Adapting Word Prediction to Subject Matter without Topic-labeled Data

Keith Trnka
University of Delaware

Outline

- background - AAC, word prediction, evaluation methods
- progression of word prediction methods
  - topic modeling with Switchboard
  - mixed-domain training
  - mixed-domain topic modeling, without labeled topics

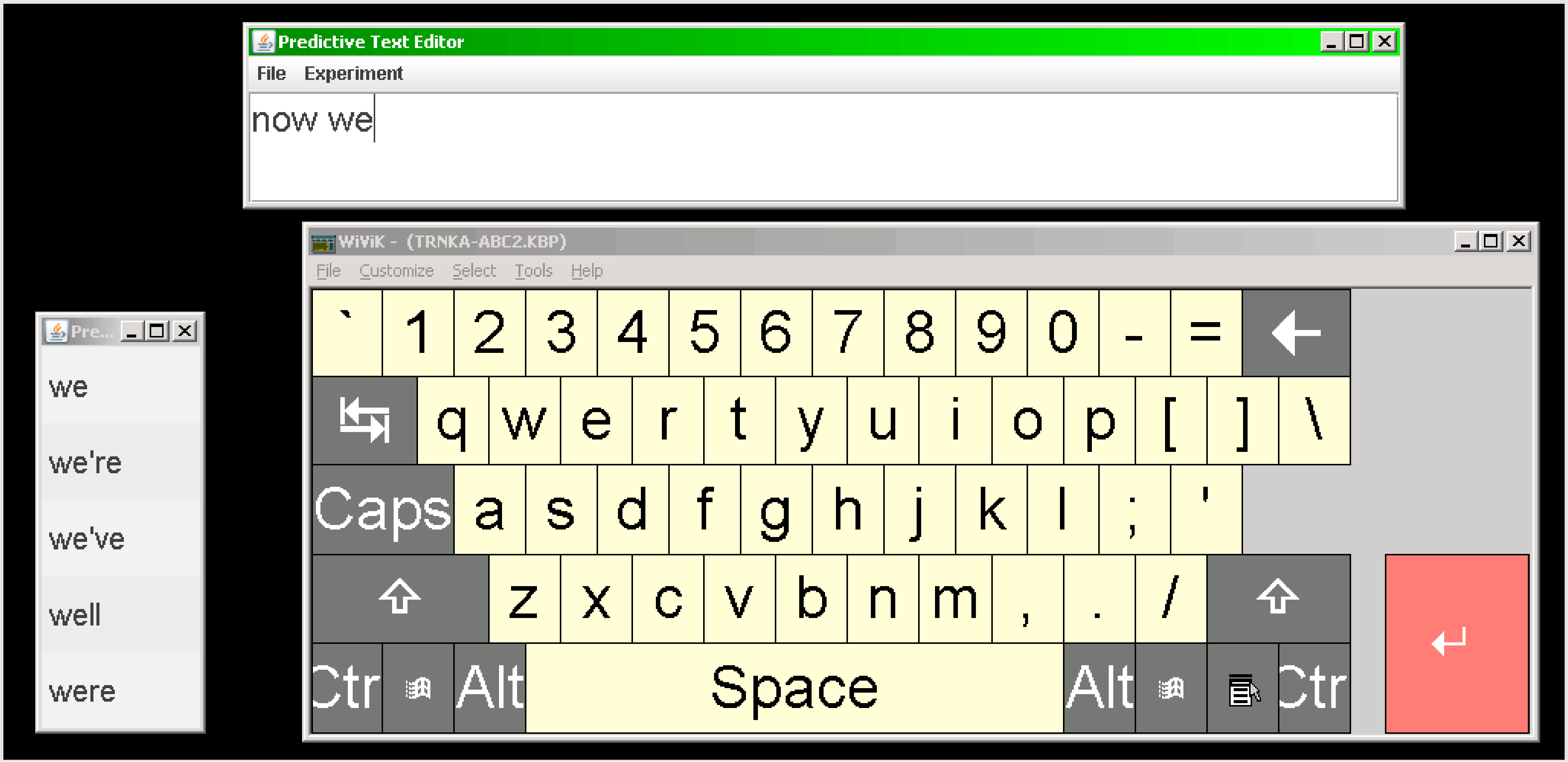
AAC Background

- Augmentative and Alternative Communication (AAC)
- communicating with speech and/or motor impairments
- AAC devices
  - high-tech devices – word/letter/phrase/icon input, speech synthesis output
- the communication rate divide and fatigue

