

Corpus Studies in Word Prediction

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

Kathleen F. McCoy
University of Delaware
Newark, DE 19716
mccoy@cis.udel.edu

ABSTRACT

Word prediction can be used to enhance the communication rate of people with disabilities who use Augmentative and Alternative Communication (AAC) devices. We use statistical methods in a word prediction system, which are trained on a corpus, and then measure the efficacy of the resulting system by calculating the theoretical keystroke savings on some held out data. Ideally training and testing should be done on a large corpus of AAC text covering a variety of topics, but no such corpus exists. We discuss training and testing on a wide variety of corpora meant to approximate text from AAC users. We show that training on a combination of in-domain data with out-of-domain data is often more beneficial than either data set alone and that advanced language modeling such as topic modeling is portable even when applied to very different 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

Experimentation, Measurement

Keywords

word prediction, statistical methods, language modeling, corpora

1. INTRODUCTION

A fundamental problem in the field of Augmentative and Alternative Communication (AAC) is that the communication rate of AAC users is far below the communication rate of speech. AAC devices are often electronic devices that take word and letter input and produce speech. The slow speed of typing creates a communication divide which can

cause communication partners to lose interest or attempt to dominate the conversation. Word prediction is an application of Natural Language Processing (NLP) to AAC devices that allows words to be predicted and selected for fewer keystrokes.

Our word prediction system relies on statistical methods, where the basis for the word prediction system is a language model that has been trained on a large corpus of data. Such a model, which has traditionally been used in such tasks as speech recognition, then forms the basis of predicting the next word of input based on what the user has already typed. Generally such systems are trained on a corpus, and then evaluated by calculating theoretical keystroke savings on some held-out data. This evaluation then drives further research by focusing on correcting poor predictions. Ideally a word prediction system should be trained and tested on language that is similar to the expectations for actual use. However, no such corpora exist for AAC users. Instead, we must use other corpora for training and testing.

However, the results of evaluation are heavily dependent on the characteristics of the texts used for training and testing. One of the often noted factors in evaluation is the effect of the number of words used to train a language model [11]. Language models built from larger corpora tend to perform much better, particularly on words that are infrequent. However, another determinant of performance is how well the language training data reflects the actual language the system is to be run on. Statistical systems tend to perform poorly when they are applied to language very different from the training texts [4, 23, 15]. Thus, important questions include: What text should be used to evaluate a word prediction method? And what text should be used to compute the statistics required by language models?

In this paper we investigate corpus issues in developing language models that form the basis for word prediction systems. We vary the corpus used in training and testing to show how trigram models and topic models are affected by the differences between the text used in training and testing.

Traditionally, most researchers have performed what we call *in-domain evaluation* [11, 10, 7, 5, 12] — a corpus of text is split into training and testing sections, where the training section is used to build the ngram language model and the testing section is used to evaluate the quality of the predictions. Splitting a corpus into training and testing sets is the most common means of evaluation in NLP because it gives a “fair” evaluation of each method of language model development. This allows two different methods to be compared by holding constant the training and testing data so that any

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ASSETS'07, October 15–17, 2007, Tempe, Arizona, USA.
Copyright 2007 ACM 978-1-59593-573-1/07/0010 ...\$5.00.

differences in the methods can be attributed to language model development (such as [13, 2, 17, 6]). However, this evaluation gives only a vague indication of end-user benefit from various techniques.

Since the actual use of the system might be with language data that is quite different from the training corpus, some researchers have performed *out-of-domain evaluation* [22, 3, 14] — evaluating their predictions on text not from the training corpus. This approach gives a more reasonable estimate of real-world performance, especially in situations where the actual user text is likely to differ from the training corpus. The same approach is sometimes used in the more general field of language modeling to validate that the performance improvement on in-domain data still applies when tested out-of-domain (e.g., [2]). However, out-of-domain testing isn't always a fair evaluation to verify that new techniques are working. Poor results in the out-of-domain testing corpus may be a result of differences in language use between the training and the testing corpus and may not carry over to the actual domain of use. For instance, trying to apply a method like topic modeling to a testing corpus containing mostly general conversations is unlikely to show much difference between different variations in topic modeling, while the real-world difference may be highly significant.

Some researchers have also performed *mixed-domain evaluation* where some of the training data is from the same corpus as the testing data, but much of it is from other corpora. The incorporation of recency information into the prediction method is an example of this [23] — the dynamically updated collection of user text is in-domain and the baseline model is out-of-domain. Mixed domain evaluation has also been used in a part-of-speech (POS) framework, where the sequences of POS tags are trained using a very small in-domain corpus and the word-based probabilities are trained using a larger out-of-domain corpus [4].

