Automatic Speech Recognition
Continuous Word Recognition using Monophone and Triphone Hidden Markov Models and Neural Networks

s1886258 , s1789342
The University of Edinburgh
School of Informatics

Abstract— **This paper aims at exploiting efficient ways of performing automatic speech recognition tasks based on the KALDI toolkit on two datasets; the TIMIT corpus and ASR_ALL, a dataset of recordings from students of the University of Edinburgh. We examine speech recognition performance on monophone and triphone models and then, extend our experimentation by implementing a hybrid DNN-HMM model. In the seek of improving WER on recognition results, we try to reduce speech variability by applying gender identification and then, speaker adaptation on our data. We prove that WER can improve after these moderations however, there are certain limitations which impede improvements.**

Index Terms— **Speech Recognition, GMM-HMM, hybrid, DNN-HMM, gender-identification, speaker adaptation**

I. INTRODUCTION

Automatic Speech Recognition (ASR) is the process of transcribing speech to text; transforming an acoustic signal into a sequence of words without necessarily having solid understanding of either the meaning or the intention of the context of the transcribed speech. The main focus in this paper is to carry out continuous word recognition by training monophone and then, triphone models based on Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) on the TIMIT dataset as well as, ASR19_ALL data set, which contains recordings by 51 speakers, with 20 utterances per speaker, created by the University of Edinburgh.

A standard approach to speech recognition is to create statistical models for each word in the vocabulary and through pattern recognition, recognize speech [3]. However, this applies to relatively small vocabulary sizes (up to several hundred words). In order to deal with larger vocabulary sizes, the most widely adopted method is the use of phonemes, units of speech, which help represent all spoken sounds in a language. In ASR, phonemes represent the words in a language.

Speech modelling based on phonemes, is represented by the Hidden Markov Model. The feature vectors used in HMMs are extracted from the speech signal and represent the changes across these features. Each phoneme is represented by a 3-state HMM, where each state includes a continuous density Gaussian Model. However, in order for the models generated to be able to capture contextual information, triphone models are built and the states are tied with decision trees.

It has been observed that tied-state triphones improve recognition accuracy. However, the latest approach combines an HMM with Gaussian Density Models with Neural Networks, which are generally, able to capture more information without the computational complexities of more statistical models. The use of a hybrid DNN-HMM model proves to improve the accuracy of recognition performance. However, it is computationally more expensive and falls under the same limitations as traditional approaches.

II. MONOPHONE MODEL

A monophone model is a phone-based acoustic model which contains information about the individual sound units that make up words in a language. Monophone models are trained based on HMMs, where each phoneme is represented by one HMM with Gaussian Probability Density Functions (PDFs).

A. Gaussian Mixture Components

An acoustic observation is a 39-dimension feature vector; it includes MFCCs, deltas and delta-deltas. Multivariate Gaussians allow for the probability assignment on these feature vectors. Gaussian Mixture Models are a weighted mixture of multivariate Gaussians, which avoid modelling features as a single normal distribution, since it can be too strong as an assumption.

In speech recognition system, the standard metric used for evaluation is ***word error rate*** (WER), which computes how much a hypothesized string differs from the reference one.

First, we aim at investigating how the number of Gaussian Mixture Components influences WER. During the parameter tuning, we aim to find the optimal number of Gaussians for the minimizing WER. The search for Gaussian mixture components ranges from 1 mixture component to 20000 (grid-search), as usually large speech recognition systems have 30000 GMMs each with 32 components (cite slides?). The results are summarized in Figure 1.

Based on the WER, the optimal number of Gaussian mixture components is 10000, which reduces WER to

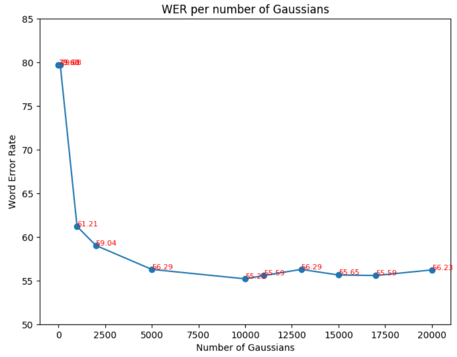

Fig. 1. WER against the number of Gaussian Mixture Components

55.21. More components than 10000 leads to an increase of the WER which is an indicator of over-fitting to the training data; the model starts assigning too much probability mass around the observed data points and not enough mass to the unseen values that may occur in the test set. Additionally, we also consider the train time. As can be seen in Figure 2, it has a linear relationship with the number of Gaussian components; as the number of the components increases, so does the time required for training. Basic acoustic modelling computations make

