Computational Linguistics 1

CMSC/LING 723, LBSC 744

Kristy Hollingshead Seitz
Institute for Advanced Computer Studies
University of Maryland

Lecture 21: 17 November 2011

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.
  - wbfpl (or wbf, wbfp, wbfpl)
- 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

major type	minor type	%
numeric	number	26%
	year	7%
	ordinal	3%
alphabetic	as word	30%
	as letters	12%
	as abbrev	2%

NSW Classification

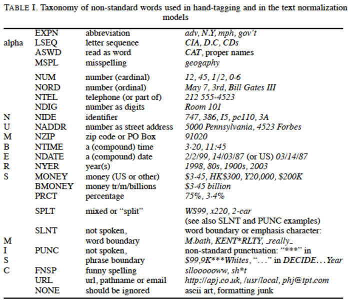


TABLE I. Taxonomy of non-standard words used in hand-tagging and in the text normalization models

	EXPN	abbreviation	*adv, N.Y, mph, gov't*
alpha	LSEQ	letter sequence	*CIA, D.C, CDs*
	ASWD	read as word	*CAT*, proper names
	MSPL	misspelling	*geogaphy*
	NUM	number (cardinal)	*12, 45, 1/2, 0·6*
	NORD	number (ordinal)	*May 7, 3rd, Bill Gates III*
	NTEL	telephone (or part of)	*212 555-4523*
	NDIG	number as digits	*Room 101*
N	NIDE	identifier	*747, 386, I5, pc110, 3A*
U	NADDR	number as street address	*5000 Pennsylvania, 4523 Forbes*
M	NZIP	zip code or PO Box	*91020*
B	NTIME	a (compound) time	*3·20, 11:45*
E	NDATE	a (compound) date	*2/2/99, 14/03/87* (or US) *03/14/87*
R	NYER	year(s)	*1998, 80s, 1900s, 2003*
S	MONEY	money (US or other)	*$3·45, HK$300, Y20,000, $200K*
	BMONEY	money tr/m/billions	*$3·45 billion*
	PRCT	percentage	*75%, 3·4%*
	SPLT	mixed or "split"	*WS99, x220, 2-car* (see also SLNT and PUNC examples)
	SLNT	not spoken,	word boundary or emphasis character:
M		word boundary	*M.bath, KENT*RLTY, _really_*
I	PUNC	not spoken,	non-standard punctuation: "***" in
S		phrase boundary	*$99,9K***Whites*, "…" in *DECIDE…Year*
C	FNSP	funny spelling	*slloooooww, sh*t*
	URL	url, pathname or email	*http://apj.co.uk, /usr/local, phj@tpt.com*
	NONE	should be ignored	ascii art, formatting junk

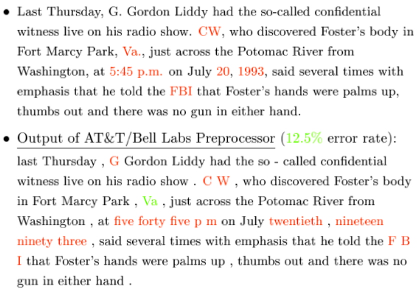
Motivation for Text Normalization

- 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

Text Normalization: (Previous) State-of-the-Art

- TTS
  - Table lookup
  - Specialized rules
    - St. → *Saint* if following word is capitalized
  - Trainable models for particular (classes of) ambiguous cases
- ASR
  - Use TTS "pre-processors"
  - Specialized ad-hoc scripts
    - e.g., the LDC ARPA Continuous Speech Recognition text-conditioning tools

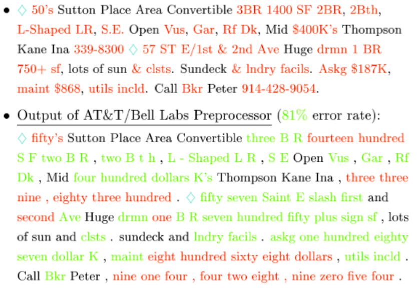
AT&T's Text Normalizer: Example 1



- Last Thursday, G. Gordon Liddy had the so-called confidential witness live on his radio show. CW, who discovered Foster's body in Fort Marcy Park, Va., just across the Potomac River from Washington, at 5:45 p.m. on July 20, 1993, said several times with emphasis that he told 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 (12.5% error rate):
  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 Marcy Park , Va , just across the Potomac River from Washington , at five forty five p m on July twentieth , nineteen ninety three , said several times with emphasis that he told the F B I that Foster's hands were palms up , thumbs out and there was no gun in either hand .