Realistically, the question of in-domain vs. out-of-domain testing is a spectrum — training on phone calls between family and testing on face-to-face communication between family is closer to in-domain testing than training on phone calls between family and testing on newspaper articles. There are various dimensions along which corpora can be similar or different, but we can group the dimensions roughly into topic and style, where topic covers the content of communication and style describes variations such as formality, speech repairs, opinionated vs. objective communication, common vs. uncommon word choices, etc.

In addition to the problem of selecting similar or dissimilar text for training and testing, developing a system for AAC users is hampered by the lack of substantial corpora of AAC text. We feel that non-AAC conversations will use longer sentences with many more speech repairs than AAC text. Therefore, evaluations on non-AAC text may not be representative of AAC users. The three ways in which we address this problem are 1) to select corpora that are similar to AAC text, 2) to transform text to be more AAC-like when possible, and 3) to construct a small corpus of AAC text.

In this work, we show that a large amount of out-of-domain training data is more beneficial in statistical word prediction than a small amount of very similar language. Furthermore, the combination of in-domain training data with a much larger amount of out-of-domain data is more useful than either data set alone, even when the two train-

ing sets are combined naïvely. Beyond this, we show how language modeling improvements can still hold even when the training and testing languages are very different — we apply a topic model from [21] both in-domain and out-of-domain and show that the topic model significantly improves keystroke savings despite the topical differences in training and testing corpora.

Section 2 will give an overview of highly related work. Section 3 describes the corpora and the preprocessing used to make them more AAC-like. Section 4 describes our general evaluation framework and our ngram models. Section 5 presents the evaluation of domain-varied testing with a trigram model and Section 6 presents the evaluation of a topic model with out-of-domain training. Section 7 presents analyses of the vocabulary of the corpora in an effort to better characterize the text as well as explain the trends of Sections 5 and 6. Section 8 discusses the major findings and implications. Sections 9 and 10 summarize future work and conclude.

2. RELATED WORK

Our work is most similar to Wandmacher and Antoine's work in developing adaptive ngram models to combat differences between training and testing data [23]. They trained a trigram model on 5.6 million words of French news text and applied that language model to word prediction for different types of text — newspaper, scientific, literature, speech, and email. They keep the training text constant throughout their studies. In contrast, we vary our training text and experiment on English word prediction. While they are primarily interested in comparing testing on news text with out-of-domain text, we are primarily interested in comparing results from testing on the same data but with different training sets. The goal of their work is the development of an adaptive language model that lessens the loss of performance with out-of-domain testing, whereas our goal is to compare mixed-domain performance to in-domain and out-of-domain performance and to evaluate topic modeling when applied out-of-domain.

Other related work includes [2], which showed that a Latent Semantic Analysis approach to speech recognition (somewhat similar to topic modeling) still offers performance improvement when tested on out-of-domain material. [11] varied ngram order and training text size rather than domain, validating intuitions regarding the relation between training text size and keystroke savings. [20] seeks to answer the same general question as this work (how will word prediction help real users?) but studies the connection between keystroke savings and communication rate.

3. CORPORA

AAC devices are used for a wide variety of communication needs — everything from spontaneous conversation to preplanned speeches to homework assignments and technical articles. For this reason, we feel that a variety of texts should be used to evaluate word prediction. However, obtaining a large corpus of AAC text can be difficult. To address this issue, we have assembled a variety of corpora.

Because spoken conversation is arguably the most common use for an AAC device, we first assembled several corpora of spoken English and performed cleanup processing to remove speech repairs, bringing the text closer to what an

AAC user might say. Because AAC devices can be used for writing as well, we also assembled a small collection of emails from AAC users and included a corpus of written text to get a rough idea of any differences in word prediction in written vs. spoken text. Each of the corpora came with their own formatting conventions: most were in all lower case (except for proper names) but some (primarily the written corpora) used a capital letter to start each new sentence. Some omitted punctuation while others included commas and other punctuation symbols within a sentence. To best reflect the primary usage of AAC devices in speech, we reformatted the corpora to reflect a standard style. Therefore, “I” and contractions such as “I’ll” were capitalized in all corpora. The first word of each sentence was converted to lowercase unless it was a known named entity, which would remain capitalized. Also, punctuation between words in a sentence was removed. We feel that these changes make the collection of corpora more AAC-like and facilitate a more fair evaluation of word prediction. The summary word counts of each corpus are shown in Table 1 — the overall collection is roughly half spoken and half written.