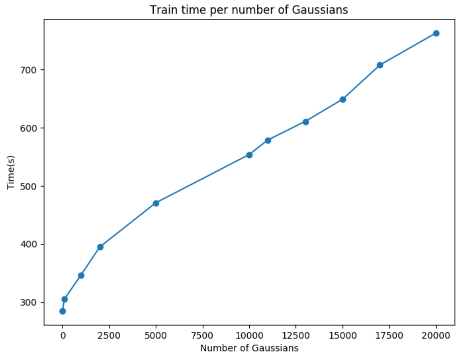

Fig. 2. Train time against the number of Gaussian Mixture Components

use of probabilities. However, due to numeric underflow, which results from the multiplication of many small probability values, as well as the need for computational speed, adding log-probabilities over multiplying plain probabilities is preferred. This way exponentiation is also, avoided [1]. Figure 3 shows the log-likelihood in terms of train/test time.

During testing time, the log-likelihood increases alongside with the number of Gaussian components while during train time, after 10000 Gaussian components, the log-probabilities start decreasing, indicating that they start fitting the data.

Based on the above observations on WER, train time and log-likelihoods, the optimal number of Gaussian components chosen for the next experiments are 10000 Gaussians.

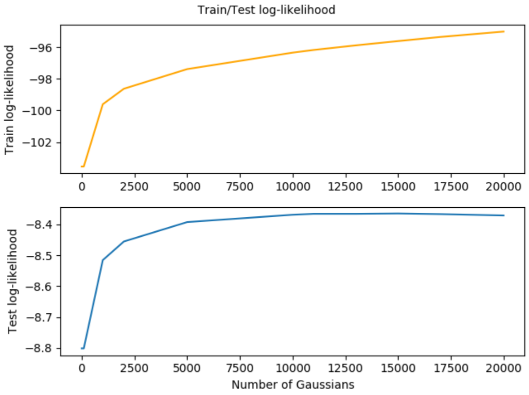

Fig. 3. Log-likelihood during train/test time against the number of Gaussians

B. Dynamic Features and Cepstral Mean/Variance Normalization

MFCC vectors include 39 features. The cepstrum extraction, via the inverse DFT, results in 12 cepstral features per frame. Energy is the 13th feature and is important for phone detection, as it is correlated with the identity of phones. Since speech signal is not constant from frame to frame, in order to capture the changes across phonemes, extra features can be added, that are related to these changes. These dynamic features are 12 deltas and 1 energy delta coefficient alongside with 12 double-delta cepstral coefficients and 1 energy delta-delta coefficient. The deltas represent the changes between frames across each cepstral/energy feature while the delta-deltas represent the changes between frames across each delta feature [1].

Environmental variation can affect the performance of ASR systems, since they tend to be sensitive to the contaminated speech signal from noisy environments, like convolutional noise due to channel variation[1]. Effective ways to increase noise robustness is cepstral mean/variance normalization (CMN/CVN). Particularly, the cepstral coefficients are linearly transformed to have the same segmental statistics, zero mean and unit variance [4].

We investigated how these dynamic features and CMN/CVN influence the WER on the monophone model. Table 1 shows the differences in WER when certain and/or all features are enabled.

Features Enabled	WER
MFCCs + ΔMFCCs + Δ^2MFCCs + CMN/CVN	**55.78**
MFCCs + ΔMFCCs + Δ^2MFCCs	**57.96**
MFCCs + CMN/CVN	**81.53**
MFCCs + ΔMFCCs + CMN/CVN	**59.62**

TABLE I

EFFECT ON WER BASED ON THE USE OF DIFFERENT FEATURES

Based on the results presented in Table 1, it is clear that the use of dynamic features alongside with mean normalization are necessary for covering the different aspects that can affect the speech recognition procedure. The WER increases a lot when no delta features are included or when deltas are not accompanied with delta-deltas. Variance normalization affects WER but not substantially, WER decreased 2.18% when variance normalization was included. This could imply that noise variance was not big in the particular experiments. The WER, however, worsens substantially, when mean normalisation is not performed, which suggests that this step is indeed needed to mitigate the effects of noise (channel variation).

