# CS4740 Natural Language Processing

- **Last classes**
  - Word-sense disambiguation
- **Today**
  - WSD assignment (review)
  - 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/PPS 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 Question Answering systems
  - 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/ VBD.
  - The boat floated/ VBD down the river.
  - 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