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
<i>Total spoken</i>	spoken	3,902,380
Slate	written	4,178,543

Table 1: Word counts for each corpus (see corpus-specific sections below for more details)

3.1 Conversational Speech Transcriptions

The primary target of research in word prediction is conversational usage of AAC devices. The real-time nature of spoken conversations creates a communication rate divide which word prediction attempts to lessen. However, AAC devices are used for a wide variety of conversation. We evaluate word prediction on transcriptions of spoken conversation as well as on written text and email.

Ideally, we would like to evaluate word prediction with conversational AAC text, but thus far, such a corpus has been unavailable. Instead, we will evaluate word prediction on several conversational speech texts that have speech repairs removed in an effort to bring the corpora closer to what an AAC user would type.

3.1.1 Speech repair removal

Transcribed conversational text is characterized by frequent speech repairs. Many speech repairs are the result of “getting ahead of oneself”, such as suggested by [18]. However, AAC communication rate is more limited by the speed of producing words rather than the speed of planning a message. For this reason, speech repairs were removed when they could be easily identified. We follow the work of [8] in processing simple speech repairs such as backchannels (e.g., uh, um), repetitions, and limited cases of word replacements.

We used pauses, abandonment of words, and backchannels as candidates for being an editing signal, depending on what was annotated in each corpus. For example, one corpus (Switchboard) clearly marked words as abandoned, whereas in other corpora we relied on commas and multiple periods to signal pauses. Each sentence would have candidate editing signals identified and then lexical pattern matching was performed on the words to the left and right of the potential editing mark. In the case of abandoned words, exact matching wasn’t required. However, some speech repairs (e.g., “I I would ...”) were not signaled. Therefore, we created special processing for single-word repetitions: repeated lowercase words were considered speech repairs unless they appeared in an exceptions list. Repeated uppercase words were considered *legitimate repetitions* unless they appeared in an exceptions list, which primarily contained derivations of “I”. Any backchannels that remained after speech repair removal were filtered out. The resulting “cleaned” text was far easier for the authors to read and is much closer to what we think an AAC user would have said.

3.1.2 Switchboard

The Switchboard corpus is a collection of 2,438 English phone conversations recorded by Texas Instruments using a variety of speakers and topics [19]. Participants indicated which of the predefined topics they were comfortable discussing and the experimental software connected two subjects to speak about a particular topic, given by a prompt such as “Find out what kind of fishing the other caller enjoys...” After the speech repair cleanup, there are roughly 2.9 million words in Switchboard — more than any other corpus of speech that we used. Although the task-focused nature of Switchboard makes it slightly unrealistic of unprompted day-to-day conversations, we feel that the large size of Switchboard outweighs any small dissimilarities.

3.1.3 SBCSAE

The Santa Barbara Corpus of Spoken American English (SBCSAE) [16] is a collection of 60 recorded conversations, which are predominantly face-to-face communications and have been collected to sample a wide variety of speakers. As an example of the variety found in SBCSAE, it contains a social conversation held over lunch, a conversation on a ranch, and a church sermon. Although SBCSAE spans a wide variety of topics, it contains a mere 237,191 words — roughly 8% of the size of Switchboard. However, the natural nature of the text in SBCSAE is a step closer to the conversational communication of AAC devices.

3.1.4 Micase

The Michigan Corpus of Spoken Academic English (Micase) is a collection of university-setting spoken English. Several example conversations are advisor-advisee discussions or moderated class discussions. We obtained a portion of the Micase corpus though the second release of the American National Corpus (ANC) [1]. Special processing was added for parentheticals and quotations to focus the ngram model on the proper conditioning information. The Micase data we used contained 545,411 words across 50 conversations. Although the text isn’t representative of most day-to-day speech, it should be representative of word prediction performance for other highly specialized conversations, such as speech in the workplace.

3.1.5 Callhome

The Callhome corpus is represented in part in the ANC corpus, and contains 24 telephone conversations between friends and family. This free-form conversation is very representative of day-to-day communication and is very appropriate to approximate AAC user text. Callhome contains a mere 48,407 words, but like SBCSAE, although the text is relatively small, it is a valuable approximation of day-to-day AAC user conversation.

3.1.6 Charlotte

The Charlotte Narrative and Conversation Collection (Charlotte) is a collection of 93 narratives, conversations, and interviews centered around an area in North Carolina, USA, available as part of the ANC corpus. Charlotte contains 187,597 words, about 80% of the size of SBCSAE. This corpus is very similar in its conversational nature to Callhome and SBCSAE, and is therefore useful despite its small size.

