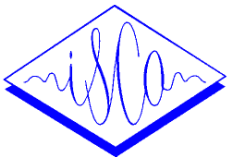

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

BROADCAST NEWS TRANSCRIPTION IN MANDARIN

Langzhou Chen, Lori Lamel, Gilles Adda and Jean-Luc Gauvain

Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
{clz,lamel,gadda,gauvain}@limsi.fr

ABSTRACT

In this paper, our work in developing a Mandarin broadcast news transcription system is described. The main focus of this work is a port of the LIMSI American English broadcast news transcription system to the Chinese Mandarin language. The system consists of an audio partitioner and an HMM-based continuous speech recognizer. The acoustic models were trained on about 24 hours of data from the 1997 Hub4 Mandarin corpus available via LDC. In addition to the transcripts, the language models were trained on Mandarin Chinese News Corpus containing about 186 million characters. We investigate recognition performance as a function of lexical size, with and without tone in the lexicon, and with a topic dependent language model. The transcription character error rate on the DARPA 1997 test set is 18.1% using a lexicon with 3 tone levels and a topic-based language model.

1. INTRODUCTION

It is well known that radio and television broadcast shows contain different types of speech from the acoustic and linguistic points of view, from prepared speech to spontaneous speech, from clean speech to speech with background noise or music, and with wideband and narrowband data. All of these varieties increase the complexity of automatic broadcast news transcription [6]. The goal of the DARPA Hub4 task is to improve speech recognizer performance on this type of inhomogeneous, real-world data. LIMSI has developed an American English broadcast news transcriber, and which has been successfully ported to the French and German languages. Unlike most Western languages, Mandarin Chinese is a character-based language, where there are approximately 8000 frequent characters and about 400 syllables. Mandarin is also a tone-based language, with five different tones associated with the syllables. This paper describes our work in porting the LIMSI American English system to Mandarin Chinese. Porting consists of creating appropriate language-specific components, namely the acoustic models, the pronunciation dictionary and the language models. This process was greatly aided by the availability of training resources at the LDC [1], and manually transcribed evaluation test data from DARPA sponsored benchmark tests [11, 12] and other reported results on this data [3, 13].

The LIMSI transcription system has two main phases: audio partitioning and speaker independent continuous speech recognition [8]. The first phase serves to partition the continuous data stream into homogeneous acoustic segments, assigning appropriate labels with each segment. The second phase carries out word recognition, where the system determines the sequence of words in the segment, associating start and end times and optionally a confidence measure with each word.

Data partitioning, developed for the American English system, is based on an iterative maximum likelihood segmentation/clustering procedure using Gaussian mixture models and agglomerative clustering [5]. In contrast to partitioning algorithms that incorporate phoneme recognition, this approach is language-independent, and the same models are used to partition English, French, German and Mandarin data. The result of the partitioning process is a set of speech segments with speaker, gender and telephone/wide-band labels. The cluster labels of the partitioned data are used during transcription to carry out unsupervised cluster-based adaptation. The partitioning result on this data has not been carefully evaluated, however, the partitioner does not appear to make too many errors. We did notice some confusions in the gender labels, which do not affect the recognition performance as the acoustic models are sex-independent.

The speaker-independent, large vocabulary, continuous speech recognizer makes use of continuous density hidden Markov models (HMMs) with Gaussian mixtures for acoustic modeling [7]. In our Mandarin system, all of the acoustic and language model training data used to develop the system were distributed by the LDC [1].

The acoustic training material consists of about 24 hours of manually transcribed broadcasts recorded from 3 sources: Voice of America (VOA), People's Republic of China Television (CCTV) International News programs and Commercial radio based in Los Angeles (KAZN-AM), along with manual transcriptions. In order to be robust with respect to the varied acoustic conditions, the acoustic models are trained on all data types: clean speech, speech in the presence of background noise or music, or transmitted over noisy channels.

The language model training data comes from the Mandarin Chinese News Corpus which contains about 186 million characters of text from the Renmin Ribao newspaper (People's Daily), the Xinhua newswire service, and scripts from China Radio International, and on the manual transcriptions (460k characters) of the acoustic training data. The language model is obtained by interpolating multiple models trained on data sets with different linguistic properties (newspaper and newswire texts, speech transcriptions).

