Adapting Word Prediction to Subject Matter without Topic-labeled Data

Keith Trnka University of Delaware
advised by Kathy McCoy

Problem

- alternative communication, slow communication rate

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

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

How can we improve keystroke savings?

Why topic adaptation?

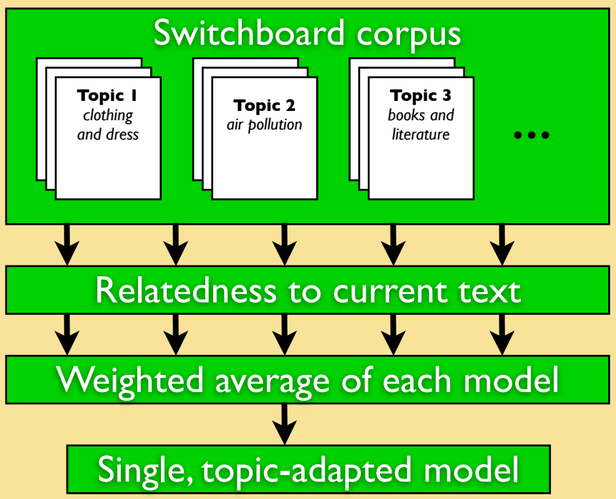
- topic can substantially affect word choice:

how long would it be in the

world
winter
city
area
United
future
mountains

- existing devices tend to use word prediction more for **fringe words**, which tend to be topical words

Previous work - labeled topic modeling



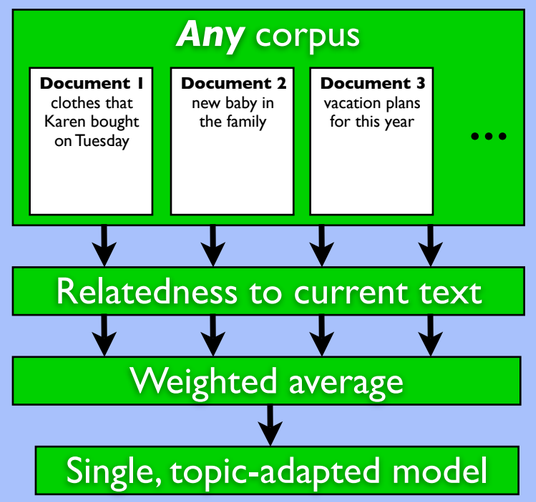
Testing corpus	Switchboard trigram	Switchboard topic
AAC Email	43.25%	**43.53%**
Callhome	49.33%	**49.52%**
Charlotte	49.64%	**50.07%**
SBCSAE	43.49%	**43.90%**
Micase	46.52%	**46.99%**
Switchboard	60.35%	**61.48%**

- compare each topic to current conversation
- weight based on similarity of keyword usage
- take a weighted average of all the topic models

Advantages compared to basic ngrams
- higher keystroke savings (even when tested on other corpora)
- predictions are appropriate for the overall topic (e.g., cooking)

Disadvantages compared to basic ngrams
- need training corpus with topic labels (limited to training on Switchboard)

Current work - unlabeled topics



- compare each ***document*** to current conversation
- weight based on similarity of keyword usage
- take a weighted average of all the document models

- each document is treated as if it were a very specific topic
- tested using a mixture of corpora for training data (below)

Testing corpus	Switchb. topic	Mixed trigram	Mixed topic
AAC Email	43.53%	52.18%	**53.14%**
Callhome	49.52%	52.14%	**53.39%**
Charlotte	50.07%	53.50%	**53.92%**
SBCSAE	43.90%	47.78%	**48.84%**
Micase	46.99%	51.46%	**53.13%**
Switchboard	**61.48%**	59.80%	61.17%

Advantages
- works with any kind of corpus
- can add more and more data (even general-purpose data) with little negative effect

Future work

- can use this method to integrate user text into the language model
  - difficult to perform controlled evaluations
- use the same methodology with a different similarity score
  - e.g., syntactic similarity
- improving the similarity score to provide guarantees that there will be no loss in keystroke savings by adding more training texts

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