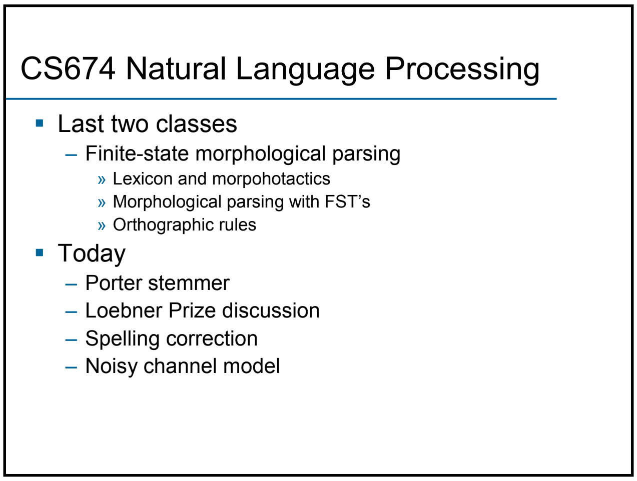
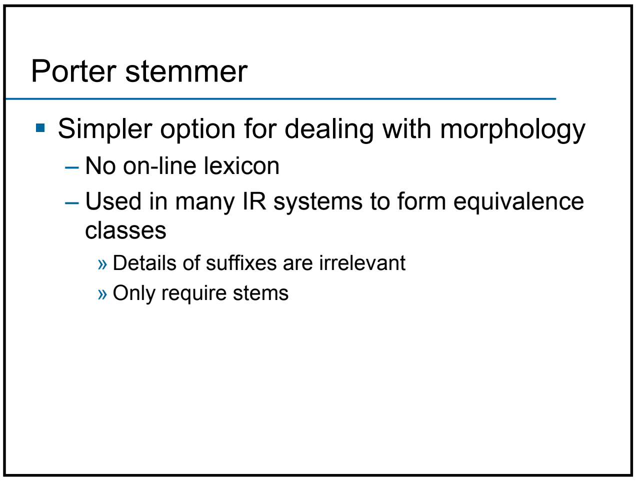


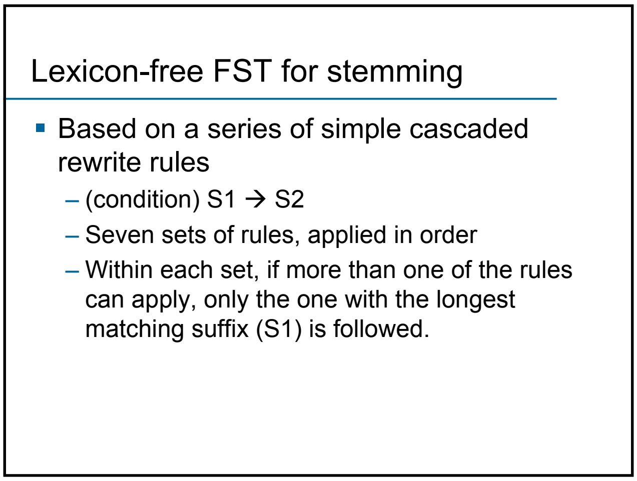
CS674 Natural Language Processing

- Last two classes
  - Finite-state morphological parsing
    - Lexicon and morpohotactics
    - Morphological parsing with FST's
    - Orthographic rules
- Today
  - Porter stemmer
  - Loebner Prize discussion
  - Spelling correction
  - Noisy channel model



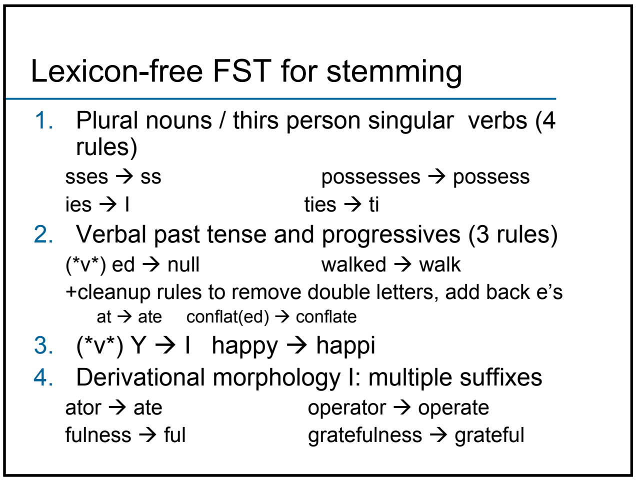
Porter stemmer

- Simpler option for dealing with morphology
  - No on-line lexicon
  - Used in many IR systems to form equivalence classes
    - Details of suffixes are irrelevant
    - Only require stems



Lexicon-free FST for stemming

- Based on a series of simple cascaded rewrite rules
  - (condition) S1 → S2
  - Seven sets of rules, applied in order
  - Within each set, if more than one of the rules can apply, only the one with the longest matching suffix (S1) is followed.



