Language modeling with limited domain data

Alexander I. Rudnicky

School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

ABSTRACT

Generic recognition systems contain language models which are representative of a broad corpus. In actual practice, however, recognition is usually on a coherent text covering a single topic, suggesting that knowledge of the topic at hand can be used to advantage. A base model can be augmented with information from a small sample of domain-specific language data to significantly improve recognition performance. Good performance may be obtained by merging in only those n-grams that include words that are out of vocabulary with respect to the base model.

1. Introduction

Current language modeling practice requires access to a substantial amount of text from a target domain in order to create a reliable language model. For the North American Business (CSR NAB) domain, 227M words were available. Of necessity models based on large corpora cover a diversity of material and are fairly general in nature. In practice, a given sequence of input utterances (say a dictation) will stick to a particular topic, whether it's a news story, a business letter or other coherent text. It is reasonable for a recognition system to try to take advantage of this in some fashion. A number of techniques address this issue. Caching [4] temporarily increases the likelihood of n-grams occuring in the recent past. Triggering [6] allows clusters of related words to have their likelihood increased, given the occurrence of a word known to be associated to them.

These techniques address the problem of improving recognition accuracy through a process of selection: components of an existing model best suited to the current input are selectively potentiated. Such approaches presuppose that the relevant information is already present in the model and simply needs to be identified. Other proposals, such as varying the weights within a mixture of broad-domain models [5] are also selection procedures, though they operate at a much coarser level of representation.

We are more interested in the case where necessary information is not already present in a model and must somehow be added. Such a situation is most likely to arise in practical contexts such as dictation and appear in combination with other problems, such as identifying and assimilating new words. Note that difficulties arise not only because a truly new topic was encountered but also due to practical limitations in the size of the language model that can be implemented. Low-frequency topics, though present in the domain, may have been eliminated from an on-line model. Similarly for domains such as news, shifts occur in contemporary topics of discourse which cannot be anticipated through the collection of historical data.

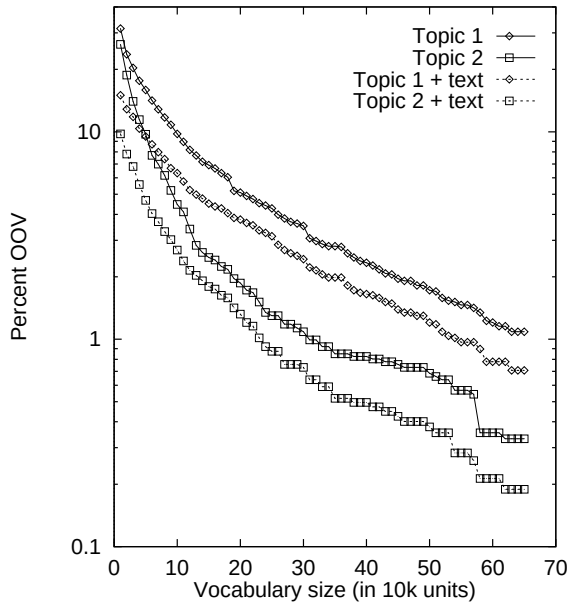
Techniques have been proposed for incremental expansion of models (for example [2]) and indeed commercial dictation systems offer the option to add words to an existing language model. In this paper we are concerned with an intermediate case, where users are able to offer the system some guidance, a small amount of text related to the topic of a subsequent dictation. In many cases, this reflects the actual circumstances of dictation. For example if I am writing a paper on speech recognition, I should be able to provide several previous papers of my own as well as papers on the specific topic by others. The recognition system should be able to take advantage of this sample to augment its (base) vocabulary as well as assimilate some of the n-gram structure. We have to assume that the language of the base and domain are sufficiently similar such that the former can provide serviceable scaffolding while the latter fills in details. The question at hand is whether such an approach is workable.

