Word Prediction Techniques for User Adaptation and Sparse Data Mitigation

Keith Trnka
trnka@cis.udel.edu

PhD Proposal - May 5th, 2008

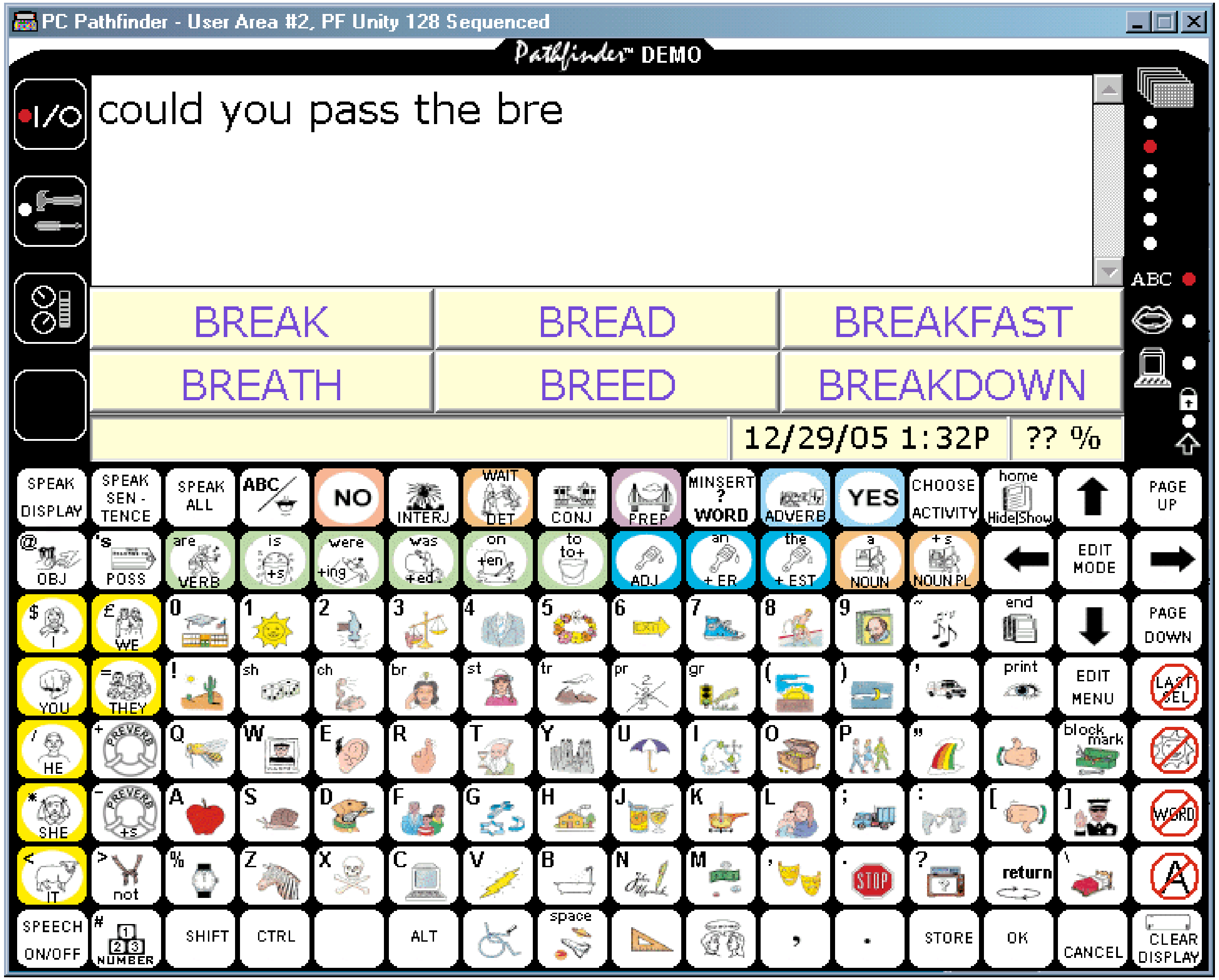
The problem

background

- **problem:** severe speech impairment, often limited motor control
- **solution:** Augmentative and Alternative Communication (AAC) devices

