Admin

- Assignment 2

Simplified View of Linguistics

- Phonology: /waddyasai/
- Morphology: /waddyasai/ \rightarrow what did you say
- Syntax: what did you say \rightarrow say subj obj
- Semantics: say subj obj you what \rightarrow \{ I, x, say(you, x) \}
- Discourse: what did you say \rightarrow what did you say
Morphology

- What is morphology?
  - study of the internal structure of words
  - morphology: word's jumping

- Why might this be useful for NLP?
  - generalization (runs, running, runner are related)
  - additional information (it's plural, past tense, etc)
  - allows us to handle words we've never seen before
  - smoothing?

New words

  - 300K unique words

- New words seen on Dec 31
  - compounds: prenatal-care, publicly-funded, channel-switching, …
  - New words:
    - dumbbells, groveled, fuzzier, oxidized, ex-presidency, puppetry, boulderlike, over-emphasized, antiprejudice

Morphology basics

- Words are built up from morphemes
  - stems (base/main part of the word)
  - affixes
    - prefixes
      - precedes the stem
    - suffixes
      - follows the stem
    - infixes
      - inserted inside the stem
    - circumfixes
      - surrounds the stem
  - Examples?

Morpheme examples

- prefix
  - circum- (circumnavigate)
  - dis- (dislike)
  - mis- (misunderstood)
  - com-, de-, dis-, in-, re-, post-, trans-, …

- suffix
  - -able (movable)
  - -ance (resistance)
  - -ly (quickly)
  - -ion, -ness, -ate, -ful, …
Morpheme examples

- *infix*
  - *-fucking-* (cinder-fucking-rella)
  - more common in other languages
- *circumfix*
  - doesn’t really happen in English
  - *a-* *-ing*
    - a-running
    - a-jumping

Agglutinative: Finnish

talo ‘the-house’
talo-ni ‘my house’
talo-ssa ‘in the-house’
talo-ssa-ni ‘in my house’
talo-i-ssa ‘in the-houses’
talo-i-ssa-ni ‘in my houses’

kaup-pa ‘the-shop’
kaup-pa-ni ‘my shop’
kaup-a-ssa ‘in the-shop’
kaup-a-ssa-ni ‘in my shop’
kaup-o-i-ssa ‘in the-shops’
kaup-o-i-ssa-ni ‘in my shops’

Stemming (baby lemmatization)

- Reduce a word to the main morpheme

Stemming example

This is a poorly constructed example using the Porter stemmer.

This is a poorly constructed example using the Porter stemmer.

http://maya.cs.depaul.edu/~classes/ds575/porter.html
(or you can download versions online)
Porter’s algorithm (1980)

- Most common algorithm for stemming English
  - Results suggest it’s at least as good as other stemming options
- Multiple sequential phases of reductions using rules, e.g.
  - sses → ss
  - ies → i
  - ional → ate
  - ional → tion
- [http://tartarus.org/~martin/PorterStemmer/](http://tartarus.org/~martin/PorterStemmer/)

What is Syntax?

- Study of structure of language
  - Examine the rules of how words interact and go together
- Rules governing grammaticality
- I will give you one perspective
  - no single correct theory of syntax
  - still an active field of research in linguistics
  - we will often use it as a tool/stepping stone for other applications

Structure in language

The man __all the way home.

- what are some examples of words that can/can’t go here?

The man __all the way home.

- why can’t some words go here?
Structure in language

The man flew all the way home.

- Language is bound by a set of rules
- It’s not clear exactly the form of these rules, however, people can generally recognize them
- This is syntax!

Syntax != Semantics

Colorless green ideas sleep furiously.

- Syntax is only concerned with how words interact from a grammatical standpoint, not semantically

Parts of speech

What are parts of speech (think 3rd grade)?

