Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures

Ivan Bulyko, Mari Ostendorf
Department of Electrical Engineering
University of Washington, Seattle, WA 98195.
{bulyko,mo}@ssli.ee.washington.edu

Andreas Stolcke
SRI International
Menlo Park, CA 94025.
stolcke@speech.sri.com

Abstract

Sources of training data suitable for language modeling of conversational speech are limited. In this paper, we show how training data can be supplemented with text from the web filtered to match the style and/or topic of the target recognition task, but also that it is possible to get bigger performance gains from the data by using class-dependent interpolation of N-grams.

1 Introduction

Language models constitute one of the key components in modern speech recognition systems. Training an N-gram language model, the most commonly used type of model, requires large quantities of text that is matched to the target recognition task both in terms of style and topic. In tasks involving conversational speech the ideal training material, i.e. transcripts of conversational speech, is costly to produce, which limits the amount of training data currently available.

Methods have been developed for the purpose of language model adaptation, i.e. the adaptation of an existing model to new topics, domains, or tasks for which little or no training material may be available. Since out-of-domain data can contain relevant as well as irrelevant information, various methods are used to identify the most relevant portions of the out-of-domain data prior to combination. Past work on pre-selection has been based on word frequency counts (Rudnicky, 1995), probability (or perplexity) of word or part-of-speech sequences (Iyer and Ostendorf, 1999), latent semantic analysis (Bellegarda, 1998), and information retrieval techniques (Mahajan et al., 1999; Iyer and Ostendorf, 1999). Perplexity-based clustering has also been used for defining topic-specific subsets of in-domain data (Clarkson and Robinson, 1997; Martin et al, 1997), and test set perplexity has been used to prune documents from a training corpus (Klakow, 2000). The most common method for using the additional text sources is to train separate language models on a small amount of in-domain and large amounts of out-of-domain data and to combine them by interpolation, also referred to as mixtures of language models. The technique was reported by IBM in 1995 (Liu et al, 1995), and has been used by many sites since then. An alternative approach involves decomposition of the language model into a class n-gram for interpolation (Iyer and Ostendorf, 1997; Ries, 1997), allowing content words to be interpolated with different weights than filled pauses, for example, which gives an improvement over standard mixture modeling for conversational speech.

Recently researchers have turned to the World Wide Web as an additional source of training data for language modeling. For "just-in-time" language modeling (Berger and Miller, 1998), adaptation data is obtained by submitting words from initial hypotheses of user utterances as queries to a web search engine. Their queries, however, treated words as individual tokens and ignored function words. Such a search strategy typically generates text of a non-conversational style, hence not ideally suited for ASR. In (Zhu and Rosenfeld, 2001), instead of downloading the actual web pages, the authors retrieved N-gram counts provided by the search engine. Such an approach generates valuable statistics but limits the set of N-grams to ones occurring in the baseline model.

In this paper, we present an approach to extracting additional training data from the web by searching for text that is better matched to a conversational speaking style. We also show how we can make better use of this new data by applying class-dependent interpolation.

2 Collecting Text from the Web

The amount of text available on the web is enormous (over 3 billion web pages are indexed via Google alone) and continues to grow. Most of the text on the web is non-conversational, but there is a fair amount of chat-like material that is similar to conversational speech though often omitting disfluencies. This was our primary target when extracting data from the web. Queries submitted to Google were composed of N-grams that occur most frequently in the switchboard training corpus, e.g. "I never thought I would", "I would think so", etc. We were searching for the exact match to one or more of these N-grams within the text of the web pages. Web pages returned by Google for the most part consisted of *conversational* style phrases like "we were friends but we don't

actually have a relationship" and "well I actually I I really haven't seen her for years."

We used a slightly different search strategy when collecting topic-specific data. First we extended the baseline vocabulary with words from a small in-domain training corpus (Schwarm and Ostendorf, 2002), and then we used N-grams with these new words in our web queries, e.g. "wireless mikes like", "I know that recognizer" for a meeting transcription task (Morgan et al, 2001). Web pages returned by Google mostly contained technical material related to topics similar to what was discussed in the meetings, e.g. "we were inspired by the weighted count scheme...", "for our experiments we used the Bellman-Ford algorithm...", etc.

The retrieved web pages were filtered before their content could be used for language modeling. First we stripped the HTML tags and ignored any pages with a very high OOV rate. We then piped the text through a maximum entropy sentence boundary detector (Ratnaparkhi, 1996) and performed text normalization using NSW tools (Sproat et al, 2001).

