Adapting Word Prediction to Subject Matter without Topic-labeled Data

Keith Trnka
University of Delaware
Newark, DE 19716
trnka@cis.udel.edu

ABSTRACT

Word prediction helps to increase communication rate when using Augmentative and Alternative Communication devices. Basic prediction systems offer topically inappropriate predictions for the context, thus we adapt the predictions to the topic of discourse. However, previous work has relied on texts that are grouped into topics by humans. In contrast, we avoid this restriction by treating each document as a topic. The results are comparable to human-labeled topics and also the method is applicable to unlabeled text.

Categories and Subject Descriptors: I.2.1 [Applications and Expert Systems]: Natural language interfaces; I.2.7 [Natural Language Processing]: Language models

General Terms: Algorithms, Experimentation

1. INTRODUCTION

Augmentative and Alternative Communication (AAC) systems help people with speech impairments to communicate by speaking words for them. The person using the system selects letters and words to enter sentences to be spoken. However, many AAC users also have reduced motor control, causing communication with AAC devices to be slow.

Word prediction reduces the effort of producing text and increases communication rate by predicting the desired words and allowing the user to select them for only one keystroke. The list of predicted words is generated using a *language model*: a technique from the field of natural language processing that estimates the likelihood of a word in a sequence of words. The most common type of language model is an *ngram model*: this method processes a large amount of text to count the number of times each word occurs after the list of the previous $n-1$ words. The process of estimating the counts from text is called *training* and evaluating the resulting model is called *testing*.

AAC devices are used for a wide variety of topics. Devices are used in school, in the home, in recreation, and for many other topics. However, standard ngram models are best suited for the predominant topics and styles of training data. If the ngram model is trained on school books, communicating in school may be very easy, but communicating in the home may be very difficult. Our high-level research goal is to develop language models that are suitable for the wide variety of topics and styles encountered with AAC devices.

ASSETS'08, October 13–15, 2008, Halifax, Nova Scotia, Canada.
ACM 978-1-59593-976-0/08/10.

In this work, we focus on adapting to the keywords of the current text to provide more topically-relevant predictions.

Several researchers have adapted word prediction to the topic of discourse. Trnka and McCoy [4] trained separate ngram models for each topic in a corpus and tuned the overall model to the most similar topics. However, the topics were defined manually, whereas a larger collection of text may not be labeled for topic. Wandmacher and Antoine [5] performed Latent Semantic Analysis (LSA) on the training data to automatically learn the relatedness of words and tune the predictions to related words. However, LSA can be very computationally intensive and also is difficult to reliably update with additional text. Matiasek et al. [3] learned trigger pairs from the training text — pairs of words which tend to appear together in the same document. The predictions were tuned to match previously used words. However, the trigger pairs contributed very little to actual performance. Li and Hirst [1] utilized a combination of trigger pair methods and semantic relatedness. However, their work focuses on short-distance relations rather than matching words to the overall theme of the document.

We propose an extension to [4] by treating each document as a topic, similar to information retrieval approaches to topic adaptation [2]. In this way, the adaptation matches the (partial) current document to other documents in training and tunes the predictions to similar training documents. This approach has three advantages over prior work: 1) the only training requirement is a collection of documents 2) topic adaptation can be as specific or general as desired, based on the similarity score and 3) user texts can be easily integrated into the training data after being used for testing.

2. LANGUAGE MODELS

We adapt the language model to the topic of discourse by focusing the model on the most similar documents from training. The adaptation roughly follows this equation:

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

where $P(d \mid h)$ is the estimated relevance of any document d in training to the (partial) current document h, implemented as a normalized similarity score. This weight is higher for documents that use similar keywords to the current document. $P(w \mid w_{-1}, w_{-2}, d)$ is a trigram model measured from only document d by dividing the number of times the word triple (w_{-2}, w_{-1}, w) occurs in document d by the number of times the word pair (w_{-2}, w_{-1}) occurs.

The relevance scores for each document $P(d \mid h)$ initially start with the value $1/docs$. Therefore the initial predictions are the same for our adaptive model compared to an unadaptive trigram model. As the user types words, the relevance scores will increase for documents using the same keywords and decrease for all other documents. Thus the model will be weighted to favor documents that seem similar to the document the user is currently typing.

We evaluate the topic model by comparison to an appropriate baseline — an unadaptive trigram model.

