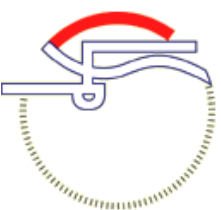


Normalization of “Non-Standard” Words

Final Presentation, Aug 18 1999

Splitting Tokens

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

Splitting Tokens: Motivation

- Whitespace-separated tokens are not good enough.
- Typos and Scanos and Adverts, oh my!
- Examples:

carpet-bagger, 11/2, No1, xjack

- The data show that about 6.1% of non-standard words should be split. (This is in the multi-domain case.) In the classified ad domain, where non-standard words account for 27.9% of all tokens, 1.7% of all tokens should be split.

Rule-based Approach

- Splitting can be construed as a parsing problem.
- Define classes to group and classes to split.
- Use regular expressions to match “interesting” tokens and substrings of tokens.
- Implemented both in Perl (fast but ugly); and as a weighted finite-state transducer (slow but elegant). All results pertain to the former.

Coarse-grained Results

How well can we predict *when* to split a given token?

	nantc	ads	pc110	rfr	total
Recall	98.78	99.23	94.30	97.39	98.71
Precision	82.15	64.40	75.31	96.55	67.26
Accuracy	99.04	96.59	95.88	99.27	96.97

Coarse-grained Errors

Misses:

ESANDWICH, 3400sq.ft, xjack, 11/2

False positives:

1-3pm, w/d, R-Ariz, PC-110

Fine-grained Results

How well can we predict *how* to split a given token?

	nantc	ads	pc110	rfr	total
Split Correct	92.83	93.71	88.38	93.91	93.16
Total Correct	98.79	96.25	95.21	98.86	96.63

Fine-grained Errors

GUESS	TRUTH
w/ walk - in	w/ walk-in
RE978 - 851 - 0048	RE 978-851-0048
1500 m Ah	1500 mAh