Corpus Studies in Word Prediction

Keith Trnka
SIGAI 5/14/2007
University of Delaware

Goals and non-goals

- This talk is:
  - sharing research
  - feedback to improve a paper submission
- This talk is not:
  - preliminary conference talk

Outline

- Background - AAC, word prediction
- Motivation
- Choosing Corpora
- Testing Methods
- Domain Variations in Baseline and Topic Models
- Vocabulary Analysis

Background

- Augmentative and Alternative Communication (AAC)
- Communication rate divide between AAC and non-AAC
- Electronic devices - letter-by-letter entry with word prediction, specialized interface for “core” words

Background

Example AAC device

Motivation

- How does word prediction affect actual users?
  - How does keystroke savings affect communication rate?
    - NAACL 2007 paper
  - How does the difference between split training/testing data and actual usage affect our expected performance?
    - This paper/presentation

Research Questions

- How is word prediction affected when tested out-of-domain?
  - *Why* is it affected?
  - Can we fix it?
  - How about topic modeling?
- How should we do out-of-domain testing?

Vijay - What is a domain? It hasn't been defined.

Kathy - Should motivate it better by explaining that other researchers only do in-domain testing.

Problems

- AAC corpora don't exist (or are miniscule)
  - The expected performance loss is higher
- Solutions
  - Make a small AAC corpus
  - Use corpora that are conversational
  - Transform corpora to be AAC-like

Choosing Corpora

- Main target: conversational speech transcriptions
  - Switchboard, Santa Barbara, Micase, Charlotte, Callhome
  - Remove speech repairs

Switchboard

- Telephone transcriptions
- Pairs of people were selected by matching topics each person agreed to talk about
- ~2.9m words in 2,438 conversations
- Utility: somewhat artificial, but big and conversational

Santa Barbara

- Santa Barbara Corpus of Spoken American English (SBCSAE)
- Natural, face-to-face conversations
- Often several speakers
- 237,191 words over 60 conversations
- Utility: ideal content, just not AAC. Small size.

Micase

- Michigan Corpus of Spoken Academic English (Micase)
- Spoken transcriptions from University of Michigan
- Advisor-advisee meetings, discussion/study groups, etc
- 545,411 words across 50 conversations
- Utility: useful to approx. specialized conversation

Callhome

- Telephone conversations between friends and family
- 48,407 words across 24 conversations
- Utility: near-ideal, just not AAC. Very small size.

Charlotte

- Charlotte Narrative and Conversation Collection
- narratives, conversations, and interviews in a county in North Carolina
- 187,597 words across 93 conversations
- Utility: ideal, just not AAC. Small size.

AAC Email

- public online email list archive, selected emails from known AAC users
- Some emails are conversational, but many are policy-opinion texts or help service
- 27,710 words across 117 emails (only 2 people)
- Utility: ideal, except for size

Slate Magazine

- Online magazine like a newspaper
- 4.2m words across 4,531 articles
- Utility: useful to approx. written text and general English

Cleanup

- Conversation cleanup
  - Removed many speech repairs
- Email: removed signatures
- Slate: removed titles
- Written/formal: un-embedded parentheticals, quoatations

Corpora Summary

Corpus	Medium	Word count
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	*spoken*	*3,902,380*
Slate	written	4,178,543

Testing Methods

- Communication rate requires users and is tedious
  - dependent on keystroke savings (NAACL 2007)
- Keystroke savings: $\frac{KS = chars - keystrokes}{chars}$
- Word prediction vs. word completion
- Fringe words only

Testing Methods

- Cross-validation
  - 11-fold cross-validation
  - In run 0, selected set 0 from the testing corpus, all non-0 sets from training corpora

Testing Methods

- Domain variations
  - In-domain (training = testing)
  - Out-of-domain (training = all corpora but testing)
  - Mixed-domain (trained on all corpora)

Testing Example: In-domain

Corpus	Set 1	Set 2	Set 3	Set 4	Set 5
BNC					
ANC					
WSJ					
NYT					

green = training, red = testing

Cross validation run 1 on BNC

Testing Example: In-domain

Corpus	Set 1	Set 2	Set 3	Set 4	Set 5
BNC					
ANC					
WSJ					
NYT					

green = training, red = testing

Cross validation run 2 on BNC

Testing Example: In-domain

Corpus	Set 1	Set 2	Set 3	Set 4	Set 5
BNC					
ANC					
WSJ					
NYT					

green = training, red = testing

Cross validation run 1 on ANC

Testing Example: Out-of-domain

Corpus	Set 1	Set 2	Set 3	Se	
BNC					
ANC					
WSJ					
NYT					

green = training, red = testing

Cross validation run 1 on BNC

Vijay - Why not include the Set 1 from the OOD corpora? Also, that would negate the need for any sort of cross-validation - you'd just test on a whole corpus and train on the rest.