The remainder of this paper is organized as follows. In the next section the recognition lexicon is described. Sections 3 and 4 describe the acoustic and language models. The decoding procedure is briefly summarized in Section 5, and experimental results are given in Section 6.

2. RECOGNITION LEXICON

Our lexicon is based on the Mandarin lexicon distributed by LDC, with some modifications to the descriptive phone symbol

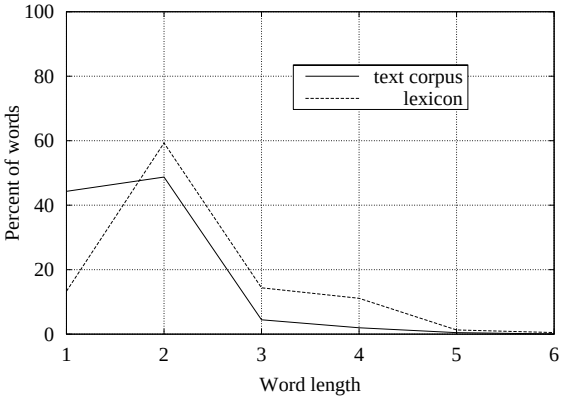

Figure 1: Percent of lexical items as a function of word length in the 50k lexicon and in the training texts.

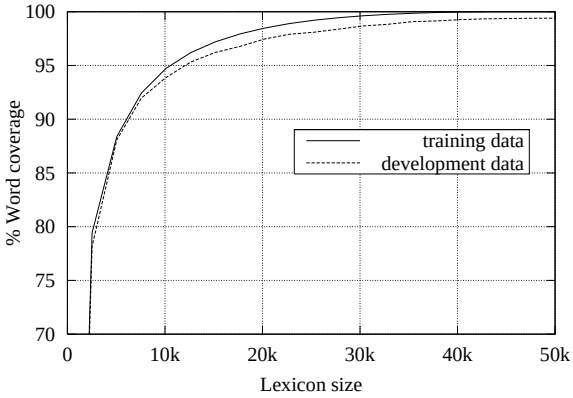

Figure 2: Word coverage vs. lexicon size for training corpus and developing corpus

set and additional lexical entries. The LDC lexicon was developed for use in the Hub5 LVCSR (Large Vocabulary Conversational Speech Recognition) task, and contains a total of 44,405 words with phonemic transcriptions, tone markers (5 levels) and additional information on the morphology and frequency of occurrence in the Xinhua newswire texts and the Hub5 CallHome corpus. The additional entries consist of about 4300 frequent Chinese characters and 3000 words selected from the Mandarin News text corpus. The full lexicon contains a total of 50,590 entries. Most of the additional words added to the LDC lexicon are proper names of people and places that occur frequently in news, but are much less common in conversational speech. These additional words are selected using the maximum match method, which is the the most popular method for Chinese word segmentation. It matches the text in a sentence with the longest item in the lexicon, so as to determine a complete parse of the sentence. This method, which is simple and fast, is as follows: The training texts are first segmented using the entries in the LDC lexicon, which resulted in a word stream where the unknown words (i.e., those words not in the LDC lexicion) are segmented into individual characters. The single character words are then concatenated to form new two-character words if the frequency of occurrence and the mutual information of two character sequence is greater than a threshold. These new two-character words are added to the lexicon and the same procedure is applied to segment the texts using the new lexicon, to form three-character words by combining a single-character word and a two-character word according to the same frequency and mutual information criteria. This procedure was carried out iteratively to obtain additional words from 2 characters to 6 characters in length. All the additional words were manually verified to ensure to remove character strings without any semantic meaning. The result of the process was a list of 3000 additional words.

Because Mandarin Chinese is a character-based language, any Chinese text can be covered by a character string. This means that if all individual characters are included in the recognition lexicon, there is no problem of out-of-vocabulary items. In the extreme case, the recognition vocabulary can contain only characters (there are only about 8000 frequent characters in Mandarin), and character-based language models can be estimated. However, including some multi-character words in the lexicon can significantly reduce the character perplexity.