3.2 AAC Email Corpus

AAC user text has been difficult to obtain, but one resource we found was a publicly available AAC user mailing list archive. We surveyed emails from this archive and collected emails from AAC users. The resulting corpus contains 117 emails and 27,710 words. Like several other corpora, this data is useful despite its small size because it's a test directly applicable to the target, AAC users. Email processing presented new challenges for cleanup processing. Signature text was removed, as an email user only types the text once, not for each email sent. Quoted emails in replies were also removed. In addition, parentheticals and quotations were extracted like with Micase.

3.3 Slate Magazine

Slate Magazine is an online publication covering a wide range of topics, similar to a newspaper. The ANC project contains 4,531 articles from Slate published in a span of 4 years. At 4,178,543 words, Slate is the largest corpus in this study. We feel that it serves as an approximation of one kind of written AAC text as well as a general-purpose corpus of English. The cleanup processing for Slate was similar to the AAC Email Corpus with the exception of small adjustments to make articles in Slate more natural.

4. METHODS

First, we will give an intuitive explanation of how ngram modeling works and how it is used for word prediction. Then we will present our evaluation methods, including the keystroke savings metric, usage of cross-validation for more reliable results, and the way we will vary domains in evaluation. Finally, we will present the trigram model used to study domain-varied evaluation and the topic model used to evaluate the robustness of the technique.

4.1 Ngrams in Word Prediction

The premise behind using ngram models for word prediction is the idea that a word is primarily dependent on the previous few words. In word prediction, when a sequence of words in a sentence is seen, the ngram model asks the question "What words have I seen in training that followed these ones?" To do this, an ngram model is built from some training data by recording how often each word follows a sequence of words. The number of words in the sequence of

prior words determines the *order* of the ngram model. If only the previous two words are considered, then it is a 2nd-order Markov model and is called a trigram model. Similarly, a 1st-order model is called a bigram model and a model that ignores the previous words is called either a unigram model or frequency model.

In word prediction, a trigram model would generate a list of predictions of words that followed the previous two words, sorted in descending order by probability (which is computed from the frequency). However, the number of possible pairs of two previous words is very large, and it is unlikely that all possible pairs will have been seen in training. Therefore, some sort of fallback strategy is necessary for the situation in which the combination of the previous two words has never been seen. A common strategy in language modeling is to use Katz' backoff [9] to generally say that a bigram model should be consulted if the trigram model is unable to give predictions and a unigram model should be consulted if even a bigram model is unhelpful. In fact, Katz' backoff goes further: some amount of probability of the words to follow a sequence of two words is held out to redistribute to words that followed the previous one word (and likewise, some probability is held out to redistribute to word without any prior words in the worst case). The amount of probability held out is in proportion to the reliability of the distribution.

4.2 Evaluation

Word prediction is evaluated by how many keystrokes it saves over manually typing a piece of text. Although an end user is more concerned with communication rate (words per minute), communication rate has been shown to increase with increased keystroke savings [20].

We evaluate prediction methods using the keystroke savings offered by 5 predictions.¹ Keystroke savings is computed using the formula below, where *chars* is the number of characters in the text, including spaces and newlines. *keystrokes* is the minimum number of key presses required to enter the text using word prediction, including the keystroke to select a prediction from the list and a key press at the end of each utterance. For example, suppose a user is typing "the car." Non-predictive text entry requires 4 key presses to type "the" and a space afterwards. If a prediction system could guess "the" before typing the "t" and the user selected it, the system has achieved 75% keystroke savings for that word.

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

Due to the trend of some notable AAC manufacturers to provide a static interface for the very small set of vocabulary deemed *core words* (e.g., "the", "am", "me") and word prediction for all other words, we limit our evaluation to non-core words.

Results on the smaller corpora (all except Switchboard and Slate) are measured using 11-fold cross-validation. This ensures that we have much more reliable comparisons of keystroke savings and also prevents overfitting the data.

¹Researchers have evaluated word prediction methods with many different prediction windows (ranging from 1 to 20). A list of 5 words seems to be the most common [5, 14]. In our past experience [21], we found that the differences between methods using the same ngram order were roughly the same regardless of window size.

