Speech Enhanced Multi-span Language Model

A. Nayeeemulla Khan and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036, India
Email: {nayeem,yegna}@cs.iitm.ernet.in

Abstract

To capture local and global constraints in a language, statistical n-grams are used in combination with multi-span language models for improved language modelling. Use of latent semantic analysis (LSA) to capture the global semantic constraints and bigram models to capture local constraints, is shown to reduce the perplexity of the model. In this paper we propose a method in which the multi-span LSA language model can be developed based on the speech signal. Reference pattern vectors are derived from the speech signal for each word in the vocabulary. Based on the normalised distance between the reference word pattern vector and the pattern vector of a word in the training data, the LSA model is developed. We show that this model in combination with a standard bigram model performs better than the conventional bigram + LSA model. The results are demonstrated for a limited vocabulary on a database for the Indian language, Tamil.

1. Introduction

In every language there exist dependencies in the usage of words which could be syntactic, semantic or pragmatic. Local level constraints are captured by means of statistical n-gram models. N-gram models are unable to predict long range dependencies as this requires a large value of n, making the parameter estimates of the model unreliable due to the limited training data available. To model long range dependencies, equivalence classes on the n-gram history [1] and structured language models [2] are useful for limited domains. In less constrained domains they are not as useful. Trigger based language models [3] are also potential ways in which long range dependencies can be captured. But trigger pair selection is a complex task, with different pairs displaying different behaviors. Use of latent semantic analysis to capture long range dependencies has been shown to be effective. In combination with n-gram models it results in a substantial reduction in perplexity [4][5]. In conventional language models no knowledge of the language is used. The data being modelled could as well be a sequence of arbitrary symbols. It is essential to use knowledge sources available to enhance the performance of the statistical language models.

One application of statistical language models is in speech recognition. The use of speech knowledge, prosodic constraints, large span semantic and local syntactic constraints, when integrated with the speech recogniser would improve the performance of the recogniser. In this paper we propose a method in which the semantic constraints in terms of the co-occurrence of words in a document are captured indirectly from the speech signal in the latent semantic analysis (LSA) framework. We show that the speech enhanced LSA language model performs better than the n-gram and the hybrid n-gram + LSA model. The reduction in perplexity for a test set is used to measure the performance of the model.

The paper is organised as follows, the next section briefly illustrates the technique of LSA. Section 3 describes the development of the speech enhanced multi-span language model. In Section 4 the database used is described. Section 5 details the evaluation of the model, followed by discussion of the results in Section 6. We summarise the study in Section 7.

2. Latent semantic analysis

A brief overview of related work on LSA relevant to this study as described in [4][5][6] is presented here. LSA is an algebraic technique that can be used to infer the relationship among words by means of the co-occurrence of the words in identical contexts. Given a set of N documents from a text corpus $\mathcal{T}$, with a vocabulary $\mathcal{V}$ of M words, it specifies a mapping between the discrete sets $\mathcal{V}$ and $\mathcal{T}$ and a continuous vector space $\mathcal{S}$. A document could arbitrarily be a sentence, a paragraph or a larger unit of text.

A matrix W containing the co-occurrence statistics between words and documents is constructed. Here word order is ignored unlike conventional n-gram modelling. Each element of W is weighed by the normalised word entropy and scaled for the document length. The element (i,j) of W is given by

$$w_{i,j} = (1-\epsilon_i)\frac{c_{i,j}}{n_j}$$

$$\epsilon_i = -\frac{1}{\log N}\sum_{j=1}^{N}\frac{c_{i,j}}{t_i}\log\frac{c_{i,j}}{t_i}$$

where $c_{i,j}$ is the number of times word w_i occurs in document d_j, n_j is the total number of words present in d_j, $t_i = \sum_j c_{i,j}$, and ϵ_i is the normalised entropy of w_i in the entire corpus $\mathcal{T}$. The matrix W can be approximated by its order-R singular value decomposition (SVD).

$$W \approx \hat{W} = USV^T$$

This results in three matrices $U_{M\times R}$, $S_{R\times R}$ and $V_{N\times R}$. U and V are column orthonormal and S is a diagonal matrix. This transformation to the lower dimensional space captures major structural association between the words and the documents and removes noise. It also provides a R dimensional representation for both the words and the documents. Based on information retrieval and language modelling studies [5], values of R in the range of 100 to 300 seems to work reasonably. The R-dimensional scaled representation of the word and document vector is given by u_iS and v_jS where u_i and v_j are the corresponding rows of U and V. Any new document (test document) d can be considered as an additional column of the matrix W. Its representation v in the reduced dimensional space is given by $v = d^TU$.

For language modelling given such a representation, and a distance metric in the R-dimensional space it is possible to combine the standard n-grams and the LSA model to derive a hybrid n-gram + LSA language model as detailed in [4][5]. In the following sections we detail the construction of the W matrix from the acoustic signal. Using this speech based W matrix we develop the speech enhanced hybrid n-gram + LSA model.

3. Speech enhanced multi-span language model

The block diagram of the proposed system for construction of matrix W is shown in Figure 1. Availability of a database segmented in terms of words is assumed. The duration of a word segment is variable. To find the closeness of a pattern vector representing a word, to other words of the vocabulary, it is desirable to have fixed dimensional pattern vectors for all the words in the vocabulary. From the speech signal corresponding to a word segment, for every frame of 15 msec and a frame shift of 1 msec we derive 13 dimensional mel frequency cepstral coefficients. The euclidean distance between adjacent pairs of feature vectors is computed for all the frames corresponding to the word segment. Depending on the number of frames needed to construct the desired pattern vector, frames are added/dropped. If the number of frames in the word segment is less than the desired, then the frame with the minimum euclidean distance is replicated. For a word segment with larger number of frames than desired, a frame is dropped if its euclidean distance to its neighbor is minimum among the distances computed. This is repeated until the desired number of frames is obtained. It is assumed that there is minimal distortion/loss in adding/dropping the above frames. The selected frames are concatenated to form the fixed dimensional pattern vector representing the word. The resulting pattern vector is large (390 to 572 dimensions). Comparing pattern vectors in such a high dimension space is not preferable. It has been shown that non-linear compression of large dimensional pattern vectors of speech using AANN models does not degrade the speech recognition performance [7]. We use AANN models to compress the large dimensional pattern vector into 40 to 100 dimensions. Thus a reduced dimension pattern vector is derived for each word segment in the entire training data. Pattern vectors corresponding to a word in the training set are used to derive a mean pattern vector, which serves as the reference pattern vector for that word in the vocabulary. One such reference pattern vector is derived for each word in the vocabulary.

For every word segment in a training document (speech file), a compressed pattern vector is derived as explained. The euclidean distance between this pattern vector and all the reference pattern vectors in the vocabulary is determined. The resulting distances are normalised between zero and one. The membership defined as (1 - normalised distance) indicates how close the current pattern vector is to each of the reference pattern vectors in the vocabulary. If this membership is above a certain threshold then the appropriate element $w(i,j)$ is incremented by the membership value. The elements of W are also scaled for the length of the document (No. of words) and weighted by the entropy of the term. Thus the W matrix is derived from the acoustic signal.

4. Database

The database used for the study is the Indian language speech corpus [8]. TV news bulletins from Doordarshan for Tamil language were collected. Speech pertaining to the news reader was manually transcribed and segmented into words representing around 4 hours of speech. Among these bulletins 23 are spoken by females and 10 by males. For this task the database was partitioned manually into news stories belonging to 8 different categories. The details of the database in terms of news stories are shown in Tables 1 and 2. There are no standard text corpus of news bulletins or news wire corpora in Tamil language. As it is preferable to use bigram models trained on data pertaining to the domain of use, we used a bigram

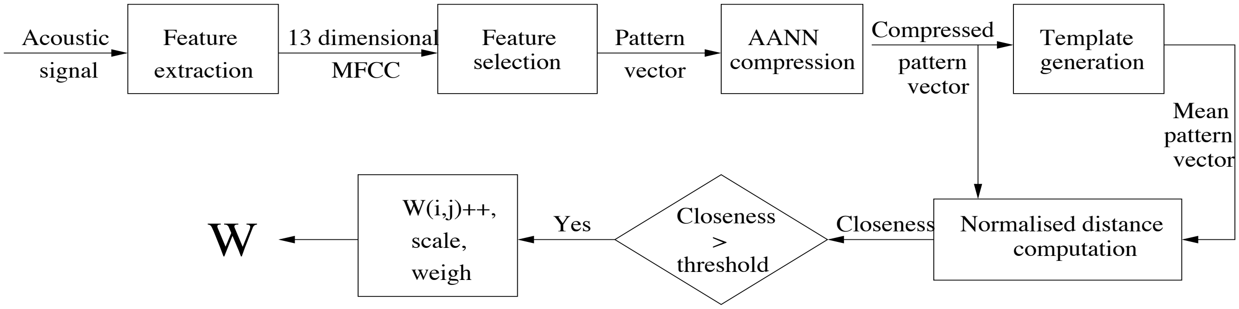

Figure 1: Block diagram for construction of W in the proposed speech enhanced LSA language model

model derived from the limited training data for integration with the LSA model.

Table 1: Description of the database in terms of stories

Story category	No. of documents	
	Training set	Test Set
Economics	44	2
Events	104	23
Others	124	24
Politics	104	4
Sports	34	3
War	163	32
Weather	6	1
World politics	64	3
Total	643	92

Table 2: Database statistics

	Training set	Test set
No. of Docs	643	92
No. of Words	26,380	3,706
Min. No. of Words	6	8
Max. No. of Words	159	136
Avg. No. of Words	41	40

5. Experimental evaluation

From the transcription of the training data we chose a limited vocabulary of 1,278 words (inclusive of the unknown word tag <UNK>), that had at least 4 occurrences in the training data. For the 643 training documents a W matrix of size 1278× 643 is created. The average duration of these words in the database is 431 msec. Assuming a frame shift of 10 msec, 44 frames are chosen using the procedure mentioned in Section 3 to represent the word. The feature vectors concatenated together resulted in pattern vectors of 572 dimension. Such pattern vectors are derived for each word in the training data. The pattern vectors are non-linearly compressed to a smaller dimension (60, 80 or 100) using AANN models. The structure of the AANN model is 572L 858N kN 858N 572L, where L represents a linear unit, N represents a non-linear unit and k is the dimension of the desired compressed pattern vector. All the word pattern vectors in the training data are used to train the AANN model. The model in trained for 200 epochs. The compressed feature vector is obtained from the compression layer of the trained AANN model. These compressed feature vectors are used in the construction of the W matrix. As the structure of the AANN model is large and the training patterns limited, the AANN model may not generalise well. An alternative compact representation of a word using only 30 frames concatenated to form a 390 dimensional pattern vector was also employed. These 390 dimension pattern vectors were compressed to 40, 60 or 80 dimension using an AANN model represented by 390L 585N kN 585N 390L. Using these compressed pattern vectors of different dimensionalities, appropriate W matrices were constructed. The integrated bigram + LSA model was derived as described in [4][5]. We report results for pattern vectors compressed to 60 dimension.

To test the performance of the language model the perplexity of the speech enhanced hybrid bigram + LSA model was found for the test data of Table 1. During testing based on the transcription of the speech document, in a manner similar to the standard LSA language model, for every word in the test document, the appropriate vectors in the reduced dimensional space $u_iS^{1/2}$ and $vS^{1/2}$ are used for computation of the probabilities. The out of vocabulary rate for the test set is very high (41%) due to the limited vocabulary chosen, and the fact that the data pertains to news bulletins. The out of vocabulary words were ignored in perplexity computation.

6. Results

The performance of the speech enhanced hybrid bigram + LSA model is shown in Table 3 for different thresholds of membership values, and a SVD order of 75 (optimal order balancing reconstruction error and noise suppres-

Table 3: Perplexity of the speech enhanced bigram + LSA model for different pattern representation and membership threshold, for a SVD order of 75

Word pattern representation	Membership greater than	Perplexity
Compressed from 572 to 60	0.98	227
	0.97	231
	0.96	231
	0.92	233
Compressed from 390 to 60	0.98	199
	0.97	196
	0.96	195
	0.92	197

Table 4: Comparison of performance of three different language models. LSA models use SVD of order 250

Model	Perplexity	Improvement over bigrams
Bigram	234	-
Bigram + LSA	199	15%
Speech enhanced bigram + LSA	185	21%

sion). If the threshold is high (0.98) the W matrix is similar to its text based counterpart in its sparseness. As the threshold is lowered, more elements of the W matrix are filled, which is like smoothing. The performance of the model improves marginally. For lower thresholds the performance is likely to deteriorate. This behaviour is observed for both the representations of the word pattern vectors. The performance of the model using patterns vectors compressed from 390 to 60 dimension is better than the model using pattern vectors compressed from 572 to 60 dimension.

The performance comparison of the three different language models is shown in Table 4. The perplexity of the speech enhanced hybrid n-gram model is better than the standard bigram model by 21% and shows an improvement of 6% over the conventional text based bigram + LSA model for a SVD order of 250. Order 250 is chosen due to its better performance over SVD order 75.

7. Summary

In this study we proposed an approach for developing a speech enhanced multi-span language model. We have shown that the performance of the system is better than the text based bigram + LSA model for the limited vocabulary of words. No use of word level and document level smoothing [5] is made, which would further reduce the perplexity of the model. Different parameters of the system like word pattern vector representation, order of compression of the pattern vector, membership threshold, SVD order and scaling factor for the LSA probabilities are not optimised. Doing so may improve the performance of the language model. This method of indirect incorporation of the speech information may be a small step towards using speech level constraints in language models for better speech recognition performance. One limitation in extending the study is the lack of a large speech corpus of the size required for language modelling, segmented in terms of words for Indian languages.

8. References

[1] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[2] C. Chelba and F. Jelinek, "Recognition performance of a structured language model," in *Proc. 6th Eur. Conf. Speech Commun. Technol.*, Budapest, Hungary, Sept. 1999, vol. 4, pp. 1567–1570.

[3] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger-based language models: A maximum entropy approach," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Minneapolis, USA, Apr. 1993, vol. 2, pp. 45–48.

[4] N. Coccaro and D. Jurafsky, "Toward better integration of semantic predictors in statistical language modeling," in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, Dec. 1998, pp. 2403–2406.

[5] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proc. IEEE*, vol. 88, no. 8, pp. 1279–1296, Aug. 2000.

[6] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Process*, vol. 25, pp. 259–284, 1998.

[7] S. V. Gangashetty, C. Chandra Sekhar, and B. Yegnanarayana, "Dimension reduction using autoassociative neural network models for recognition of consonant-vowel units of speech," in *Proc. Fifth Int. Conf. Advances in Pattern Recognition*, ISI Calcutta, India, Dec. 2003, pp. 156–159.

[8] A. Nayeemulla Khan, Suryakanth V. Gangashetty, and S. Rajendran, "Speech database for Indian languages- A preliminary study," in *Proc. Int. Conf. Natural Language Processing*, NCST, Mumbai, India, Dec. 2002, pp. 295–301.