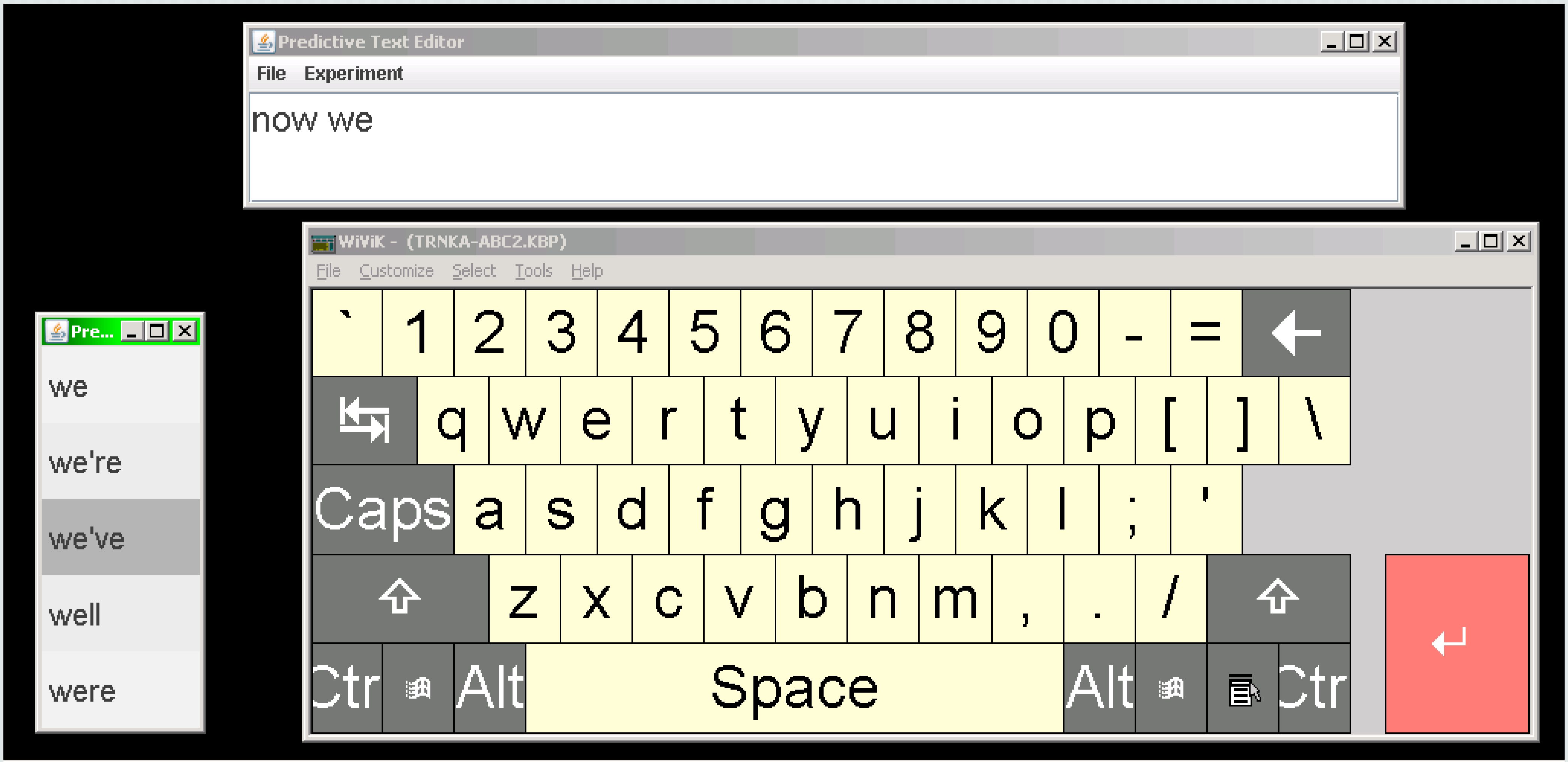
Word Prediction in AAC

- NLP technique to reduce the number of keystrokes
- predict the word currently being typed on the basis of:
  - the part of the word typed so far (can be no letters)
  - a language model (tells the likelihood of every word given the previous few words and possibly other inputs)