The text corpus was segmented using the 50k lexicon. Figure 1 shows the distribution of the lexical entries as a function of the number of characters in the lexicon and in the training corpus. The lexical coverage of both the training and development corpora as a function of word frequency is shown in Figure 2. Two character words are the most frequent in the lexicon, and account for over 60% of all word occurences. It can be seen that the 20% most frequent words cover almost 95% of all word occurrences in the training corpus and 94% of the development corpus. With the 40% most frequent words, a lexical coverage of over 97% is obtained for both corpora. These curves suggest that it could be interesting to use a mixed lexicon containing both characters and a set of frequent words.

Three additional lexicons were developed, each of which is a subset of the full 50k lexicon.

1. A 7k lexicon that is composed only of characters
2. A lexicon composed of all frequent characters and the 20% most frequent multi-character words. This lexicon contains about 15k entries.
3. A lexicon containing all frequent characters and the 40% most frequent multi-character words. This lexicon contains about 24k entries.

The lexica are represented using a base set of 39 phones, including specific phones for silence, breath noise, a filler vowel and a filler consonant. (This is a slightly modified set of phones compared to the original LDC lexicon which distinguised 38 phone units.) Two pronunciation lexica were developed differing only in whether or not different tones are distinguished. For pronunciations with tone, we distinguish only 3 tones: flat (tones 1 and 5), rising (tones 2 and 3), and falling (tone 4).

3. ACOUSTIC MODELING

The recognition feature vector contains 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band every 10ms. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment-cluster basis using cepstral mean removal and variance normalisation. Thus each cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficents and the log energy, along with the first and second order derivatives.

The acoustic models are triphone-based context-dependent, phone models, which may be dependent of not on the position of the phone in the word. For trigram decoding, depending on the models set have about 3000-5000 phone contexts (with about 4000 tied-states and 16 Gaussians per mixture) for a total of about 50k Gaussians. Larger tied-state models are used for 4-gram decoding, with about twice as many contexts, and 32 Gaussians per state. Different acoustic model sets of similar size were built for the two sets of pronunciations, with tone and without tone.

We have not explictly modeled the tone information in the feature vector, but model separately phones with different tones. After state tying, the number of Gaussians of the acoustic models with and without tone have roughly the same number of parameters, however, the number of context-dependent models is typically larger when tone is represented. A decision tree was designed for state-tying. A sef of 73 questions which concern the distinctive features of phone, the different tones, and the neighboring phones have been used to control the tying.

4. LANGUAGE MODEL

The language models were trained on the text corpora distributed by the LDC [1] containing about 186 million characters, and the transcriptions of the acoustic training data, 460k characters. The text data come from three sources:

- 1994-1996 China Radio International: 86.7M characters
- 1991-1996 People's Daily: 89.2M characters
- 1994-1996 Xinhua News Agency 9.9M characters

A portion of the transcripts was set aside for use as development texts. This portion, corresponding to the 1997 development test data, contains a total of 14.6k characters.

Prior to estimating the language models, the character stream was segmented into words using the maximum match method. For each lexicon, character based and word based 4-gram language models were trained on the text data, and then interpolated with a 3-gram LM trained on the transcriptions of the audio material. The perplexity of the development text data using the word based interpolated model is 192.6. For reference, a 4-gram LM trained only on the text data gives a word perplexity of 722 and the 3-gram trained on the audio transcripts yields a perplexity of 385 on the development text set.

We carried out some experiments with topic dependent language models [10]. Compared with a general language n-gram model, topic dependent models are trained using topic-specific corpora. These models therefore have a better predicting ability for one of the specific domains than a general model. We divided the training corpus into different topic-dependent subcorpora and trained different language models on these to obtain a more accurate language model.

