LANGUAGE MODELING STRUCTURES IN AUDIO TRANSCRIPTION FOR RETRIEVAL OF HISTORICAL SPEECHES

Mikko Kurimo[1,3], *Bowen Zhou*[2,4], *Rongqing Huang*[2,5], *John H.L. Hansen*[2,5]

[1] Neural Networks Research Centre, Helsinki University of Technology, Finland Mikko.Kurimo@hut.fi
[2] Robust Speech Processing Group, Center for Spoken Language Research,
University of Colorado at Boulder, USA {huangr,jhlh}@cslr.olorado.edu

ABSTRACT

In this paper we apply speech recognition for automatic transcript generation for spoken document retrieval. The transcripts are used to compute an index for an archive of historical speeches and to provide the index, speech, and transcripts available for query based retrieval and browsing. In addition to acoustic variability, the task is challenging, because it covers a broad spectrum of different speaking styles and use of language. Language modeling is important for speech recognition to determine the prior probabilities of the compared word and sentence candidates in decoding. Various large text corpora are available in electronic format for language model training, but the open question is what and how should we include to improve the audio transcripts of this task. In this work we compare large overall language models to focused ones trained on selected subsets of the data, and to combinations between both. With respect to the potential index terms, improvements were obtained for transcripts that did not fit well to the scope of the large overall language model.

1. INTRODUCTION

In recent years, there has been rapid growth and interest in converting traditional library holdings and other large archives of text and audio into digital libraries. In practice, this means that books, audio and video tapes are being digitized and their content extracted by OCR (optical character recognition) and ASR (automatic speech recognition) systems. The extracted text-like transcripts are further transformed into browsable and searchable formats that can be accessed by large audiences all over the world via internet and mobile devices. However, out of all text, image, audio, and video sources only the information from clearly typewritten text sources without images and any special formulas can still be extracted almost error-free. For video and non-speech audio sources it is not even straight-forward to exactly specify what should be transcribed. For the spoken audio recordings studied in this paper, it is clear that the transcriptions should mainly decode the speech into text.

In order to improve the accessibility of spoken audio, it is important to develop methods to enhance the correctness of the obtained transcripts. This not only helps to manually browse and spot interesting parts from data archives, but in general, to improve the possibilities to automatically index the recordings and retrieve answers to the posed queries. The special target in this work is to enhance the models for the language used in the audio material at hand.

The motivation of language modeling is to improve speech recognition accuracy by recovering the contextually more suitable words from acoustically confusable utterances. In addition to introducing correct priors to words and word sequences in the audio, an important task is also to specify the correct search vocabulary, because often the rare content words, such as proper names, are very important to characterize the retrieved information. Unfortunately, those words are often misrecognized, because they may be too infrequent to have proper co-occurrence statistics with other words.

Using the automatic speech transcripts in audio indexing has recently become an important application of ASR, see e.g. [1, 2]. In addition to motivating the research it has also led to new frameworks of performance evaluation such as the TREC spoken document retrieval track [3] and public demonstrations of the state-of-art via the internet [1].

The specific topic of this paper is to build and evaluate differently structured large vocabulary language models (LMs) with respect to the training data and transcription task at hand. The data is broadcast audio recordings spanning over seven decades. Some of the time blocks and topics can be adequately covered by our training material, but most of them poorly or not at all. The relevant modeling issues for this work are how to smooth the existing LMs [4], how to adapt them with new data [5], and how to extend the LM and its vocabulary [8, 6, 7] We report experiments for utilizing the relatively small amounts of historical language data retrieved, e.g. by scanning books with OCR, to adapt large LMs trained mostly with news data from the 1990's that is much easier to obtain in electronic format. Our goal was to evaluate the potential improvement in the audio transcriptions obtained by enhancing the modern language models with old language data.

The framework of this paper is the National Gallery of Spoken Word (NGSW) project [2] which aims at transcribing the historically significant recordings of the 20th century into a searchable and browsable format to be accessed by the SpeechFind web browser [9]. The project involves a large variety of methodological and research challenges, such as the different and evolved recording conditions during the last century, a number of very different speakers and speaking styles, and various disturbing factors ranging from overlapping speakers to music and noise. From the point of this paper the project is especially interesting, because of the variation of the vocabulary and language based on the evolvement

[3] Thanks to Nokia Foundation for a visiting researcher scholarship.

[4] Bowen Zhou was with RSPG-CSLR, Univ. of Colorado. He has since joined IBM T.J. Watson Research Center, NY, USA zhou@us.ibm.com

[5] RSPG-CSLR was supported by NSF DLI-2 Cooperative Agreement No. IIS-9817485

[1] http://speechbot.research.compaq.com http://SpeechFind.colorado.edu

of styles and topics throughout the century.

