Adaptive Language Modeling for Word Prediction

Keith Trnka University of Delaware
advised by Kathy McCoy

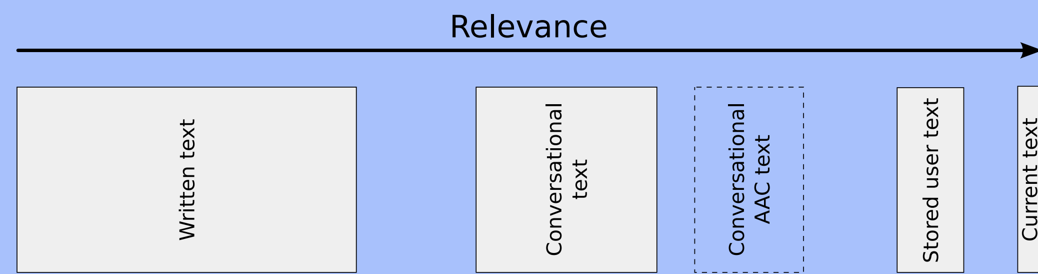
Background

- alternative communication, slow communication rate

- word prediction speeds up communication rate
- evaluation: keystroke savings

$$KS = \frac{keys_{normal} - keys_{prediction}}{keys_{normal}} \times 100\%$$

Motivation



- ngrams sensitive to training data
- multiple uses for AAC devices
- need good (relevant) training data
- ***adapt to the current text to get the most out of training data***

email excerpt

Switchboard is really low .
NNP VBZ RB JJ .

This could reflect that we chose
DT MD VB IN PRP VBD

a good **corpus** originally , maybe that
DT JJ NN RB , RB IN

the **cleanup** was more consistent
DT NN VBD RBR JJ

(I do n't think it 's any
-LRB- PRP VBP RB VB PRP VBZ DT

more **advanced** than the others ,
RBR JJ IN DT NNS ,

but I think I spent far more time on it
CC PRP VBP PRP VBD RB JJR NN IN PRP

paper excerpt

The **self-test** **analysis** is affected
DT JJS NN VBZ VBN

by both the size of the **corpus**
IN DT DT NN IN DT NN

as well as the **diversity** of the **corpus**
IN RB IN DT NN IN DT NN

, which explains the **trend** with **Switchboard**
, WDT VBZ DT NN IN NNP

: **participants** in the **corpus** collection
: NNS IN DT NN NN

were **restricted** to one of roughly 70 **topics**
VBD VBN TO CD IN RB CD NNS

, most of which are **represented**
, JJS IN WDT VBP VBN

in every **set** of **Switchboard** .
IN DT NN IN NNP .

Adapting to Match the Topic

$$P(w \mid h) = \sum_{t \in topics} P(t \mid h) * P(w \mid h, t)$$

Topic Granularity

- granularity of topic labels: the size of topics; specific or general topics
- medium-grained: human-annotated, typical clusters (e.g., clothing, weather, jobs)
- fine-grained: document as topic, IR-like modeling (e.g., seasonal clothing at work)
- coarse-grained: corpus as topic, very high-level (e.g., news, chit-chat)
- evaluation (with domain variations)

	In-domain	Out-of-domain	Mixed-domain
Trigram baseline	60.35%	53.88%	59.80%
SWB (medium)	61.48% (+1.12%)	-	-
Documents (fine)	61.42% (+1.07%)	54.90% (+1.02%)	61.17% (+1.37%)
Corpora (coarse)	-	52.63% (-1.25%)	60.62% (+0.82%)

Topic Identification

- current document representation: frequency, recency, inverse topic freq.
- similarity scores: cosine (best), Jacquard, Naïve Bayes (worst)
- polarizing the scores for more discrimination $sim'(t,h) = \frac{sim(t,h) - min_{t'}(sim(t',h))}{max_{t'}(sim(t',h)) - min_{t'}(sim(t',h))}$
- smoothing to prevent non-zero scores for sparse topics $sim'(t,h) = \frac{sim(t,h) + \gamma * min_{t' \mid sim(t',h)>0}(sim(t',h))}{max_{t'}(sim(t',h)) + \gamma * min_{t' \mid sim(t',h)>0}(sim(t',h))}$
- stemming helps with sparse topics (+0.2%) but hurts for normal topics (-0.1-0.2%)

Topic Application

- using trigrams
- smooth/backoff after interpolation - interpolating frequencies
- rescaling the frequency distribution for smoothing (+0.2-0.4%)

$$\sum_w f'_{topic}(w \mid h) = \alpha * \sum_w f_{topic}(w \mid h) = \sum_{t \in topics} \sum_w f(w \mid h, t)$$

- binning frequencies for smoothing
- smoothing extremely sparse conditional distributions on-demand

$$\frac{f(w \mid h)}{f(w \mid h) + \lambda} \times \frac{f(w \mid h)}{f(h)}$$

- modeling *h* and *t* independently (-0.6-1.2%)

$$P_{hybrid}(w \mid h) = P(w \mid w_{-2}, w_{-1}) * \left(\sum_{t \in topics} P(t \mid h) * P(w \mid t) \right)^{\alpha}$$

