Agenda

- HW5 due today
- Schedule changes
- Questions, comments, concerns?
- More text normalization
  - Tokenization
  - Abbreviations
  - Misspellings

Examples of Non-Standard Words (NSW)

- Numbers
  - 132 goats
  - 132 Park St
  - extension 132
- Acronyms/Letter sequences
  - NATO
  - UFO
- Abbreviations
  - Blvd.
  - St.
  - wdbf (or wdb, wdbf, wdbf)
- Mixed examples
  - msdos, cdrom
  - named
  - w/11", w/mahog.

Challenges with Text Normalization

- Genre/topic dependence
  - named is probably the ordinary word named in most cases; probably name D (= name daemon) in discussions of internet domain servers
  - BA is probably bath(room) in real-estate classifieds; probably just B A in most other contexts
- Enumeration and selection
  - BA: bathroom, B A
  - lv: living (Formal lv rm), leave (lv msg)
  - (Think of this as a kind of pronunciation modeling problem)
- What to expand
  - Do you read IMHO as in my humble opinion or I M H O?

Distribution of Examples

- In NANTC (North American News Text Corpora) from 121,464 NSWs

<table>
<thead>
<tr>
<th>major type</th>
<th>minor type</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>numeric</td>
<td>number</td>
<td>26%</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>7%</td>
</tr>
<tr>
<td></td>
<td>ordinal</td>
<td>3%</td>
</tr>
<tr>
<td>alphabetic</td>
<td>as word</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>as letters</td>
<td>12%</td>
</tr>
<tr>
<td></td>
<td>as abbrev</td>
<td>2%</td>
</tr>
</tbody>
</table>
Motivation for Text Normalization

- Text-to-Speech (TTS)
- Universal Access
- Speech Recognition (ASR)
  - Increase the useful set of textual training materials for ASR systems
    - e.g., Internet Relay Chat (IRC) for conversational LM training
  - Improved pronunciation dictionaries for ASR
  - Named Entity Recognition
    - Many named entities are referred to with acronyms (e.g., GWB?); expand acronyms into their full renditions
  - Parsing
  - Information Extraction
  - Machine Translation

AT&T's Text Normalizer: Example 1

- Last Thursday, G. Gordon Liddy had the so-called confidential witness live on his radio show, C.W., who discovered Foster’s body in Fort Myer Park, Va., just across the Potomac River from Washington, at 5:15 pm on July 20, 1963, and several times with emphasis that he said the FBI that Foster’s hands were palms up, thumbs out and there was no gun in either hand.
- Output of AT&T/Bell Labs Preprocessor (4.2% error rate):

  
  AT&T's Text Normalizer: Example 2

  - 569 Sutton Place Ave, SPT 1409 ST HH, 15th, L-Shaped LR, 2 pl Open Cary, Gar, R3, Mid four hundred dollars X’s Thompson Kase box 503-8000 $
  - Output of AT&T/Bell Labs Preprocessor (4.2% error rate):
    - 569 Sutton Place Ave, SPT 1409 ST HH, 15th, L-Shaped LR, 2 pl Open Cary, Gar, R3, Mid four hundred dollars X’s Thompson Kase box 503-8000 $

A More General Approach

Treat as a language modeling problem:

1. Robust expansion model to enumerate possible ways of reading a NSW
   - Assumes an NSW has been identified, but this could also be part of the task
2. Language model to select among the alternatives

Two components of text normalization

- Given a string of characters in a text, predict a set of "normal" words that might correspond to the text sequence
  - Assume the "non-standard" words have been identified, but identifying these could be part of the task
  - A reasonable set of possible normal words
  - Can also apply to word sequences
  - Select the correct "normal" word, given a particular context
Text Normalization Components

• Expansion
  - 123 = one hundred (and) twenty three
  - 123 = one twenty three
  - 123 = one two three

• Selection
  - 123 goats → one hundred twenty three goats
  - 123 Park St → one twenty three Park street
  - extension 123 → extension one two three

Concrete Example from English

Consider a machine that maps between digit strings and their reading as number names in English.

30,294,005,179,018,903.56 → thirty quadrillion, two hundred and ninety four trillion, five billion, one hundred seventy nine million, eighteen thousand, nine hundred three, point five six

FSTs for Text Normalization: Digit to Number-Name Translation

• Factor digit string:
  - 123 → 1 · 10² + 2 · 10¹ + 3

• Translate factors into number names:
  - 10² → hundred
  - 2 · 10¹ → twenty
  - 1 · 10¹ + 3 → thirteen

• Languages vary on how extensive these lexicons are
  - Some (e.g., Chinese) have very regular (hence very simple) number name systems;
  - Others (e.g., Urdu/Hindi) have a large set of number names with a name for almost every number from 1 to 100.
  - Each of these steps can be accomplished with FSTs

