

Normalization of “Non-Standard” Words

Final Presentation, Aug 18 1999

Introduction

Team Members: Richard Sproat AT&T Labs

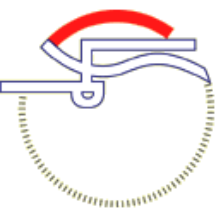
Alan Black CMU

Stan Chen CMU

Shankar Kumar CLSP/JHU

Mari Ostendorf Boston University

Christopher Richards Williams College



WS 99

Synopsis

- Nearly all speech and natural language technologies must deal, in one way or another, with real text.
- Real text is *messy*: it is hard to “normalize”.

Real Text is Messy

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 & Indry facils. Askg \$187K, maint \$868, utils incld.
Call Bkr Peter 914-428-9054.

Text Normalization Gets no Respect

- Text normalization is viewed as an unpleasant chore, not as a research area.
- There are exceptions to this generalization but they are rather focussed in their scope, and don't address the whole problem.

Treat the Problem with Respect and you'll get Better

Results

- Extant text normalizers (TTS systems, LDC's text-conditioning tools) work well, if at all, only on the kinds of text they were designed for. They perform abysmally on other types of text.
- We will present techniques that:
 - Perform better, across the board, than previous approaches
 - Are more readily extensible to new domains
 - Combine a detailed understanding of the scope of the problem with robust techniques that address its various aspects.