The problem

background



PC Pathfinder - User Area #2, PF Unity 128 Sequenced

could you pass the bre

The problem

background

- **problem:** low communication rate with AAC devices (with motor impairment, 10 wpm or less)
  - **subproblem:** communication divide with speech (<10 wpm vs. 150-200 wpm)
  - **subproblem:** fatigue from continual use
- **solution:** word prediction in AAC

  (Alm et al., 92), (Boggess, 88), (Carlberger et al., 97), (Carlberger, 98), (Garay-Vitoria and Gonzàlez-Abascal, 97), (Garay-Vitoria and Abascal, 06), (Hunnicutt and Carlberger, 01), (Lesher and Rinkus, 02), (Lesher et al., 02), (Matiasek et al., 02), (Väyrynen, 05), (Wandmacher and Antoine, 06), (Wandmacher and Antoine, 07) and many more...

The problem

part 1

- **problem:** simplistic word prediction methods
  - unigrams + recency is common in devices
  - trigrams are the top-end (e.g., WordQ)
  - **subproblem:** users tend to ignore poor predictors (or worse, distractions)
- evidence that a better predictor will increase communication rate

  in contrast to (Koester and Levine, 94), (Venkatagiri, 93), (Anson et al., 04)

(NAACL, 2007) (Telehealth/AT, 2008)

The problem

part 2

- **background:** ngrams sensitive to training data
    - highly influenced by:
        - amount of training data
        - similarity of training and testing data
- **problem:** no conversational AAC corpora, only AAC written corpora is very small
- many target domains (e.g., school, homework, email)

The proposal

high level

- Design an adaptive language model that:
    - utilizes **all training data**
    - focuses on the **most similar data**
        - may be only a small amount of similar data; data sparseness must be addressed

(ACL-SRW, 2008) (ISAAC, 2008)

The proposal

details

- Weight chunks of training data based on:
    - topical similarity to current text
    - stylistic similarity to current text
- Utilize current text (cache or recency modeling)

(ACL-SRW, 2008) (ISAAC, 2008)

Outline

- Word prediction evaluation methods
- Topic modeling
- Proposed future work
  - Style adaptation
  - Cache modeling
  - Combining topic, style, and cache

Evaluation methods

keystroke savings

- percentage of keystrokes avoided using word prediction

$$KS = \frac{keystrokes_{\text{normal}} - keystrokes_{\text{with prediction}}}{keystrokes_{\text{normal}}} * 100\%$$

- affected by the number of predictions (window size) - typically 1-10

Evaluation methods

simulated usage

- How many keys does predictive entry take?
    - simulate a perfect user
    - direct selection
    - automatically add a space after a word is selected

Evaluation methods

interpreting keystroke savings

- How do we interpret keystroke savings?
    - Minimum - zero (the desired word was never predicted)
    - Maximum - much less than 100% (each word takes at least one keystroke)
        - 80-86% max for our tests (varies by corpus)

(ACL short paper, 2008)

Evaluation methods

corpora

- Which corpora for training and testing?
  - conversational text is the focus
  - AAC devices are also used for writing, email, etc, so also use a variety of corpora

Evaluation methods

corpora

Corpus	**Medium**	**Words**
AAC Email	email	27,710
Callhome	spoken	48,407
Charlotte	spoken	187,587
SBCSAE	spoken	237,191
Micase	spoken	545,411
Switchboard	spoken	2,883,774
Total spoken		3,902,380
Slate	written	4,178,543

Evaluation methods

corpora

- Which corpora for training and testing?
- Domain-varied evaluation
  - In-domain
  - Out-of-domain
  - Mixed-domain

(ASSETS, 2007)

