COLLECTION AND EVALUATION OF BROADCAST NEWS DATA FOR ARABIC

Mohamed Afify[1] *Ossama Emam*[2]

[1]Department of Information Technology
Faculty of Information and Computer, Cairo University
[2] Human Language Technologies Group , IBM Egypt

ABSTRACT

This paper focuses on presenting a general methodology for acquiring and automatically segmenting broadcast news data from the web. It was shown that it is possible starting from a relatively small corpus of about 10 hours to segment automatically about 30 hours of data. This step is important because manual segmentation of broadcast news data is generally very tedious and time consuming. In addition to the data collection proposal we show the development of an initial recognition system. We present an automatic procedure for creating vowelizations for Arabic words. This is again important because most available Arabic transcriptions lack vowelization, which is crucial for creating phonetic transcription. The performance of our system is initially 36% error rate.

1. INTRODUCTION

Transcription of broadcast news is currently attracting a lot of attention in the speech research community. This interest is fueled by the importance of the transcription itself, for example, for creating captions, and also as a front end for more sophisticated text processing algorithms as topic classification or named entity identification. It is known that building a successful speech recogniton system requires the availability of appropriate corpra for building acoustic and language models. This paper addresses the collection of such corpra for Arabic Language and their evaluation in a broadcast news transcription system. This data collection effort is important because inspite of the recent interest in Arabic broadcast news transcription e.g.[1, 2] the amount of available data is still sparse compared to other languages as English. While broadcast news data(both audio and transcription) can be found on the web the available audio files are very long[1] and this causes problems in subsequent processing. It is usually preferred to chunk the corresponding files into mangeable segments according to some criterion. In addition, data chunking is also useful in identifying speaker turns and acoustic quality of the recordings for performing more sophisticated processing as speaker or environment adaptive training. In broadcast news data collection this step is usually performed manually and the data is equipped with additional files indicating the speaker turns and quality of recordings. This can be found for example with broadcast news data available from Linguistic Data Consortium for English. The latter manual segmentation stage is very time consuming and requires extensive human effort. For this reason we performed manual segmentation for only a subset of the corpus (about 10 hours), and developed an automatic segmentation procedure for the rest of the data(about 30 hours)[2]. The goal is to break the data into mangeable pieces for subsequent model building steps, and this automatic segmentation does not guarantee segmenting the data into meaningful speaker turns for example. However, the main motivation is to circumvent the costly manual segmentation procedure, and explore whether this automatic process would lead to satisfactory performance. Thus, the work done in this paper may be better viewed as a methodology to quickly set up broadcast news transcription systems from unsegmented data.

In order to perform automatic segmentation we need to bootstrap an initial system as will be discussed in the paper. This was done using the manually segmented data and starting from a telephony based initial model[3]. In the light of the previous discussion we can summarize the data collection process as follows. First the audio data and corresponding text transcriptions are downloaded and initially processed, then 10 hours are manually segmented according to speaker turns and a system is bootstrapped from telephony models. This system is tested for verification using test data, and then used to

[1]Typically these correspond to one hour shows.
[2]Although we have a total of 66 hours we processed only 40 hours up to this writing.
[3]The downloaded data has 8K sampling rate and this is the reason for using a telephony model.

automatically segment the rest of the data, as will be discussed in the paper. Finally a system is built using the whole data and its performance is evaluated.

The rest of the paper is organized as follows. Section 2 describes data collection and preparation. Bootstrapping an initial system, and automatic segmentation of the speech corpus are given in Sections 3, and 4 respectively. Finally we conclude in Section 5.

2. DATA COLLECTION AND PREPARATION

The web currently contains a lot of broadcast news data and their transcription. This is a valuable source of data which we decided to use in constructing our data base. We focused on coversational programmes as they are more challenging from speech recogniton point of view and they substantially deviate from the read corpra that we already have. The data consists of 66 programmes, each of about one hour length, together with their transcriptions. Thus we have about 66 hours of audio data, and about 300K words of text data. No streaming was done while downloading the speech, and thus no packet loss was experienced. The acquired speech data is compressed using a standard windows format which is decompressed using the windows media player. The used compression is lossy. On one hand this may lead to performance degradation, and on the other hand this is interesting to study the use of compressed data for speech recognition or to complement other larger data bases. The decompressed audio resulted in 8K sampled data, the reason may be that many audio compression techniques mainly distort high frequencies and it will not be very useful to keep wide band speech after decompression.

The collected audio data consists of 66 files, each of about one hour length. These long files can not be directly processed by our training tools. Thus, we called for manual segmentation of a susbet of 10 files to bootstrap a reasonable system and then perform the segmentation automatically. The manual segmentation is done based on speaker turns, where the beginning and ending of each speaker's speech are marked together with their transcription, in addition to identifying music segments and other types of noises. The output of the segmentation is saved in text files which are then automatically processed to create the segmented speech data. To highlight the difficulty of this manual step, the segmentation of these 10 hours took more than one month by a human transcriber. This step resulted in about 10 hours of speech together with their transcription for system bootstrapping. In addition we isolated and segmented about 15 minutes development set consisting of about 3K words.