4.3 Domain Variations

The overall goal of this research is to evaluate how users will benefit from word prediction in practice. The narrower goal of this paper is to investigate the effects of training data — in practice, an AAC device is using a language model from a different topic or style to predict words in their real conversation. We use three tests to evaluate this for each corpus: in-domain, out-of-domain, and mixed-domain training. In-domain training uses the same corpus for both the training and testing sets. Out-of-domain training uses the training sets of all corpora except the corpus used for testing. Mixed-domain training uses the training sets of all corpora and evaluates on the testing set of each corpus. In-domain training is the most common means of evaluating word prediction (e.g., [11, 10, 7, 5, 12]), so it forms a baseline to which other training sets can be compared. In-domain performance is determined primarily by the size of the corpus and the intrinsic complexity of the corpus (where this includes the corpus’ self-similarity). Out-of-domain training shows the expected degradation (or improvement) of performance resulting from using word prediction in a realistic scenario. The difference between in-domain and out-of-domain performance should be proportional to the difference in training data size and the differences in language between in-domain and out-of-domain text. Mixed-domain training approximates what a high-performance system might do: incorporate user text back into training in addition to a large out-of-domain data set. Mixed-domain performance should be scrutinized with respect to *both* out-of-domain and in-domain performance, as it subsumes both training sets.

4.4 Trigram Predictions

Trigram modeling with backoff is a standard technique for language modeling and is a common baseline in word prediction research [11, 5, 12]. The vocabulary seen in training is filtered by the prefix of the word that has been entered so far (if any) and then this list is sorted in decreasing order of probability. The most likely W words are used to build the prediction window, in this case $W = 5$. If the language model can’t produce 5 predictions, then the remaining predictions are filled in alphabetical order from a dictionary of about 200,000 words.²

4.5 Topic-adapted Predictions

Topic-based word prediction has been studied and found to improve keystroke savings for in-domain training/testing [10, 12, 21], however, an AAC device is normally used on language different than the training text. Ideally, we would like to train language models on topics that are hand-crafted. However, of the corpora we used, only Switchboard is labeled for topic. We apply an approach like Trnka et al.’s Method A [21] using trigrams for evaluating a topic model trained on Switchboard. The dictionary used for trigram predictions was also used for topic-adapted predictions as a final step of backoff.

5. TRIGRAM PREDICTIONS ACROSS DOMAINS

We evaluated word prediction using a standard trigram model with in-domain training and out-of-domain training for each corpus, shown in the first two columns of Table 2.

²Taken from the Yet Another Word List distribution.

Out-of-domain training doesn’t perform as poorly as we expected. In most cases the larger size of the out-of-domain data was beneficial and increased the keystroke savings over in-domain training. Interestingly, a notable exception to this rule is the AAC Email Corpus which performs about a percent worse using a much larger amount of out-of-domain data (almost 300 times the in-domain training data). Note as well that when the in-domain training corpus is quite large (e.g., Switchboard and Slate), in-domain outperforms out-of-domain most likely because the out-of-domain training doesn’t offer much in the way of training data size over the much more similar in-domain training data.

We also evaluated an estimation of what a user-adaptive model might do — we used both the in-domain and out-of-domain training texts for mixed-domain training, as shown in the last column of Table 2.

Corpus	Training domain		
	In	Out	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%

Table 2: Keystroke savings of in-domain vs. out-of-domain vs. mixed-domain training. The maximum keystroke savings for each row is shown in bold.

Mixed-domain training shows that even a simplistic mix of a small amount of in-domain data with a large amount of out-of-domain data can increase keystroke savings. The most notable increase here was found in the AAC corpus which improved (3.3% – 4.3%) over both in-domain and out-of-domain training. Callhome, Charlotte, SBCSAE, and Micase also save more keystrokes using a mix of training data over either training set alone. The larger corpora, Switchboard and Slate, show a performance loss with mixed training over the in-domain models — the out-of-domain data “distracts” the language model from the in-domain data. This distraction is a similar trend for all corpora with in-domain training, however, the in-domain trigram models for Switchboard and Slate were already fairly reliable, whereas the in-domain trigram models were much less reliable for the much smaller corpora. The performance improvement on the AAC Email Corpus in particular is astonishing considering it contributes such a small fraction of the probability mass of the learned model.

6. TOPIC MODELING ACROSS DOMAINS

Switchboard is the only corpus in this study that has topic labels, so to approximate a general-purpose topic-labeled corpus, we trained a topic model on Switchboard and compared it to a baseline trigram model also trained on Switchboard, shown in Table 3. This test can be viewed as in-domain training for Switchboard and out-of-domain training for all other corpora.

Topic modeling improves performance for all testing corpora, even though the topics in Switchboard aren’t necessarily well represented in other corpora. Although the change

Corpus	Trigram	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%

Table 3: Keystroke savings of a trigram model vs. trigram topic model trained on Switchboard.

is small, the improvement for each testing set is significant at $\alpha = 0.05$. The results lend support that an AAC device using a topic model in a real-world setting will still see some additional keystroke savings even despite the topics of the training data being independent of the testing domain.