Topic modeling

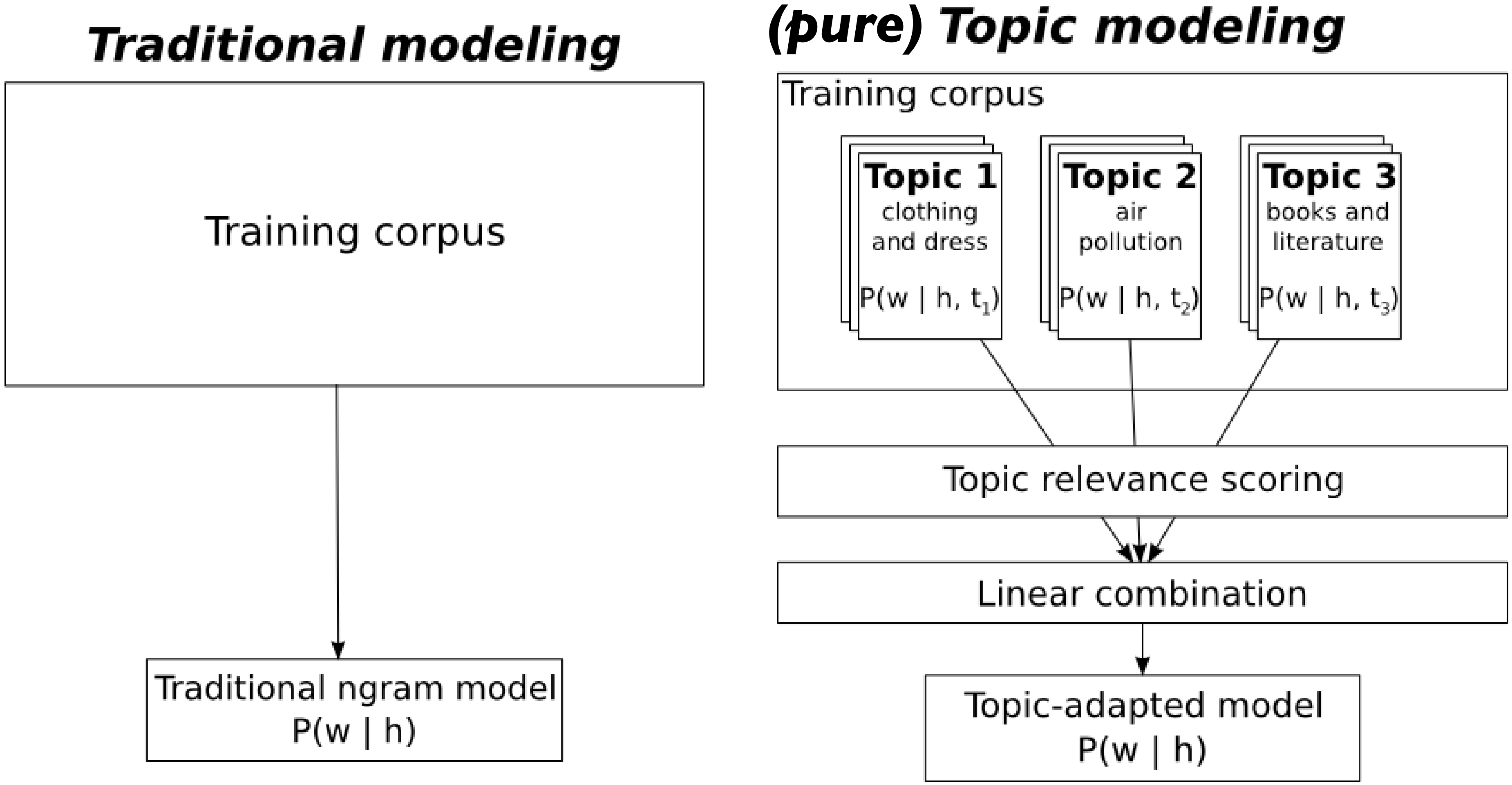
goals

- seamlessly adapt to the topic of conversation
- boost similar training data, but allow all training data to contribute
- worst case, the model should degrade to the baseline model

(ISAAC, 2006)

Topic modeling

overview



Traditional modeling

Training corpus

Traditional ngram model
P(w | h)

(pure) Topic modeling

Training corpus

Topic 1
clothing and dress
$P(w \mid h, t_1)$

Topic 2
air pollution
$P(w \mid h, t_2)$

Topic 3
books and literature
$P(w \mid h, t_3)$

Topic relevance scoring

Linear combination

Topic-adapted model
P(w | h)

(ISAAC, 2006)

Topic modeling

weighting each topic

$$\underset{\substack{\uparrow\\\text{Overall}\\\text{model}}}{P_{topic}(w \mid h)} = \sum_{t \in topics} \underset{\substack{\uparrow\\\text{Relevance/similarity score}}}{P(t \mid h)} * \underset{\substack{\uparrow\\\text{Ngram model}\\\text{for topic t}}}{P(w \mid h, t)}$$