Slide from Richard Sproat, JHU/CLSP Workshop'99

AT&T's Text Normalizer: Example 2



- ◇ 50's Sutton Place Area Convertible 3BR 1400 SF 2BR, 2Bth, L-Shaped LR, S.E. Open Vus, Gar, Rf Dk, Mid $400K's Thompson Kane Ina 339-8300 ◇ 57 ST E/1st & 2nd Ave Huge drmn 1 BR 750+ sf, lots of sun & clsts. Sundeck & lndry facils. Askg $187K, maint $868, utils incld. Call Bkr Peter 914-428-9054.
- Output of AT&T/Bell Labs Preprocessor (81% error rate):
  ◇ fifty's Sutton Place Area Convertible three B R fourteen hundred S F two B R , two B t h , L - Shaped L R , S E Open Vus , Gar , Rf Dk , Mid four hundred dollars K's Thompson Kane Ina , three three nine , eighty three hundred . ◇ fifty seven Saint E slash first and second Ave Huge drmn one B R seven hundred fifty plus sign sf , lots of sun and clsts . sundeck and lndry facils . askg one hundred eighty seven dollar K , maint eight hundred sixty eight dollars , utils incld . Call Bkr Peter , nine one four , four two eight , nine zero five four .

Slide from Richard Sproat, JHU/CLSP Workshop'99

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*

FSTs for Text Normalization: Digit to Number-Name Translation

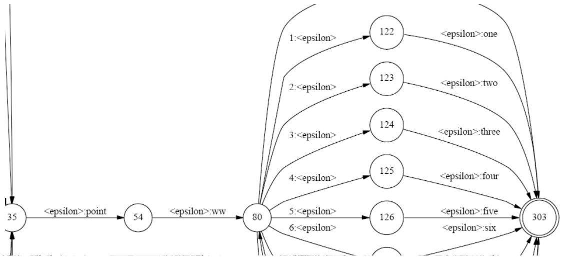
- Factor digit string:
  - *123* → $\mathbf{1 \cdot 10^2 + 2 \cdot 10^1 + 3}$
- Translate factors into number names:
  - $\mathbf{10^2}$ → *hundred*
  - $\mathbf{2 \cdot 10^1}$ → *twenty*
  - $\mathbf{1 \cdot 10^1 + 3}$ → *thirteen*
- Languages vary on how extensive these lexicons are
  - Some (e.g. 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

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

566 states and 1492 arcs

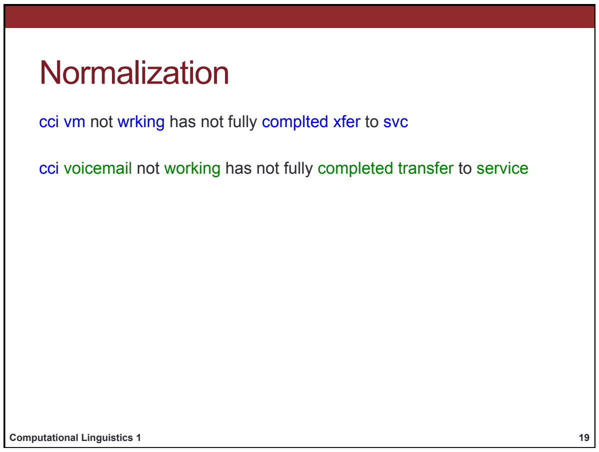


Task: Expand Abbreviations

- `CUST RCVD LTTR CNCRNG LOCAL SRVC`
- `VISIT NECESSARY BUT CST STILL HAS PAC BELL SRV 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 at&t s/w on comp he is going to be moving to PA in a mth and wants to know if he can reactivate this acct`
- `1st att, left mssg for CB from Lynda, will wait for call`
- `CUST REQUESTD CHANGE IN HUNTING, FOLLOW ORDER. NO CSR FOUND. CUST WITH RESELLER ALEGIANCE.`

Define "Abbreviations"