2. METHODS

2.1 Spoken document retrieval system

To understand the framework and motivations of the language modeling experiments presented in this paper, it is important to keep in mind the whole process of spoken document retrieval, which is briefly reviewed in this section.

The source audio tracks usually have rather heterogeneous content, so in addition to transcription it should be split into homogeneous segments, e.g., speech separated from music and other audio that will not be transcribed. Ideally, the speech should also be segmented into parts, where the conditions, speakers, and topics remain constant, in order to both help the speech recognizer to adapt efficiently and get the speech indexed into semantically coherent segments. The features describing the signal are obtained from the spectrograms by different transformations and normalizations in order to carefully remove noise and irrelevant information. The SpeechFind system relies here on the conventional MFCC features and their derivatives which are used in the audio segmentation with an iterative Bayesian information criterion based on T2-statistics [9, 10].

The actual speech recognition operates by generating probabilities of phonemes and words for the observed sequence of speech frames and then searching for the most probable sequence of words that could have generated the observations. The central piece of the recognizer is a decoder which takes the probabilities given by the acoustic and language models, and the word-to-phoneme rules from pronunciation dictionary and finds the output word sequence or network. The acoustic models of SpeechFind are conventional mixture Gaussian density hidden Markov models and default language models are smoothed back-off word trigrams. The Sphinx3 system [2] has been used as a decoder in the experiments reported in this paper, but it is currently being replaced by CSLR's own Sonic system [11].

The word sequence output of the speech recognizer forms a raw "dirty" transcript of the analyzed audio segment. Using the so-called bag-of-words content representation we can transform the transcripts into inverted file index for retrieval. This basic index is further enhanced by filtering the content words and word stems that carry most of the content information and adding some new index words by document expansion on relevant text material. Similarly, we can perform stemming, stopping, and expansion to each query for improved retrieval precision and recall as in [9]. For the speech indexing and retrieval, it is most important to correctly recognize the content words. The readability of the transcripts is helpful for browsing the audio, but it is unlikely that an audio segment with incorrect content words will ever be retrieved.

2.2 Statistical language modeling

Statistical language models play a key role in a spoken document retrieval system. In general, for optimal retrieval performance the index term representation of the audio content should be based on the words into which both the queries and the speech data can be easily mapped. For example, the proper names are often the main content of the queries, but they are very difficult to recognize from speech, because their variable pronunciations and few occurrences in any training data makes the exact modeling impossible. In practice, the storage size and trainability of statistical language models and the realizability of the search task in the decoder, usually limits the applicable vocabulary for the language models to, say, 65,000 words, which makes the out-of-vocabulary (OOV) words very difficult to recognize.

For very large vocabulary tasks most of the rule-based language models are impractical and higher-order n-gram models must rely heavily on smoothing. Because of the diversity of the audio material in this task, the models will frequently be applied in different domains that require quick adaptation. Other important language modeling techniques include the training and interpolation of separate language models specialized in certain important time blocks and discussion topics or contexts. One promising language modeling framework for this is the automatically focusing language model [12].

A special topic that has recently gained remarkable attention in the NGSW project is how to use the large text archives that can be accessed for optimal language modeling performance. Even if we in principle could just train a single language model with all the existing text material in libraries and the internet, this would not make sense or even be practically possible. In the current project we have relied on the expertise of librarians to preselect books and other digital resources that represent relevant text styles and vocabulary for the general 20th century speeches and Chicago Roundtable discussions of the 1940's. In this paper we evaluate a language modeling framework where one huge monolithic n-gram language model is trained using as much text data as possible and then combined with smaller specialized n-gram models to obtain best matches to the material.

3. EXPERIMENTS

To evaluate the suitability of the proposed language models, we conducted experiments on recognition accuracy of speech samples and language modeling accuracy of relevant text samples. As language models we trained one for broadcast news (BN) using HUB4 broadcast news transcriptions, one using the same augmented with North American news texts (News), one for old texts (Old) using Gutenburg archives 1900-1920 texts and Chicago Roundtable 1940s texts, and finally, one including almost all the above (All). All the models were standard back-off trigrams with interpolated Kneser-Ney smoothing [13, 14]. A 65,347-word vocabulary was selected for the LMs mainly based on the most common words in the BN and News corpora, but with some additions from the smaller old corpora, too. The baseline LM was the one used previously [9]. It is an optimized BN back-off trigram model which corresponds probably closest to our new BN model, although the training tools and data have had some changes.

The idea in the following experiments was to evaluate how much a small, but focused text material actually helps, if we have an overall language model trained with a substantial amount, but not very well matching data. Another topic was to evaluate, if it is better just to merge the new observed n-grams to the top of the old ones, or to use them to train a more specialized language model and interpolate between the word probabilities given by the two models, or to perform

[2] See http://www.speech.cs.cmu.edu/sphinx/index.html, for more information about Sphinx speech recognition system by CMU

both.

