Temporal Adaptation of Language Models

E. W. D. Whittaker

Compaq Cambridge Research Laboratory
Cambridge MA 02142 USA

Abstract

In this paper, the implications of recognising speech from news broadcasts that change on a daily basis are investigated from the perspective of audio indexing. First, vocabulary coverage is found to be of great importance. In daily newspapers and daily news broadcasts it is observed that half the total number of unique words that occur over time are used on only one day and are never used again. The extent to which vocabularies can be adapted is examined and it is found that it is difficult to obtain substantially increased vocabulary coverage using external sources of time-dependent data. The second implication of time-varying data is the ability to effectively adapt the language model used in the speech recogniser so as best to match the speech that is being recognised. Language model adaptation using different methods of combining external sources of time and domain dependent data are investigated using both fixed and adapted vocabularies. It is found that the improvements obtained recognising ten consecutive shows of the radio programme *Marketplace* do not generally justify the effort involved in adapting the language models especially if the baseline language model is well trained.

1. Introduction

In this paper, we investigate methods for temporal language model (LM) and vocabulary adaptation. Compaq's Cambridge research laboratory has been involved for several years in the area of audio indexing tasks within the "SpeechBot" system [1] which has been publicly accessible for over a year and indexes over 10,000 hours of audio available on the internet. Despite the relatively poor speech recognition performance of the system, good retrieval precision is still achieved on the top 20 documents returned.

A significant proportion of the audio sources that we are aiming to index in the near future changes its content on a daily basis. This data might be archived and cover many years or it might be contemporary but in either case, the content changes every day. This is particularly applicable to news broadcasts, for example. In addition to the obvious issues of improving recognition performance on such data by adapting the LM and the acoustic models, there are the difficulties of vocabulary selection for maximising coverage for the audio retrieval task we are concerned with.

A substantial amount of research has already been performed on various LM adaptation methods for different scenarios, all with varying degrees of success [2, 3, 4, 5]. We present three simple methods for adapting LMs to temporally changing data. Using external time-dependent data sources we investigate two different methods of combining temporally relevant data and one method of combining domain-dependent data. We evaluate performance on two weeks' worth of the half-hour financial-news radio show *Marketplace*. *Marketplace* was chosen due to its inherently time-varying nature and the ready availability of data for both perplexity and recognition experiments.

2. Experimental setup

For every *Marketplace* show from 1st January 1996 to 31st December 2000 a partial transcript of the show was obtained from the *Marketplace* web-site [6]. Each partial transcript amounted to around 1500 words, or approximately 6 minutes of speech, on average. *Marketplace* is only broadcast on the five working days of every week so when we refer to consecutive shows in this paper, we are ignoring the absence of broadcasts at the weekend.

Text from the Los Angeles Times newspaper (LATWP) from 1st January 1994 to 30th April 1996 and Reuters financial (REUFF) and Reuters newswire (REUTE) text from the same time period was preprocessed and used as the baseline LM training data. The most frequent 65k words from this data were chosen to be the baseline vocabulary and the baseline LM was built using Katz back-off with Good-Turing discounting, after discarding all singleton bigrams and trigrams[1]. The above three text sources were chosen firstly because they constituted a substantial amount of training data (150 million words) and secondly because data from the three sources was available on the days of interest for which *Marketplace* test text and test acoustic data was available. Text for the temporal adaptation experiments was drawn from the same three sources for the appropriate days in the period 1st May 1996 to 31st December 1997.

3. Vocabulary variation

The first set of experiments investigated the properties of the vocabulary of words used in the corpus of partial daily *Marketplace* transcripts for the five months from 1st January 1996 to 31st May 1996. On average, the size of the daily vocabulary was around 650 words while the vocabulary for the five month corpus was 11.5k words. A plot of the number of words that occur on any given number of days throughout this period is shown in Figure 1. From the figure we see that a large number of words (around 5500) occurred on only one day over the time period investigated and never occurred again. It should be noted that this number is approximately half the vocabulary size of the entire corpus. A smaller proportion occurred on two days only (around 1800) and so on. At the other end of the scale, a small subset of words occurs in the shows practically every day. These are mainly function words and the ever-present 'David Brancaccio' for example. From an audio indexing point of view it is likely that the words that occur every day are of little interest while the daily singleton and doubleton words are potentially of most interest—an examination of the latter showed that

[1] It should be noted that the performance of the baseline LM did not differ substantially from the performance of one built using a similarly sized corpus of broadcast news transcripts.