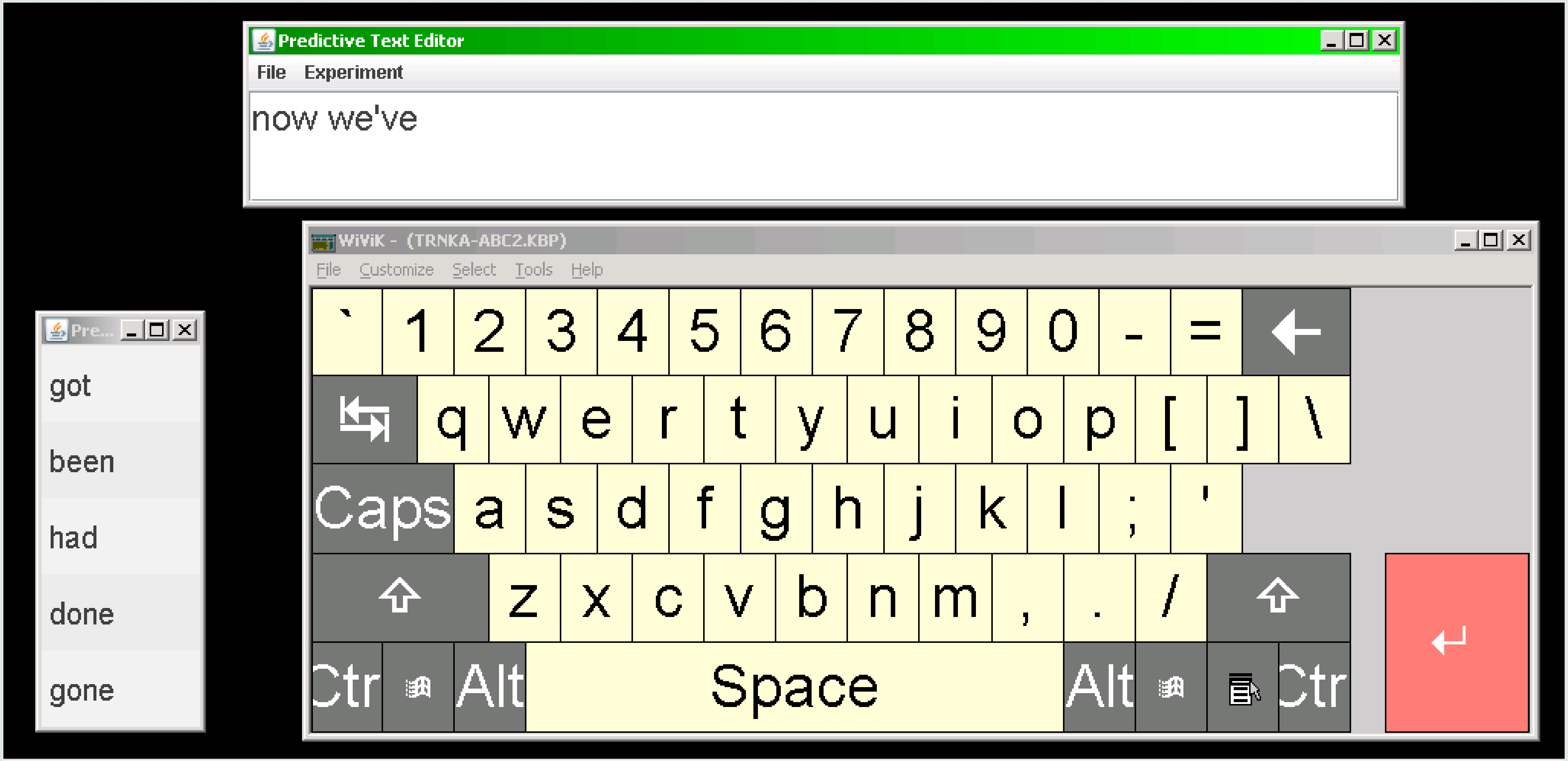
AAC Background



AAC Background



AAC Background



Predicting Words

- Steps
  - Filter the vocabulary by the prefix
  - Compute the probability of all matching words given the context (previous words)
  - Sort the list
  - Present the top W words in order