3. SYSTEM BOOTSTRAP

In order to setup an initial speech recognition system we need to identify a language model vocabulary, and a lexicon that contains the phonetic transcription of each word, and to build acoustic and language models. These steps will be discussed below.

3.1. Building language model

The number of unique words in the text data was found to be about 50K words. These are taken as the language model vocabulary. Arabic is a highly inflectional language where many prefixes and suffixes are applied to modify stems. Thus chunking words into prefixes, stems, and suffixes improves the coverage but increases the perplexity. In this work no word segmentation was tried, and other alternative segmentation schemes remain as a subject of future work. After identifying the language model vocabulary, we build a standard trigram language model on the transcriptions using IBM language model building tools, and only the 300K word text transcription. This led to a perplexity of about 500. We also attempt to augment the data with other text corpra that we already have but no improvement in perplexity or recognition rate was observed. This may be due to the large difference between these corpra and the broadcast news data that we have. Again this point needs further exploration.

3.2. Building the lexicon

We identified the same 50K word of the language model vocabulary as the acoustic vocabulary, and also used a set of 38 phonemes as the basic phonetic set. Phonetic transcription is challenging for Arabic because it requires short vowels which are generally not present in standard Arabic writing including the transcriptions we have. To construct these short vowels we used a combination of lookup in our existing lexicon together with some stemming techniques, where words are segmented into prefixes and stems and the vowelized stems are looked up in a dictionary, then the voweliztion is concatenated back into the vowelized word. When this procedure failed we resorted to a statistical tool. This tool is based on searching the maximum likelihood sequence of vowels as follows

$$V^* = \text{argmax}\, P(V|L)P(V) \quad (1)$$

where V is the vowel sequence, and L the letter sequence. The vowel probability is calculated from a standard trigram on the vowels, and the vowel given letter probabilites are limited to the two left and right letters. The search is implemented using an A^* algorithm which gives the N-best vowel sequence candidates. The probabilites are trained from the vowelized stem lexicon that we already have. Detailed derivation, and systematic

Model	Error
Telephony	60.7
Broadcast-10 hour	36.4
Broadcast-22 hour	36.1

Table 1: Error rate (in %) on development set using both telephony and broadcast acoustic models.

evaluation of this statistical vowelizer will be presented in future publications, and is documented in [4].

After finding the voweliztion using either stemming and lookup, or the outlined statistical method the whole lexicon was checked manually to correct gross errors. Once the vowelization is obtained phonetic transcription becomes straight forward and almost a one-to-one procedure which is performed using our phonetic transcription tool. It is interesting to note here that recent efforts in Arabic speech recognition, e.g. [1], use the letters as the basic phonetic set, and hence eliminate the need for vowelization, and greatly facilitate system construction. The systematic comparison between these approaches remain an interesting topic for future work.

3.3. Building acoustic model and system evaluation

The acoustic data is obtained using a standard front end which uses 13 MFCC coefficients (including the C_0), and their first and second order derivatives leading to a 39 dimension feature space. Also cepstral mean normalization is applied on the utterance level. As the data is 8k sampled we start from a telephone speech model that we already have for the Arabic. The model is built using standard decision tree clustering with context questions and has about 3K leaves and 50K prototypes. The resulting speech recognition system uses ranks [3] and is built using tools provided by IBM Human Language Technologies. The word error rate of this telephony acoustic model on the development set is about 60%, which is very high for practical applications.

Using the initial model and the manually segmented 10 hour speech data and their transcription we want to create a new acoustic model for broadcast news. We tried different adaptation strategies as tree based maximum likelihood linear regression(MLLR) and maximum a posteriori (MAP) based adaptation, and also retraining the acoustic model. IBM adaptation and acoustic model building tools were used in this step. After some initial experiments we found that retraining the models was the best in our setting and reduced the error rate to about 36%. The resulting model has fewer leaves, about 1.5K, and fewer prototypes, about 15K, than the initial model due to the limited amount of training data. Table 1 summarizes the results of this section, where the last line corresponds to the model trained after adding automatically segmented data as will be discussed in Section 4.

4. AUTOMATIC SEGMENTATION AND MODEL REFINEMENT

Manual segmentation of broadcast data is a time consuming and tedious process. For this reason we decided to resort to automatic segmentation of the rest of the acoustic data that we have. This automatic segmentation process, which assumes the existence of a reasonable initial model, will be described below.

