Part of speech tagging

COMP90042 Lecture 11

THE UNIVERSITY OF MELBOURNE

Authorship attribution revisited

- Training data:
  - "The lawyer convinced the jury." → Sam
  - "Ruby travelled around Australia." → Sam
  - "The hospital was cleaned by the janitor." → Max
  - "Lunch was served at 12pm." → Max
- "The bookstore was opened by the manager." → ?
- Similar **structure** (passive voice).
  - Not captured by simple BOW representations.
- How to ensure a computer knows/learns this?

Information extraction (teaser)

- 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).
- These are examples of **parts-of-speech** (POS).

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 (*yesterday*)

POS Closed classes (for English)

- Prepositions (*in*, *on*, *with*, *for*, *of*, *over*,…)
  - Regular (e.g. ***on*** *the table*)
  - Particles (e.g. *turn it* ***on***)
- 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 (for English)

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

Ambiguity

- Many word types belong to multiple classes
- 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
- Juvenile Court to Try Shooting Defendant
- Teachers Strike Idle Kids
- Ban On Soliciting Dead in Trotwood
- Eye Drops Off Shelf

Tagsets

- A compact representation of POS information
  - Usually ≤ 4 capitalized characters
  - 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

JJ adjective

DT determiner

IN preposition

MD modal

RP particle

TO *to*

VB verb

RB adverb

CD cardinal number

PRP personal pronoun

CC coordinating conjunction

WH wh-pronoun

Penn treebank derived tags

NN: NNS (plural, *wombats*), NNP (proper, *Australia*), NNPS (proper plural, *Australians*)

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

JJ: JJR (comparative, *nicer*), JJS (superlative, *nicest*)

RB: RBR (comparative, *faster*), RBS (superlative, *fastest*)

PRP: PRP$ (possessive, *my*)

WH: WH$ (possessive, *whose*), WDT(*wh*-determiner, *who*), WRB (*wh*-adverb, *where*)

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 ./.

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 classification
- POS tags can offer word sense disambiguation
  - E.g. *cross/NN* *cross/VB* *cross/JJ*
- Can use them to create larger structures (parsing)

Automatic Taggers

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

Rule-based tagging

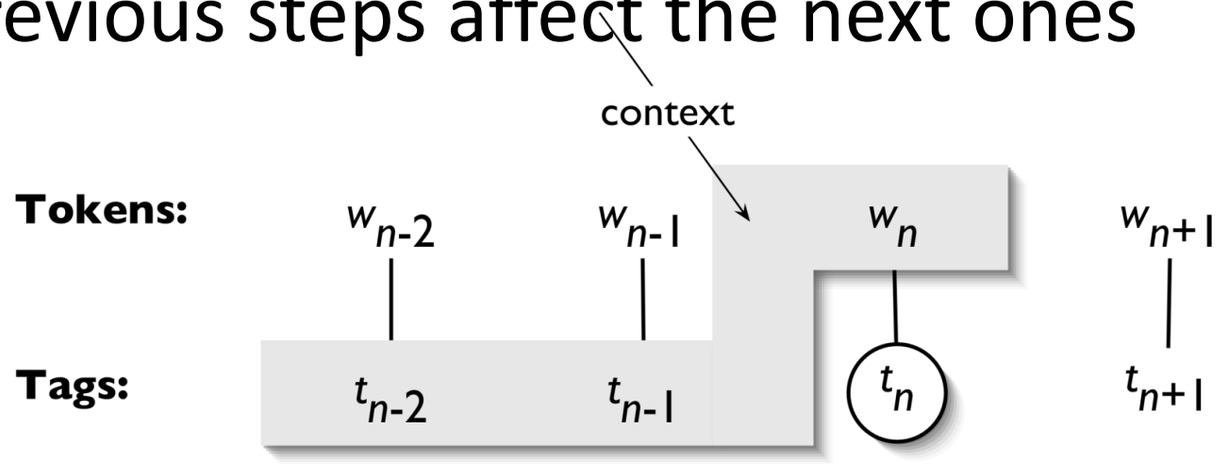
- Typically starts with a list of possible tags for each word
  - From a lexical resource, or a corpus
- Often includes other lexical information, e.g. verb *subcategorisation* (its arguments)
- Apply rules to narrow down to a single tag
  - E.g. If DT comes before word, then eliminate VB
  - Relies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

- 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
- Among the best sequential models
  - 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, treat previous tag(s) 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
- Closely related to CRFs

Unknown words

- Huge problem in morphologically rich languages (e.g. Turkish)
- Can use *hapax legomena* (things we've seen only once) to best guess for things we've never seen before
- 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 8.1-8.3, 8.5.1