C. Speech Recognition

Based on the experiments in A and B, the optimal parameters for running a speech recognition experiment on the monophone model are MFCCs with dynamic features and mean normalization applied with 10000 Gaussians. We tested our own voice recordings (MyAudio/ASR_OWN-1 and MyAudio/ASR_OWN-2 as a single set).

The results of the experiment are shown in Table 2 below. Although we picked the parameters yielded as optimal based on the previous experiments, the WER is not satisfactory as well as, the test log-likelihood, which has also, decreased. This could be due to the nature of the monophone model itself. It represents the acoustic parameters of a single phoneme and is not able to capture any contextual dependencies. It could also be due to the nature of the test and train data itself. For that reason, we aim at modelling a triphone model in the next experiment.

	Resutls
WER	**72.77**
Test log-likelihood	**-9.21614**

TABLE II
SPEECH RECOGNITION RESULTS ON MONOPHONE MODEL

III. Triphone Model

A triphone is a model of a whole phone. It takes into account the context of three phones; the previous, the current and the next phone. It is a context-dependent phone model that represents the phone variance in the context of a left and right phoneme. However, since a complete set of such phones would be too large to handle, due to the problem of training data sparsity, which refers to the restricted language-specific phone combinations, it is preferred to use a smaller set of triphones that have been clustered automatically [1].

A. Clustered Triphones

Clustering triphones is the procedure of tying subphones, the states within an HMM, which share the same context within a cluster [5]. Tied states share the same Gaussians which leads into needing to train one Gaussian for this tied state. This procedure is accomplished through a decision tree, in which the trained Gaussian monophone models are cloned into triphone models that are then, clustered and tied into triphones and are finally, expanded into GMMs [1].

We investigated how the number of clusters alongside with the number of Gaussian mixture components affect WER, while looking for the parameters that yield the lowest WER. Regarding the number of clusters, considering the fact that phonemes may vary considerably depending on their position in a word, ideally, each phoneme will have three HMM states. Then, the model will decide if it should allocate one HMM state to more specific allophones of a particular phoneme. These allophones are usually referred to as subphones or leaves.

The exact number of subphones and Gaussians however, is often decided based on heuristics. This depends on how much data is available and/or on what the model wants to achieve. The final number of leaves will be less than the number of leaves we provide, due to clustering; the leaves are merged as long as the final likelihood, after tree splitting is more than it was before splitting.

The number of clusters we experimented with is in the range (500,5000) and the number of Gaussians (5000,30000). As can be seen from the heatmap in Figure 4, the local optimum parameters for the minimization of WER are:

· number of clusters: 2000
· number of Gaussians: 15000

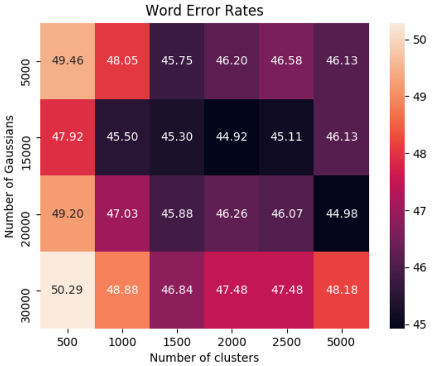

Fig. 4. Effect of different number of clusters and Gaussians on Word Error Rate

To choose the final optimal parameters, we also considered the log-likelihoods and the training times similar to that in the monophone models. Table 3 shows the results: Based on the table, we see that 1000, 1500, 2000 and 2500 with 15000 Gaussians each have the best results considering the log-likelihoods. Making a tradeoff between the WER, log-likelihoods and train time, we choose 2000 Clusters with 15000 Gaussians as our optimal parameters.

