

# TEXT NORMALIZATION WITH VARIED DATA SOURCES FOR CONVERSATIONAL SPEECH LANGUAGE MODELING

*Sarah Schwarm<sup>†</sup> and Mari Ostendorf<sup>‡</sup>*

<sup>†</sup>Dept. of Computer Science      <sup>‡</sup>Dept. of Electrical Engineering  
University of Washington  
Seattle, WA 98195. USA

{sarahs,mo}@ssl.i.ee.washington.edu

## ABSTRACT

Collecting sufficient language model training data for good speech recognition performance in a new domain is often difficult. However, there may be other sources of data that are matched in terms of topic or style, if not both. This paper looks at the use of text normalization tools to make these data more suitable for language model training, in conjunction with mixture models to combine data from different sources. We specifically address the task of recognizing meeting speech, showing a small reduction in word error rate over a baseline language model trained from conversational speech data.

## 1. INTRODUCTION

Language model adaptation is a major challenge for speech recognition. N-gram language modeling typically requires large quantities of in-domain training data, i.e. data that matches the task in both topic and style. For conversational speech applications, this is often unrealistic; topics change frequently, and collecting conversational style training data is time-consuming and expensive. Thus, the ability to adapt an existing language model to a new domain is desirable.

Our work is motivated by the ICSI/UW Meeting Recorder project [1]. The goal of this project is to develop a system for automatically transcribing meeting speech. In real meetings and many other potential ASR applications, topics are constantly changing, so it will be impossible to have enough transcribed speech training data for any given topic. Our goal is to modify language models trained on general speech data to be useful for a particular topic. Although adding in-domain training data is an effective means of improving language models [2], adding out-of-domain data is not always successful. In particular, use of text sources in training language models for conversational speech can sometimes degrade recognition performance [3].

Previous LM adaptation efforts include adding unigram probabilities from data for the target domain to an existing class bigram [4], using part-of-speech conditioning for weighting the out-of-domain data [5], and selectively weight-

ing out-of-domain data according to its relevance to the in-domain task [3]. Mixture models have also been used to combine LMs from different corpora [6, 7]. A major limitation of these techniques with respect to our current task is that they require some in-domain training data, whereas here we have little or no target training data for a new topic.

This work is based in part on the hypothesis presented in [5] that similarities and differences between domains can be described in terms of topic and style, where topic is largely conveyed by content words and style can be characterized mainly by patterns of POS usage. We assume that small amounts of style-specific and topic-specific training data are available, but there is no data that matches the task in both topic and style. This is a realistic assumption for the meeting task; for a novel topic, transcribed meeting speech on that topic is not available, but style-specific data is present in the form of a corpus of meetings on other topics. Topic-specific data comes from text sources related to the meeting. In our approach, we use text normalization to make text sources suitable for language modeling for a conversational speech task, then combine different text and speech sources using a mixture language modeling framework. The remainder of the paper is organized as follows. Section 2 presents an analysis of style differences between the corpora used in this study. Section 3 details our approach to the task, followed by experimental results in Section 4. We summarize our findings and describe future directions in Section 5.

## 2. STYLE DIFFERENCES BETWEEN CORPORA

### 2.1. Corpora

Our target task – recognition of speech in meetings – uses data collected by ICSI [1]. Meetings in the corpus are regularly scheduled group meetings at ICSI, i.e. real meetings that would occur even if they were not being recorded for this project. This corpus comprises our test data (five meetings from the meeting recorder group), held-out data (four other meetings from this group) and style-specific data from

Corpus	Size (# words)
Published Text	32,006
Email	64,171
Meetings	162,258

**Table 1.** Size of corpora for training supplemental language models.

meetings on other topics used as supplemental training data.

We consider three categories of supplemental data: “published” text, which consists of papers and web pages relating to the meeting recorder research group; email, including archived mailing list messages sent to the meeting group mailing list and to two other speech-related mailing lists at ICSI;<sup>1</sup> and speech from meetings of groups other than the group represented in the test set. The supplemental meeting speech matches in style but not topic,<sup>2</sup> while the published and email text is assumed to be topic-specific training data. Table 1 lists the size of each supplemental corpus.

