# CS674 Natural Language Processing

- **Last classes**
  - Noisy channel model
  - N-gram models
- **Today**
  - Part-of-speech tagging
    - introduction

# Part of speech tagging

"There are 10 parts of speech, and they are all troublesome."

-*Mark Twain*

- POS tags are also known as word classes, morphological classes, or lexical tags.

- Typically much larger than Twain's 10:
  - Penn Treebank: 45
  - Brown corpus: 87
  - C7 tagset: 146

# Part of speech tagging

- **Assign the correct part of speech (word class) to each word/token in a document**

  "The/DT planet/NN Jupiter/NNP and/CC its/PRP moons/NNS are/VBP in/IN effect/NN a/DT mini-solar/JJ system/NN ,/, and/CC Jupiter/NNP itself/PRP is/VBZ often/RB called/VBN a/DT star/NN that/IN never/RB caught/VBN fire/NN ./."

- **Needed as an initial processing step for a number of language technology applications**
  - Answer extraction in QA
  - Base step in identifying syntactic phrases for IR systems
  - Critical for word-sense disambiguation (WordNet apps)
  - Information extraction
  - …

# Why is p-o-s tagging hard?

- **Ambiguity**
  - He will race/VB the car.
  - When will the race/NOUN end?
  - The boat floated/VBN down the river **sank**.

- **Average of ~2 parts of speech for each word**

- **The number of tags used by different systems varies a lot. Some systems use < 20 tags, while others use > 400.**

# Hard for Humans

- **particle vs. preposition**
  - He talked *over* the deal.
  - He talked *over* the telephone.
- **past tense vs. past participle**
  - The horse *walked* past the barn.
  - The horse *walked* past the barn fell.
- **noun vs. adjective?**
  - The *executive* decision.
- **noun vs. present participle**
  - *Fishing* can be fun.

To obtain gold standards for evaluation, annotators rely on a set of tagging guidelines.

From Ralph Grishman, NYU

# Penn Treebank Tagset

| Tag | Description | Example | Tag | Description | Example |
|---|---|---|---|---|---|
| CC | Coordin. Conjunction | *and, but, or* | SYM | Symbol | +,%, & |
| CD | Cardinal number | *one, two, three* | TO | "to" | *to* |
| DT | Determiner | *a, the* | UH | Interjection | *ah, oops* |
| EX | Existential 'there' | *there* | VB | Verb, base form | *eat* |
| FW | Foreign word | *mea culpa* | VBD | Verb, past tense | *ate* |
| IN | Preposition/sub-conj | *of, in, by* | VBG | Verb, gerund | *eating* |
| JJ | Adjective | *yellow* | VBN | Verb, past participle | *eaten* |
| JJR | Adj., comparative | *bigger* | VBP | Verb, non-3sg pres | *eat* |
| JJS | Adj., superlative | *wildest* | VBZ | Verb, 3sg pres | *eats* |
| LS | List item marker | *1, 2, One* | WDT | Wh-determiner | *which, that* |
| MD | Modal | *can, should* | WP | Wh-pronoun | *what, who* |
| NN | Noun, sing. or mass | *llama* | WP$ | Possessive wh- | *whose* |
| NNS | Noun, plural | *llamas* | WRB | Wh-adverb | *how, where* |
| NNP | Proper noun, singular | *IBM* | $ | Dollar sign | $ |
| NNPS | Proper noun, plural | *Carolinas* | # | Pound sign | # |
| PDT | Predeterminer | *all, both* | " | Left quote | (' or ") |
| POS | Possessive ending | *'s* | " | Right quote | (' or ") |
| PP | Personal pronoun | *I, you, he* | ( | Left parenthesis | ( [, (, {, <) |
| PP$ | Possessive pronoun | *your, one's* | ) | Right parenthesis | ( ], ), }, >) |
| RB | Adverb | *quickly, never* | , | Comma | , |
| RBR | Adverb, comparative | *faster* | . | Sentence-final punc | (. ! ?) |
| RBS | Adverb, superlative | *fastest* | : | Mid-sentence punc | (: ; ... – -) |
| RP | Particle | *up, off* | | | |

# Among easiest of NLP problems

- **State-of-the-art methods achieve ~97% accuracy.**
- **Simple heuristics can go a long way.**
  - ~90% accuracy just by choosing the most frequent tag for a word (MLE)
  - To improve reliability: *need to use some of the local context.*
- **But defining the rules for special cases can be time-consuming, difficult, and prone to errors and omissions**

# Approaches

1. **rule-based**: involve a large database of hand-written disambiguation rules, e.g. that specify that an ambiguous word is a noun rather than a verb if it follows a determiner.
2. **probabilistic**: resolve tagging ambiguities by using a training corpus to compute the probability of a given word having a given tag in a given context.

   - HMM tagger, Maximum Likelihood Tagger
3. **hybrid corpus-/rule-based**: E.g. transformation-based tagger (Brill tagger); learns symbolic rules based on a corpus.
4. **ensemble methods**: combine the results of multiple taggers.