B. Speech Recognition

Based on the results that have been configured in the previous task, we again, run a speech recognition

	Train Log-Likelihood	Test Log-likehood	Training Time(s)
500 Clusters, 5000 Gaussians	-97.36	-8.39	151
1000 Clusters, 5000 Gaussians	-97.49	-8.41	149
1500 Clusters, 5000 Gaussians	-97.64	-8.43	169
2000 Clusters, 5000 Gaussians	-97.77	-8.45	176
2500 Clusters, 5000 Gaussians	-97.88	-8.47	182
5000 Clusters, 5000 Gaussians	-97.93	-8.48	182
500 Clusters, 15000 Gaussians	-95.48	-8.35	206
1000 Clusters, 15000 Gaussians	-95.49	-8.37	199
1500 Clusters, 15000 Gaussians	-95.55	-8.38	198
2000 Clusters, 15000 Gaussians	-95.61	-8.39	206
2500 Clusters, 15000 Gaussians	-95.66	-8.41	211
5000 Clusters, 15000 Gaussians	-95.70	-8.42	260
500 Clusters, 20000 Gaussians	-94.86	-8.36	279
1000 Clusters, 20000 Gaussians	-94.85	-8.37	265
1500 Clusters, 20000 Gaussians	-94.89	-8.39	230
2000 Clusters, 20000 Gaussians	-94.93	-8.40	226
2500 Clusters, 20000 Gaussians	-94.97	-8.41	245
5000 Clusters, 20000 Gaussians	-95.00	-8.42	243
500 Clusters, 30000 Gaussians	-93.81	-8.38	329
1000 Clusters, 30000 Gaussians	-93.76	-8.39	270
1500 Clusters, 30000 Gaussians	-93.78	-8.40	273
2000 Clusters, 30000 Gaussians	-93.80	-8.42	268
2500 Clusters, 30000 Gaussians	-93.81	-8.43	257
5000 Clusters, 30000 Gaussians	-93.83	-8.44	306

TABLE III

EFFECT OF NUMBER OF CLUSTERS AND GAUSSIANS ON TRAIN AND TEST LOG-LIKELIHOODS AND TRAIN TIME (IN SECONDS)

experiment, trained on the TIMIT corpus and tested on our recorded utterances. The results are as shown in Table 4 below.

	Resutls
WER	**79.94**
Test log-likelihood	**-9.28272**

TABLE IV

SPEECH RECOGNITION RESULTS ON TRIPHONE MODEL

The results show an increase in WER and decrease in the likelihood. This could be due to the training parameters, which are tuned for the TIMIT dataset. We investigated the test ASR19_ALL corpus to which the test utterances belong to, and conclude that these utterances lie in a different domain as compared to TIMIT. Thus the results could also be attributed to the out-of-domain utterances of the test corpus. The results are however worse than the monophone model. This could be due to the fact that, triphone models increase the amount of unseen data since not all triphones would be seen in the training data. Clustering helps solve this problem to some extent, but it tends to be domain dependent and does not solve the problem entirely. We investiated the decoded utterances and observe that, indeed, the number of <UNK>s in the triphone recognition is far higher than the monophone recognition leading to a worse WER.

IV. EXPERIMENTS

Speech is not constant and is affected by many factors that can distort its quality and cause a lot of variability, which is an obstacle for a successful speech recognition task. Speech variability can be clustered in four main categories: the task domain, speaker characteristics, speaking style and the recognition environment [2].

Different speakers have different characteristics regarding their speech production; different anatomy and physiology [5], age and gender differences as well as, language and accent differences. The gender of the speaker is one source of speech variability. Although many speech recognition systems are tuned to represent statistics over many speakers, to compensate the variability that is inherent in the speech signal, gender identification can provide a more fine-tuned system and improve the recognition performance.

For that reason, we have decided to run a gender dependent speech recognition experiment to test how gender knowledge can improve the recognition performance. In general, speech recognition systems are designed to be speaker-independent due to the problem of data sparsity concerning each user. However, speaker-dependent systems (SD), when implemented, have shown to yield better results than the speaker-independent (SI) ones. In order for this to be accomplished, we need to reduce the variability of the training data and try making them as similar as the data used for testing [1].

