

Part-of-speech tagging

A simple but useful form of
linguistic analysis

Many slides adapted from slides by Chris Manning

Parts of Speech

- Perhaps starting with Aristotle in the West (384–322 BCE), there was the idea of having parts of speech
 - a.k.a lexical categories, word classes, “tags”, POS
- It comes from Dionysius Thrax of Alexandria (c. 100 BCE) the idea that is still with us that there are 8 parts of speech
 - But actually his 8 aren’t exactly the ones we are taught today
 - Thrax: noun, verb, article, adverb, preposition, conjunction, participle, pronoun
 - School grammar: noun, verb, adjective, adverb, preposition, conjunction, pronoun, interjection

Open class (lexical) words

Nouns

Proper

IBM
Italy

Common

cat / cats
snow

Verbs

Main

see
registered

Adjectives *old older oldest*

Adverbs *slowly*

Numbers

122,312
one

... more

Closed class (functional)

Determiners *the some*

Conjunctions *and or*

Pronouns *he its*

Modals

can
had

Prepositions *to with*

Particles *off up*

... more

Interjections *Ow Eh*

POS Tagging

- Words often have more than one POS: *back*
 - The back door = JJ
 - On my back = NN
 - Win the voters back = RB
 - Promised to back the bill = VB
- The POS tagging problem is to determine the POS tag for a particular instance of a word.

POS Tagging

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

- Input: Plays well with others
- Ambiguity: NNS/VBZ UH/JJ/NN/RB IN NNS
- Output: Plays/VBZ well/RB with/IN others/NNS
- Uses:
 - Text-to-speech (how do we pronounce “lead”?)
 - Can write regexps like (Det) Adj* N+ over the output for phrases, etc.
 - As input to or to speed up a full parser
 - If you know the tag, you can back off to it in other tasks

Penn
Treebank
POS tags

POS tagging performance

- How many tags are correct? (Tag accuracy)
 - About 97% currently
 - But baseline is already 90%
 - Baseline is performance of stupidest possible method
 - Tag every word with its most frequent tag
 - Tag unknown words as nouns
 - Partly easy because
 - Many words are unambiguous
 - You get points for them (*the*, *a*, etc.) and for punctuation marks!

How difficult is POS tagging?

- About 11% of the word types in the Brown corpus are ambiguous with regard to part of speech
- But they tend to be very common words. E.g., *that*
 - I know *that* he is honest = IN Preposition or Subordinating conjunction
 - Yes, *that* play was nice = DT
 - You can't go *that* far = RB
- 40% of the word tokens are ambiguous

Deciding on the correct part of speech can be difficult even for people

- Mrs/NNP Shaefer/NNP never/RB got/VBD around/RP to/TO joining/VBG
particle
- 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

Part-of-speech tagging

A simple but useful form
of linguistic analysis

Part-of-speech tagging revisited

A simple but useful form
of linguistic analysis

Sources of information

- What are the main sources of information for POS tagging?
 - Knowledge of neighboring words
 - Bill saw that man yesterday
 - NNP NN DT NN NN
 - VB VB(D) IN VB NN
 - Knowledge of word probabilities
 - *man* is rarely used as a verb....
- The latter proves the most useful, but the former also helps

More and Better Features → Feature-based tagger

- Can do surprisingly well just looking at a word by itself:
 - Word the: the → DT
 - Lowercased word Importantly: importantly → RB
 - Prefixes unfathomable: un- → JJ
 - Suffixes Importantly: -ly → RB
 - Capitalization Meridian: CAP → NNP
 - Word shapes 35-year: d-x → JJ
- Then build a maxent (or whatever) model to predict tag
 - Maxent $P(t|w)$: 93.7% overall / 82.6% unknown

How to improve supervised results?

- Build better features!

PRP VBD ^{RB} IN RB IN PRP VBD .
They left as soon as he arrived .

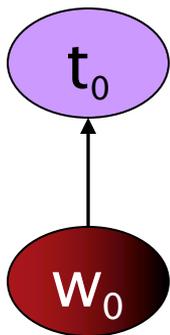
- We could fix this with a feature that looked at the next word

JJ
NNP NNS VBD VBN .
Intrinsic flaws remained undetected .

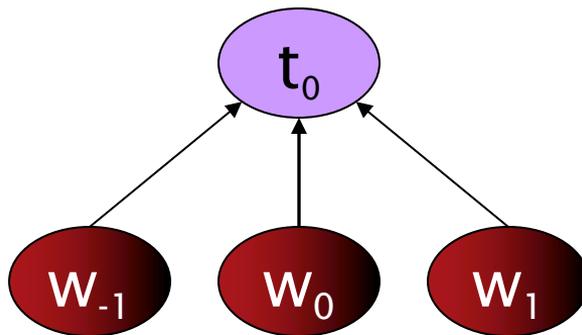
- We could fix this by linking capitalized words to their lowercase versions

Tagging Without Sequence Information

Baseline



Three Words



Model	Features	Token	Unknown
Baseline	56,805	93.69%	82.61%
3Words	239,767	96.57%	86.78%

Using words only in a straight classifier works as well as a basic (HMM or discriminative) sequence model!!

Overview: POS Tagging Accuracies

- Rough accuracies:

- Most freq tag:

~90% / ~50%

- Maxent $P(t|w)$:

93.7% / 82.6%

- Trigram HMM:

~95% / ~55%

- MEMM tagger:

96.9% / 86.9%

- Upper bound:

~98% (human agreement)

Most errors
on unknown
words

Summary of POS Tagging

One profits from models for specifying dependence on **overlapping features of the observation** such as spelling, suffix analysis, etc.

An MEMM allows integration of rich features of the observations and considers dependence with the previous word's tag, but can suffer strongly from assuming independence from following observations; this effect can be relieved by adding dependence on following words.

This additional power (of the CRF, Structured Perceptron models) has been shown to result in improvements in accuracy

The **higher accuracy** of discriminative models comes at the price of **much slower training**

Part-of-speech tagging revisited

A simple but useful form
of linguistic analysis