## 2.2. Style

The style of conversational speech differs greatly from written text. This difference can be characterized in part by variations in part-of-speech usage patterns, as illustrated in [3] with a comparison of Switchboard, Broadcast News, and Wall Street Journal data. Table 2 provides an analysis of selected word categories in our data, to illustrate differences in the corpora. (Note that filled pauses and back-channels have non-zero probability in the text data, because the group studies conversational speech and so sometimes words like “uh-huh” and “uh” are discussed.) The pattern of more pronouns in speech and more nouns in written text is consistent with that observed in [3]. Like Switchboard, meetings often include casual, conversational speech. Based on the patterns seen here, we can classify Switchboard and the meeting corpus as more stylistically similar, while published text and email are more closely matched in topic but not style. However, meetings have different styles – e.g. formal committee meetings differ from research group brainstorming sessions – and not all styles are represented in our data.

## 3. APPROACH

In many conversational speech tasks, there will never be enough transcribed speech training data for task-dependent

<sup>1</sup>Ideally we would just use the meeting recorder emails, but there was very little text available (only about 4,000 words) so we chose to augment this with messages from the somewhat more general speech lists.

<sup>2</sup>Due to the nature of the corpus (primarily meetings that occurred at ICSI), there is some speaker overlap between the training data and test data, so speaker-specific dependencies may be inadvertently captured by this approach.

POS	Unigram Frequency(%)			
	SWBD	Mtgs	Pub.	Email
Pronouns	10.4	7.8	2.4	3.5
Nouns	17.9	18.0	30.6	28.6
Filled Pauses	1.7	2.5	$10^{-2}$	$10^{-3}$
Backchannels	1.6	1.3	$10^{-2}$	$10^{-3}$

**Table 2.** Frequency of selected word types in Switchboard, meeting data, published text sources and email, demonstrating differences between these domains.

language model training. To make use of different data sources, we use text normalization of written text, with mixture techniques to combine text and conversational speech language models, as described below.

### 3.1. Text Normalization

The meeting data is transcribed speech and therefore may be used directly for language model training with good results. However, text is unlike speech in a variety of ways. In particular, written text also includes numbers (e.g. 101, 1/2, VII, \$3M), abbreviations (e.g. mph, gov’t), acronyms (e.g. IBM, ICSI), and other “non-standard words” (NSWs) which are not written in their spoken form. In order to effectively use this text for language modeling, these items must be converted to their spoken forms. This process has been referred to as text conditioning or normalization and is often used in text-to-speech systems. We hypothesized that text normalization would be a valuable step in transforming text data for conversational speech language modeling.

A set of text conditioning tools are available from the Linguistic Data Consortium (LDC) [8]. The LDC tools perform text normalization using a set of ad hoc rules, converting numerals to words and expanding abbreviations listed in a table. A more systematic approach to the NSW normalization problem is introduced in [9], referred to here as the NSW tools. These tools use models trained on data from several categories (news text, a recipes newsgroup, a PC hardware newsgroup, and real-estate classified ads). The NSW tools perform well in a variety of domains, unlike the LDC tools which were developed for business news.

The NSW tools are built on a taxonomy of 23 categories, including numeric and alphabetic labels. The alphabetic labels include: ASWD, indicating that a token should be said as a word; LSEQ, meaning that a token is read as a sequence of individual letters; and EXPN, indicating that a token is an abbreviation that should be expanded to its full form. Other tokens refer to different types of numbers (e.g. dates, money, cardinal, ordinal). The text normalization process involves first splitting complex tokens using a simple set of rules, and then classifying all tokens as one of the 23 cat-

Source	Phrase
Original	“the 2001 hub5 evals”
LDC tools	“the two thousand , one h. u. b. fi ve evals”
NSW tools	“the two thousand and one hub fi ve evals”

**Table 3.** Example output from the LDC and NSW tools.

egories using a decision tree. After a token is classified, it is expanded according to type-dependent predictors. We used the NSW tools as tuned on data from the PC hardware newsgroup, since this was the most similar domain to our task of recognizing technical research group meetings. We also added 52 domain-specific abbreviation expansions after examining the output of the tools when used on our topic-specific text.