Parts of speech

Parts of speech are constructed by grouping words that function similarly:
- with respect to the words that can occur nearby
- and by their morphological properties

The man ______ all the way home.

ran  forgave  ate  drove  hid  learned  hurt
integrated  programmed  shot  showed  slept  understood  voted
washed  warned  walked  spiked  survived  read  recorded
Parts of speech

- What are the English parts of speech?
- 8 parts of speech?
  - Noun (person, place or thing)
  - Verb (actions and processes)
  - Adjective (modify nouns)
  - Adverb (modify verbs)
  - Preposition (on, in, by, to, with)
  - Determiners (a, an, the, what, which, that)
  - Conjunctions (and, but, or)
  - Particle (off, up)

English parts of speech

- Brown corpus: 87 POS tags
- Penn Treebank: ~45 POS tags
  - Derived from the Brown tagset
  - Most common in NLP
  - Many of the examples we’ll show us this one
- British National Corpus (C5 tagset): 61 tags
  - C6 tagset: 148
  - C7 tagset: 146
  - C8 tagset: 171

Brown tagset


English Parts of Speech

- Noun (person, place or thing)
  - Singular (NN): dog, fork
  - Plural (NNS): dogs, forks
  - Proper (NNP, NNPS): John, Springfields
  - Personal pronoun (PRP): I, you, he, she, it
  - Wh-pronoun (WP): who, what
- Verb (actions and processes)
  - Base, infinitive (VB): eat
  - Past tense (VBD): ate
  - Gerund (VBG): eating
  - Past participle (VBN): eaten
  - Non 3rd person singular present tense (VBP): eat
  - 3rd person singular present tense: (VBZ): eats
  - Modal (MD): should, can
  - To (TO): to its eat
English Parts of Speech (cont.)

- Adjective (modify nouns)
  - Basic (JJ): red, tall
  - Comparative (JJR): redder, taller
  - Superlative (JJS): reddest, tallest
- Adverb (modify verbs)
  - Basic (RB): quickly
  - Comparative (RBR): quicker
  - Superlative (RBS): quickest
- Preposition (IN): on, in, by, to, with
- Determiner:
  - Basic (DT): a, an, the
  - WH-determiner (WDT): which, that
- Coordinating Conjunction (CC): and, but, or,
- Particle (RP): off (took off), up (put up)

Closed vs. Open Class

- **Closed class** categories are composed of a small, fixed set of grammatical function words for a given language.
  - Pronouns, Prepositions, Modals, Determiners, Particles, Conjunctions
- **Open class** categories have large number of words and new ones are easily invented.
  - Nouns (Googler, futon, iPad), Verbs (Google, futoning), Adjectives (geeky), Adverb (chompingly)

Part of speech tagging

- Annotate each word in a sentence with a part-of-speech marker
- Lowest level of syntactic analysis

John saw the saw and decided to take it to the table.

**NNP** VBD DT NN CC VBD TO VB PRP IN DT NN

Ambiguity in POS Tagging

- I like candy.
  - **VBP** (verb, non-3rd person, singular, present)
- Time flies like an arrow.
  - **IN** (preposition)

Does “like” play the same role (POS) in these sentences?
Ambiguity in POS Tagging

- I bought it at the shop around the corner.
  - IN (preposition)
- I never got around to getting the car.
  - RP (particle... on, off)
- The cost of a new Prius is around $25K.
  - RB (adverb)

Does “around” play the same role (POS) in these sentences?

---

Ambiguity in POS tagging

- Like most language components, the challenge with POS tagging is ambiguity.
- Brown corpus analysis:
  - 11.5% of word types are ambiguous (this sounds promising)
  - 40% of word appearance are ambiguous
  - Unfortunately, the ambiguous words tend to be the more frequently used words

---

How hard is it?

- If I told you I had a POS tagger that achieved 90% would you be impressed?
  - Shouldn’t be… just picking the most frequent POS for a word gets you this
- What about a POS tagger that achieves 93.7%?
  - Still probably shouldn’t be… only need to add a basic module for handling unknown words
- What about a POS tagger that achieves 100%?
  - Should be suspicious… humans only achieve ~97%
  - Probably overfitting

---

POS Tagging Approaches

- Rule-Based: Human crafted rules based on lexical and other linguistic knowledge
- Learning-Based: Trained on human annotated corpora like the Penn Treebank
  - Statistical models: Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF), log-linear models, support vector machines
  - Rule learning: Transformation Based Learning (TBL)
- The book discusses some of the more common approaches
- Many publicly available:
  - [http://www.coli.uni-saarland.de/~thorsten/tnt/](http://www.coli.uni-saarland.de/~thorsten/tnt/)