3. EVALUATION METHODS

The quality of the language model is evaluated using *keystroke savings*, which measures the percentage reduction in the number of keystrokes to produce a text compared to letter-by-letter entry. For example, if a system only requires one keystroke to produce the letters "the", then the system has achieved 66.7% keystroke savings for those characters. In larger evaluations, a computer program simulates a person typing text(s) with word prediction, assuming the prediction list is 5 words long. The system selects the desired word as soon as possible and compares the number of keystrokes required with word prediction to the number of characters in the text.

One problem in evaluating word prediction systems is that the results are highly affected by the texts used for training and testing. In real-world usage, the quality of the system will be dependent on the similarity of the training texts to actual usage. For example, if the language model is trained on weather forecasts, the system will provide very high keystroke savings for discussing the weather, but the system may perform very poorly for discussing a different domain such as new movies. To control for the effects of different corpora (collections of documents), we use the domain-varied evaluation of [4], which systematically varies the texts used for training and testing. Each corpus is used individually for testing, shown in the first column of both tables. We primarily use collections of speech transcriptions common in natural language processing, but also test on a small collection of emails from AAC users.

The results are not dependent on only the testing data, but also the training data. The major impact of the training data is the similarity to the testing data, so we vary the similarity of the training data for each testing set. The in-domain evaluations in Table 1 split each corpus into training and testing portions, similar to most research evaluations. This represents the highest similarity of training and testing data, and is overly optimistic for assessing real-world performance. The mixed-domain evaluations in Table 2 perform the same split as in-domain evaluation, but also add texts from all other corpora into training. This test is more realistic of real-world performance, where we have some highly similar training data (potentially logged user data), but the training data is predominantly different than actual usage.

4. RESULTS

The results of in-domain evaluation are shown in Table 1. Topic modeling improves keystroke savings on all corpora but Callhome with statistical significance. Although there is a decrease for Callhome, the result is not statistically significant. Trnka and McCoy [4] applied human-annotated topics in the same framework, but the only comparable result is the evaluation on Switchboard — topic modeling with human-annotated topics produced 61.48% keystroke savings compared to 61.42% when documents are treated as topics.

Table 1: In-domain evaluation

Testing corpus	Trigram	Document-topic
AAC Emails	48.92%	49.38% (+0.45%)
Santa Barbara	42.30%	42.57% (+0.28%)
Callhome	43.76%	43.72% (-0.05%)
Charlotte	48.30%	48.60% (+0.30%)
Micase	49.00%	49.55% (+0.56%)
Switchboard	60.35%	61.42% (+1.07%)

Table 2: Mixed-domain evaluation

Testing corpus	Trigram	Document-topic
AAC Emails	52.18%	53.14% (+0.96%)
Santa Barbara	47.78%	48.84% (+1.06%)
Callhome	53.14%	53.39% (+0.26%)
Charlotte	53.50%	53.92% (+0.42%)
Micase	51.46%	53.13% (+1.67%)
Switchboard	59.80%	61.17% (+1.37%)

The results of mixed-domain evaluation are shown in Table 2. Topic modeling improves keystroke savings on all corpora with statistical significance. The topic modeling results are the best results to date for all corpora except Switchboard, demonstrating the ability of the model to focus on the best documents in the training data while still leveraging all available data when appropriate.

5. CONCLUSIONS AND FUTURE WORK

We have improved on previous methods in topic adaptation by treating each document as a topic. The results of in-domain evaluation on Switchboard are comparable to using human-labeled topics. The mixed-domain evaluation demonstrates the ability of the method to balance the highly relevant in-domain texts with the large amount of additional general-purpose texts. This allows topic modeling to be applied with any corpus, not only corpora with topic labels.

6. REFERENCES

[1] J. Li and G. Hirst. Semantic knowledge in word completion. In *ASSETS*, pages 121–128, 2005.

[2] M. Mahajan, D. Beeferman, and X. D. Huang. Improved topic-dependent language modeling using information retrieval techniques. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 541–544, 1999.

[3] J. Matiasek and M. Baroni. Exploiting long distance collocational relations in predictive typing. In *EACL-03 Workshop on Language Modeling for Text Entry*, pages 1–8, 2003.

[4] K. Trnka and K. F. McCoy. Corpus Studies in Word Prediction. In *ASSETS*, pages 195–202, 2007.

[5] T. Wandmacher and J. Antoine. Methods to integrate a language model with semantic information for a word prediction component. In *EMNLP-CoNLL*, pages 506–513, 2007.