Style adaptation with a part-of-speech model for word prediction

Keith Trnka

SIG-AI 2010-5-3

keywords

- same genre as previous talk:
  - augmentative and alternative communication (AAC)
  - word prediction
  - adaptive language modeling
  - part of speech
- new focus: **style**

word prediction

- quick summary
  - predict or complete a word while typing
  - generate predictions with a language model and filter the list by any typed letters
  - evaluate in keystroke savings

motivation

- training data is likely to be a mixture of different topics and styles
  - adapt to focus on the most relevant training data
- topic adaptation was successful, this might be too

style

- what is style?
  - formal vs informal
  - academic vs business vs legal
  - spoken vs written

style

- how can we identify a style computationally?
  - style as a lexical feature
  - grammatical aspects
    - more complex constructions, etc

style

- our focus: grammatical features
  - will use a part-of-speech ngram model
  - adapt transition probabilities to style

part of speech ngrams

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

part of speech ngrams

- the history comes from Viterbi alg:

$$P(w \mid h) = \sum_{(tag_1,\ldots,tag_{-2},tag_{-1}) \in candidates} P(tag_1, \ldots, tag_{-2}, tag_{-1}) * \sum_{tag \in POS(w)} P(tag \mid tag_{-2}, tag_{-1}) * P(w \mid tag)$$

style model

- as with topic, weight parts of training data by stylistic similarity:

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

style model

- questions
  - what is a style?
  - how to do style similarity?
  - how about any of the topic modeling tweaks?

style model

- what's a style?
  - corpus
  - document

similarity scores

- topic modeling: cosine similarity
- style modeling
  - cosine not really appropriate (how to combine transition tri/bi/unigrams?)
  - P(style | text) = probability of text using the transition probs. from the style

similarity scores

- notable departure from topic - we need to smooth the transition models before interpolating
- may as well just combine probabilities

first evaluation

- mixed-domain training: training data of all corpora is mixed, evaluate individually on each
- developmental evaluations: just mixing two small corpora

first evaluation

Corpus	No style	Style
AAC Email	47.755%	48.069% (+0.314)
SBC	44.714%	44.518% (-0.196)

first evaluation

- what went wrong?
  - most weights ended up as 55% vs 45% at most
  - SBC is much larger than AAC Email
  - this type of modeling implicitly favors smaller corpora

size weighting

- multiply style similarity by corpus size

size weighting

Corpus	No style	Style (size weight)
AAC Email	47.755%	47.768% (+0.013)
SBC	44.714%	44.711% (-0.003)

is size the problem?

- testing using two larger corpora: Micase and Switchboard

is size the problem?

Corpus	No style	Style (size weight)
Micase	49.120%	49.240% (+0.120)
Switchboard	53.392%	53.296% (-0.096)

is size the problem?

- still a tendency towards the smaller corpus, but the benefit remains small

are scores conservative?

- scores are pretty conservative without size-weighting
- can polarize the scores like topic modeling
  - scale max weight to 1
  - scale min weight to 0
  - add a fraction of min weight to everything
  - normalize to get probabilities

are scores conservative?

$$w'_i = \left(\frac{w_i - min}{max - min} \right) + 0.5 * min$$

polarized weights

Corpus	No style	Style (polarized)
AAC Email	47.755%	47.891% (+0.136)
SBC	44.714%	44.710% (-0.004)

more corpora

Corpus	No style	+style	+polarize
AAC Email	48.090%	+0.174	+0.041
SBC	46.063%	+0.193	+0.099
Callhome	50.098%	+0.217	+0.133
Charlotte	50.589%	+0.149	+0.129
Micase	49.726%	+0.048	+0.069
Switchboard	52.555%	+0.068	+0.119
Slate	49.127%	*N/A*	*N/A*

document as style

- topic modeling:
  - bad: corpus as topic
  - good: document as topic
  - very good: clusters/human-annotated topics

document as style

- in small tests (AAC Email+SBC)
  - better than baseline
  - worse than corpus as style
  - polarization was beneficial
- advantage: can do even faster tests (don’t need mixed-domain testing)

document as style

- should do well on corpora with distinct subsets (e.g., two authors in AAC Emails)

probs vs freqs

- topic modeling: we combined frequencies instead, then did smoothing after
  - seamless degradation to the baseline model (uniform weights)
  - more accurate smoothing
- cons for style: slower (cause we already smooth the style models to score)

probs vs freqs

- evaluation on AAC Emails, doc. as style

Test	KS
baseline	47.017%
style/probs	46.677% (-0.340)
style/freqs	47.058% (+0.041)

inverse style frequency

- topic modeling: inverse topic frequency very beneficial
- inverse style frequency
  - measure a separate ISF distribution for tri/bi/unigram transition probabilities

inverse style frequency

Test	KS
baseline	47.017%
style/freqs	47.058% (+0.041)
style/freqs+ISF	47.083% (+0.066)
style/freqs+ISF+pol.	47.081% (+0.064)

conclusions

- most lessons from topic translate:
  - IDF/ITF/ISF
  - polarizing
  - frequencies
- some new lessons:
  - style modeling primarily for multi-corpus training