In order to train topic dependent language models, we needed to solve the problem of text corpus clustering. Our clustering algorithm is based on a list of topic dependent keywords. A keyword is a word that is representative of a particular topic. A keyword usually occurs frequently in articles related to the keyword topic, and seldomly occurs in others articles. Assuming that we have a corpus which is composed of m articles $A_1, A_2, ..., A_m$, a word w_i is declared a keyword after looking at the distribution of the values $\Pr(A_j|w_i)$ for all A_j (using the relative frequencies $N(w_i, A_j)/N(w_i)$ as estimates). Based on this histogram the articles are separated in two disjoint classes by maximizing the difference between the class means. We assume that w_i is a keyword if this difference is greater than an empirically determined threshold. Using this approach a list of about 3000 keywords was identified. A topic was then associated to each keyword. This has been done semi-automatically. In our system 8 topics were determined, corresponding to very broad categories: international politics, national politics, economics, sports, legal issues, history, arts & leisure, and science & technology.

Based on the keyword and topic lists the corpus clustering procedure is as follows:

1. The entire text corpus is first segmented into stories. If an article is longer than 3000 words, we try to detect topic change points in the article, by comparing the similarity of the texts on the two sides of each potential change point. This is based on a vector space model where the cosine of the angle between the two vectors is computed, and local maxima are chosen as change points. Potential changes points are proposed every sentence, using a window of 40 sentences. For shorter articles we assume that the article is made of only one story.
2. The articles (or stories) are then associated to the different topics. If $\Pr(A_j|w_i)$ is greater than a threshold, then A_j is associated with the topic associated to keyword w_i. If more than one keyword is located in the article then the story is associated with multiple topics.
3. Finally a language model is trained for each topic using all the stories associated to that topic. The resulting LMs are then interpolated with a LM trained on the transcriptions of the acoustic training data. The interpolation weights are estimated with the EM algorithm on the development set.

The word perplexity of the topic mixture models on the development data is 177.3, which is significantly lower than 192.6 the perplexity of the general Mandarin 4-gram model.

5. DECODING

Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation [6, 4].

Step 1: Initial Hypothesis Generation - This step generates initial hypotheses which are then used for cluster-based acoustic model adaptation. This is done via one pass cross-word trigram decoding with gender-specific sets of position-dependent tied state triphones and a trigram language model.

Step 2: Word Graph Generation - Unsupervised acoustic model adaptation is performed for each segment cluster using the MLLR technique [9]. A word graph is generated for each segment in a one pass trigram decoding using position-dependent, tied-state triphones (16 Gaussians per state) and the trigram language model used in step 1.

Step 3: Final Hypothesis Generation - The final hypothesis is generated after a second MLLR adaptation using the word graphs, a 4-gram model and a 32-Gaussian version of the acoustic models used in step 2.

6. EXPERIMENTAL RESULTS

The transcription system was evaluated on the 1997 NIST Hub4 Mandarin evaluation data containing 1h of speech. The test data come from the same sources as the training data, that is

Lexicon size	*7k*	*15k*	*24k*	*50k*
Word Perplexity	3411	244	209	192
Character perplexity	61	37	29	26
4-gram	27.8	21.2	20.5	20.3
Rel.gain		24%	3%	1%
4-gram, tones	25.8	19.8	19.8	19.4
Rel.gain		23%	0%	2%

Table 1: Transcription character error rate (%) for different lexicons with and without tone using a 4-gram language model. Normalized word and character perplexities of the development text set based on different lexica and 4-gram language models.

	word	*Character Error Rate*	
Language model	*perplexity*	*40 phones*	*40 phones+3tones*
3-gram SI	201.4	23.3	21.7
3-gram	201.4	20.4	19.3
4-gram	192.6	-	18.5
4-gram topic	177.3	-	18.1

Table 2: Character error rates for different system configurations.

two radio sources VOA, KAZN-AM (Los Angeles), and on TV sources CCTV (Beijing) [11].

First a set of experiments were carried out to explore the relationship of the size of lexicon (in terms of the number of multiple character words) and the character recognition performance. Similar position-independent acoustic models were trained for each lexicon, and a 4-gram LM was used for each system. The results are summarized in Table 1 in terms of character error rate, obtained with a 3-pass decoding carried out in under 10xRT [4]. Concerning the lexicon size, better performance is obtained with the larger lexica. A gain of over 20% is observed with the 20k lexicon which includes the 20% most frequent multi-character words. The addition of more words yields much smaller gains. There is a corresponding decrease in the normalized word and character perplexities [2] also shown in the table. Including tone information in the lexicon improves the recognizer performance by about 5% relative.