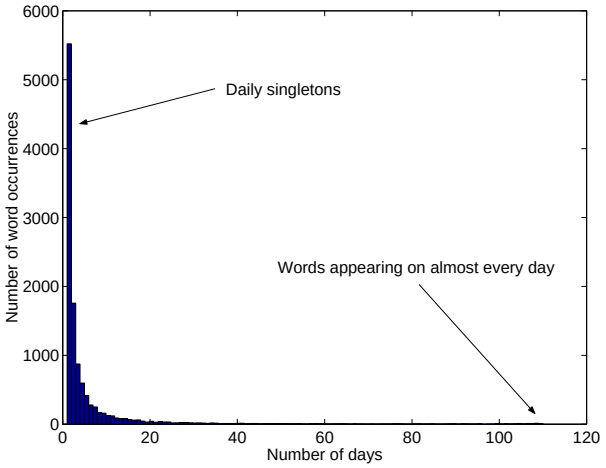

Figure 1: Number of words that occur on any given number of days over five months of *Marketplace*.

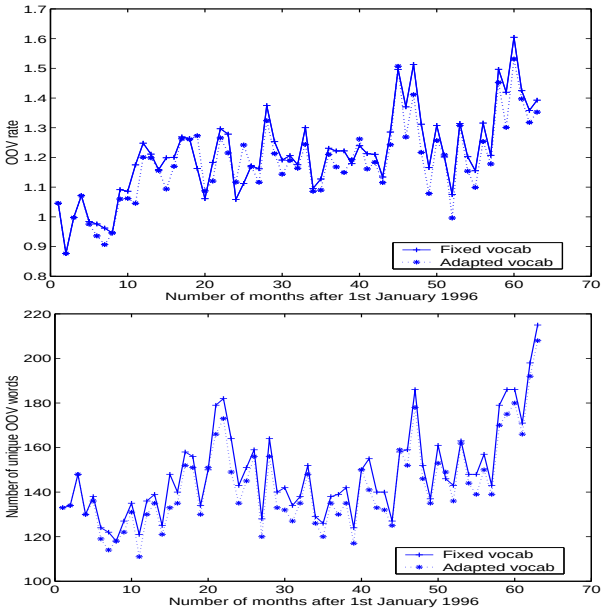

Figure 2: Monthly variation in OOV rate and unique OOV words over five years of partial *Marketplace* transcripts.

many of them are proper nouns for example. These characteristics are not restricted to the *Marketplace* data. It was also found that very similar trends were obtained using two years' worth of New York Times newspapers texts. Of the total 300k unique words used over two years in the daily Internet edition, 140k occurred on only one day. However, without an external source of data to hint that such words will be used or without an extremely large vocabulary we cannot hope to capture all such spontaneous occurrences.

4. Vocabulary adaptation

Since our primary interest is audio indexing we also investigated how we can improve vocabulary coverage using external time-dependent data sources. In Figure 2 the out-of-vocabulary (OOV) rate and the number of unique OOV words with respect to the baseline LM vocabulary (which contained data up to 30th April 1996) is plotted over five year's worth of the partial *Marketplace* transcripts. The OOV rate expresses the percentage of words in the data that are not in the vocabulary. Each data point corresponds to the OOV rate (or unique number of OOV words) over 20 consecutive partial daily transcripts i.e. approximately one month's worth of the partial transcripts. This choice was arbitrary but reduced the noise of the plot. It is clear from the upper plot that the OOV rate with respect to the fixed vocabulary deteriorates gracefully over time. The mean *daily* OOV rate using the baseline 65k vocabulary is 1.22%.

Vocabulary adaptation is performed by updating the word-frequency statistics of the corpus, which progressively accumulates each new day's external time-dependent data from the LATWP, REUFF and REUTE sources. The most frequent 65k words are selected from the word frequency list formed at the end of the month for which the OOV rate is to be computed. The external time-dependent data was available from 1st May 1996 to 31st December 1997 which corresponds approximately to 4 and 24 months after 1st January 1996 in the plots. Adapting the vocabulary in the above manner reduces the OOV rate for each monthly period by 6.1% relative (0.07% absolute) on average. The number of unique OOV words is reduced by 7.5% relative, or around 12 words on average. The mean *daily* OOV rate using the adapted vocabulary is 1.19%. For the period where adaptation data was available, the mean daily OOV rate was 1.15% compared to 1.17% with the fixed baseline vocabulary.

