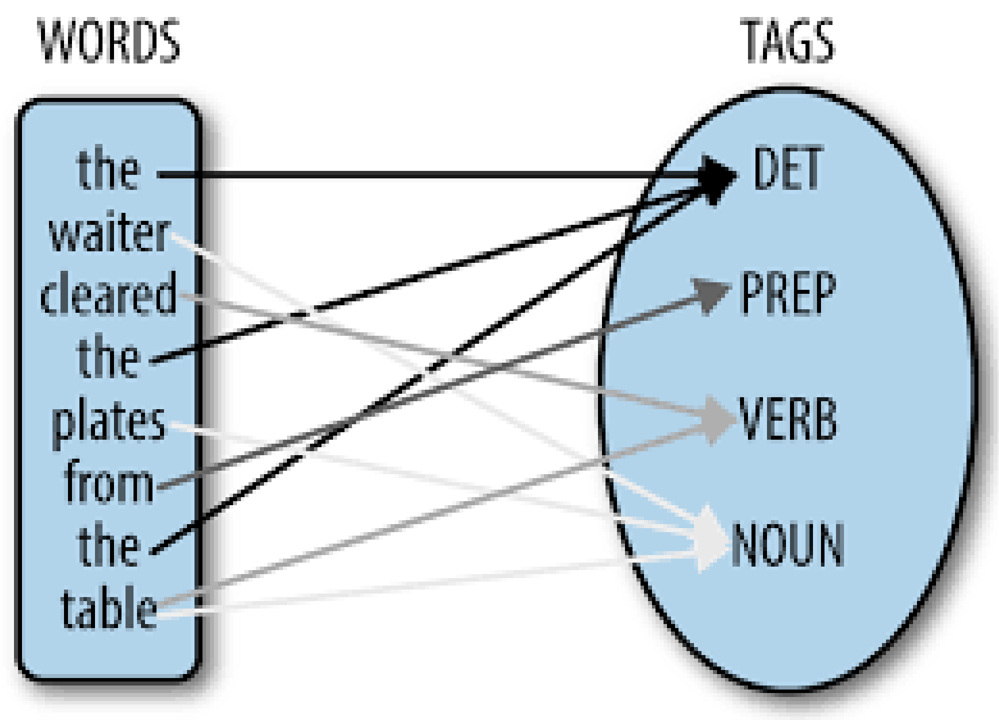


COMP90042 LECTURE 5

PART OF SPEECH TAGGING

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.” -> ?
- pollev.com/wsta
- Text “WSTA” to 0 427 541 357, then “S” or “M”.

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.” -> Max
- Why?
- pollev.com/wsta
- Text “WSTA” to 0 427 541 357, then your answer.

AUTHORSHIP ATTRIBUTION REVISITED

- "The coffee shop was opened by the manager." -> Max
- "The hospital was cleaned by the janitor." -> Max
- 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).

POS 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*)

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 (transitive; e.g. *on the table*)
  - Particles (intransitive; 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 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	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*		

PENN TREEBANK DERIVED TAGS

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

VB: VBP (base, *eat*), VB (infinitive, *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 ./.

AUTOMATIC TAGGERS

- Rule-based taggers
  - Hand-coded
  - Transformation-based (Brill)
- Statistical taggers
  - Unigram tagger
  - Classifier-based taggers
  - N-gram taggers
  - Hidden Markov Model (HMM) taggers

HAND-CODED RULES

- 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

TRANSFORMATION-BASED TAGGING

- Requires a tagged training corpus
- First, apply unigram tagger to get an initial tagging
- Then, sequentially learn rules to correct tags
  - Possible rules are generated from a small set of templates
    - Eg. Convert X to Y if previous tag is Z
  - Test the effect of all possible rules on current tagging
  - Apply rule that most improves tagging accuracy
    - E.g. **NN VB PREV-TAG TO**
- Accurate and very fast