3.1 Evaluation by speech recognition

To evaluate the speech recognition error rates resulting from the development of language models, we applied the same "6 decades" (1950–2000) data set as in [9], the same segmentation, preprocessing, features, acoustic models, speaker adaptation, and pronunciation dictionary. The data contains 3.8 hours of audio samples from the past 6 decades with significant variation of recording technology, conditions, and noise (see Table 3). One hour of additional old speech data from 1940s was included to monitor the effect of the new LMs on topically and stylistically matching conditions. Because the acoustic models were originally trained from broadcast news speech data, it is clear that even with adaptation the match of language, speaking style, and other characteristics can not be perfect.

We performed the speech decoding experiments with the Sphinx3 recognizer using one language model at a time to obtain the different transcripts. The interpolated language models were prepared by computing a new language model out of the two components by equal interpolation weights. The performance can be optimized by tuning the interpolation weights based on a relevant development or adaptation data. However, if suitable tuning data does not exist, there is no time for separate optimization for each topic and style, or the decoder is not equipped for interpolation, this pre-interpolation with the default weights is all one can use. If there is a possibility to adapt the weights online, there is usually a chance to focus on specific LMs as in [12], as well.

In addition to the conventional word error rate (WER) we evaluated as well the term error rate (TER). TER is often used to measure the quality of speech transcripts for speech retrieval applications [1]. Term errors are computed in each audio segment and the total is counted for each time block in Table 3. TER is defined as the difference of two word histograms (the recognition result H and the correct transcription R) after stemming (word suffixes excluded) and stopping (common function words excluded) both word sequences (summation t is over all resulting terms):

$$\mathrm{TER} = \sum_t |R(t) - H(t)| / \sum_t R(t) * 100\% \; . \qquad (1)$$

3.2 Evaluation by language modeling accuracy

For less complicated, but more approximative, evaluation of the language model performance, we measured the modeling accuracy by computing the average of the inverse of next word prediction probability, a.k.a. perplexity, on some left-out text data. While the perplexity, or its negative logarithm a.k.a. entropy, measures well the accuracy of the LM, it lacks the indication of whether the accuracy improvements concern the discrimination of the acoustically close rival hypothesis or something else that matters less for the recognition task.

The perplexity evaluations were done by the SRILM toolkit [14] that was used to train the language models, as well. From the left-out texts that are close to certain LMs we sampled a transcription from 1940s (CTT) and several from 1990s (BN). In addition to these we took the transcriptions of U.S. Presidential inauguration speeches spanning over one hundred years (Pres.) and the reference transcripts from our evaluation audio data (6 decades).

4. RESULTS AND DISCUSSIONS

LM	training #words	Average perplexity			
		CTT	Pres.	BN	6 decades
eval. #words		7.9K	132K	2.1M	30K
Baseline		486	509	*201*	*240*
OOV%		2.1	0.8	1.7	1.3
1.BN	168M	410	444	204	271
2.News	724M	421	452	208	289
3.All	730M	410	420	207	286
4.Old	5.7M	*275*	*339*	774	475
OOV%		1.6	0.8	1.6	1.3
3+4 ip		*230*	*280*	215	243
1+3 ip		384	399	*193*	263

Table 1: The evaluation of the four language models and their interpolations (ip) by four sets of texts. "6 decades" averages over the speech reference transcripts from 1950–2000.

Decade	ref. #words	Baseline LM		3+4 ip LM	
		OOV	perpl.	OOV	perpl.
1940	2068	0.5	258	0.6	177
1950	6241	1.5	325	1.5	280
1960	2142	2.2	384	2.5	343
1970	4434	0.8	132	0.8	151
1980	3330	0.9	177	0.8	194
1990	5951	1.7	280	1.8	285
2000	7530	0.9	237	0.9	285

Table 2: More detailed description and evaluation of the speech reference transcripts from different decades.

Decade	Audio mins	SNR dB	Baseline LM		3+4 ip LM	
			WER	TER	WER	TER
1940	14	10	73	59	68	55
1950	52	34	39	42	36	33
1960	17	20	37	33	37	29
1970	35	21	26	34	26	24
1980	27	21	60	42	60	35
1990	47	14	48	82	48	78
2000	50	28	59	75	60	75

Table 3: Description and evaluation of the audio transcription tests. Word (WER%) and term (TER%) error rates are shown for the two LMs using the same acoustic models.

From Table 1 we see that for most evaluations the lowest perplexity is obtained by interpolating the all-material LM and the most specific LM (Old or BN, respectively). However, the interpolation between the large News model instead of the all-material would give almost equal results.