5. Language model adaptation

Next we turned our attention to adapting the probabilities in the LM. Adaptation took the form of incorporating data from external sources in three different ways: i) progressively adding external time-dependent data to the baseline LM training data and rebuilding the dictionary and LM; ii) building a LM from each day's external time-dependent data and interpolating it with LMs built with other days' data and the baseline LM; and iii) building a LM using the first five months of partial *Marketplace* transcripts together with the baseline LM training data. The three methods are described in more detail below and the results of recognition experiments are presented.

The single pass recognition system used continuous density 3-state HMMs with 6000 tree clustered states, each modelled by a mixture of 16 Gaussians, which were trained using the 1996 Hub4 training data provided by LDC, from which all *Marketplace* data had first been removed. For the recognition experiments, the test set comprised thirteen consecutive *Marketplace* shows from 29th May to 14th June which had been removed from the acoustic training data. The first three days' shows (29th to 31st May) were used for development testing and the remaining ten days' shows were used for evaluation.

5.1. Accumulated data

The simplest method of combining the available data from the LATWP, REUFF and REUTE sources was to start with the baseline LM training data and accumulate the external time-dependent data for the days up to and including that for the day of the show to be recognised. With such a method there is no obvious way of determining what the contribution of newly added data should be, i.e. what importance it should be assigned. It might be that the data for the current day would be better added twice

	Development set			Evaluation set										
Date of show	29/5	30/5	31/5	3/6	4/6	5/6	6/6	7/6	10/6	11/6	12/6	13/6	14/6	Overall
Baseline	23.2	28.6	24.4	27.7	32.7	30.1	26.2	31.6	24.3	27.6	23.2	27.8	34.8	28.6
Accumulated	23.1	28.5	24.1	27.6	32.5	29.7	26.3	31.7	24.5	27.1	23.0	27.5	34.4	28.4
Interpolated	22.3	28.5	24.0	26.9	32.9	29.5	25.1	31.5	23.8	26.2	22.1	26.5	33.9	27.8
Baseline+mktplc	22.9	27.5	23.8	26.8	32.4	29.7	25.5	31.0	24.0	27.0	22.4	27.1	33.7	27.9

Table 1: Recognition performance (% word error rate) of baseline and adapted LMs on 13 consecutive *Marketplace* shows in 1996.

or a fractional number of times but this is difficult to determine. Consequently, each new day's data was added only once to the accumulated training data. The new vocabulary was found by taking the top 65k words from the adapted training data, and an updated LM built for that day. The recognition results on each of the 13 shows are presented in the third row of Table 1.

5.2. Interpolated component language models

Interpolating component LMs required that the same vocabulary was used to build all component LMs, so the vocabulary used in the baseline LM was chosen. A 'daily' trigram LM was then built using adaptation data from LATWP, REUFF and REUTE for every day over the five month period from 1st January to 31st May 1996, including days on which there were no *Marketplace* broadcasts. The amount of adaptation data available for each day was only around 300k words so it was decided to retain all bigrams and trigrams in the component daily LMs after carefully preprocessing the data. The adapted LM, P^t_{adapt} for some day t was formed by linearly interpolating the baseline LM, P_b with ten component 'daily' trigram LMs, P_{t-i} built on the current and the nine previous days' adaptation data, where none of the days preceded 1st January 1996 as follows:

$$P_t^{adapt}(w_i \mid w_{i-2}, w_{i-1}) = \alpha_b \cdot P_b(w_i \mid w_{i-2}, w_{i-1}) + \sum_{j=0}^{9} \alpha_{t-j} \cdot P_{t-j}(w_i \mid w_{i-2}, w_{i-1}).$$

The weights α for the eleven models were determined by minimising the perplexity of the above model on the current day's partial *Marketplace* transcript using the E-M algorithm. The average weights are shown for the daily LMs in Figure 3; the baseline LM had the highest weight of 0.76 (not shown). It is interesting to note that the weights for each component daily LM decrease as the adaptation data used to build each component LM gets 'older'. This trend continues for the five previous days' components until there is an increase for the LM built using adaptation data from exactly one week prior, after which the weights continue to decrease again. In addition to there being a correlation between the weight assigned to a LM and the amount of data used to build it, there also appears to be a correlation between a LM's temporal appropriateness for the task and its assigned weight. Almost identical, and more pronounced, results were obtained with experiments conducted on New York Times data. Such a phenomenon confirms what we might already have suspected, that particular news items and journalists will appear on certain days of the week and hence the style and vocabulary will be periodic. It is precisely these phenomena that we are aiming to capture with this method.

