Abbreviated Text Input

Stuart M. Shieber and Ellie Baker

Harvard University / Division of Engineering and Applied Sciences

Abstract

We address the problem of improving the efficiency of natural language text input under degraded conditions (for instance, on PDAs or cell phones or by disabled users) by taking advantage of the informational redundancy in natural language. Previous approaches to this problem have been based on the idea of *prediction* of the text, but these require the user to take overt action to verify or select the system's predictions. We propose taking advantage of the duality between prediction and *compression*. We allow the user to enter text in compressed form, in particular, using a simple stipulated abbreviation method that reduces characters by about 30% yet is simple enough that it can be learned easily and generated relatively fluently. Using statistical language processing techniques, we can decode the abbreviated text with a residual word error rate of about 3%, and we expect that simple adaptive methods can improve this to about 1.5%. Because the system's operation is completely independent from the user's, the overhead from cognitive task switching and attending to the system's actions online is eliminated, opening up the possibility that the compression-based method can achieve text input efficiency improvements where the prediction-based methods have not.

The problem

Text entry under degraded conditions

- Growth of portable and embedded computing and telecom devices
- Interaction under environmental hindrances
- Disabilities

The question

Can we leverage language processing technology to improve the speed and accuracy of text entry even while using slow or otherwise impoverished input media?

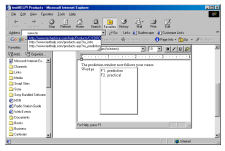
The standard approach: Prediction

Exploiting statistical redundancy through prediction

- e.g., The Reactive Keyboard (Darragh and Witten, 1992)
- `indifference, glanced back to where he had witnessed her` (27/57)

Advantages: Reduced keystrokes

Disadvantages: Cognitive load from task switching (Goodenough-Trepagnier et al., 1986)



Our approach: Compression

Prediction and compression are duals

Lempel-Ziv (gzip) as an (impractical) text input method

- \+ Excellent bitrate
- \+ Eliminates task switching
- \+ Lossless
- – Adds cognitive load from compression method

Making the approach practical

- ***Problem 1:*** A *human-centered* compression method, i.e., low human cognitive load
- ***Problem 2:*** ...and computer decodable
- ***Problem 3:*** ...with good bitrate
- ***Problem 4:*** ...and low error rate
- ***Problem 4:*** Computer must aid in error correction

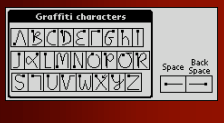
Taxonomy of compression methods

- Word level (e.g., compansion)
- Character level
  - Natural compression
    - Nonabbreviatory (probably oxymoronic)
    - Abbreviatory
      - *If u cn rd ths, u cn gt a gd jb.*
  - **Stipulated compression**
    - Nonabbreviatory (e.g., Vanderheiden, 1987)
    - Abbreviatory

Sidebar:
Natural versus stipulated methods
Replace freeform natural process with a stipulated process

- Model can be designed for ease and accuracy of decoding
- No requirement to acquire data for modeling
- Must be simple, learnable, "natural"

Case study: Grafitti



Stipulated abbreviation method

- Drop all vowels after first character.
- Drop all but one of repeated consonants.

An example

Original
Chinese military officials have boarded a grounded U.S. surveillance plane and removed equipment from it despite U.S. protests. In a signal that a standoff between the two nations is not likely to end soon, Pentagon sources told CNN today that China had begun removing sensitive eavesdropping equipment from the plane. Meanwhile, U.S. and Chinese diplomats were meeting on Hainan Island.
(387 characters)

Natural abbreviation
Chnse mltry offcls hv brdd a grnded US srvllnce pln nd remvd eqpt fr it dspte US prtsts. In a sgnl tht a stdoff btw th tw ntns is n lkly t nd son, Pntg srcs tld CNN tdy tht Chn hd bgn remvg snstv evsdrpng eqpt fr th pln. Mnwhl, US nd Chns diplmts wr mtg on Hainan Islnd.
(270 characters)

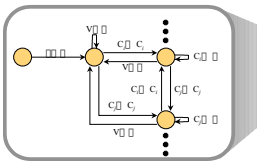
Stipulated abbreviation
Chns mltry ofcls hv brdd a grndd U.S. srvlnc pln and rmvd eqpmnt frm it dspt U.S. prtsts. In a sgnl tht a stndf btwn th tw ntns is nt lkly t end sn, Pntgn srcs tld CNN tdy tht Chn hd bgn rmvng snstv evsdrpng eqpmnt frm th pln. Mnwhl, U.S. and Chns dplmts wr mtng on Hainan Islnd.
(279 characters)

Automated reconstruction
Chinese military officials have **bearded** a grounded U.S. surveillance **plan** and removed equipment from it despite U.S. protests. In a signal that a standoff between the two nations is not likely to end soon, Pentagon sources told **canon** today that China had begun removing sensitive eavesdropping equipment from the **plan**. Meanwhile, U.S. and Chinese diplomats were meeting on *Hainan* Island.

The computational mechanism

Statistical natural-language processing via a cascade of weighted finite-state transducers

- General
- Flexible
- Adaptable



‹an›‹example›‹of›‹NUM›‹words›
Language Model
‹an›‹example›‹of›‹NUM›‹words›
Spelling Model
an_example_of_‹NUM›_words
Compression Model
an_exmpl_of_‹NUM›_wrds
Unknowns Model
an_exmpl_of_5_wrds

Empirical performance

LM training corpus: *1.8M words of Wall Street Journal*
Test corpus: *5099 words (28,045 characters) of WSJ*
Performance: *26.5% compression (cf 60.5% gzip)*

	Disabbreviation		*Keypad*	*Both*
Model	*errors*	*rate*	*rate*	*rate*
uniform	2586	50.7%	41.3%	79.7%
unigram	310	6.1	12.5	28.7
bigram	177	3.5	5.4	13.3
trigram	155*	**3.0**	5.0	12.1

* About half of the residual errors are second occurrences (or later). Adaptive methods thus have excellent potential for error reduction.