Future: style adaptation

- POS tags and pairs across styles

POS unigrams

Email					Papers				
POS	f	p	d	p1/p2	POS	f	p	d	p2/p1
Frequent and different POS tags									
PRP	699	0.0797399041752224	1.28	4.5649	RB	598	0.0385458295732886	0.52	0.5846
RB	578	0.0659365731234314	0.52	1.7106	VB	480	0.0309397963130076	0.56	0.5627
VB	482	0.054985169974903	0.56	1.7772	VBN	478	0.0308108804950367	0.61	1.8756
VBP	301	0.034337211955281 8	0.50	1.6699	VBP	319	0.02056207296635 3	0.50	0.5988
VBD	202	0.0230435774583619	0.58	1.8240	PRP	271	0.0174680933350522	1.28	0.2191
VBN	144	0.0164271047227926	0.61	0.5332	VBD	196	0.0126337501611448	0.58	0.5483
-RRB-	98	0.0111795573807894	0.74	2.1680	:	90	0.00580121180868893	0.61	0.5353
:	95	0.0108373260323979	0.61	1.8681	-RRB-	80	0.0051566327188346	0.74	0.4613
-LRB-	88	0.0100387862194844	0.67	1.9967	-LRB-	78	0.0050277169008637 4	0.67	0.5008
RP	60	0.00684462696783025	1.56	8.1683	RBR	60	0.00386747453912595	0.66	1.9943
PRP$	56	0.00638831850330824	0.57	1.8020	PRP$	55	0.00354518499419879	0.57	0.5549
SYM	49	0.00558977869039471	1.87	28.9066	FW	31	0.00199819517854841	0.55	1.7516
WP	30	0.00342231348391513	1.12	3.5396	EX	16	0.00103132654376692	1.05	0.3117
EX	29	0.00330823636778462	1.05	3.2077	WP	15	0.000966868634781488	1.12	0.2825
RBR	17	0.00193931097421857	0.66	0.5014	RP	13	0.000837952816810623	1.56	0.1224
FW	10	0.00114077116130504	0.55	0.5709	SYM	3	0.000193373726956298	1.87	0.0346

POS bigrams

Email					Papers				
tags	f	p	d	p1/p2	tags	f	p	d	p2/p1
Frequent and different POS bigrams									
NN IN	242	0.027606662103582	0.53	0.5788	NN IN	740	0.0476988526492201	0.53	1.7278
PRP VBP	227	0.0258955053616245	1.13	3.5870	NN NN	619	0.0398994456619827	0.67	2.0101
TO VB	200	0.0228154232261008	0.58	1.8152	JJ NN	576	0.0371277555756091	0.61	1.8813
NN NN	174	0.0198494182067077	0.67	0.4975	IN NN	409	0.0263632847750419	0.57	1.8055
JJ NN	173	0.0197353410905772	0.61	0.5316	NN VBZ	270	0.0174036354260668	0.69	2.0616
IN PRP	149	0.0169974903034451	1.31	4.7945	NNS IN	236	0.0152120665205621	0.60	1.8521
IN NN	128	0.0146018708647045	0.57	0.5539	NN NNS	230	0.0148253190666495	0.99	2.9536
PRP MD	121	0.013803331051791	1.31	4.7588	TO VB	195	0.0125692922521593	0.58	0.5509
PRP VBD	111	0.01266255989048 6	1.21	4.0926	VBN IN	186	0.0119891710712904	0.76	2.2361
VB IN	99	0.0112936344969199	1.20	3.9820	IN JJ	175	0.0112801340724507	0.77	2.2473
RB VB	[illegible]	[illegible]	0.98	2.9169	NNS .	150	0.00966868634781488	0.88	2.5684
. PRP	[illegible]	[illegible]	1.01	3.0234	. DT	144	0.00928193889390228	0.92	2.7122
PRP VBZ	[illegible]	[illegible]	1.34	5.0565	IN NNP	120	0.0077349490782519	0.52	1.6951
NN VBZ	[illegible]	[illegible]	0.69	0.4851	PRP VBP	112	0.00721928580636844	1.13	0.2788
RB JJ	[illegible]	[illegible]	0.63	1.9260	. IN	110	0.00709036998839758	1.06	3.2713
, PRP	[illegible]	[illegible]	0.62	1.8980	NNS VBP	100	0.00644579089854325	0.77	2.2602
NNS IN	[illegible]	[illegible]	0.60	0.5399	IN NNS	99	0.00638133298955782	0.60	1.8646
VBP RB	[illegible]	[illegible]	0.97	2.8921	, DT	98	0.00631687508057239	1.19	3.9553
, CC	[illegible]	[illegible]	0.72	2.1303	NNP NN	98	0.00631687508057239	0.63	1.9094
CC PRP	[illegible]	[illegible]	1.77	16.5181	NN MD	93	0.00599458553564522	1.07	3.2843
VB PRP	[illegible]	[illegible]	1.72	13.1470	NN TO	86	0.0055433801727472	0.64	1.9437
RB RB	49	0.00558977869039471	0.87	2.5506	NNS ,	82	0.00528554853680547	0.71	2.1061
VBN IN	47	0.0053616244581337	0.76	0.4472	NNP NNP	74	0.00476988526492201	1.11	3.4844

Top 10 Instances	
say that	7
work on	5
talk about	5
know if	3
show that	2
think about	2
go into	2
explain that	2
see that	2
look at	2

- style granularity: treat each corpus as a style (e.g., Switchboard, Micase)
- style identifications: cosine similarity of POS tags and pairs
- style application: condition transition probabilities of POS ngram model

$$P_{style}(w \mid h) = \sum_{s \in styles} P(s \mid h) * \sum_{tag \in POS(w)} P(tag \mid tag_{-1}, tag_{-2}, s) * P(w \mid tag)$$

Funded by US Department of Education grant H113G040051, under the National Institute on Disability and Rehabilitation Research program