CS674 Natural Language Processing

- Last class
  - Porter stemmer
  - Loebner Prize discussion
- Today
  - Spelling correction
  - Noisy channel model
  - Bayesian approach to spelling correction

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 such that *P (word|observation)* is highest

$$\hat{w} = \arg\max_{w \in V} P(w \mid O)$$

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)*

$$P(x \mid y) = \frac{P(y \mid x) \quad P(x)}{P(y)}$$

$$\hat{w} = \arg\max_{w \in V} \frac{\overbrace{P(O \mid w)}^{\text{likelihood}} \quad \overbrace{P(w)}^{\text{prior}}}{P(O)}$$

Applying the Bayesian Method to Spelling

- Focus on non-word errors
- First suggested by Kernighan et al. (1990)
- Two-stage approach
  - Propose candidate corrections
  - Score the candidates

Proposing candidate corrections

- Simplifying assumption: the correct word will differ from the misspelling by a ***single*** insertion, deletion, substitution, or transposition
  - Handles most spelling errors in human typed text
- Generate the candidates by applying any single transformation that results in a word in an on-line dictionary

Candidate corrections for *acress*

Error	Correction	Transformation: Correct Letter	Error Letter	Position (Letter #)	Type
acress	actress	t	–	2	deletion
acress	cress	–	a	0	insertion
acress	caress	ca	ac	0	transposition
acress	access	c	r	2	substitution
acress	across	o	e	3	substitution
acress	acres	–	2	5	insertion
acress	acres	–	2	4	insertion

Score the corrections

- Let *c* range over the set *C* of candidate corrections
- Let *t* represent the typo
- Select the most likely correction:

$$\hat{c} = \arg\max_{c \in C} \overbrace{P(t \mid c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}$$

Computing the prior

- $P(c) = \frac{C(c)}{N}$
- Problem: counts of 0
- Solution: *smoothing*

$$P(c) = \frac{C(c) + 0.5}{N + 0.5|V|}$$

Resulting prior probabilities

c	freq(c)	P(c)
actress	1343	.0000315
cress	0	.000000014
caress	4	.0000001
access	2280	.000058
across	8436	.00019
acres	2879	.000065

Computing the likelihood

- Computing the likelihood term P(t|c) exactly is an unsolved problem
- Can estimate its value
  - The most important factors predicting an insertion, deletion, transposition are simple local factors
- Simple method: estimate the number of times that a single-letter error occurs in some large corpus of errors
  - E.g. estimate P(*acress* | *across*) using the number of times that *e* was substituted for *o*