Automatic segmenation is based on a Viterbi alignment program which is specially designed for segmentation of long speech files in the IBM Human Language Technologies group. It takes as input the correct transcription and a decoded script with ending times[4]. The basic idea is a combination of text alignment, that is used in scoring programs for speech recognition, and classical Viterbi decoding where the times of the correctly decoded words are used to anchor Viterbi alignments of segments of the file. The main flow of the alignment can be summarized as follows

1. Decode the input file while enabling the output of the end times of the recognized words.

2. Enter the output of the decoder and the correct transcription together with the acoustic data into the segmentation program. The output is the segmentation of the speech data together with their transcription[5].

After performing the above segmentation step for all available speech files, the resulting sub-files and their transcriptions are added to the training data and used to construct a new acoustic model. The above steps can be iterated, if desired, using the new resulting model. Up to this writing we have performed the above segmentation procedure for about 30 hours of speech corrsponding to 30 different broadcast shows. It is important to point out that the accuracy of the

[4]The decoder can optionally output the decoding times for a recognized script.

[5]Notice that this segmentation is completely data driven and is not related by any means to speaker turns for example.

decoding is crucial for the proper operation of the segmentation process. When we have poor decoding results the segmentation process will tend to produce very long chunks, because it anchors on correctly decoded words, which will be dropped in further processing by the training tools. In our experiments the used 30 hours resulted, after excluding very short and very long segments, in a total of about 12 hours of speech. This leads to a total of about 22 hours used in training the final model. Of course reiterating with the hopefully better models may lead to keeping more data at each iteration, and hence improve performance.

An important point about the decoding is also worth mentioning here. It possible to use a general language model and the full vocabulary in the decoding step above. We tried this in initial experiments but this leads to relatively poor recognition score and very long decoding times which makes the segmentation process very slow and leads to failing of many segments as discussed above. As a solution to this problem we construct a mini voacbulary and language model for each programme consisting of the words found only in the programme and trained only on the programme transcription. This leads to faster decoding and better recognition accuracy where perplexities of about 10 are typically obtained. On the other hand this may be less robust, because inspite of decoding on the same show a poor acoustic match of the correct test may make the decoding fail. This was observed in some of our experiments. Thus, creating simple language models which are also robust remains as an issue for further investigation.

5. SUMMARY AND CONCLUSION

In this paper we consider both data collection, and initial system setup for Arabic broadcast news recognition. For data collection we use the web which is a rich source of broadcast data and their trancriptions, and identify the problem of long data files which are usually prohibitive for most existing training tools. To address this problem we propose to manually segment a small corpus, about 10 hours for the current work, and develop an iterative segmentation procedure which automatically cuts the data into smaller chunks. While broadcast news segmentation is mainly based on meaningful cues as speaker turns and quality of recordings the proposed algorithm is completely data driven and does not relate to any subjective measure. Further investigation of adding useful knowledge to this segmentation procedure may help in improving its performance. When the automatically segmented data was added to the original manually segmented data no appreciable improvement in recognition performance was observed. This can be attributed to the lack of knowledge in the segmentation process, and in part to building show based language models which are not robust especially for poor quality recordings. We plan to investigate these issues in future work and also maintain a manual segmentation effort which proved to lead to promising results. For example 10 hours of manually segmented data lead to huge improvement over a baseline telephony model.

For setting up an initial speech recognition system we focused on constructing vowelization of the usually unvowelized Arabic words. The vowelization uses a mix of table lookup together with some stemming techniques and a statistically based method. Creating vowelization is crucial for generating phonetic transcription in Arabic. The used methodology contrasts, and should be compared to, recent efforts in Arabic speech recognition[1] where the letter sequence is used as a phonetic transcription. In addition to the vowelization issue we considered word segmentation for building the language model vocabulary. In this work we did not consider any word segmentation. However, this issue should be further revisited to balance the perplexity-coverage tradeoff. Apart from the above considerations the developed system follows standard conventions in speech recognition.

The best performance obtained is 36% error rate. While this is reasonable as a start we plan to improve on this issue by adding more language model data, we build the langauge model from only 300K words, and also add manually segmented acoustic data, whenever available, to building the acoustic model. This is expected to greatly enhance the performance, and we can revisit the automatic segmentaion process with the improved acoustic models to fully automate the system development.

6. REFERENCES

[1] J. Billa et al., "Audio indexing of Arabic broadcast news," in Proceedings of ICASSP 2002, Orlando, Florida.

[2] K. Kirchoff et al., "Novel approaches to Arabic speech recognition-final report from the JHU summer workshop," Tech. report, John-Hopkins University, 2002.

[3] L.R. Bahl, P.V. deSouza, P.S. Gopalakrishnan, D. Nahamoo, and M. Picheny, "Robust methods for context dependent features and models in a continuous speech recognizer," in Proc. ICASSP'94, Adaleide, Australia, April 1994.

[4] M. Afify, "Automatic vowelization of Arabic text," unpublished research report, IBM Cairo Technology Development Center, December 2002.