

Normalization of “Non-Standard” Words

Final Presentation, Aug 18 1999

A Classifier for Non Standard Words

Team Members: Richard Sproat AT&T Labs

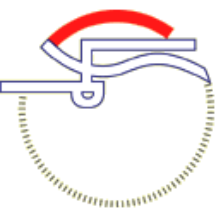
Alan Black CMU

Stan Chen CMU

Shankar Kumar CLSP/JHU

Mari Ostendorf Boston University

Christopher Richards Williams College



WS 99

A CLASSIFIER FOR NON STANDARD WORDS

- Problem Statement:

Assign tags [Eg. EXPN,NUM,NTIME] to NSW tokens

- CART Model with features to predict all the NSW Tags

Features may be :

- Domain Independent :
all upper case , has no vowels , has punctuation etc
- Domain Specific :

Features for alphabetic tokens : A sub-classifier by itself

Statistical Formulation: Sub-Classifier for alphabetic tokens

- NSW tags \mathbf{t} for alphabetic tokens (observations) \mathbf{o}
NATO : ASWD, **PCMCIA** : LSEQ, **frple** : EXPN
- $$p(\mathbf{t}|\mathbf{o}) = \frac{p(\mathbf{o}|\mathbf{t})p(\mathbf{t})}{p(\mathbf{o})}, \mathbf{t} \in ASWD, LSEQ, EXPN$$
- Trigram Letter Language Model (LLM)

$$p(\mathbf{o}|\mathbf{t}) = \prod_{k=1}^n p(\mathbf{l}_k | \mathbf{l}_{k-2} \mathbf{l}_{k-1})$$

The Trigram LM $p(\mathbf{o}|\mathbf{t})$ trained from tokens labeled as NSW tag \mathbf{t}

Sub-classifier features for Alphabetic Tokens

Letter Language Model Features for the Full CART Model

- p_{max} = Maximum probability of an alphabetic category
- t_{max} = Most probable alphabetic tag
- $diff$ = Difference between 1-best and 2-best probabilities $p(t|\mathbf{o})$

Token	$p(\text{ASWD} \mathbf{o})$	$p(\text{LSEQ} \mathbf{o})$	$p(\text{EXPN} \mathbf{o})$	p_{max}	t_{max}	$diff$
mb	0.0001	0.0038	0.9962	0.9962	EXPN	0.9924
Grt	0.0024	0.0000	0.9976	0.9976	EXPN	0.9952
NBA	0.0017	0.9983	0.0000	0.9983	LSEQ	0.9966
Cust	0.5456	0.0000	0.4544	0.5456	ASWD	0.0912

Distribution of Alphabetic NSWs

Percentages	nantc	ads	pc110	rfr
alphabetic	54.52	38.65	40.31	31.75
ASWD	81.50	29.73	64.33	70.61
LSEQ	12.30	12.46	20.88	2.25
EXPN	5.68	55.60	11.32	25.88
misc.	0.52	2.21	3.47	1.26

The Supervised Training and Evaluation Paradigm for the Alphabetic Tag Classifier

- Training Data : Labeled Alphabetic NSW data from a domain
- Letter Language Model Statistics
 - ASWD LLM : From tokens labeled as *ASWD*
 - LSEQ LLM : From tokens labeled as *LSEQ*
 - EXPN LLM : From tokens labeled as *EXPN*
- Evaluate 3 Letter LMs on test data & output all the LLM features to the CART tree

Accuracy: Supervised Training of the Tag Classifier on Alphabetic Tokens

- Baseline : Label All Tokens with most frequent tag in training data
- Uniform : $p(\mathbf{o}|\mathbf{t})p(\mathbf{t})$ where $p(\mathbf{t})$ uniform for all tags
- Unigram : $p(\mathbf{o}|\mathbf{t})p(\mathbf{t})$ where $p(\mathbf{t})$ from unigram frequencies in training data

	nantc	ads	pc110	rfr
Baseline	83.9[ASWD]	80.53[EXPN]	63.77[ASWD]	69.98[ASWD]
Uniform	88.92	98.5	90.83	97.36
Unigram	95.72	98.74	92.27	97.92

Unsupervised Training and Evaluation Paradigm in the Alphabetic Tag Classifier

- Training : Unlabeled data from a specific domain
- Training Lists for 3 Letter LMs
 - ASWD LLM: CMU dictionary with words more than 4 characters
 - EXPN LLM: Extract possible EXPNs using simple heuristics : words without vowels, word-period-lowercase letter, plural forms
 - LSEQ LLM: Extract possible LSEQs from news domain using simple heuristics : alternating letters and periods
- Evaluate 3 Letter LMs on test data & output all the LLM features to the CART tree

Accuracy: Unsupervised Training of the Tag Classifier on Alphabetic tokens

Accuracy	nanc	ads	pc110	rfr
unsupervised	92.98	87.90	68.90	92.06
supervised	95.72	98.74	92.27	97.92

Classifier Results : Tag Error Rate on all NSWs

Accuracy	nantc	ads	pc110	rfr
supervised:No LLM Feats	2.3	7.3	9.1	2.7
supervised:All LLM feats	1.9	6.5	8.2	3.2

Cross Domain testing : Tag Error Rate

- Train nantc tree with supervised alphabetic features
- Test other domains with unsupervised alphabetic features

	nantc	ads	pc110	rfr
Accuracy	3.5	21.6	41.2	9.7