Comparisons of the obtained LM accuracy improvements between the different evaluation sets (Table 1) seem to suggest the following: 1. For such a well-modeled data set as BN there is not much further improvements obtainable by adding more either related training data (News) or unrelated data (Old). 2. For a data set, such as Presidents, that is mostly quite different from the main training data (topic, style, and

age), any additional interpolated small LM can improve performance quite much.. 3. If the evaluation data is related to a small but well-matching LM, such as CTT, the interpolation is especially useful.

The 65,000 word LM vocabulary was the same for all the models (the old baseline had slightly different vocabulary, though). However, since it was selected mostly based on BN and News data, further improvement would be likely if it could be adapted to the vocabulary of the evaluation corpus to ensure the inclusion of the right n-grams. If the decoder were allowed to increase the vocabulary size significantly, this could improve the modeling of the rare words, but then more confusions would be expected, as well, because of the larger lexicon. In these experiments we applied rather conventional trigrams with interpolated Kneser-Ney smoothing. It has been suggested [4] that, especially, when there is not enough training data available, the further language model developments, such as higher-order n-grams, n-gram caching, clustering, semantic models etc., can still improve the performance significantly.

As can be seen from the Table 3, the speech recognition WER does not seem to be much affected by any of the LM improvements. However, for the speech transcriptions that we aim to produce and for speech retrieval, the recovery of the rare content words that are likely to act as good index terms is far more important than the overall WER which is most affected by the statistically much more frequent common function words. That is why measures, such as TER, are more interesting, because the focus is in the potential index terms. Tables 2 and 3 show that the large overall LM interpolated with the small old text LM improves the perplexity and TER for the older time blocks, whereas for the more recent time blocks the modern BN model is better. This seems to suggest that the focusing has an important effect for the speech transcripts as well as for the LM accuracy. The absolute differences between the decades are mainly not due to the language differences, but to the differences in the speaking conditions as explained in [9].

5. CONCLUSIONS

In this work we evaluated ways to use new and existing language modeling resources for the transcription of historical speeches. The experiments indicate that a Promising results were obtained by combining the large overall LM with LMs that are with much less data but more focused to the topics and styles of the specific decade. We tested the models by measuring the average perplexity on independent text data and the average word and term error rate on independent speech samples. The improvement on the whole system obtained by improving just the language models is visible but not very significant. It is noteworthy, however, that the focused language training data obtained for the transcription task is rather small and matches well only to a portion of the evaluation data. In future, when more relevant training text and speech for the historical speeches are produced, more thorough evaluations can be performed to specific topics and time blocks.

6. ACKNOWLEDGEMENTS

We wish to thank Michael Seadle, Rick Peiffer, and the entire Vincent Voice Library at Michigan State Univ. for their assistance in providing text and audio material for evaluations in this paper. The authors would also like to thank Andreas Stolcke at SRI and Rita Singh at CMU for help in different language model formats.

REFERENCES

[1] S.E. Johnson, P. Jourlin, G.L. Moore, K. Sparck Jones, and P.C. Woodland, "The Cambridge university spoken document retrieval system," in ICASSP 1999, pp. 49–52.

[2] J.H.L. Hansen, J. Deller, and M. Seadle, "Engineering challenges in the creation of a national gallery of the spoken word: Transcript-free search of audio archieves," in *IEEE and ACM JCDL-2001: Joint Conference on Digital Libraries*, 2001.

[3] J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," 2000.

[4] J. T. Goodman, "A bit of progress in language modeling," *Computer Speech and Language*, vol. 15, no. 4, pp. 403–434, 2001.

[5] J. R. Bellegarda, "An overview of statistical language model adaptation," in *ISCA ITRW workshop on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, August 2001, pp. 165–174.

[6] P. Geutner, M. Finke, and P. Scheytt, "Adaptive vocabularies for transcribing multilingual broadcast news," in ICASSP 1998.

[7] V. Siivola, M. Kurimo, and K. Lagus, "Large vocabulary statistical language modeling for continuous speech recognition," in EUROSPEECH 2001, pp. 737–747.

[8] N. Bertoldi and M. Federico, "Lexicon adaptation for broadcast news transcription," in *ISCA ITRW workshop on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, August 2001, pp. 187–190.

[9] B. Zhou and J.H.L. Hansen, "Speechfind: An experimental on-line spoken document retrieval system for historical audio archives," in ICSLP 2002.

[10] B. Zhou and J.H.L. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian information criterion," in ICSLP 2000, pp. 714–717.

[11] B. Pellom and K. Hacioglu, "Sonic: The university of Colorado continuous speech recognizer," Technical report TR-CSLR-2001-01, University of Colorado, Boulder, Colorado, March 2001.

[12] M. Kurimo and K. Lagus, "An efficiently focusing large vocabulary language model," in ICANN 2002, pp. 1068–1073.

[13] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modeling," *Computer Speech and Language*, vol. 8, pp. 1–38, 1994.

[14] A. Stolcke, "SRILM - an extensible language modeling toolkit," in ICSLP 2002.