Language Modeling

- Language models provide the probability of a word in context
- Trigram models are a typical language model, focusing on previous two words:

$$P(w \mid w_{-1}, w_{-2})$$

Language Modeling

- Where do the probabilities come from?
  - train the model by estimating the probabilities from a collection of text (corpus)

Evaluating Word Prediction

- Communication rate influenced by keystroke savings
- Measure keystroke savings on testing text

$$KS = \frac{chars - keystrokes}{chars} \times 100\%$$

- Simulated ideal user
- Simulated user interface – 5 predictions

Corpora

- **Switchboard** (2.8M words) – telephone conversations
- **Micase** (545K words) – university-setting conversation
- **SBCSAE** (237K words) – mostly face-to-face conversation
- **Charlotte** (188K words) – speech from the Charlotte, NC area
- **Callhome** (48K words) – telephone conversations between friends and family
- **AAC Email Corpus** (28K words) – public mailing list archive, filtered by AAC users

Trigram baseline

- How much keystroke savings do we expect?
  - test with a trigram baseline
  - trained on Switchboard, tested on each corpus

Trigram baseline
keystroke savings

Testing corpus	Switchboard trigram
AAC Email	43.25%
Callhome	49.33%
Charlotte	49.64%
SBCSAE	43.49%
Micase	46.52%
Switchboard	60.35%

Trigram baseline

- Problem: trigram predictions not always appropriate for the overall text, catered to just the previous two words
  - adapt to the topic of the overall text

Topic Modeling

- Goal: seamlessly adapt the predictions to the topic
  - Build a separate trigram model for each topic in Switchboard
  - Combine the topic models using a weighted average
  - Weights based on similarity to the conversation

Topic Modeling

- How should we weight each topic?
  - topical similarity of the current (partial) text to each topic
  - similarity of keyword distributions weighted by:
    - frequency of use
    - recency of use (topics tend to evolve)
    - how well the keywords discriminate between topics

Topic Model Evaluation

keystroke savings

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%**

Mixed-domain training

- How can we do better?
  - train on more texts
  - use both similar data (some from the same corpus) and general-purpose data (texts from other corpora, plus some written texts)

Mixed-domain evaluation

keystroke savings

Testing corpus	Switchboard topic	Mixed trigram
AAC Email	43.53%	**52.18%**
Callhome	49.52%	**52.14%**
Charlotte	50.07%	**53.50%**
SBCSAE	43.90%	**47.78%**
Micase	46.99%	**51.46%**
Switchboard	**61.48%**	59.80%

Fine-Grained Topic Modeling

- How can we improve over mixed-domain training?
  - add topic adaptation with mixed-domain training
- Problem: limited to training on Switchboard (split by topics)
- Solution: treat <u>each document</u> as a topic
  - removes training limitations
  - called <u>fine-grained topic modeling</u>

Fine-Grained Topic Modeling

mixed-domain evaluation - keystroke savings

Testing	Switchb. topic	Mixed trigram	Fine 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%

Conclusions

- each document can be treated as a very specific topic to avoid the need to split a corpus into topics
  - much more flexible
- improvement comparable to labeled data
- very useful for mixed-domain data
- future: allows easy incorporation of user texts into topic model

Fine-Grained Topic Modeling

- How to test it?
  - in-domain - train/test on different parts of the same corpus
  - mixed-domain - similar to in-domain, but with data from other corpora added into training

Fine-Grained Topic Modeling

in-domain evaluation

Testing corpus	Trigram baseline	Fine topic
AAC Email	48.92%	**49.38%**
Callhome	**43.76%**	43.72%
Charlotte	48.30%	**48.60%**
SBCSAE	42.30%	**42.57%**
Micase	49.00%	**49.55%**
Switchboard	60.35%	**61.42%**