Testing Example: Out-of-domain

Corpus	Set 1	Set 2	Set 3	Set 4	Set 5
BNC					
ANC					
WSJ					
NYT					

green = training, red = testing

Cross validation run 1 on ANC

Testing Methods

- Domain variations
  - Most researchers do in-domain testing and then testing on one out-of-domain corpus
    - Answers the question: How will it perform on real data?
    - **Doesn't** answer the question: How will it perform on real data relative to what an in-domain model would've done?

Testing Methods

- Domain variations
  - Out-of-domain **training** vs out-of-domain **testing**
  - Factors for OOD testing performance:
    - How similar are the training/testing corpora?
    - How difficult is the testing corpus?

Language Modeling and Word Prediction

- Basic method (inefficient):
  - 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 Models

- Standard baseline - trigrams with backoff
  - Dictionary from YAWL as final backoff
- Topic modeling
  - Using topics annotated in Switchboard:

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

Language Models

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

- First part - estimated with unigram cosine similarity (other tricks too)
- Second part - a separate trigram model for each topic
- Implementation - computational and sparseness problems, so use frequency-based interpolation:

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

Trigram predictions

Corpus	In-domain	Out-of-domain
AAC Email	48.92%	47.89%
Callhome	43.76%	52.95%
Charlotte	48.30%	52.44%
SBCSAE	42.30%	46.97%
Micase	49.00%	49.62%
Switchboard	60.35%	53.88%
Slate	53.13%	40.73%

$$\frac{KS = chars - keystrokes}{chars}$$

Trigram predictions

Corpus	In-domain	OOD	Mixed
AAC Email	48.92%	47.89%	**52.18%**
Callhome	43.76%	52.95%	**53.14%**
Charlotte	48.30%	52.44%	**53.50%**
SBCSAE	42.30%	46.97%	**47.78%**
Micase	49.00%	49.62%	**51.46%**
Switchboard	**60.35%**	53.88%	59.80%
Slate	**53.13%**	40.73%	53.05%

$$\frac{KS = chars - keystrokes}{chars}$$

Topic-based predictions

- Common criticism: topics are domain-specific, so they won't translate to user benefit
- Topic is only annotated in Switchboard
  - Baseline: trigram model trained on Switchboard
  - Topic: trigram topic model trained on Switchboard

Topic-based predictions

Corpus	Baseline	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%***
Slate	39.17%	**39.78%**

Statistically significant!

Three factors in testing

- Size of training data
  - Easily measurable
- Similarity of training and testing data
  - Can be placed on the scale (ID, MD, OOD)
- Intrinsic difficulty of testing data
  - Hasn't been isolated *yet!*

Vocabulary Analysis

- Common analyses
  - Named entities
  - Out of vocabulary words (OOVs)

Named Entities

- Percentage of named entites
- Tends to follow written vs. spoken trend
- Switchboard: topic effect?
- Slate: news effect?

Corpus	NE
AAC Email	8.92%
Callhome	8.23%
Charlotte	6.59%
SBCSAE	5.67%
Micase	3.12%
Switcboard	2.10%
Slate	12.03%

Out-of-vocabulary words

- Problem: need a reference vocabulary
- Solutions
  - Google unigram model trained on ***1 trillion words***
  - Self-test OOV using cross-validation

Google OOVs

- Percentage of words in the corpora that didn't appear in the Google LM
- Captures how specific each corpus is
- Somewhat correlated with OOD performance

Corpus	OOV	OOD KS
AAC	0.81%	47.89%
Callhome	0.38%	52.95%
Charlotte	0.37%	52.44%
SBCSAE	0.77%	46.97%
Micase	1.35%	49.62%
Switcboa	0.22%	53.88%
Slate	2.12%	40.73%

Self-OOVs

- Percentage of words in the corpora that didn't appear in the training sets of the same corpus using cross-validation
- Good measure of difficulty, but compounded by size
- Somewhat correlated with in-domain performance

Corpus	OOV	ID KS
AAC	8.48%	48.92%
Callhome	6.86%	43.76%
Charlotte	4.49%	48.30%
SBCSAE	5.76%	42.30%
Micase	4.40%	49.00%
Switcboa	0.52%	60.35%
Slate	1.96%	53.13%

Intrinsic difficulty

- How else could we determine the intrinsic difficulty of a corpus? (future work)
  - In-domain testing using fixed-size training data
  - Very large general-purpose LM performance/ perplexity (e.g. Google)

Open problems

- What is an appropriate approximation of an AAC corpus?
- What to do when topics aren't annotated in the training corpus? (half-solved)
- How should the separating of performance into factors affect research?

Conclusions

- The AAC Email corpus exhibits some similarities to non-AAC text, so the approximation is decent
- A larger amount of out-of-domain text is equivalent to a smaller amount of in-domain text for word prediction
- A combination of in-domain and out-of-domain text is good (Wandmacher and Antione, LREC 2006)
- Developed a preliminary framework for domain-varied training

Questions?