A. Data Preparation

Women and men have different phonetic characteristics, such as different vocal tracts, which can cause distinct phonetic changes in speech production. In this task, we use gender information to distinguish between these phonetic differences, separate the data and train two main models based on them. [1]. The datasets used in this experiment are the TIMIT corpus with 3696 utterances, in which gender information per speaker is already provided, and ASR19_all dataset, which includes 1020 recorded utterances from 51 students of the University of Edinburgh, 20 utterances per student. Since this dataset is relatively small, we hand-labelled the utterances based on the speakers as either male, -m, or female, -f, and used it for testing. The resulting TIMIT datasets had 1088 female utterances and 2688 male utterances. The resulting ASR19_OWN data set had 480 female utteranes and 540 male utterances. The TIMIT was used for training. Both datasets were separated using `separate_gender.sh`.

After the gender is determined, the models that contain the determined gender are used for the recognition process.

B. Neural Network-Hidden Markov Model

We typically train a Neural Network-Hidden Markov Model (DNN-HMM) from the labeled frames, phoneme-to-audio alignments,generated by the GMM-HMM system. This indicates that the DNN system is directly affected by the features extracted from the already trained GMM-HMM model and the quality of them. A DNN serves as a classification tool which takes as input audio features and outputs classified, phoneme labels. Therefore, the input nodes of the neural network system will correspond to the dimensions of the input features while the output ones, to the decision tree leaves' labels.

The dimensions of the hidden layers are not constrained by the GMM-HMM model, whose number and size are heuristically decided during the experiment. Before starting the main experiment, we tuned the hyperparameters of the neural network on the TIMIT corpus in order to find the local optima. We initialized the neural network with 2 hidden layers and increased them during training, reaching 10 hidden layers. The number of hidden dimensions was exponentiated, starting from 256 hidden dimensions and reaching 2048 hidden dimensions. Figure 5 shows the effect of each hyperparameter combination in terms of WER. Figure 6 shows the duration of each experiment's training.

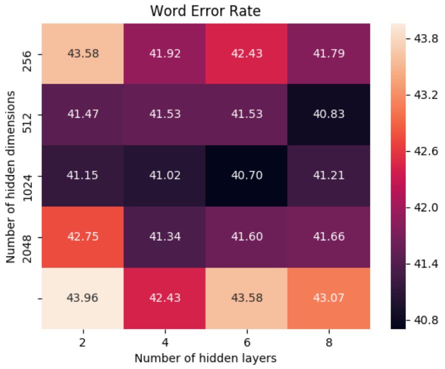

Fig. 5. Effect of tuning neural network's hyperparameters on WER

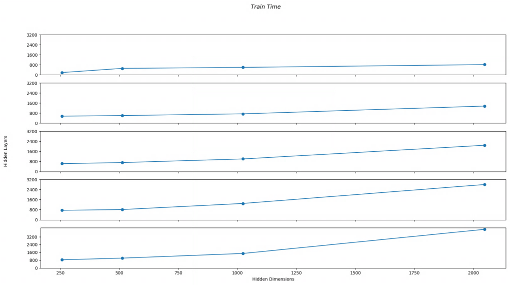

Fig. 6. Train Time of different hypereparameter combination

After tuning the different hyperparameters, we select the combination that yields the best WER in order to train the baseline model. The local optima found during tuning, which minimized WER at 40.70, are:
·Number of Hidden Layers: 6
·Number of Hidden Dimensions: 1024

C. Baseline Model

Based on these best hyperparameters found in the previous task and taking the models, monophone and triphone, that yielded the best results, we train the baseline model and test it on the gender-specified ASR_ALL dataset. The results are shown in Table 4 below.

Baseline System	Test Set	WER
GMM-HMM	ASR19_ALL - Female	71.86
GMM-HMM	ASR19_ALL - Male	65.83
Hybrid	ASR19_ALL - Female	83.03
Hybrid	ASR19_ALL - Male	80.96

TABLE V
BASELINE RESULTS

D. Gender Dependent Acoustic Modelling

We implemented the model by training two GMM-HMM and two DNN-HMM models on each of the gender-separated TIMIT datasets. During test time, we used the gender-separated ASR19_ALL dataset with the corresponding gender-based system for decoding. The results of each task are shown in Table 5 below.