Topic modeling

pure and hybrid modeling

- measuring a full-fledged ngram model for each topic - ***pure topic modeling***
- **Problem:** more sparseness in posterior compared to baseline (due to larger parameter space)
- **Potential solution:** measuring an impoverished model for each and combining with baseline - ***hybrid topic modeling***

(IUI, 2006)

Topic modeling

pure topic modeling

$$P(w \mid h) = \sum_{t \in topics} P(t \mid h) * P(w \mid w_{-1}, w_{-2}, t)$$

Overall model

Relevance/similarity score

Trigram model for topic t

Topic modeling

hybrid topic modeling

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

Overall model

Baseline trigram model

Relevance/ similarity score

Unigram model for topic t

Topic modeling

pure vs. hybrid modeling

- Evaluating pure vs. hybrid modeling (W=2-10)
    - Baseline trigram: 51.1% - 62.5% savings
    - Theoretical limit: 82.6% savings
    - Pure modeling (bigrams) +1.2-1.5% savings
    - Hybrid modeling (trigrams) - +0.3-0.4% savings
- A spectrum of hybridization

(IUI, 2006)

Topic modeling

issues in essentials

- smoothing after interpolation
- dealing with floating-point frequencies for smoothing methods
- re-scaling the distribution to reflect true total frequencies
- Good-Turing smoothing too finicky, used robust approximation

Topic modeling

assessing relevance of a topic

$P(t \mid h)$

Topic modeling

assessing relevance of a topic

- Problems in assessing topical relevance
  - How to represent the current document?
  - How to represent each topic?
  - How to compare the two representations?
- Compare unigram-like distributions

Topic modeling

assessing relevance of a topic

- Representing the current document
  - frequency
  - recency
  - topical salience

Topic modeling

assessing relevance of a topic

- frequency

words	Kathy	shared	an	office	with	Kathy	in	g—
weights	1	1	1	1	1	1	1	

term	normalized score	unnormalized score
Kathy	0.286	2
office	0.143	1
an	0.143	1

Topic modeling

assessing relevance of a topic

- frequency + recency

words	Kathy	shared	an	office	with	Kathy	in	g—
weights	λ^6	λ^5	λ^4	λ^3	λ^2	λ^1	λ^0	
$\boldsymbol{at}\ \lambda = 0.95$	0.735	0.774	0.815	0.857	0.903	0.95	1	

term	normalized score	unnormalized score
Kathy	0.279	1.685
office	0.142	0.857
an	0.135	0.815

Topic modeling

assessing relevance of a topic

- topical salience - Inverse Topic Frequency (ITF)
- frequency + recency + ITF + stopword removal

words	Kathy	shared	an	office	with	Kathy	in	g—
weights	$\lambda^3 * ITF(\text{Kathy})$	$\lambda^2 * ITF(\text{shared})$	0	$\lambda^1 * ITF(\text{office})$	0	$\lambda^0 * ITF(\text{Kathy})$	0	
example	5.165	4.256	0	2.649	0	6.024	0	

term	normalized score	unnormalized score
Kathy	0.618	11.189
office	0.146	2.649
an	0	0

Topic modeling

assessing relevance of a topic

- Relevance functions
  - cosine $$sim_{cosine}(t,c) = \frac{\sum_{w \in t \cap c} f_t(w) * f_c(w)}{\sqrt{\sum_{w \in t} f_t(w)^2} * \sqrt{\sum_{w \in c} f_c(w)^2}}$$
  - Jacquard $$sim_{Jacquard}(t,c) = \frac{\mid t \cap c \mid}{\mid t \cup c \mid}$$
  - Naïve Bayes $$P(t \mid h) = P(t) * \prod P(w \mid t)^{f_h(w)}$$
- Normalize to probability distribution

$$P(t \mid h) \approx \frac{sim(t,c)}{\sum_{t'} sim(t',c)}$$

Topic modeling

assessing relevance of a topic

- Evaluating similarity scores
  - cosine: +0.273 to +1.124% over baseline
  - Jacquard: -0.162 to +0.336% over baseline
  - Naïve Bayes - -2.265% to -0.947% over baseline (with scaling even worse)
- Additional tuning - score scaling, smoothing, stemming

Topic modeling

what’s in a topic?