The perplexity of the adapted LMs on the first five months of partial *Marketplace* transcripts using both the daily optimised and the averaged weights is shown in Figure 4. The daily optimised weights applied to the component LMs are the optimal

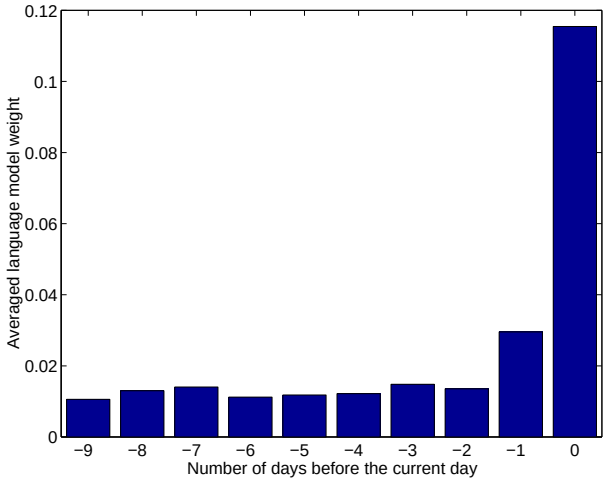

Figure 3: Averaged component daily LM weights.

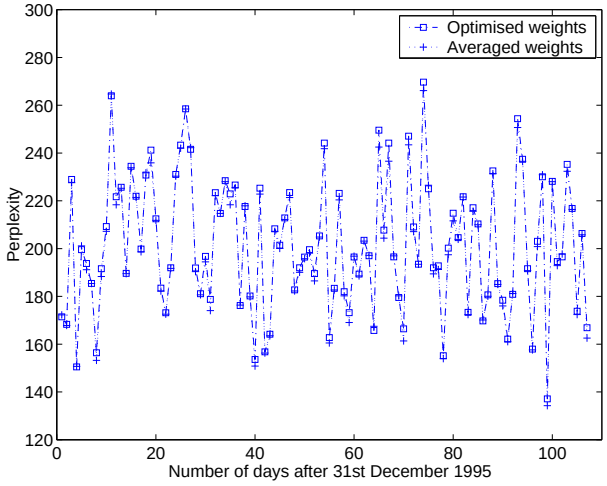

Figure 4: Daily perplexity of adapted LMs using optimised and averaged weights.

weights determined for the previous day's partial transcript (optimised on Friday's transcript in the case of testing on Monday's transcript). We see that the perplexity of the interpolated model using the averaged weights closely follows that of the model using the daily optimised weights. The average daily reduction in perplexity of the adapted LM over the baseline model is 6.6% when optimised weights are used and 7.4% when the averaged

weights are used. The averaged weights can therefore be used as a smoother estimate of the optimised weights without the need to determine new sets of weights for each day. It is unclear after how much time, if ever, these weights would become suboptimal.

For each day's show in the recognition test set an interpolated LM was built using the baseline LM and the appropriate daily LMs with the corresponding averaged weights determined in the perplexity experiment above. The component LMs were merged into a single back-off LM that was used in the recogniser. The merged LM is an approximation of the linear interpolation of component LMs since any backed-off estimates are not strictly a linear interpolation of the component LMs' probabilities. However, this was necessary for recognition purposes and has been shown not to be detrimental to performance. The recognition results using the interpolated models are given in the fourth row of Table 1.

5.3. Baseline plus Marketplace data

Since *Marketplace* data was available for adaptation, it is only fair to evaluate what the effect of adding the first five months of partial *Marketplace* transcripts to the baseline LM training data would be. The adaptation here is more domain than temporal adaptation. Again there is no obvious way to weight the contribution that the adaptation data should be given in the LM, so it was simply added once to the baseline LM training data. The top 65k words in the adapted training data formed the vocabulary and a trigram LM was built using the adapted training data in an identical manner to that for the original baseline LM. The recognition results are given in the fifth row of Table 1.