Table 3 shows the output of both tools on a phrase from the published text corpus. The NSW tools correctly split the token “hub5” into its spoken form and also handled 2001 in a more natural way. Of course, not all sentences have perfect transcriptions, though a brief inspection suggests that the NSW tools have fewer errors.

### 3.2. Language Models

Our baseline language models were those used in the SRI March 2000 Hub-5 recognition system. The baseline for the first recognition pass was a backoff bigram model consisting of 35K unigrams and 1.3M bigrams. The baseline model for n-best rescoring was an unpruned trigram model with 4.8M bigrams and 11.5M trigrams. Both models were mixtures built from individual n-gram models trained on data from the Switchboard, CallHome, and Broadcast News corpora. The baseline models and our supplemental models described below use multi-words, lexical entries that contain multiple words, e.g. you\_know, a\_couple\_of. More details on these models are available in [7]. For our work, we modified the baseline models by adding 252 new unigram entries (corresponding to new words in the vocabulary, taken from our supplemental sources) and renormalizing the model. This made no difference to baseline recognition performance, but it did affect language model perplexity as described in section 4.

We also built supplemental language models trained on the text sources described above. For the published text, we built two versions to compare the effects of the LDC tools and the NSW tools. All supplemental models are built using Witten-Bell discounting. (We also tried Good-Turing and Ney discounting; Witten-Bell resulted in models with lower perplexity.) We used interpolated mixture models combining the baseline model with one or more supplemental models for our recognition experiments. The interpolation weights were estimated automatically in order to mini-

Vocabulary	# Add'l Words	OOV rate (%)
Baseline	0	2.27
+ Published Text	99	1.97
+ Email	134	1.90
+ Meetings	98	2.10
+ All	252	1.79

**Table 4.** OOV rates on meeting test data using baseline vocabulary alone and supplemented with words from other sources.

mize perplexity on a held-out set of meeting data. We used the SRI Language Modeling Toolkit [10] to train language models and estimate mixture weights, but tuned the weights of the full combination by hand, further optimizing perplexity on the held out set.

## 4. EXPERIMENTAL RESULTS

For the results reported here, the test set consists of five meetings (approximately 5 hours / 74,000 words) from one group. We exclude speakers who are not native speakers of American English, as in [1].

For our recognition experiments, we used a modified version of the SRI’s large-vocabulary conversational speech recognition system from the March 2000 Hub-5 evaluation. The original SRI system is described in [7], with modifications for the meeting task, including downsampling in order to use the telephone-band acoustic models from the Hub-5 system [1]. We replaced the language model used in the Hub-5 domain with the models described above. The first-pass recognizer used a bigram LM to generate n-best lists with  $n = 1000$ , followed by a rescoring pass using a trigram LM. The oracle error rate for the n-best lists was 27.3%.

As a preliminary step, we looked at the effect of adding words from supplemental sources to the baseline vocabulary (originally 35K words). For each source, we selected words that occurred at least 5 times in the supplemental data but were not in the baseline vocabulary. Results are shown in Table 4. In all cases, the rate of occurrence of out-of-vocabulary (OOV) words was reduced. Words from email and published text (the topic-specific sources) produced a greater reduction than words from other meetings. Not surprisingly, adding words from all three sources yielded the greatest reduction. There was no significant difference for text normalized using the LDC tools vs. the NSW tools.

Perplexity results for the supplemental language models are presented in Table 5. Many of these numbers are higher than might be expected for the following reasons. First, the baseline perplexity is high because of the new words that were added without training data. Second, the supplemental models are trained on small amounts of text; the worst case

LM	Perplexity		
	Unigram	Bigram	Trigram
Baseline	1607	1077	1027
Published Text (LDC)	2544	3344	3343
Published Text (NSW)	2176	2977	2980
Email (NSW)	1237	1543	1578
Meetings	601	538	540