$t \in topics$

Topic modeling

granularity

- What's in a topic?
- Granularity - how general or specific a topic is
  - general topics tend to include many documents
  - specific topics tend to include few documents

Topic modeling

granularity

- **medium-grained** - typical human-annotated topics (e.g., sports, weather, politics, food)
- **fine-grained** - treating each document as a topic (may be very, very specific)
- **coarse-grained** - treating each corpus as a topic (very, very general topics, like news vs. academic texts)

Topic modeling

medium-grained evaluation

Testing corpus	Trigram	Medium-grained topic	Significance
AAC Emails	43.25%	43.53% (+0.27%)	$p < 0.05$
Santa Barbara	43.49%	43.90% (+0.41%)	$p < 0.001$
Callhome	49.33%	49.52% (+0.19%)	$p < 0.005$
Charlotte	49.64%	50.07% (+0.43%)	$p < 0.001$
Micase	46.52%	46.99% (+0.47%)	$p < 0.001$
Switchboard*	60.35%	61.48% (+1.12%)	$p < 0.001$
Slate*	39.17%	39.78% (+0.61%)	$p < 0.001$

Topic modeling

granularity

- Evaluation of topic granularity
  - **Medium-grained**
    in-domain: +1.12% over baseline
    out-of-domain: +0.19% to +0.61% over baseline
  - **Coarse-grained**
    out-of-domain: -1.25% to -0.24% over baseline
    mixed-domain: -0.30 to +1.11% over baseline
  - **Fine-grained**
    in-domain: -0.05% to +1.07% over baseline
    out-of-domain: -0.11% to +1.07% over baseline
    mixed-domain: +0.26% to +1.67% over baseline

Topic modeling

related work

- Topic modeling
  (Lesher and Rinkus, 02), (Seymore and Rosenfeld, 97), (Seymore et al., 98), (Chen et al., 98), (Mahajan et al., 99), (Florian and Yarowsky, 99), (Clarkson and Robinson, 97), (Iyer and Ostendorf, 99), (Adda et al., 99), (Kneser and Steinbiss, 93)
- Latent Semantic Analysis models
  (Wandmacher and Antoine, 07), (Wandmacher et al., 07), (Bellegarda, 98), (Bellegarda, 00)
- Trigger pair models
  (Li and Hirst, 05), (Li, 06), (Matiasek et al., 03), (Gong, 07), (Lau et al., 93), (Rosenfeld, 94), (Rosenfeld, 96)

Topic modeling

summary

- Combating data sparseness - pure vs. hybrid, smoothing after interpolation, adjusting smoothing for the task
- Similarity scores - cosine with tweaks worked best
- Granularity - human-annotated topics worked best, but fine-grained is promising

Style modeling

motivation

- In-domain vs. out-of-domain is a question of both topic and ***style***
- AAC devices are used for many different genres (e.g., conversation, speeches, homework, papers)

Style modeling

background

- example: written vs. spoken language
  (Biber, 88)
- Several researchers use a three-level model
  (Hovy, 88), (Kessler et al., 97), (Michos et al., 96)
  - top level - genre labels (e.g., news broadcast)
  - middle level - features (e.g., formality, conciseness)
  - low level - realization (e.g., pronoun usage)

Style modeling

corpus study

- Corpus study of style
  - collected 8.8k words of my earlier research emails, 15.5k words of my papers
  - compared various distributions of the two corpora to computationally show stylistic differences

Style modeling

corpus study

- compared distributions, focusing on words/tags/pairs that were frequent but very different
  - word unigrams
  - part of speech (POS) unigrams
  - coarse-grained POS bigrams

Style modeling

corpus study

- word unigrams
  - pronouns much more likely in email (e.g., you, I, it)
  - contractions more likely in email
  - certain specific nouns more likely in papers (e.g., model, language, training)

Style modeling

corpus study

- POS unigrams
  - pronouns and wh-pronouns more common in emails
  - particle verbs more common in emails (e.g., write up, figure out)
  - comparative adverbs, foreign words more common in papers