7. VOCABULARY ANALYSIS

The results of domain-varied evaluation are strongly affected by the amount of training data as well as the similarity of training and testing data, but other factors are at work — Slate, for example, performed much more poorly than Switchboard under an in-domain test despite its much larger size. It also performed particularly poorly under an out-of-domain test. Even the smaller corpora exhibited substantial variation in performance. In this section, we study the vocabulary of each corpus independently of the training set in order to explain some of the differences in keystroke savings.

7.1 Named Entities

We measured the percentage of uppercase words in each of the corpora, shown in Table 4. The major trend is that more written forms of communication (i.e., Slate, AAC Email) tend to have more named entities. Slate in particular has many named entities, likely because it discussed current events, which are often centered about a named entity. The AAC Emails tend to have more named entities than the spoken corpora, likely due to a similar trend regarding current events. Switchboard has relatively few named entities, which we feel is due to the topic-prompted nature of the conversations, such as discussing care for the elderly or gardening. The high percentage of named entities in Slate may explain why out-of-domain training on Slate performed so poorly in comparison to in-domain training. It may also partially explain why out-of-domain training on AAC Emails offered no benefit over in-domain training, even with 300 times the training data. In practice, this huge performance hit due to named entities might be avoided through named entity caching as in [12].

7.2 Infrequent Vocabulary (OOVs)

Another factor in the trends with keystroke savings is the specialization of each corpus. For instance, Slate, AAC Email, and Micase are all very specialized, using words and structure uncommon in colloquial English. The specialization of each corpus can be estimated by measuring the percentage of words that are out of vocabulary (OOV) with respect to a large, general-purpose vocabulary. We measured the OOVs in reference to the vocabulary from the

Corpus	Named Entities
AAC Email	8.92%
Callhome	8.23%
Charlotte	6.59%
SBCSAE	5.67%
Micase	3.12%
Switchboard	2.10%
Slate	12.03%

Table 4: Percentage of named entities

Web 1T 5-gram Version 1 language model [24], an ngram model built from roughly 1 trillion words by Google and filtered by frequency cutoffs. The percentage of OOV words for each corpus with respect to this language model is shown in Table 5.

Corpus	OOV Words
AAC Email	0.81%
Callhome	0.38%
Charlotte	0.37%
SBCSAE	0.77%
Micase	1.35%
Switchboard	0.22%
Slate	2.12%

Table 5: OOVs with respect to a large dictionary

The first trend is that the vast majority of words appear in a large word list — over 99% for most corpora. By comparison, Wandmacher and Antoine [23] found OOV percentages of 2%–16% with respect to a 5.6 million word newspaper training corpus, where the smallest OOV percent was found for speech and the largest OOV percent for scientific text. The same trend is shown here — Micase uses very specialized vocabulary which doesn’t occur in the large word list, much like the scientific text [23] used. Slate also uses an uncommon vocabulary, sometimes in specialized columns or regarding events relevant only to a particular day. The vocabulary of AAC Emails is somewhat specialized, dealing with very specific technical and political issues, and is reflected in the amount of OOVs. Similarly, some of the speech in SBCSAE is very specialized.

A similar measure, the diversity of a corpus, can be performed by testing the corpus’ vocabulary against itself. For this test, we use 11-fold cross-validation to measure the OOVs of each corpus with respect to itself, essentially measuring how diverse the vocabulary of each corpus is.

Corpus	OOV Words
AAC Email	8.48%
Callhome	6.86%
Charlotte	4.49%
SBCSAE	5.76%
Micase	4.40%
Switchboard	0.52%
Slate	1.96%

Table 6: OOVs using cross-validation

Switchboard shows the lowest percentage of OOVs in the self-test, whereas the larger Slate corpus shows a much higher amount of OOVs. The self-test analysis is affected by both the size of the corpus as well as the diversity of the corpus, which explains the trend with Switchboard: participants in the corpus collection were restricted to one of roughly 70 topics, most of which are represented in every set of Switchboard. Additionally, Switchboard topics were evenly distributed across the sets, which further reduces the variation in vocabulary. On the other hand, Slate was not restricted to such a small set of topics.

8. DISCUSSION

Our first major finding is that a much larger amount of out-of-domain language is more beneficial than a smaller amount of in-domain language for training language models. This is especially the case when there is some overlap between the topic and style of the training and testing corpora, such as the smaller spoken corpora in Table 2. The specific amount of out-of-domain training data needed to match in-domain performance is likely a function of both the amount of in-domain training data and the similarity between training and testing data. The language model requires fewer words to match in-domain training when those sequences of words are similar to the in-domain material in topic and style, such as the similarity between the spoken corpora. In practice, the architect of an AAC system could build an ngram model from a huge amount of general text, a reasonable amount of text from the same style or topic, and a very small amount of highly similar text (where similarity is between the actual user text and training text).