6. Discussion

The problem of OOV words is one that is potentially serious for audio indexing tasks and from the results obtained in Sections 3 and 4 it appears that using external sources of time-dependent data does not significantly reduce the problem. Although, by minimising the OOV rate we maximise the number of words that we are able to recognise, capturing the 'spontaneous' daily singletons still presents a problem. This is especially true if we conjecture that these are the most indexable words. There is also the question of query OOVs which has not been addressed. This is difficult to assess without examples of the words people will actually use to query a system. Alternative methods to brute-force approaches to adapting the vocabulary may be more productive. Query expansion is one solution that has been shown to increase retrieval performance even if the word that is used in the query was not in the vocabulary in the first place. Another promising solution is phone-based recognition of speech for audio indexing [7]. Query words are mapped to a phonetic representation which is then searched for. A combination of both word and phone-based recognition is likely to optimise the tradeoff between retrieval performance and search speed.

Two methods of combining external sources of time-dependent data for temporally adapting LMs and one simple domain adaptation method have been described. In general, all three methods improved recognition accuracy over the baseline LM, showing that, as expected, there is useful information in data that is temporally close to data we are trying to recognise. This was exhibited in the average values of the weights of the interpolated component models where a decay was observed in the importance assigned to data that was older with respect to the current day. On top of this there was an additional periodicity that highlighted the relevance of data from exactly a week prior. The two methods that used *Marketplace* data either directly (Section 5.3) or to determine the contribution of other data (Section 5.2) both outperformed the method of simply adding temporally close data with no regard to what the contribution of each day's adaptation data should be (Section 5.1).

Although the interpolated LM was shown to perform best, the overall improvements in recognition performance in all cases were small. It is possible that more complex adaptation methods such as using log-linear interpolation [2] would yield better results. However, it is well known that, in general, improvements in the quality of the speech recognition transcript do not improve retrieval performance much (see for example [1]). Therefore, it is unclear whether the effort involved in building a new adapted LM for recognising each new day's data is justified by the improvements obtained, particularly if the baseline LM is well trained and performs well to start with.

7. Conclusion

In this paper, we have investigated the implications of recognising speech from news broadcasts in which the content varies on a daily basis. We have seen that many words which are potentially of value for audio indexing purposes inevitably fall outside the fixed vocabulary of the speech recognition system. It was also found to be difficult to remedy this problem using external sources of time-dependent data. Two methods of combining external sources of time-dependent data and one method of combining domain dependent data were shown to improve the recognition performance on *Marketplace* data. However, largely because the baseline LM was well trained, the improvements were found to be small. Overall, the results suggested that it is probably better to choose the vocabulary and build a one-off LM based on as much data as one has available, especially data that is both temporally and domain relevant to the recognition task.

8. Acknowledgements

The author wishes to thank Bhiksha Raj for his help with building acoustic models and the speech group at CRL for their helpful discussions on the work in this paper.

9. References

[1] J-M. Van Thong, D. Goddeau, A. Litvinova, B. Logan, P. Moreno, and M. Swain, "SpechBot: a Speech Recognition based Audio Indexing System for the Web," in *Proceedings of RIAO*, Paris, April 2000.

[2] D. Klakow, "Log-linear Interpolation of Language Models," in *Proceedings of ICSLP*, Sydney, Australia, 1998.

[3] R. Rosenfeld, *Adaptive Statistical Language Modelling: A Maximum Entropy Approach*, Ph.D. thesis, School of Computer Science, Carnegie Mellon University, 1994, CMU-CS-94-138.

[4] A. Berger and R. Miller, "Just-in-time Language Modelling," in *Proceedings of ICASSP*, Seattle, USA, 1998.

[5] P. R. Clarkson, *Adaptation of Statistical Language Models for Automatic Speech Recognition*, Ph.D. thesis, Cambridge University, 1999.

[6] "`http://www.marketplace.org`".

[7] K. Ng and V. Zue, "Phonetic Recognition for Spoken Document Retrieval," in *Proceedings of ICASSP*, Seattle, USA, 1998.