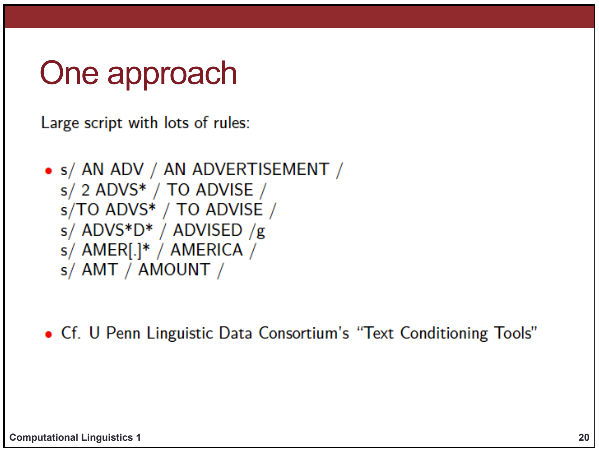
- Any word that is shortened from its normal spelling, but that should be read as if it were spelled in full
- Under this definition:
  - *cust* and *mth* are abbreviations since they are clearly to be read *customer, month*
  - *NATO, UN, CSR* are not abbreviations since they are standardly read as 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 wrking has not fully complted xfer to svc

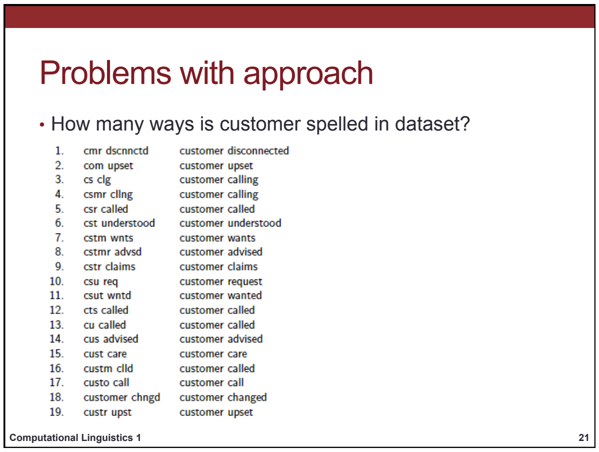
cci voicemail not working has not fully completed transfer to service



One approach

Large script with lots of rules:

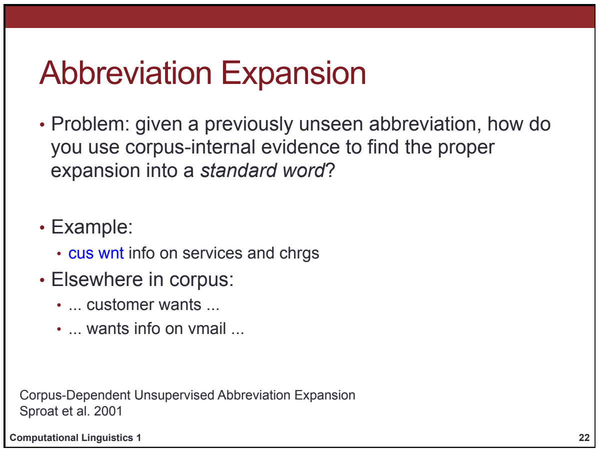
- s/ AN ADV / AN ADVERTISEMENT /
  s/ 2 ADVS* / TO ADVISE /
  s/TO ADVS* / TO ADVISE /
  s/ ADVS*D* / ADVISED /g
  s/ AMER[.]* / AMERICA /
  s/ AMT / AMOUNT /
- Cf. U Penn Linguistic Data Consortium's "Text Conditioning Tools"



Problems with approach

- How many ways is customer spelled in dataset?

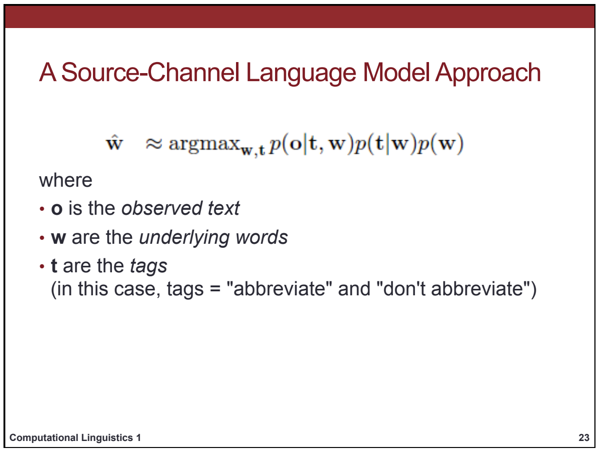
1.	cmr dscnnctd	customer disconnected
2.	com upset	customer upset
3.	cs clg	customer calling
4.	csmr cllng	customer calling
5.	csr called	customer called
6.	cst understood	customer understood
7.	cstm wnts	customer wants
8.	cstmr advsd	customer advised
9.	cstr claims	customer claims
10.	csu req	customer request
11.	csut wntd	customer wanted
12.	cts called	customer called
13.	cu called	customer called
14.	cus advised	customer advised
15.	cust care	customer care
16.	custm clld	customer called
17.	custo call	customer call
18.	customer chngd	customer changed
19.	custr upst	customer upset