The second major finding of this work is the next logical step — a combination of some similar (in-domain) data and much dissimilar (out-of-domain) data improves keystroke savings over training on either set alone. The improvement is particularly large for specialized corpora such as AAC Email or Micase (see Table 2). More general corpora, such as Callhome, mostly benefit from the increased amount of training data, rather than the combination of a reliable out-of-domain language model combined with an unreliable in-domain model, as the general ngram distribution is measured well in the out-of-domain model and improved only somewhat by adding in-domain data. Corpora such as Charlotte and SBCSAE are something of a middle ground — like all of the small corpora, the combination of a little in-domain data with a lot of out-of-domain data produces the best language model — but the increase in keystroke savings neither follows the small improvement of Callhome nor the large improvement of AAC Email. In fact, the benefit gained from combining some in-domain data with much out-of-domain data is proportional to how specific or general the corpus is. In other words, a corpus with very neutral style and/or very general topics is likely to exhibit many of the same sequences of words as in a large, general-purpose collection of text (such as our out-of-domain models). This trend doesn't apply to Switchboard and Slate because the size of the out-of-domain text is not as large relative to the amount of in-domain text. If the amount of out-of-domain text were much larger, we expect that the same trends would be shown for Switchboard and Slate as well as the smaller corpora. We also expect that the data from the various corpora could be combined more effectively using a linear combination of

probabilities from each corpus' ngram model with weights optimized for data held out from training and testing.

The third contribution of this work is the study of topic modeling across domains — the additional keystroke savings offered by dynamically adapting the language model to the topic of conversation is still realized even for very different texts. The improvement due to topic modeling when testing out-of-domain is most likely due to two factors: 1) some of the topics of Switchboard may have occurred in other corpora and 2) topic modeling may have adapted using slightly relevant topics in Switchboard and “weeded out” obviously dissimilar topics. The second effect is the product of our topic modeling implementation — no lone topic is selected for language modeling, but all are allowed to contribute in proportion to their similarity. This method allows weak similarity to effectively fine-tune the language model to the current topic so long as the current conversation is remotely similar to some training data. For example, one topic in Switchboard discusses baseball. If this topic model were used in a production AAC device, a user talking about a football game may use a few generic sports words, which would weight the baseball topic more highly, which in turn would boost the prediction of baseball words (which may include other generic sports terms). Similarly, the keywords in a conversation about football are unlikely to have any overlap with a topic model about the federal budget, which in turn would depress words related to the budget. This adaptability allows topic modeling to potentially fine-tune its predictions to a previously unseen topic of conversation, provided that the topic of conversation is remotely similar or dissimilar to some of the topics in Switchboard.

Additionally, we analyzed the vocabulary of each corpus and found that the relative in-domain performance of word prediction follows a trend similar to a vocabulary self-test (i.e., corpus size and self-similarity). On the other hand, out-of-domain performance is more correlated with an OOV test with respect to a very large dictionary (i.e., vocabulary specificity). A named entity test was primarily useful to explain the loss in performance of the Slate corpus between in-domain and out-of-domain training. The vocabulary tests suggest that keystroke savings may be highly related to the frequency of named entities and OOVs. In fact, [12] demonstrated that keystroke savings on content words and spoiled non-content words could be increased by roughly 8% using a recency model for named entities. However, there is likely more potential in directly addressing the problem of named entities and OOVs in word prediction.

9. FUTURE WORK

We plan to expand this study to include a very large, general-purpose language model, such as the *Web 1T 5-gram Version 1* language model, available from LDC [24]. The heterogeneous nature of the web should cause this model to be a suitable general-purpose model to serve as the base language modeling component of an interpolated model.

Another general-purpose resource is Wikipedia, which contains information on a multitude of topics. We envision using Wikipedia as a large data source on which to train topic-based methods. The studies in out-of-domain topic modeling could also be improved by applying a fine-grained topic model such as in [13] or [17] or by first applying automatic clustering on training texts such as in [6].

While in this paper, we have shown how topic modeling can be used to adapt a language model (even when trained on very different topics), our future plans include adapting a language model to the style of the discourse. We hypothesize that style adaptation will be especially beneficial when testing the language model on different styles of text, such as would be found with an AAC device used for both conversation and email. We hope that adaptation will allow a single, complex language model built from multiple different styles to adapt to many different situations, ranging from spoken to written language, formal to informal language, and many other dimensions of stylistic variation.