Table 2 summarizes some experimental results comparing different language models for the 50k lexicon and running slower (35xRT instead of 10xRT). The first line gives the error rates after the first decoding pass with a trigram language model and speaker-independent acoustic models. All other entries include cluster-based acoustic model adaptation using the MLLR technique. Including 3 levels of tone information in the lexicon is seen to improve the recognizer performance by about 5% relative. Both of the 4-gram language models give an improvement over the 3-gram language model. A larger reduction in error is obtained for the topic 4-gram (6% relative) compared to the standard 4-gram (4% relative).

7. CONCLUSION

We have described our broadcast news transcription system for Mandarin Chinese, which is a port of the LIMSI American English broadcast news transcription system using linguistic resources available from the LDC. Since Mandarin is a tone-based language, two sets of pronunciations lexicons have been developed, with tone and without tone. Similar sized sets of context-dependent, tied state HMMs were built to assess the importance of representing tone in the lexicon. Chinese being a character based language, we also investigated the relationship between the size of the lexicon and the performance of the recognition system. Four lexica of different sizes have been developed, containing different number of multi-character words. Our experiments showed that multi-character words are very important to system performance, and that the main gain comes by adding the most frequent 10k words. Topic dependent language models, which are more accurate than general language models also gave a small improvement in system performance. The character error rate of the recognizer is 18.1% for system with a lexicon including tone information on the 1997 ARPA Hub4 test data.

REFERENCES

[1] The Mandarin Chinese News Corpus, the 1997 Mandarin Broadcast News Speech and Transcripts corpus, and the Mandarin lexcion, distributed by LDC.

[2] G. Adda, M. Adda-Decker, J.L. Gauvain, L. Lamel, "Text normalization and speech recognition in French," *Proc. Eurospeech'97*, pp. 2711–2714, Rhodes, Greece.

[3] X. Guo, W. Zhu, Q.Shi, "The IBM LVCSR System Used for 1998 Mandarin Broadcast News Transcription Evaluation"*Proc. DARPA Broadcast News Workshop*, pp. 179–182, Herndon, VA, Feb. 1999.

[4] J.L. Gauvain, L. Lamel, "Fast Decoding for Indexation of Broadcast Data," *Proc. ICSLP-2000*.

[5] J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and Transcription of Broadcast News Data," *Proc. ICSLP'98*, **5**, pp. 1335–1338, Sydney, Australia, Dec. 1998.

[6] J.L. Gauvain, L. Lamel, G. Adda, M. Jardino, "The LIMSI 1998 HUB-4E Transcription System," *Proc. DARPA Broadcast News Workshop*, Herndon, VA, pp. 99–104, Feb. 1999.

[7] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *Proc. DARPA Speech Recognition Workshop*, pp. 56–63, Feb 1997.

[8] J.L. Gauvain, L. Lamel, G. Adda, "Recent Advances in Transcribing Television and Radio Broadcasts," *Proc. Eurospeech'99*, pp. 1463–1466, Budapest, Sep. 1999.

[9] C.J. Legetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, Vol. 9, pp. 171–185, 1995.

[10] S.C. Lin, C.L. Tsai, L.F. Chien, K.J. Chen, L.S. Lee, "Chinese Language Model Adaptation Based on Document Classification and Multiple Domain-Specific Language Models", *Proc. Eurospeech'97*, pp. 1463–1466, Rhodes, Greece, Sep. 1997.

[11] D.S. Pallett, J.G. Fiscus, A. Martin, M.A. Przybocki, "1997 Broadcast News Benchmark Test Results: English and Non-English," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 5–12, Landsdowne, VA, Feb. 1998.

[12] D.S. Pallett, J.G. Fiscus, J. Garofolo, A. Martin, M.A. Przybocki, "1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures," *Proc. DARPA Broadcast News Transcription Workshop*, pp. 5–12, Herndon, VA, Feb. 1999.

[13] P. Zhan, S. Wegmann, L. Gillick, "Dragon systems' 1998 broadcast news transcription system for Mandarin" *Proc. DARPA Broadcast News Workshop*, pp. 183–186, Herndon, VA, Feb. 1999.