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 wnt info on services and chrgs
- Elsewhere in corpus:
  - ... customer wants ...
  - ... wants info on vmail ...

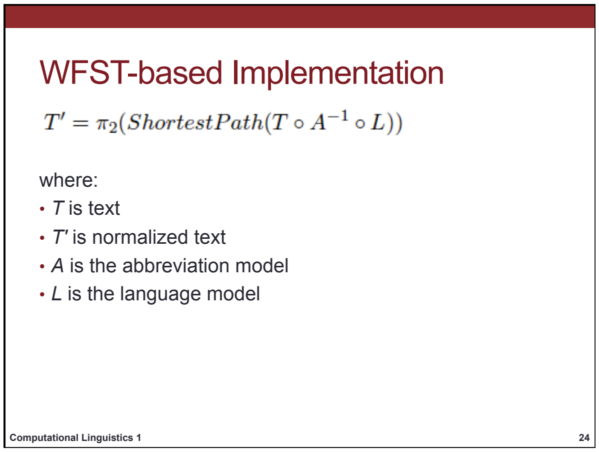
Corpus-Dependent Unsupervised Abbreviation Expansion
Sproat et al. 2001



A Source-Channel Language Model Approach

$$\hat{\mathbf{w}} \approx \mathrm{argmax}_{\mathbf{w},\mathbf{t}}\, p(\mathbf{o}|\mathbf{t},\mathbf{w})p(\mathbf{t}|\mathbf{w})p(\mathbf{w})$$

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

$$T' = \pi_2(ShortestPath(T \circ A^{-1} \circ L))$$

where:
- *T* is text
- *T'* is normalized text
- *A* is the abbreviation model
- *L* is the language model

Processing Steps

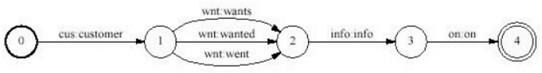
- Pre-process text ("splitter")
- Collect possible abbreviations and their possible expansions; use a stoplist of things not to expand
- Train a language model on "clean" text
- Normalize text

Language Model Training

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

WFST-based Implementation

$$T' = \pi_2(ShortestPath(T \circ A^{-1} \circ L))$$



Example Normalizations

cst cld 2 hv cllr id blck rmvn snt local form
customer called 2 have caller id block rmvn sent local form

cst clld to verify insde wre / i cncled his near mve on accident / cst now wnts to ploc to anther cmpny
customer called to verify inside wire / i cancelled his near move on accident / cst now wants to ploc to anther cmpny

cust no lnger wnts ld on acct
customer no longer wants ld on account

xplnd chrgs .. cust stated he w/ pay 26.45 & then w/ cancel his srvc w/ att
explained charges .. customer stated he will pay 26.45 & then will cancel his service with 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

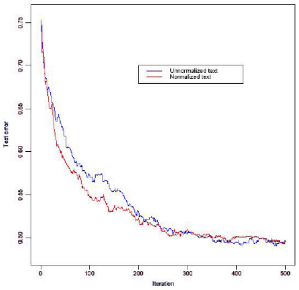
- Classify UNE-P RAMP comments into 26 different categories:

  Account Inquiry, Adjustments, Billing, CSS, Cancellation, Carrier Selection, Complaint, Disconnect, Features, Hot Button, Inside Wire Maintenance, Installation Issues, Long Distance, Marketing Incentive, Misdirect, Move, Order Redirect, Order Status, Other, OutPLOC, Repair, Sale, Snowbirds, Unmatched, Voicemail, Worldnet.

- Use BoosTexter (Schapire and Singer, 2000)

- Train on 39K examples; test on 1000

Utility of Text Normalization



Agenda: Summary

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