System	Test Set	WER
GMM-HMM	ASR19_ALL - Female	67.38
GMM-HMM	ASR19_ALL - Male	84.55
Hybrid	ASR19_ALL - Female	81.88
Hybrid	ASR19_ALL - Male	91.89

TABLE VI
RESULTS FROM GENDER DEPENDENT MODELS ON ASR_ALL DATASET

Comparing the results, after training the gender-specified datasets, with the Baseline, it can be seen that the WER in the Female dataset has improved while the WER in the Male dataset has deteriorated. Moreover, the Hybrid model does not improve WER neither on the Baseline nor on the Gender-dependent dataset. This could be explained from the availability of the data in each dataset. After splitting, we had lesser data in each set, causing the model to not be fully trained. Another factor is the immense variation in the accents in the test data as compared to the TIMIT data set. The test data consists of diverse accents, whereas the TIMIT is based on American accents. The split training sets become even less representative, further degrading the WER.

E. Speaker Adaptation

Although we have a limited availability of speaker dependent data, we have a sufficient amount of data from different speakers. We aim at making our models adaptive to all kinds of speaker variations in order to reduce error and increase the accuracy of speech recognition performance.

There are various kinds of adaptation techniques available, where the features are normalized in order to avoid mismatches between training and recognition. Maximum Likelihood Linear Regression (MLLR) is one normalization technique used in speaker adaptation systems. It adapts the Gaussians of the acoustic model to a small amount of data from a new speaker by applying a linear transformation on the means [1]. Another technique is the Constrained Maximum Likelihood Regression (cMLLR) or also called, feature-space MLLR (fMLLR), which normalizes both the mean and the covariance of the data to better fit the speaker.

We used the fMLLR technique to adapt and normalize the acoustic features to match speaker variability. The results are shown in Table 6 below.

System	Test Set	WER
Baseline	ASR_ALL	40.51
Hybrid	TIMIT	80.95
Hybrid	ASR19_ALL	82.13

TABLE VII

Results after applying fMLLR technique for speaker adaptation to Gender Dependent models on ASR_all and TIMIT datasets

After applying feature normalization to the data, we notice a slight improvement to WER, indicating that speaker-adaptive systems can benefit the performance of speech recognizers. More training data available and probably, trying other adaptation methods, could yield better WER results.

V. Conclusion

We carried out continuous word recognition experiments on monohpone and triphone Gaussian Mixture Models - Hidden Markov Models and have observed that Word Error Rate decreases when we consider more contextual information, tied-phones, about the language. This is expected as, triphones are able to capture more linguistic variation that could help disambiguate between meanings and sounds.

We extended the experiments by creating a hybrid model that combines Neural Networks and Hidden Markov Models combined with fMLLR Speaker Adaptation technique alongside, with gender-identification. These parameters aimed at reducing speech variability and deal with the data available in the best way possible. We have proven that Word Error Rate, although there were many data limitations, had a slight improvement, indicating the importance of having well-classified data and flexible models.

VI. Appendix

Scripts used for experiments in section IV:

Gender Dependent Systems:
Data Preparation:
separate_gender.sh
separate_gender_gcloud.sh,separate_gender_gcloud_b.sh
–> gcloud versions

Hyperparamter Tuning of Neural Networks :
exp_task3_1_TuneNN.sh
Baseline GMMHMM :
exp_task3_1_GmmHmm_Baseline_GD.sh
Baseline Hybrid :
exp_task3_1_NN_Baseline_GD.sh
Gender Dependent GMMHMM:
exp_task3_1_GmmHmm_GD.sh
Gender Dependent Hybrid :
exp_task3_1_NN_TestASR_GD.sh

Speaker Adaptation:
Data Prepapartion :
prepare_asr_all.sh
prepare_asr_all_gcloud.sh–>gcloud version Baseline :
exp_task3_2_Baseline_SAT.sh
SAT hybrid :
exp_task3_2_SAT.sh

References

[1] D. Jurafsky and J. H. Martin. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

[2] S. Renals and T. Hain. *Speech Recognition*, chapter 12, pages 297–332. John Wiley Sons, Ltd, 2010.

[3] S. C. Sajjan and Vijaya. Speech recognition using monophone and triphone based continuous density hidden markov models. 2015.

[4] O. M. Strand and A. Egeberg. Cepstral mean and variance normalization in the model domain. In *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, 2004.

[5] S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 307–312, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.