3 Class-dependent Mixture of LMs

Linear interpolation is a standard approach to combining language models, where the probability of a word w_i given history h is computed as a linear combination of the corresponding N-gram probabilities from S different models: $p(w_i|h) = \sum_{s \in S} \lambda_s p_s(w_i|h)$. Depending on how much adaptation data is available it may be beneficial to estimate a larger number of mixture weights λ_s (more than one per data source) in order to handle source mismatch, specifically letting the mixture weight depend on the context h. One approach is to use a mixture weight corresponding to the source posterior probability $\lambda_s(h) = p(s|h)$ (Weintraub et al, 1996). Here, we instead choose to let the weight vary as a function of the previous word class, i.e. $p(w_i|h) = \sum_{s \in S} \lambda_s(c(w_{i-1}))p_s(w_i|h)$, where classes $c(w_{i-1})$ are part-of-speech tags except for the 100 most frequent words which form their own individual classes. Such a scheme can generalize across domains by tapping into the syntactic structure (POS tags), already shown to be useful for cross-domain language modeling (Iyer and Ostendorf, 1997), and at the same time target conversational speech since the top 100 words cover 70% of tokens in Switchboard training corpus.

Combining several N-grams can produce a model with a very large number of parameters, which is costly in decoding. In such cases N-grams are typically pruned. Here we use entropy-based pruning (Stolcke, 1998) after mixing unpruned models, and reduce the model aggressively to about 15% of its original size. The same pruning parameters were applied to all models in our experiments.

4 Experiments

We evaluated on two tasks: 1) Switchboard (Godfrey et al., 1992), specifically the HUB5 eval 2001 set having a total of 60K words spoken by 120 speakers, and 2) an ICSI Meeting recorder (Morgan et al, 2001) eval set having a total of 44K words spoken by 25 speakers. Both sets featured spontaneous conversational speech. There were 45K words of held-out data for each task.

Text corpora of conversational telephone speech (CTS) available for training language models consisted of Switchboard, Callhome English, and Switchboard-cellular, a total of 3 million words. In addition to that we used 150 million words of Broadcast News (BN) transcripts, and we collected 191 million words of "conversational" text from the web. For the Meetings task, there were 200K words of meeting transcripts available for training, and we collected 28 million words of "topic-related" text from the web.

The experiments were conducted using the SRI large vocabulary speech recognizer (Stolcke et al, 2000) in the N-best rescoring mode. A baseline bigram language model was used to generate N-best lists, which were then rescored with various trigram models.

Table 1 shows word error rates (WER) on the HUB5 test set, comparing performance of the class-based mixture against standard (i.e. class-independent) interpolation. The class-based mixture gave better results in all cases except when only CTS sources were used, probably because these sources are similar to each other and the class-based mixture is mainly useful when data sources are more diverse. We also obtained lower WER by using the web data instead of BN, which indicates that the web data is better matched to our task (i.e. it is more "conversational"). If training data is completely arbitrary, then its benefits to the recognition task are minimal, as shown by an example of using a 66M-word corpus collected from random web pages. The baseline Switchboard model gave test set perplexity of 96, which is reduced to 87 with a standard mixture CTS and BN data, reduced further to 83 by adding the web data, and to a best case of 82 with class-dependent interpolation and the added web data.

Increasing the amount of web training data from 61M to 191M gave relatively small performance gains. We "trimmed" the 191M-word web corpus down to 61M words by choosing documents with lowest perplexity according to the combined CTS model, yielding the "Web2" data source. The model that used Web2 gave the same WER as the one trained with the original 61M web corpus. It could be that the web text obtained with "Google" filtering is fairly homogeneous, so little is gained by further perplexity filtering. Or, it could be that when choosing better matched data, we also exclude new N-grams that may occur only in testing.

Table 1: HUB5 (eval 2001) WER results using standard and class-based mixtures.

LM Data Sources	Std. mix	Class mix
Baseline CTS	38.9%	38.9%
+ 150M BN	37.9%	37.8%
+ 66M Web (Random)	38.6%	38.3%
+ 61M Web	37.7%	37.6%
+ 191M Web	37.6%	37.4%
+ 150M BN + 61M Web	37.7%	37.3%
+ 150M BN + 191M Web	37.5%	37.2%
+ 150M BN + 61M Web2	37.7%	37.3%

Table 2: Meetings results (WER).

