB	adverb
DT	determiner	CD	cardinal number
IN	preposition	PRP	personal pronoun
MD	modal	CC	coordinating conjunction
RP	particle	WH	wh-pronoun
TO	*to*		

Derived Tags (Open Class)

- NN (noun singular, wombat)
  - NNS (plural, wombats)
  - NNP (proper, Australia)
  - NNPS (proper plural, Australians)

- VB (verb infinitive, eat)
  - VBP (1st /2nd person present, eat)
  - VBZ (3rd person singular, eats)
  - VBD (past tense, ate)
  - VBG (gerund, eating)
  - VBN (past participle, eaten)

Derived Tags (Open Class)

- JJ (adjective, nice)
  - JJR (comparative, nicer)
  - JJS (superlative, nicest)
- RB (adverb, fast)
  - RBR (comparative, faster)
  - RBS (superlative, fastest)

Derived Tags (Closed Class)

- PRP (pronoun personal, I)
  - PRP$ (possessive, my)
- WP (Wh-pronoun, what):
  - WP$ (possessive, whose)
  - WDT(wh-determiner, which)
  - WRB (wh-adverb, where)

Tagged Text Example

The limits to legal absurdity stretched another notch this week

when the Supreme Court refused to hear an appeal from a case that says corporate defendants must pay damages even after proving that they could not possibly have caused the harm .

Tagged Text Example

The/DT limits/NNS to/TO legal/JJ absurdity/NN stretched/VBD another/DT notch/NN this/DT week/NN
when/WRB the/DT Supreme/NNP Court/NNP refused/VBD to/TO hear/VB an/DT appeal/VB from/IN a/DT case/NN that/WDT says/VBZ corporate/JJ defendants/NNS must/MD pay/VB damages/NNS even/RB after/IN proving/VBG that/IN they/PRP could/MD not/RB possibly/RB have/VB caused/VBN the/DT harm/NN ./.

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Tag the following sentence with Penn Treebank's POS tagset:

CATS SHOULD CATCH MICE EASILY

PollEv.com/jeyhanlau569

Automatic Tagging

Why Automatically POS tag?

- Important for morphological analysis, e.g. lemmatisation
- For some applications, we want to focus on certain POS
  - E.g. nouns are important for information retrieval, adjectives for sentiment analysis
- Very useful features for certain classification tasks
  - E.g. genre attribution (fiction vs. non-fiction)
- POS tags can offer word sense disambiguation
  - E.g. *cross/NN* *cross/VB* *cross/JJ*
- Can use them to create larger structures (parsing; lecture 14–16)

Automatic Taggers

- Rule-based taggers
- Statistical taggers
  - Unigram tagger
  - Classifier-based taggers
  - Hidden Markov Model (HMM) taggers

Rule-based tagging

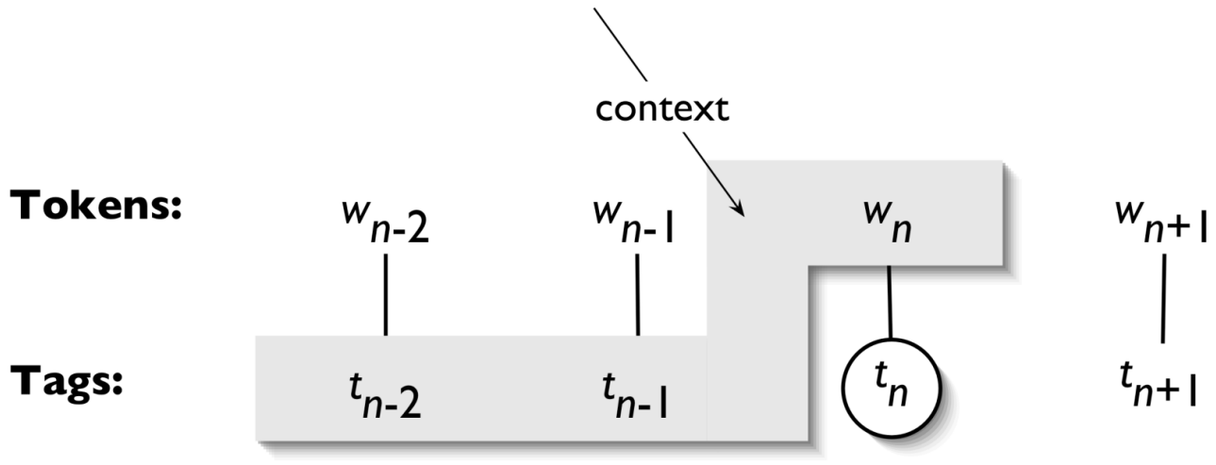
- Typically starts with a list of possible tags for each word
  - From a lexical resource, or a corpus
- Often includes other lexical information, e.g. verb *subcategorisation* (its arguments)
- Apply rules to narrow down to a single tag
  - E.g. If DT comes before word, then eliminate VB
  - Relies on some unambiguous contexts
- Large systems have 1000s of constraints

Unigram tagger

- Assign most common tag to each word type
- Requires a corpus of tagged words
- “Model” is just a look-up table
- But actually quite good, ~90% accuracy
  - Correctly resolves about 75% of ambiguity
- Often considered the baseline for more complex approaches

Classifier-Based Tagging

- Use a standard discriminative classifier (e.g. logistic regression, neural network), with features:
  - Target word
  - Lexical context around the word
  - Already classified tags in sentence
- But can suffer from error propagation: wrong predictions from previous steps affect the next ones



Hidden Markov Models

- A basic sequential (or structured) model
- Like sequential classifiers, use both previous tag and lexical evidence
- Unlike classifiers, considers all possibilities of previous tag
- Unlike classifiers, treat previous tag evidence and lexical evidence as independent from each other
  - Less sparsity
  - Fast algorithms for sequential prediction, i.e. finding the best tagging of entire word sequence
- Next lecture!

Unknown Words

- Huge problem in morphologically rich languages (e.g. Turkish)
- Can use things we've seen only once (hapax legomena) to best guess for things we've never seen before
  - Tend to be nouns, followed by verbs
  - Unlikely to be determiners
- Can use sub-word representations to capture morphology (look for common affixes)

A Final Word

- Part of speech is a fundamental intersection between linguistics and automatic text analysis
- A fundamental task in NLP, provides useful information for many other applications
- Methods applied to it are typical of language tasks in general, e.g. probabilistic, sequential machine learning

Reading

- JM3 Ch. 8-8.2