UNIGRAM TAGGER

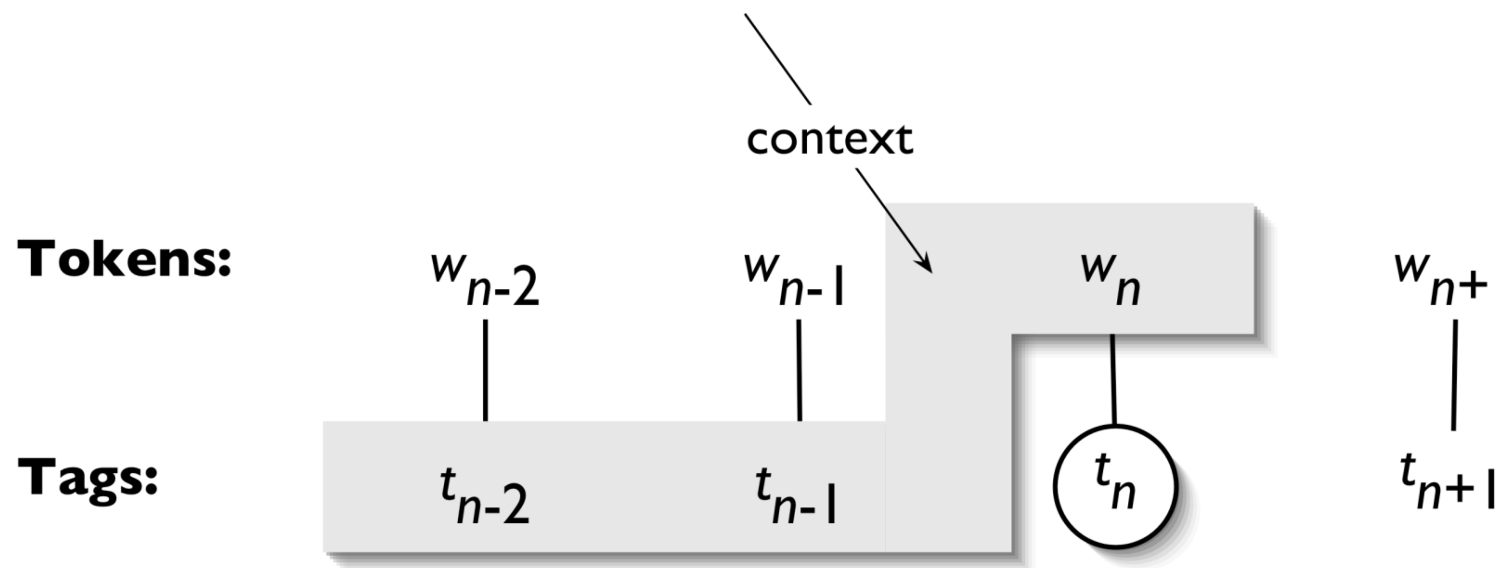
- Assign most common tag to each word type
- Requires a corpus of tagged words
- “Model” is just a look-up table
- But actually quite good, ~90% accuracy
  - Correctly resolves about 75% of ambiguity
- Often considered the baseline for more complex approaches

CLASSIFIER-BASED TAGGING

- Use a standard discriminative classifier (e.g. logistic regression), with features:
  - Target word
  - Lexical context around the word
  - Already classified tags in sentence
- Almost as good as best sequential models
  - But can suffer from **error propagation**: wrong predictions from previous steps affect the next ones (also know as **error bias** or **exposure bias**).
  - Some methods can mitigate this (imitation learning) but we will not cover them in this subject.

N-GRAM TAGGER

- Extension of unigram tagger
- Also a look-up based on corpus statistics
  - best tag for both word and previous $n-1$ tags
    - i. e. $\underset{t_n \in T}{\operatorname{argmax}} P(t_n | w_n, t_{n-1}, ..)$
  - E.g. DT *shot* $\rightarrow$ NN
- Problem: sparsity
  - Solution: backoff to *n-1* when no counts for *n*
- Also, must tag words one at a time, left to right



HIDDEN MARKOV MODELS

- A basic sequential (or structured) model
- Like *n*-gram taggers, use both previous tag and lexical evidence
- Unlike *n*-gram taggers, 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
- More on this in the next lecture…

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 morphology (look for common affixes)

A FINAL WORD

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

ADDITIONAL READING

- JM3 Ch. 10.1-10.3, 10.7
- (optional) Imitation learning tutorial
  https://sheffieldnlp.github.io/ImitationLearningTutorialEACL2017/