LM Data Sources	Std. mix	Class mix
Baseline	38.2%	
+ 0.2M Meetings	37.2%	36.9%
+ 28M Web (Topic)	36.9%	36.7%
+ Meetings + Web (Topic)	36.2%	35.9%

Results on the Meeting test set are shown in Table 2, where the baseline model was trained on CTS and BN sources. As in the HUB5 experiments, the class-based mixture outperformed standard interpolation. We achieved lower WER by using the web data instead of the meeting transcripts, but the best results are obtained by using all data sources. Language model perplexity is reduced from 122 for the baseline to a best case of 95.

We also tried different class assignments for the class-based mixture on the HUB5 set and we found that using automatically derived classes instead of part-of-speech tags does not lead to performance degradation as long as we allocate individual classes for the top 100 words. Automatic class mapping can make class-based mixtures feasible for other languages where part-of-speech tags are difficult to derive.

5 Conclusions

In summary, we have shown that, if filtered, web text can be successfully used for training language models of conversational speech, outperforming some other out-of-domain (BN) and small domain-specific (Meetings) sources of data. We have also found that by combining LMs from different domains with class-dependent interpolation (particularly when each of the top 100 words forms its own class), we achieve lower WER than if we use the standard approach where mixture weights depend only on the data source. Recognition experiments show a significant reduction in WER (1.3-2.3% absolute) due to additional training data and class-based interpolation.

References

J. Bellegarda. 1998. Exploiting both local and global constraints for multispan statistical language modeling. In *Proc. ICASSP*, pages II:677–680.

A. Berger and R. Miller. 1998. Just-in-time language modeling. In *Proc. ICASSP*, pages II:705–708.

P. Clarkson and A. Robinson. 1997. Language model adaptation using mixtures and an exponentially decaying cache. In *Proc. ICASSP*, pages II:799–802.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Proc. ICASSP*, pages I:517–520.

R. Iyer and M. Ostendorf. 1997. Transforming out-of-domain estimates to improve in-domain language models. In *Proc. Eurospeech*, volume 4, pages 1975–1978.

R. Iyer and M. Ostendorf. 1999. Relevance weighting for combining multi-domain data for n-gram language modeling. *Computer Speech and Language*, 13(3):267–282.

D. Klakow. 2000. Selecting articles from the language model training corpus. In *Proc. ICASSP*, pages III:1695–1698.

F. Liu et al. 1995. IBM Switchboard progress and evaluation site report. In *LVCSR Workshop*, Gaithersburg, MD. National Institute of Standards and Technology.

M. Mahajan, D. Beeferman, and D. Huang. 1999. Improved topic-dependent language modeling using information retrieval techniques. In *Proc. ICASSP*, pages I:541–544.

S. Martin et al. 1997. Adaptive topic-dependent language modeling using word-based varigrams. In *Proc. Eurospeech*, pages 3:1447–1450.

N. Morgan et al. 2001. The meeting project at ICSI. In *Proc. Conf. on Human Language Technology*, pages 246–252.

A. Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proc. Empirical Methods in Natural Language Processing Conference*, pages 133–141.

K. Ries. 1997. A class based approach to domain adaptation and constraint integration for empirical m-gram models. In *Proc. Eurospeech*, pages 4:1983–1986.

A. Rudnicky. 1995. Language modeling with limited domain data. In *Proc. ARPA Spoken Language Technology Workshop*, pages 66–69.

S. Schwarm and M. Ostendorf. 2002. Text normalization with varied data sources for conversational speech language modeling. In *Proc. ICASSP*, pages I:789–792.

R. Sproat et al. 2001. Normalization of non-standard words. *Computer Speech and Language*, 15(3):287–333.

A. Stolcke et al. 2000. The SRI March 2000 Hub-5 conversational speech transcription system. In *Proc. NIST Speech Transcription Workshop*.

A. Stolcke. 1998. Entropy-based pruning of backoff language models. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274.

M. Weintraub et al. 1996. LM95 Project Report: Fast training and portability. Technical Report 1, Center for Language and Speech Processing, Johns Hopkins University, Baltimore.

X. Zhu and R. Rosenfeld. 2001. Improving trigram language modeling with the world wide web. In *Proc. ICASSP*, pages I:533–536.