2. Spoke 2 in the November 1994 evaluation

Spoke 2 of the ARPA CSR evaluation[1] provides the opportunity to explore this particular paradigm. The materials provided consisted of two stories selected from the NYT wire. The criteria for selection were that a story on a particular topic should have an identifiable start time and that it generated sufficient followup articles to provide about 12-15k words of text. For purposes of the investigation, this text was divided into two parts: approximately the first 10k of raw text (to the nearest article boundary) was designated as "adaptation" material; a test set was formed from the immediately (i.e., chronologically) following stories on that same topic. Once a sufficient number of story sequences were collected, it was observed that they seemed to form two categories. Some stories were concentrated in just a few weeks while others

[1] [3] describes the design of the CSR Benchmark evaluation and defines the nomenclature of the various test components.

Figure 1: *OOV rate as a function of vocabulary size for the development set, in 1k increments.*



(producing a lower density of stories) ran on for much longer. One story from each category was placed in the development (and evaluation) sets.

In this paper we will mostly concentrate on description and analysis of the data from the `dev1` set. The two topics in this set were the death of Jackie Onassis (Topic 1) and negotiations with North Korea concerning their nuclear program (Topic 2).

Figure 1 shows the out-of-vocabulary (OOV) rates for these two topics as a function of vocabulary size, measured on the devtest data. The vocabulary is derived from the word-frequency list based on the complete CSR NAB text corpus. The two topics differ substantially in OOV rate, this possibly related to the nature of the topic. We can assess the impact of using domain-specific data by adding to the vocabulary those words found in the 10k word adaptation sample. The Figure shows the resulting OOV rates. We observe that once the base vocabulary size begins to exceed 20k words, it appears that (roughly) 10k of domain-specific text can be traded for 10k or 20k of additional words in the vocabulary, at least in terms of OOV rate.

3. Making use of sample domain text

The simplest way to make use of the adaptation data is to create a language model from it (i.e., reduce it to a set of n-grams) and combine this information with the target base model. The experiments described in this section make use of the officially distributed November 94 20k backed-off trigram language model (generated with 2-,3-gram cutoffs of 1 and 3). Models computed from the domain data have different cutoffs, as noted below. A merged base/domain model was in the form of a single backoff trigram model. Recognition performance was assessed using a 3-pass Sphinx-II decoder with separate male and female models. The test set consisted of two parts (Topic 1: 177 utterances, 3881 words; Topic 2: 155 utterances and 3922 words).

Table 1: *Decoder word error using merged models*

cutoffs	*base*	*topic 1*	*topic 2*	*mean*	*gain*
	1.0	27.4	16.5	22.0	–
0-0	0.5	26.2	17.9	22.1	−0.1
	0.95	25.3	15.8	20.6	+1.4
	0.995	25.7	15.7	20.7	+1.3
1-2	0.95	25.7	15.7	20.7	+1.3

We first examined a simple interpolation of the base and domain models. Domain models with cutoffs of 0-0 and 1-2 were used. Table 1 shows recognition word error rate for a range of interpolation weights. Using a model with cutoffs does not seem to impact recognition performance, suggesting that most of the useful information in the adaptation text was contained in 1-grams and in higher n-grams of multiple occurrence, a reasonable interpretation if the topic remains the same.

Somewhat anomalously, an equal-weight interpolation has different effects, depending on the topic. This result is difficult to interpret without confirmation from other datasets. Examination of the Topic 1 texts and errors suggests that in this case the domain model may be capturing some stylistic features of the articles. Conversely, Topic 2 is a conventional news story on a topic that is likely revisited in one form or another over time. Perhaps the base model adequately captures its long-term stylistic characteristics.

The above findings suggest that not all of the n-gram information in the domain model is necessarily useful. Perhaps only those n-grams that contain OOV words actually contribute to the improved performance seen for the merged model. We can test this by merging in only those n-grams that include an OOV word relative to the base vocabulary. If we do this, we find that comparable gains in performance can be obtained (see Table 2). This would appear to be an efficient procedure, as can be seen from the number of n-grams that are added in each case (Table 3).

Incorporating OOV words plus the n-grams they participate in is somewhat more effective than simply including the OOV words as 1-grams, though not by much (Table 4). This outcome likely reflects a very low hit rate on the actual n-grams included in the merged model.