Style modeling

corpus study

- coarse-POS bigrams
  - more modified nouns in papers
  - more ADV-N/N-ADV pairs in emails (e.g., "so I", "I just", "then I")
  - passive voice VBZ-VBN pairs more likely in papers (e.g., "has been", "is shown")

Style modeling

proposing a model

- **Focus:** stylistic differences in POS tag usage
- Plan to build on a POS ngram model (often used for Markov model POS taggers):

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

Style modeling

proposing a model

- Extending the POS ngram model like topic modeling:

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

Style modeling

expected problems

- Problems to address
  - How will we get POS tagged training text?
  - How will we determine stylistic relevance/similarity?
  - What set of styles to use?

Style modeling

expected problems

- POS tagged training text
  - **Initial plan:** use the Stanford maxent tagger (Toutanova and Manning, 00)
  - **Potential improvement:** iterate using Expectation-Maximization (EM) algorithm
- Stylistic similarity
  - **Initial plan:** use cosine with both POS unigrams and bigrams
  - **Potential improvement:** Naïve Bayes or combine separate similarities

Style modeling

expected problems

- What’s in a style?
  - **Initial plan:** treat each corpus as a style
  - **Potential alternatives:** fine-grained style modeling, automatic clustering

Style modeling

summary

- Model style using POS ngrams
- Adapt the transition probabilities based on POS tag distributional similarity
- Treat each corpus as a style

Cache-based adaptation

related work

- Two major existing approaches
  - Jelinek et al. (91) - word ngram model of most recent 1000 words
  - Kuhn and de Mori (90) - POS ngram emission probability of recent words
  - Variations: reset cache between documents or not
- Simplistic approaches - named entity cache, unigrams
  (Li and Hirst, 05), (Väyrynen, 05), (Carlberger, 98)

Cache-based adaptation

goals

- utilize the most relevant text while accounting for the extreme data sparseness
- integrate smoothly with existing model

Cache-based adaptation

the plan

- Kuhn and de Mori better address data sparseness

$$P(w \mid w_{-1}, w_{-2}) = \sum_{tag \in POS(w)} P(tag \mid tag_{-1}, tag_{-2}) * ((1 - \lambda) * P(w \mid tag) + \lambda * P_{cache}(w \mid tag))$$

- **Problem:** unknown words
  - **Initial plan:** let the Viterbi algorithm do it
  - **Potential alternative:** leverage morphology (e.g., PC-KIMMO)

Combining it all

combining topic and style

- Combining topic and style
    - Utilize the POS ngram model and adapt emission probabilities to topic

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

Combining it all

adding cache adaptations

- Adding in cache modeling

$$
\begin{aligned}
P(w \mid h) = \sum_{s \in styles} & P(s \mid h) * \\
& \sum_{topic \in topics} P(topic \mid h) * \\
& \sum_{tag \in POS(w)} P(tag \mid tag_{-1}, tag_{-2}, s) * ((1-\lambda) P(w \mid tag, topic) + \lambda * P(w \mid tag, cache))
\end{aligned}
$$

Existing contributions

- Our approach to topic modeling
- Topic modeling for word prediction
- Domain-varied evaluation
- Theoretical limits of word prediction

(IUI, 06) (ISAAC, 06) (NAACL, 07) (ASSETS, 07) (Telehealth/AT, 08) (ACL short, 08)

Expected contributions

- Style modeling
- Applying a POS-based cache to word prediction
- Tightly integrating multiple adaptive language models
- Evaluation of combination models

(ACL-SRW, 08) (ISAAC, 08)

Questions?