**Table 5.** Language model perplexity on meeting test data for baseline and supplemental LMs.

LM	Perplexity		WER	
	trigram	best	trigram	best
Baseline	1027		45.7	
+ Pub. Text	320	313	45.0	45.0
+ Email	321	313	44.7	45.0
+ Meetings	313	314	45.0	45.0
+ All	337	320	45.6	45.3

**Table 6.** Perplexity and word error rate (WER) on meeting test data for baseline and mixture LMs, where “best” indicates n-gram order.

is the published text, which consists of only 32K words, which is very sparse for a 35K word vocabulary. The version of published text normalized using the LDC tools resulted in a higher perplexity than the version from the NSW tools. Although the difference was negligible when combined with the baseline in a mixture model, we believe that if a more sophisticated combination was used, the difference would remain significant. Thus for recognition experiments, we used only the version produced with the NSW tools. Perplexities for the mixture models are reported in Table 6, including the case using only trigrams and the case using the lowest perplexity LM for other components.

Recognition results are also presented in Table 6. Email provides a slight improvement; published text and meetings result in a smaller but insignificant improvement. There is no advantage to using lower order n-grams on the sparse data sets. Recognition of new words in the test set (not in the original Switchboard LM) improves from 14% to 70% by adding data from one or more sources. The mixture of all four models used the following weights: baseline, 0.65; text, 0.05; email, 0.05; and meetings, 0.25. The full mixture does not result in improved recognition performance; surprising, but consistent with the perplexities of the models.

## 5. DISCUSSION

Our results show that it is possible to improve conversational speech recognition performance by using normalized

text as training data to compensate for the topic mismatch between a general-purpose language model and the test data for a particular task. We have also shown that using data that matches in style but not topic provides additional improvement. Although the improvements are small, the results show the feasibility of using text training data for a conversational task. In addition, there are many possible directions for improvement in this area. For example, many remaining differences between normalized text and real speech can be addressed with text transformations. Disfluencies, discourse markers and other conversational speech phenomena are lacking in text, so using text training data effectively lowers the probability of these events. Our future plans include further modification of text training data, as well as selective use of data conditioned on POS classes as in [5]. Another future direction is to develop a mechanism for picking mixture weights based on text type to completely eliminate the need for target data.

## ACKNOWLEDGEMENTS

This work was supported by IBM through the Faculty Award Program. We would also like to thank Andreas Stolcke and colleagues at ICSI for their help with recognition experiments.

## 6. REFERENCES

- [1] N. Morgan et al., “The meeting project at ICSI,” *Proc. Inter. Conf. Human Language Technology*, 2001, pp. 246–252.
- [2] R. Rosenfeld, “Optimizing lexical and n-gram coverage via judicious use of linguistic data,” *Proc. Eurospeech*, 1995, v. 3, pp. 1763–1766.
- [3] R. Iyer and M. Ostendorf, “Relevance weighting for combining multi-domain data for n-gram language modeling,” *Computer Speech & Language*, **13**(3) 267–282, 1999.
- [4] P. Witschel and H. Hoge, “Experiments in adaptation of language models for commercial applications,” *Proc. Eurospeech*, 1997, v. 4, pp. 1967–1970.
- [5] R. Iyer and M. Ostendorf, “Transforming out-of-domain estimates to improve in-domain language models,” *Proc. Eurospeech*, 1997, v. 4, pp. 1975–1978.
- [6] L. Bahl et al., “The IBM large vocabulary continuous speech recognition system for the ARPA NAB news task,” *Proc. ARPA Workshop on Spoken Language Technology*, 1995, pp. 121–126.
- [7] A. Stolcke et al., “The SRI March 2000 Hub-5 conversational speech transcription system,” *Proc. of the NIST Speech Transcription Workshop*, May 2000.
- [8] Linguistic Data Consortium, 1998, <http://morph ldc.upenn.edu/Catalog/LDC98T31.html>.
- [9] R. Sproat et al., “Normalization of non-standard words,” *Computer Speech & Language*, **15**(3) 287–333, 2001.
- [10] A. Stolcke, “The SRI Language Modeling Toolkit, version 1.1,” <http://www.speech.sri.com/projects/srilm/>, May 2001.