Table 2: *Decoder word error with filtered models*

cutoffs	*base*	*topic 1*	*topic 2*	*mean*	*gain*
0-0	0.5	25.8	15.7	20.7	+1.3
	0.95	25.5	15.8	20.6	+1.4
1-2	0.5	25.7	16.1	20.9	+1.1
	0.95	25.6	16.1	20.8	+1.2

Table 3: *Unique n-grams added to merged model*

cutoffs			*topic 1*	*topic 2*
0-0	full	2-gram	1762	792
		3-gram	6084	4068
	filtered	2-gram	1119	370
		3-gram	1679	575
1-2	full	2-gram	71	47
		3-gram	16	9
	filtered	2-gram	63	28
		3-gram	11	3

A different approach to dealing with the domain-specificity problem would be to simply increase the vocabulary size to achieve a desired decrease in the OOV rate. Table 5 shows comparable performance using a decoder based on approximately the top 59k words in the CSR NAB training corpus. Performance exceeds that of the merged-model systems; a contributing factor is the much lower OOV rates for this model (1.2% and 0.4%).

4. The November 1994 Evaluation

4.1. Test set characteristics

Two topics were presented for evaluation: a story on trade negotiations with China (Topic 04) and the initial stages of the O. J. Simpson affair (Topic 05). Figure 2 presents the analogue to Figure 1 for the evaluation set. The China story came with 10641 words of adaptation text, the Simpson story came with 11010 words. Unlike the development data, the evaluation topics show little reduction in OOV rate as words from the adaptation text are added (China, 1.5% → 1.43%; Simpson, 1.13% → 1.03%).

Table 4: *Adding 1-grams only*

model type	*topic 1*	*topic 2*	*mean*	*gain*
0.5	25.7	16.0	20.9	+1.1
0.95	25.4	16.1	20.8	+1.2

Table 5: *Word error using a 59k word language model.*

topic 1	*topic 2*	*mean*
22.2	14.2	18.2

4.2. Results

The system used for evaluation used a base vocabulary of 59166 words, derived from the word frequency list for the combined CSR NAB language training corpus (frequencies of 40 or greater); suspect words were removed. Added to this vocabulary were 247 compounded abbreviations and 124 common word phrases, as described in [1]. Multiple pronunciations and pseudo-class items expanded this to a 63019 word recognition dictionary. For Topic 04, 45 new words were added; for Topic 05, 54 were added. The decoding strategy was identical to the one described in [1]. Three different versions of the system were run for the evaluation, corresponding to the three S2 conditions:

P0 Best possible system but without knowledge of the test domain.

C1 The CMU H1P0 system[1] on this test set.

C2 Best possible system, making use of the 10k word domain adaptation sample.

Analysis of the development data suggested that the approach take for the H1P0 system, using a vocabulary based on a word-frequency list derived from only more recent Wall Street Jour-

Figure 2: *OOV function for the evaluation set*

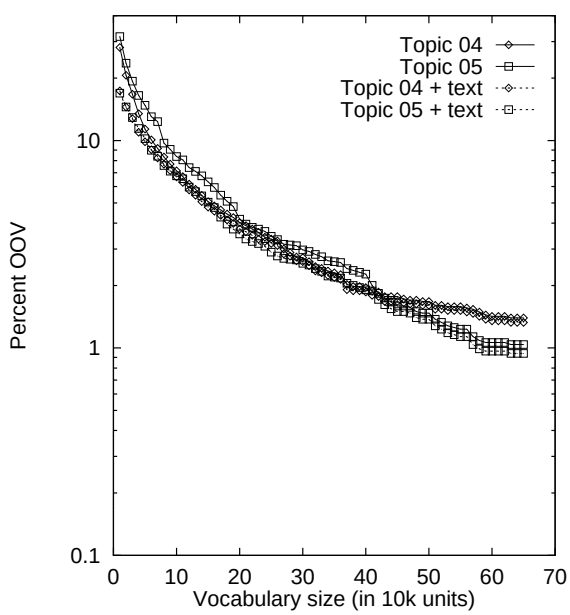


Table 6: *November 1994 Evaluation results*

	word error		
condition	*mean*	*topic 04*	*topic 05*
p0	19.4%	17.3%	21.5%
c1	19.7%	17.5%	22.0%
c2	18.5%	17.1%	19.8%

nal data (the "decay" model) was not appropriate, as lower out-of-vocabulary rates were obtained using the entire corpus. The decay model was also slightly worse that vocabularies based only on the AP wire and/or San Jose Mercury components of the corpus. Models based on only Wall Street Journal data did not do as well as the non-WSJ-only vocabulary. This vocabulary choice appears to have been the right one, as the P0 system outperformed the C1 (i.e., H1P0) system on these data (albeit by a non-significant amount).

