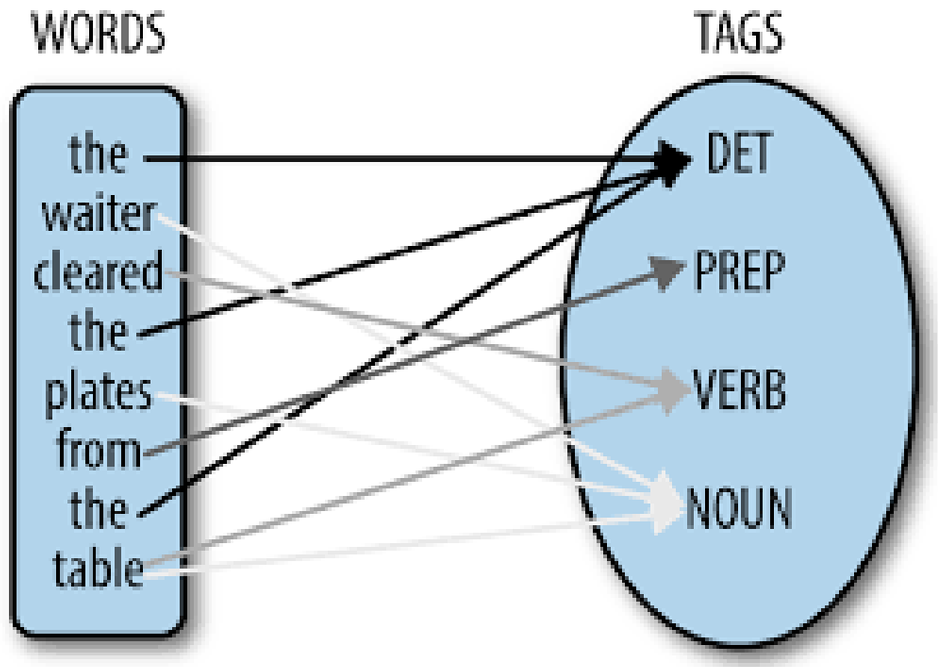


COMP90042 LECTURE 3

PART OF SPEECH TAGGING

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

OUTLINE

- Parts of speech
- Tagsets
- Automatic tagging

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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*)

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

OPEN CLASSES

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

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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*, …)

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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…

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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*		

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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*)

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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/NN 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 ./.

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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* vs *cross/VB* vs *cross*/JJ
- Can use them to create larger structures (chunk parsing)

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

AUTOMATIC TAGGERS

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

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

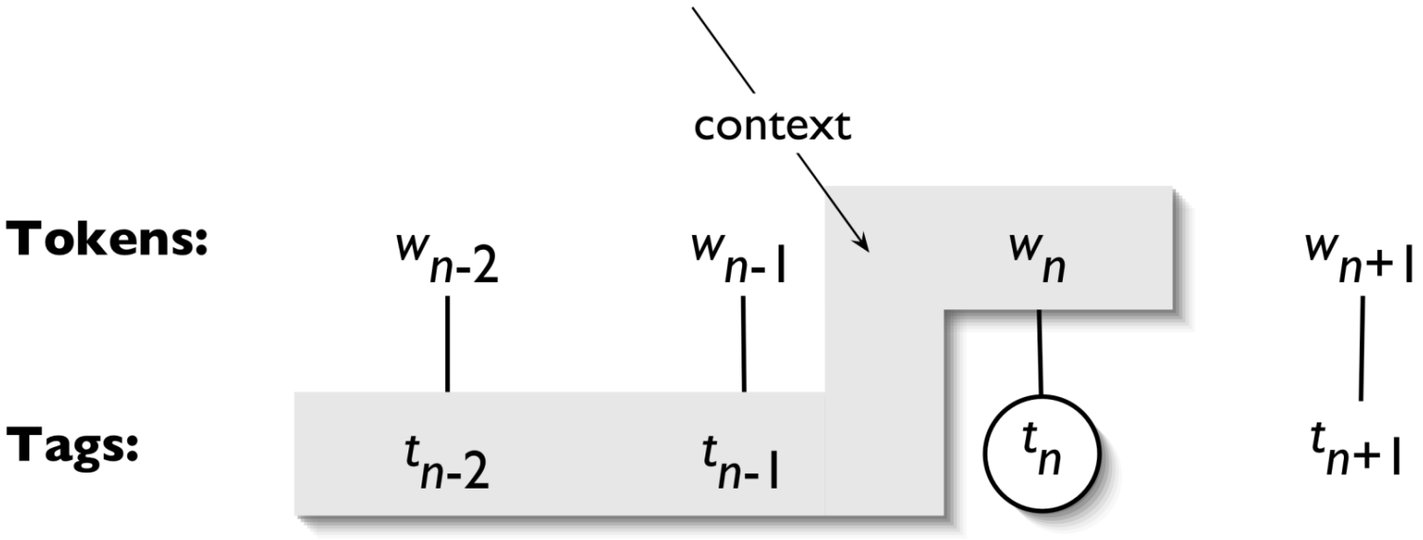
UNIGRAM TAGGER

- 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

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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. $\underset{t_n \in T}{\operatorname{argmax}} P(t_n | w_n, t_{n-1}, \ldots)$
  - E.g. DT *shot* → NN
- Problem: sparsity
  - Solution: backoff to *n-1* when no counts for *n*
- Also, must tag words one at a time, left to right



COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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…

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

CLASSIFIER-BASED TAGGING

- Use a standard discriminative classifier (e.g. logistic regression)
- Classify tag of target word based on
  - Target word
  - Lexical context around the word
  - Already classified tags in sentence
- Almost as good as best sequential models
  - And generally much faster
- MEMMs and CRFs combine sequential and classifier-based tagging

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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)

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

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

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

ADDITIONAL READING

- J&M2 Ch. 5.1-5.4, 5.6-5.8

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE