Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 30 (2012) 1081 – 1088

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Communication Technology and System Design 2011

Development & evaluation of different acoustic models for Malayalam continuous speech recognition

Cini Kurian[a], Kannan Balakrishnan[b], a*

[a,b]*Department of computer Applications , cochin University of science and Technology, cochin*

Abstract

Performance of any continuous speech recognition system is dependent on the accuracy of its acoustic model. Hence, preparation of a robust and accurate acoustic model lead to satisfactory recognition performance for a speech recognizer. In acoustic modeling of phonetic unit, context information is of prime importance as the phonemes are found to vary according to the place of occurrence in a word. In this paper we compare and evaluate the effect of context dependent tied (CD tied) models, context dependent (CD) and context independent (CI) models in the perspective of continuous speech recognition of Malayalam language. The database for the speech recognition system has utterance from 21 speakers including 11 female and 10 males. Our evaluation results show that CD tied models outperforms CI models over 21%.

Keywords: Speech recogntion ; HMM;MFCC;acoustic modelling;

1. Introduction

Speech Recognition technology has tremendous potential as it is an integral part of future intelligent devices, in which speech recognition and speech synthesis are used as the basic means for communicating with humans. It will simplify the Herculean task of typing and will eliminate the conventional keyboard [11]. This technology will support a lot in manufacturing and control applications where hands or eyes are otherwise occupied. Disabled, elderly and blind people will no longer need to be away from the internet and advanced information technology revolution [5]. Recently, there has been a large increase in the number of recognition applications; like use over telephones, including automated dialing, operator assistance, and remote data access services; such as financial services, for voice dictation

*Cini Kurian. Tel.: 91984774920; fax: 04842838682.
E-mail address: cinikurian@gmail.com

1877-7058 © 2011 Published by Elsevier Ltd.
doi:10.1016/j.proeng.2012.01.966

systems like medical transcription applications. Such tantalizing applications have initiated research in Automatic Speech Recognition (ASR) since 1950's.

Malayalam is one among the 22 languages spoken in India with about 38 million speakers. Malayalam belongs to the Dravidian family of languages and is one of the four major languages of this family with a rich literary tradition. The majority of Malayalam speakers live in the Kerala, one of the southern states of India and in the union territory of Lakshadweep. There are 37 consonants and 16 vowels in the language. It is a syllable based language and written with syllabic alphabet in which all consonants have an inherent vowel /a/. There are different spoken forms in Malayalam although the literary dialect throughout Kerala is almost uniform.[3,6]

Speech recognition is a highly complex task. The basic issue in speech recognition is dealing with two kinds of variability: acoustic and temporal [15]. Acoustic variability covers different accents, pronunciation, pitches, volume, and so on, while temporal variability covers different speaking rates. Development of a better acoustic modeling is the core task in Speech recognition research. In this work, we have used HMM for acoustic modeling. In most of the current speech recognition systems, the acoustic component of the recognizer is built by using HMM. The temporal evolution of speech is modeled by the Markov process in which each state is connected by transitions, arranged into a strict hierarchy of phones, words and sentences.

Phoneme based Hidden Markov Models (HMM)[13] are the foundation of this work. We have designed a phone set comprising of 48 phones capturing nearly all the sounds available in standard Malayalam language. For training the models we have used Baum-Welch algorithm[9] and for testing the trained model viterbi algorithms[9] are used. For processing the speech signal it has to be represented in some parametric form[2]. Since MFCC (Mel frequency cepstral coefficient) [13] is more adapted to human hearing, we have used MFCC parameterization technique in this work.

Continuous speech recognition is an area of active research for quite some time now. However when compared to languages like English or French, the state of speech research involving Indian languages is yet to gain momentum. Although some amount of effective research has been made for the development of speech recognizers in Hindi [21] and some south Indian Languages [14,16], the research scenario for Malayalam language is far from satisfactory level. A wavelet based word recognizer, [12] a number recognition system [3], and a digit recognizer [4] based on SVM are the reported works in Malayalam.

The structure of this paper is as follows. In Section 2 acoustic modeling using HMM is described. Context independent phone models are explored in Section 3. Section 4 and section 5 are devoted for Context independent and context dependent tied models respectively. The results of the experiments are discussed in section 6. And finally our conclusion is in Section 7

2. Acoustic modeling and HMM in continuous speech recognition

In this work the acoustic modeling component of the recognizer is created using Hidden Markov Model (HMM). The capability of HMM to statistically model the variability in speech is the main reason for the use of HMM in speech recognition tasks. HMM provides an elegant statistical framework for modeling speech patterns using a Markov [1] process that can be represented as a state machine as shown in Figure 1[9]. The probability distribution associated with each state in an HMM, models the variability which occurs in speech across or even different speech contexts.

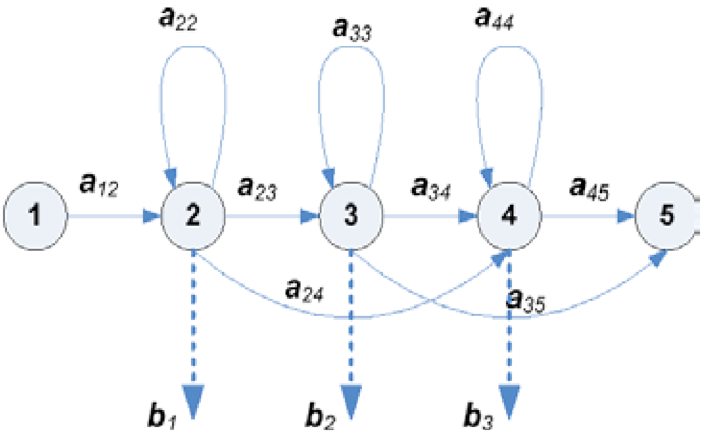

Fig. 1. Topology of a five state HMM

The speech recognition system presented here is based on principles of statistical pattern recognition. [17,18]. An unknown speech wave form is converted by a front-end signal processor into a sequence of acoustic vectors *Y = y1,y2,y3--*. Each of these vectors are a compact representation of the short-time speech spectrum covering a period of typically 10 milliseconds. The utterance consists of sequence of words , *W = w1, w2, w3 ----wn* and it is the job of the LVR system to determine the most probable word sequence , *W* , given the observed acoustic signal *Y*. To do this, Bayes' rule[9] is used to decompose the required probability *P(W/Y)* into two components , that is

$$W = argwmax\ P(W/Y) = arg\ max\ P(W)\ P(Y/W)/P(Y) \tag{1}$$

$$W = argwmaxP(Y/W)P(W) \tag{2}$$

Equation 2 indicates that to find the most likely word sequence, *W* , the word sequence that maximizes the product of *P(W)* and *P(Y/W)* must be found. The first term represents the a priori probability of observing *W* independent of observed signal, and this probability is determined by language model[10]. The second term represents the probability of observing the vector sequence *W*, and this probability is determined by the acoustic model. Figure [18] shows how these relationships might be computed. The purpose of acoustic models is to provide a method of calculating the likelihood of any vector sequence *Y* given word *W*. In principle, the required probability distribution could be located by finding many examples of each *w* and collecting the statistics of the corresponding vector sequences. However, this is unfeasible for large vocabulary system and instead , word sequences are decomposed into basic sound phones. Each individual phone is represented by an HMM. An HMM has a number of states connected by arcs. A five state HMM with simple left-right topology as illustrated in Figure 1. The entry and exist states are provided to make it easy to join models together. The exit state of one phone model can be merged with the entry state of another to form a composite HMM. This allows phone models to be joined together to form words and words to be joined together to cover complete utterance.

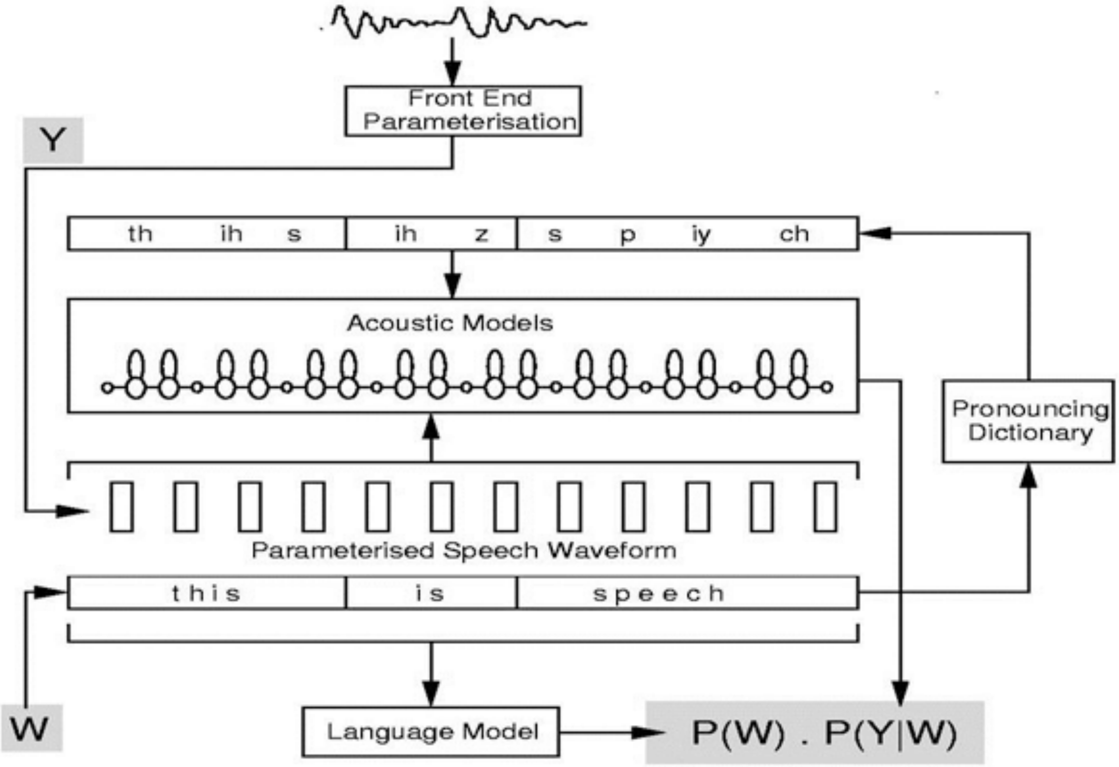

Fig. 2. Acoustic model for a speech recognition system

The acoustic model in a speech recognition engine produces the basic units of speech in the written form with respect to a particular input signal. An input signal is sliced up into overlapping timeframes of 25ms with a 10 ms overlap. Then from these individual frames, 39 MFCC [7] features are extracted. These set of features are then compared with the trained acoustic model. In this work we started by creating of single Gaussian monophone Hidden Markov Model (HMM) [18] for every phone in our phone set.

3. Context independent models

The process of creation of context independent acoustic models start with the preparation of training and testing data. This data comprises of utterance recordings by multiple speakers and the corresponding transcripts encoded using the chosen phone set for the language. These transcripts along with the recordings are fed to the training module which utilizes Baum-Welch Re-estimation [9] in order to create HMMs of all the phones occurring in the training data. The process starts with a default prototype HMM for every phone which is tuned according to the input data and transcriptions. Creation of the monophone HMMs, however require specifying the number of states prior to training. Our experiments suggested for utilizing 5-state HMMs for the purpose of acoustic modeling. However the monophone based models cannot capture the variation of a phone with respect to the context. Phones are found to vary depending on the preceding and succeeding phones and this aspect needs to be captured within the acoustic models to improve performance. Here, we have defined 71 monophones and created models for every monophone.

4. Context dependent models (Triphone based acoustic model)

There are no well defined boundaries between phonemes in continuous speech. The spectral characteristics change continuously due to the inertia of the articulators which move from the position of one phoneme to the position of the next phoneme. Also, the articulators move to the position of the next phoneme even when the current sound is being uttered. Consequently, the acoustic properties of a speech sound not only depends on the identity of the corresponding phoneme, but on the neighboring sounds as well.

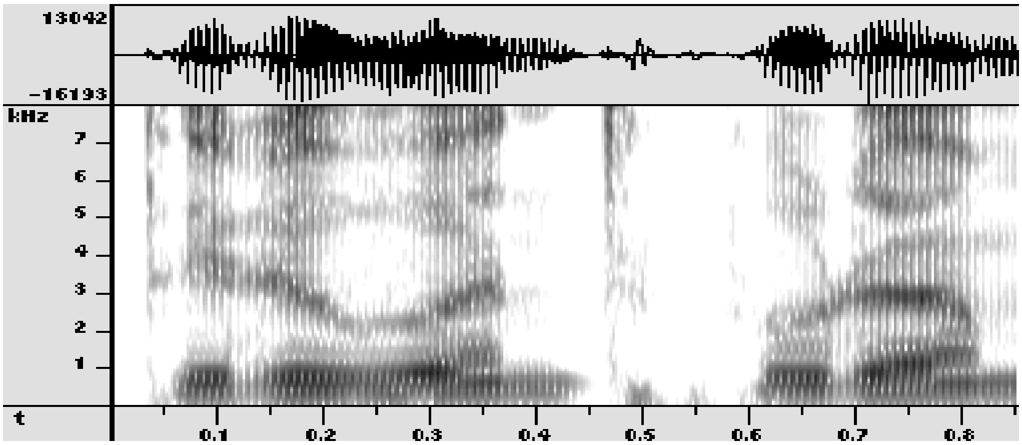

Fig. 3. Wave form and Spectrogram of the word tiruvananthapurm'

The effect of phonetic context on the spectra of phonemes is illustrated in Figure 3[15]. This Figure shows the time waveform and spectrogram of the word "tiruvananthapuram". Two occurrence of the phoneme /a/ in the word have different spectral trajectories. The temporal variations of spectra of two instances of the phonome /a/ are different due to different phonetic contexts. The second format of the vowel /a/ is increasing in the first case, whereas it is nearly steady (and decline slightly later) in the case of /a/ following /r/ . Hence contextual effect cause large variations in the way that different sounds are produced. Hence, in order to achieve good phonetic discrimination, different HMMs have to be trained for each different context. The simplest and most common approach is to use triphones, where every phone has a distinct HMM model for every unique pair of left and right neighbors.

5. Context Dependent tied models (State tying of triphones)

When triphones are used, they result in a system that has too many parameters to train. For example, in English language there are about 45 phones. And in principle, 45^3 (approximately 60,000 triphones, since all cannot occur due to phonotactic constraints of the language) be triphones need to be trained. In practice, around 10 mixture GMM model, with a 39 element acoustic vectors would require around 790 parameters per state. Hence, 60,000, 3-state triphones would have a total of 142 million parameters[17]. Here arise the problem of too many parameters and too little training data which are absolutely crucial in the design of a statistical speech recognizer. This problem is dealt with tying states [17] that are acoustically indistinguishable . This allows all the data associated with each individual state to be pooled and thereby give more robust estimates for the parameters of the tied state. This is illustrated in Figure 4. In conventional triphones, each triphone has its own private output distribution. After tying several states share distributions, the choice of which states to tie is made using decision tree algorithm. In this work we have a total of 71 monophones and after tying the triphones , we could reduce the number of states from 6222 to 1355

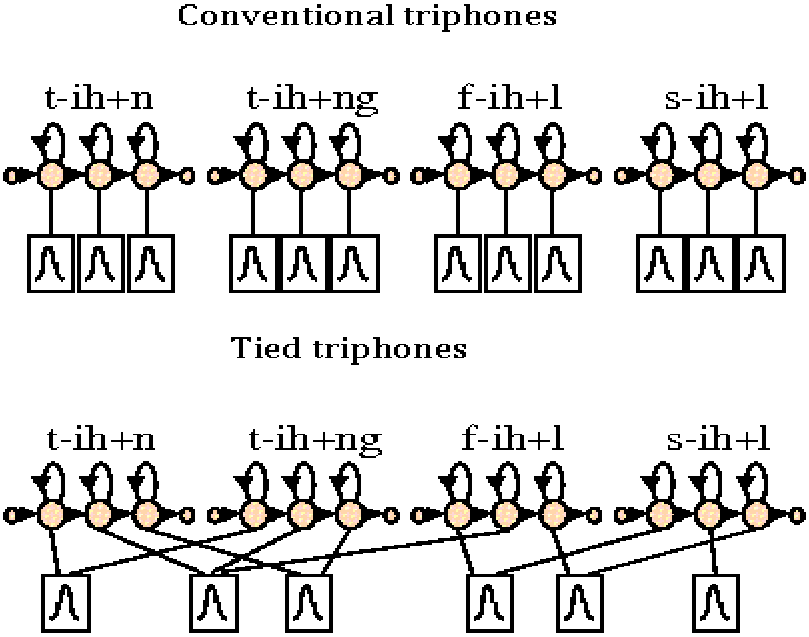

Fig 4. State typing of triphones

6. Experiments and results

We have conducted several experiments with the different types of acoustic models of Malayalam continuous speech recognizer. We have used about 420 recorded utterances from 11 female and 10 male speakers. Recording is done in normal office environment using a head set, having microphone with 70Hz to 1600Hz of frequency range. Moreover, it is done with 16 kHz sampling frequency quantized by 16 bit. The speech is saved in Microsoft wave format.

For training and testing the system, the database is divided into three equal parts- 1, 2, & 3 and the training is conducted in a round robin fashion. For each trial, 2/3rd of the data is taken for training and 1/3rd of the remaining the data is used for testing. For eg. In trial I, part 1 and part 2 of the data is given for training. Then Part 3 of the database is used for testing the trained system. In trial II, part 1 and part 3 of the data base is used for training and part II of the database is used for testing. In experiment III, part 2 and part 3 of the database is taken for training and the system is tested with part 1 of the database. The result in terms of word recognition accuracy, sentence recognition accuracy, number of words deleted, inserted, substituted are obtained from each experiment. For all the performance evaluation reports detailed in the following sections, we have adopted the above procedure and the result reported (sentence recognition accuracy) is the average of testing experiments of I, II and III.

Word Error Rate (WER) is the standard evaluation metric used here for speech recognition. It is computed by SCLITE [8], a scoring and evaluating tool from National Institute of Standards and Technology (NIST). Sclite is designed to compare text output from a speech recognizer such as hypothesis text to the original text (reference text) and to generate a report summarizing the performance. The comparison of reference to the hypothesis text is called the alignment process. Then result of the alignment process is obtained in terms of WER, SER, and number of word deletions, insertions and substitutions. If N is the number of words in the correct transcript; S, the number of substitutions; and D, the number of Deletions, then, $WER = ((S + D + I) N) / 100$ and Sentence Error rate (S.E.R) = (Number of sentences with at least one word error/ total Number of sentences) * 100

Number of Gausian Mixtures for each HMM state: Number of gausian mixtures for each HMM are varied (4,8,16) and the results are obtained as detailed in table 1. Initially we have carried out the training for context independent models. The different parameters used are: trigram language model(n=3) , Gausian mixture of 8 and 3 states per HMM. Then the testing results in terms of sentence recognition accuracy is obtained. The same procedure is repeated for CD and CD-TIED models. Then we have changed the Gausian mixture to 4 and 16 and the whole process is repeated. Table 1 shows the detailed results. . It is evident from the chart that tied-state triphone based models clearly outperform monophone based models.

Table 1. Sentence Recognition Accuracy (%)

	Context independent (CI) Models	Context dependent (CD) models	CD tied models
GMM =4	61.2	76.3	80.3
GMM=8	64.3	77.6	81.5
GMM=16	56.6	69.6	76.4

Number of HMM states per phone: Phoneme unit in our recognizer are initially modeled with 3 state left to right HMMs. 5 state per HMM for phonetic unit are also examined. The results are detailed in figure 5 for CI , CD and CD-TIED models. It is apparent from the chart that tied-state triphone based models clearly surpass monophone based models by 21%.

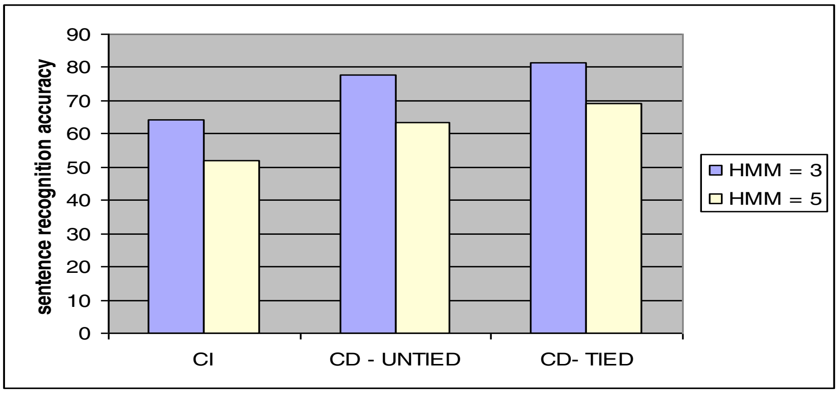

Fig. 5. Comparison of CI, CD and CD-tied models

7. Conclusion

In this paper we have developed three types of acoustic models for Malayalam continuous speech recognition system and compared the performance of recognition accuracy of speech recognizer in the event of CI,CD and CD-TIED types of acoustic modeling. From the results we have concluded that the type of acoustic modeling influence a lot in the recognition performance of the speech recognizer. CD-TIED models outperform CD models by about 21% for Malayalam continuous speech recognition. The improvement in the recognition figures indicates the adaptability of the tied-state triphone based modeling technique for the purpose of speech recognition. In continuation of this work we propose to improve the model accuracy by utilizing more information of the linguistic knowledge such as tone, prosody, and to implement more efficient approach into the acoustic modeling process.

References

[1] A. Ganapathiraju, J. Hamaker and J.Picone, “Support Vector machines for speech Recogntion,” Proceedings of the International Conferences on Spoken Language processing, Sdney,Australia, November , 1999, pp.292-296.

[2] B.Gold, N. Morgan, “Speech and audio signal processing”, John Wiley and Sons, N.Y., 2002

[3] Cini Kurian, Kannan Balakrishnan , (2009), “Speech Recognition of Malayalam Numbers”, IEEE Transaction on Nature and Biologically Inspired computing NaBIC-2009), pp 1475-1479

[4] Cini Kurian, F. Shah, A.;Balakrishnan, K. (2010), “Isolated Malayalam digit recognition using Support Vector Machines”, IEEE Transaction on Communication Control and Computing Technologies (ICCCCT-2010), pp 692 -695

[5] Cini Kurian;Kannan Balakrishnan,K ; "Natural Language Processing in India Prospects and Challanges" Proceedings of the International Conference on “Recent Trends in Computational Science 2008“(ICRTCS-2008), Kochin, India.June 11-June 13

[6] Cini Kurian , Kannan Balakrishnan K, “Automated Transcription System for MalayalamLanguage ” International Journal of Computer Applications(*IJCA*), ISSN-0975-8887, volume 19- No.5, April 2011

[7] Davis S and Mermelstein P, “Comparison of parametric representations for Monosyllabic word Recognition in continuously spoken sentences”, IEEE Trans On ASSP,vol. 28, pp.357 – 366A.

[8] Fiscus, J. (1998) Sclite Scoring Package Version 1.5, US National Institute of Standard Technology (NIST), URL - http://www.itl.nist.gov/iaui/894.01/tools/.

[9] F.Felinek, “Statistical Methods for Speech recognition” MIT Press, cambridge Massachusetts, USA, 1997

[10] Huang, X., Alex, A., and Hon, H. W. (2001). “Spoken Language Processing; A Guide to Theory, Algorithm and System Development”, Prentice Hall, Upper Saddle River, New Jersey

[11] Jurasky, D, and Martin, J.H (2007). “Speech and Language Processing : An introduction to natural language Processing, Computational linguistics, and speech recognition”, 2nd edition

[12] Krishnan, V.R.V. Jayakumar A, Anto P B (2008) , “Speech Recognition of isolated Malayalam Words Using Wavlet features and Artificial Neural Network”. DELTA2008. 4th IEEE International Symposium on Electronic Design, Test and Applications, 2008.Volume, Issue, 23-25 Jan. 2008 Page(s):240 – 243

[13] Lawrence Rabiner, Biing-Hwang Juang, “Fundamentals of Speech Recognition”, Pearson Education 2008,

[14] M Kumar., et al “A Large Vocabulary Continuous Speech recognition system for Hindi”, IBM Research and Development Journal, September 2004

[15] Samudravijaya K, “Speech and Speaker Recognition: A tutorial”, Proc. Int. Workshop on Tech. Development in Indian Languages, Kolkata, Jan 22-24, 2003

[16] Singh, S. P., et al “Building Large Vocabulary Speech Recognition Systems for Indian languges ” International Conference on Natural Language Processing, 1:245-254, 2004.

[17] S Yong, J. Odell, and P. Woodland. “Tree-Based State Tying for High Accuracy Acoustic modeling”. In Proc Human Language Technology Workshop, pages 307 312, Plainsboro NJ, Morgan Kaufman Publishers Inc, Mar. 1994