10. CONCLUSIONS

Developers and practitioners face the choice of selecting appropriate training corpora for statistical language models for AAC devices. We've shown that a large amount of dissimilar language is more useful for language model training than a much smaller amount of similar language and that there is a spectrum where less data is necessary if it uses similar language. Furthermore, we've shown that a combination of some in-domain (i.e., similar) text with a much larger amount of out-of-domain (i.e., dissimilar) text can be more beneficial than either text alone, especially when the testing domain is specialized. We've applied the domain-varied evaluation to an advanced language model, topic modeling, and found that the topic model can fine-tune itself even to very dissimilar text, improving word prediction when applied out-of-domain. Our results suggest that adaptive language models have the potential to outperform both in-domain and out-of-domain language models.

Acknowledgments

This work was supported by US Department of Education grant H113G040051. The authors would like to thank Christopher Pennington for his help with corpora and systems issues, as well as the rest of the fringe word prediction group (Amit Hetawal, John McCaw, and Debra Yarrington) for their many helpful discussions. We also thank Google for subsidizing the cost of their ngram model from LDC.

11. REFERENCES

- [1] Anc second release, 2007. Accessed from <http://americannationalcorpus.org/SecondRelease/> on 3/22/2007.
- [2] J. R. Bellegarda. Large vocabulary speech recognition with multispans language models. *IEEE Trans. On Speech and Audio Processing*, 8(1):76–84, 2000.
- [3] L. Boggess. Two simple prediction algorithms to facilitate text production. In *ANLP*, pages 33–40, 1988.
- [4] A. Copestake. Augmented and alternative NLP techniques for augmentative and alternative communication. In *ACL-97 workshop on Natural Language Processing for Communication Aids*, pages 37–42, 1997.
- [5] A. Fazly and G. Hirst. Testing the efficacy of part-of-speech information in word completion. In *EACL-03 Workshop on Language Modeling for Text Entry*, pages 9–16, 2003.
- [6] R. Florian and D. Yarowsky. Dynamic nonlocal language modeling via hierarchical topic-based adaptation. In *ACL*, pages 167–174, 1999.
- [7] G. Foster, P. Isabelle, and P. Plamondon. Word completion: A first step toward target-text mediated IMT. In *COLING*, pages 394–399, 1996.
- [8] D. Hindle. Deterministic parsing of syntactic non-fluencies. In *ACL*, pages 123–128, 1983.
- [9] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics Speech and Signal Processing*, 35(3):400–401, 1987.
- [10] G. Leshner and G. Rinkus. Domain-specific word prediction for augmentative communication. In *RESNA*, 2002.
- [11] G. W. Leshner, B. J. Moulton, and D. J. Higginbotham. Effects of ngram order and training text size on word prediction. In *RESNA*, 1999.
- [12] J. Li and G. Hirst. Semantic knowledge in word completion. In *ASSETS*, pages 121–128, 2005.
- [13] M. Mahajan, D. Beeferman, and X. D. Huang. Improved topic-dependent language modeling using information retrieval techniques. In *ICASSP*, 1999.
- [14] J. Matiassek 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.
- [15] R. Rosenfeld. Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, 88(8):1270–1278, 2000.
- [16] Santa barbara corpus of spoken american english, 2007. Accessed from <http://www.linguistics.ucsb.edu/research/sbcorpus.html> on 3/22/2007.
- [17] K. Seymore and R. Rosenfeld. Using story topics for language model adaptation. In *Proceedings of Eurospeech '97*, pages 1987–1990, Rhodes, Greece, 1997.
- [18] E. Shriberg. Disfluencies in switchboard. In *International Conference on Spoken Language Processing*, pages 11–14 (addendum), 1996.
- [19] *SWITCHBOARD: A User's Manual*, 2007. Accessed from <http://www ldc.upenn.edu/Catalog/docs/switchboard/> on 3/22/2007.
- [20] K. Trnka, D. Yarrington, J. McCaw, K. F. McCoy, and C. Pennington. The Effects of Word Prediction on Communication Rate for AAC. In *NAACL*, pages 173–176, 2007.
- [21] K. Trnka, D. Yarrington, K. McCoy, and C. Pennington. Topic Modeling in Fringe Word Prediction for AAC. In *IUI*, pages 276–278, January 2006.
- [22] P. Väyrynen. *Perspectives on the utility of linguistic knowledge in English word prediction*. PhD thesis, University of Oulu, 2005.
- [23] T. Wandmacher and J.-Y. Antoine. Training Language Models without Appropriate Language Resources: Experiments with an AAC System for Disabled People. In *LREC*, 2006.
- [24] Web 1T 5-gram Version 1, 2007. Accessed from <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13> on 3/23/2007.