Lexicon-free FST for stemming

1. Plural nouns / thirs person singular verbs (4 rules)

   sses → ss   possesses → possess
   ies → I   ties → ti

2. Verbal past tense and progressives (3 rules)

   (*v*) ed → null   walked → walk
   +cleanup rules to remove double letters, add back e's
   at → ate   conflat(ed) → conflate

3. (*v*) Y → I   happy → happi
4. Derivational morphology I: multiple suffixes

   ator → ate   operator → operate
   fulness → ful   gratefulness → grateful

Lexicon-free FST for stemming

5. Derivational morphology II: more multiple suffixes

   ful → null grateful → grate

6. Derivational morphology III: single suffixes

   ous → null analogous → analog

7. Cleanup (3 rules)

   (m>1) e → null probate → probat; rate→ rate

   dropping double letters controll→ control

Sample output

- **O'Neill Criticizes Europe on Grants**
  Treasury Secretary Paul O'Neill expressed irritation Wednesday that European countries have refused to go along with a U.S. proposal to boost the amount of direct grants rich nations offer poor countries.
  The Bush administration is pushing a plan to increase the amount of direct grants the World Bank provides the poorest nations to 50 percent of assistance, reducing use of loans to these nations.
- o'neill **criticizes** europe **grants** **treasury** **secretary** paul o'neill **expressed** **irritation** **european** **countries** **refused** US **proposal** boost direct **grants** rich **nations** poor **countries**
  bush **administration** **pushing** plan **increase** amount direct **grants** world bank **poorest** **nations** **assistance** **loans** **nations**

Loebner Prize papers: critiques

- Comments on the Turing Test
- Comments on Loebner's response
  - Inadequate
  - Subsequent runnings of the event backed some of Shieber's complaints (Chavdar)
  - Tone of the response (Doug, Chester) vs. tone of the editorial (Oren, Claire)
- Restrictions
  - for the event
    - Problematic (Doug, Chester)
  - vs. restrictions in evaluating NLP
    - Not a problem (Ves, Oren)
  - Engineering vs. science

Rest of Today

- Porter stemmer
- Loebner Prize discussion
- Spelling correction
- Noisy channel model

Detection and correction of spelling errors

- Frequency of spelling errors in human typed text varies from
  - 0.05% of the words in carefully edited newswire, to
  - 38% in difficult applications like telephone directory lookup
- Optical character recognition
  - Higher error rates than human typists
  - Make different kinds of errors, “D”→ “O”; “ri”→”n”
- On-line handwriting recognition

Types of spelling correction

- Non-word error detection
  - Detecting spelling errors that result in non-words
    - » *graffe* → *giraffe*
- Isolated-word error correction:
  - Correcting spelling errors that result in non-words
    - » Correcting *graffe* to *giraffe*, but looking only at the word in isolation

Kukich, 1992

Types of spelling correction

- Context-dependent error detection and correction
  - Using the context to help detect and correct spelling errors
  - Some of these may accidentally result in an actual word (**real-word errors**)
    - » Typographical errors
      - ◆e.g. *there* for *three*
    - » Homonym or near-homonym
      - ◆e.g. *dessert* for *desert*, or *piece* for *peace*

Kukich, 1992

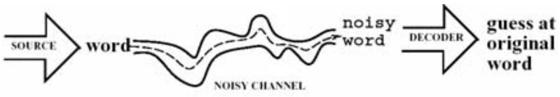
Detecting non-word errors

- Use a dictionary
- Usually include models of morphology
- For other types of spelling correction, we’ll need a model of spelling variation.

Probabilistic transduction

- surface representation → lexical representation
- sequence of letters in a mis-spelled word → sequence of letters in correctly spelled words
  - *acress → actress, cress, acres*
- string of symbols representing the pronunciation of a word in context → string of symbols representing the dictionary pronunciation
  - [er] → *her, were, are, their, your*
  - exacerbated by **pronunciation variation**
    - » *the* pronounced as THEE or THUH
    - » some aspects of this variation are systematic, like spelling error patterns

Noisy channel model



- Channel introduces noise which makes it hard to recognize the true word.
- **Goal:** build a model of the channel so that we can figure out how it modified the true word…so that we can recover it.

Decoding algorithm

- Special case of **Bayesian inference**
  - Bayesian classification
    - » Given observation, determine which of a set of classes it belongs to.
    - » Observation
      - ◆string of phones or string of letters
    - » Classify into
      - ◆words

Pronunciation example

- Given a string of phones, e.g. [ni], determine which word corresponds to this string of phones
  - Consider all words in the vocabulary, *V*
  - Select the single word, *w*, such that *P (word|observation)* is highest

Computing $P(w|O)$

- Use Bayes’ rule to transform into a product of two probabilities, each of which is easier to compute than $P(w|O)$