Task: Expand Abbreviations

- CUST RCVD LTR CNCRNG LOCAL SRVC
- VISIT NECESSARY BUT CST STILL HAS PAC BELL SRVC ON OLD TN AT RESIDENCE
- ORDERD CALLING CRDS PER CSR RQST
- Cust wanted to know if we currently had 4.95 pp Adv we do not
- cust still has att s/w on comp he is going to be moving to PA in a mth and wants to know if he can reformat this asct
- 1st att, left msg for CB from Lynda, will wait for call
- CUST REQUESTD CHANGE IN HUNTING, FOLLOW ORDER. NO CSR FOUND. CUST WITH RESELLER ALEGIANCE.

Define "Abbreviations"

• Any word that is shortened from its normal spelling, but that should be read as if it were spelled in full
- Under this definition:
  - cust and mth are abbreviations since they are clearly to be read customer, month
  - NATO, UN, CSR are not abbreviations since they are standardly read as words (“acronyms”) or sequences of letters
  - Some terms (such as LD: long distance) may have become pretty standard in the domain-specific lexicon and thus should not be treated as abbreviations
Normalization
cci vm not working has not fully compiled xfer to svc
cci voicemail not working has not fully completed transfer to service

One approach
Large script with lots of rules:
- s/ AN ADV / AN ADVERTISEMENT /
- s/ 2 ADVS* / TO ADVISE /
- s/ TO ADVS* / TO ADVISE /
- s/ ADVS*D* / ADVISED /
- s/ AMER[*] / AMERICA /
- s/ AMT / AMOUNT /
- Cf. U Penn Linguistic Data Consortium's "Text Conditioning Tools"

Problems with approach
• How many ways is customer spelled in dataset?
1. cns decnstd customer disconnected
2. cns spent customer spent
3. cns slg customer calling
4. cns stng customer calling
5. cns cmd customer cmd
6. cns cmd customer cmd
7. cns wns customer wants
8. cns wnt customer wanted
9. cns wnt customer wanted
10. cns req customer request
11. cns want customer wanted
12. cns called customer called
13. cns called customer called
14. cns asked customer asked
15. cns car customer cars
16. cns call customer call
17. cns call customer call
18. cns stng customer changes
19. cns spc customer spent

Abbreviation Expansion
• Problem: given a previously unseen abbreviation, how do you use corpus-internal evidence to find the proper expansion into a standard word?
- Example:
  - cus went info on services and chrgs
  - Elsewhere in corpus:
    • ... customer wants ...
    • ... wants info on vmail ...

A Source-Channel Language Model Approach
\[ \hat{w} = \text{argmax}_{w \in T} p(o|\hat{t}, w)p(\hat{t}|w)p(w) \]
where:
• o is the observed text
• w are the underlying words
• t are the tags
  (in this case, tags = "abbreviate" and "don't abbreviate")

WFST-based Implementation
\[ T' = \pi_2(\text{ShortestPath}(T \circ A^{-1} \circ L)) \]
where:
• T is text
• T' is normalized text
• A is the abbreviation model
• L is the language model
Processing Steps

- Pre-process text ("splitter")
- Collect possible abbreviations and their possible expansions; use a *stoplist* of things not to expand
- Train a language model on "clean" text
- Normalize text

Language Model Training

- Train a word trigram model with standard Katz backoff on "sanitized" text
  - cust business acct – tms to business office
  - <ABBR> business <ABBR> <PUNC> <ABBR> to business office
- Implement using standard LM algorithms

WFST-based Implementation

\[ T = \sigma_2(\text{ShortestPath}(T \circ A^{-1} \circ L)) \]

Example Normalizations

cust clid 2 by clid id blk rmv sm local form

- cust called 2 have caller id blk rmv sm local form
- cust clid to verify insde wre / i called his near move on accident / cust now wants to ploc to authr ccnpy
- cust called to verify inside wire / i cancelled his near move on accident / cust now wants to ploc to authr ccnpy
- cust no longer wants id on acct
- vphd chgs - cust stated he w/ pay 26.45 & then w/ cancel his svc w/ att
- explained charges & cust stated he will pay 26.45 & then will cancel his service w/ att

Is Text Normalization Useful?

- Obviously needed if you want to read the text
- May be needed for searching for a particular phrase (regardless of how it's spelled)
- Extrinsic evaluation?
  - Text classification

Text Classification Task

- Classify UNE-P RAMP comments into 26 different categories:
- Use Bosnian (Schapire and Singer, 2000)
- Train on 30K examples, test on 1000
Utility of Text Normalization

Agenda: Summary

• HW5 due today
• Schedule changes
• Questions, comments, concerns?
• More text normalization
  • Tokenization
  • Abbreviations
  • Misspellings