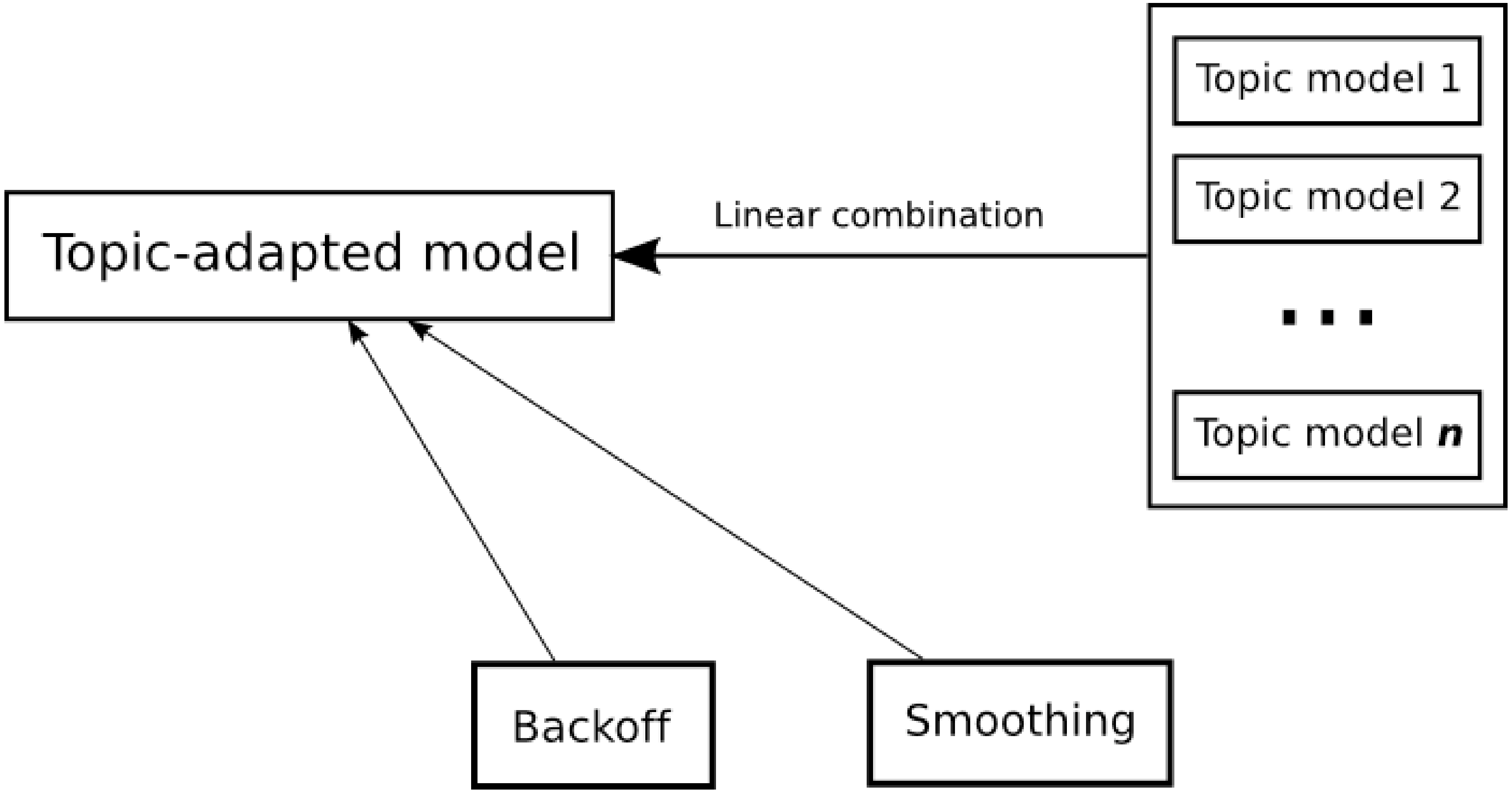
Topic modeling

issues in essentials

- Backoff and smoothing
  - smoothing before interpolation overestimates data sparseness (smoothes too much)
  - smoothing after interpolation correctly estimates sparseness

Topic modeling

issues in essentials



Topic modeling

issues in essentials

- **Problem:** interpolation turns frequencies into floating-point numbers
- **Solution:** take ceiling of all values
- **Problem:** interpolation reduces the contribution of each word occurrence
- **Solution:** scale the distribution back to its original sum +0.2-0.4% (W=2-10)

Topic modeling

issues in essentials

- Katz' backoff requires that we smooth an entire (unconditional) distribution at once
- **Problems:** too computationally demanding, Good-Turing smoothing struggles with sparse conditional distributions
- **Solution:** approximate Good-Turing smoothing in a more robust equation

$$P(w \mid h) = \frac{f(w \mid h)}{f(w \mid h) + \lambda} * \frac{f(w \mid h)}{f(h)}$$