Table 6 shows the NIST-computed error rates from the November 1994 evaluation. The difference between P0 and C1 is not significant; C2 has an error rate significantly lower than P0 and C1 for Topic 05, but does not differ for Topic 04.

The differences between topics, although their relative magnitudes mirror the ones observed for the development data, are difficult to interpret due to the inadvertent inclusion, in the China set, of an article on Mexican museums. Unfortunately, 60% of the text in this set (2481 of 4122 words) was on the Mexican topic, making Topic 04 a poor test of the hypothesis. Turning to Topic 05, it is interesting to note that substantial gains in performance were observed even when the base vocabulary size tripled in comparison to the experiments reported in section 3. Given that performance improved even with the lack of a meaningful decrease in the OOV rate, it would appear that the gain was due to improvements in 2,3-gram coverage, a result at variance with those reported for the development data. Individual topic characteristics thus appear to affect the effectiveness of domain adaptation.

We note that the overall performance on the S2 set (19.4% error) is quite poor compared to the H1 set (10.9% error): the error rate doubles. Since the same speakers were used to record both the H1 and S2 data sets, this suggests that the increase in error rate might be due to the distance between the business news language of the training corpus and the language of the S2 texts. If this proves to be the case, then the need for more powerful domain-adaptation techniques (in the absence of large training corpora) is very clear.

Table 7: *Gains in accuracy*

condition	*mean*	*topic 04*	*topic 05*
p0	+0.3	+0.2	+0.5
c1	–	–	–
c2	+1.2	+0.4	+2.2

5. Conclusions

The work reported in this paper is preliminary in nature and represents an initial exploration of domain adaptation on our part. Nevertheless, we believe that it represents a successful paradigm for domain adaptation, given the significant improvements in recognition accuracy observed. We believe that this is due to the precision afforded by the selection of adaptation text which is known to be related to the test material. Automatic methods of identifying relevant adaptation text (e.g., [7]) although promising, do not as yet provide the same degree of precision.

Model merging is an effective technique for incorporating domain-specific information into a language model and for reducing the necessary vocabulary size for a recognition system system. While knowing the identity of the new words in a domain appears to be the dominant source of useful information, structural information (n-grams incorporating these words) also appears to be of use. We were unable to identify an optimal technique for automatically computing the optimum weights for component models. It is unclear why this is the case, though the same problem has been reported by others. Possibly the combination of heterogeneous models or the small amount of data available make the process unstable.

References

1. CHASE, L., ROSENFELD, R., HAUPTMANN, A., RAVISHANKAR, M., THAYER, E., PLACEWAY, P., WEIDE, R., AND LU, C. Improvements in language, lexical and phonetic modeling in Sphinx-II. *Proceedings of the ARPA Spoken Language Technology Workshop* (1995), this volume.
2. JELINEK, F., MERCER, R., AND ROUKOS, S. Classifying words for improved statistical language models. *Proceedings of ICASSP* (1990), 621–624.
3. KUBALA, F. Design of the 1994 CSR benchmark tests. *Proceedings of the ARPA Spoken Language Technology Workshop* (1995), this volume.
4. KUHN, R. A., AND MORI, R. D. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-12*, 6 (1990), 570–583.
5. OSTENDORF, M., RICHARDSON, F., KANNAN, A., IYER, R., RONNEN, O., AND BATES, R. The 1994 BU WSJ benchmark system. *Proceedings of the ARPA Spoken Language Technology Workshop* (1995), this volume.
6. ROSENFELD, R. *Adaptive statistical language modeling: A maximum entropy approach*. PhD thesis, Carnegie Mellon University, 1994.
7. SEKINE, S., STERLING, J., AND GRISHMAN, R. NYU/BBN 1994 CSR evaluation. *Proceedings of the ARPA Spoken Language Technology Workshop* (1995), this volume.