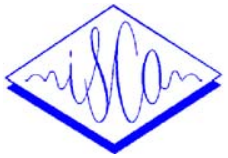

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

LANGUAGE MODELING FOR BROADCAST NEWS TRANSCRIPTION

Gilles Adda, Michèle Jardino, Jean-Luc Gauvain

Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
{gadda,jardino,gauvain}@limsi.fr

10.21437/Eurospeech.1999-351

ABSTRACT

This paper addresses the problem of language modeling for the transcription of broadcast news data. Different approaches for language model training were explored and tested in the context of a complete transcription system. Language model efficiency was investigated for the following aspects: mixing of different training material (sources and epoch); approach for mixing (interpolation vs count merging); and using class-based language models. The experimental results indicate that judicious selection of the training source and epoch is important, and that given sufficient broadcast new transcriptions, newspaper and newswire texts are not necessary. Results are given in terms of perplexity and word error rates. The combined improvements in text selection, interpolation, 4-gram and class-based LMs led to a 20% reduction in the perplexity of the LM of the final pass (3-gram class interpolated with a word 4-gram) compared with the 3-gram LM used in the the LIMSI Nov'97 BN system.

INTRODUCTION

Language modeling is a major component of any speech recognition system, aiming to capture and exploit regularities in the language. It is well known that language models provide an important contribution to the performance of speech recognizers. Developing accurate language models requires the availability of large corpora of spoken language transcriptions and related texts, tools for processing and normalizing the texts, and methods for optimally combining data from various sources and evaluating the accuracy of the resulting models. Evaluating a language model (LM) independently from a speech recognizer is difficult. The most common measure used to compare LMs is the perplexity. However, as speech recognition has addressed more heteregoneous and difficult tasks (such as the transcription of television and radio broadcasts, or conversational speech), discrepancies have been observed between LM perplexity for different language models and the resulting word error rates of the transcription system. Perplexity can be used as a first indicator of LM accuracy to avoid the CPU and memory costs of a complete recognition experiment. Once an LM appears to be interesting, our experience is that a recognition run must be carried out to access its quality. The results given in this paper provide both perplexity and word error rates for the different LMs.

In this paper we address the problem of language modeling for the transcription of broadcast news data. Different approaches for training the language model were explored and tested in the context of a complete transcription system. The following issues were addressed: the selection of training materials (source and epoch); text normalization (modeling of breath and filler words, compound words, acronyms) and wordlist selection; approaches for mixing the materials (interpolation vs count merging; and class-based language models.

THE LIMSI NOV'98 BN SYSTEM

The LIMSI Nov'98 ARPA Broadcast News transcription system [1] is an extension of our Nov'97 Hub4E system [5], using maximum likelihood partitioning and a 3-step decoding approach with acoustic model adaptation[6]. This system obtained an overall word error rate of 13.6% in the Nov'98 ARPA evaluation test. The speaker-independent large vocabulary, continuous speech recognizer makes use of n-gram statistics for language modeling and continuous density HMMs with Gaussian mixtures for acoustic modeling. Since radio and television broadcasts are comprised of segments with different linguistic and acoustic natures, the continuous stream of data is partitioned into homogeneous acoustic segments prior to word recognition. Data partitioning divides the acoustic signal into homogeneous segments, rejecting non-speech portions and associating cluster, gender and bandwidth labels with each speech segment. The partitioned data is then used for unsupervised adaptation. The acoustic models were trained on 150 hours of transcribed broadcast news data. Each context-dependent (word-independent but position-dependent) triphone model is a tied-state left-to-right CDHMM, where the tied states were obtained by means of a decision tree. Gender and bandwidth specific sets of acoustic models with 11500 tied states were used. Language models were obtained by interpolation of n-gram backoff LMs trained on different data sets: transcriptions of the acoustic training data (1.5M words), broadcast news transcriptions (200M words), newspaper and newswire texts (340M words). Word decoding is carried out in three passes for each speech segment: initial hypothesis generation using the small set of acoustic models, word graph generation, and final hypothesis generation. The initial hypothesis are used for cluster-based acoustic model adaptation using the MLLR technique. The final hypothesis is generated using a 4-gram interpolated with a category trigram model with 270 automatically generated word classes.

TRAINING MATERIALS, NORMALIZATION AND WORDLIST SELECTION

For transcription of American English Broadcast News shows, very large text databases are available for constructing language models. In this work three sources of data were used:

- NEWS: Over 700M words of news texts from various sources (newspapers and newswires from 1994 to 1998). These data, available through the LDC, consist of texts from the Los Angeles Times, New York Times, Wall Street Journal, Washington Post, Reuters News Service, and Associated Press WordStream.
- BNA: 1.5M words of accurate broadcast news transcripts of the acoustic training data. Non lexical items such as breath noise, hesitations, word fragments are transcribed.
- BNC: 200M words of commercial transcripts of various broadcast shows (from 1992 to 1998). These transcripts do not include extra-lexical events.

It can be noted that only a very small proportion of the LM data (BNA) is truly representative of the real data to be transcribed.

The BNC and NEWS training texts were processed to clean errors inherent in the texts or arising from the preprocessing tools, and transformed to be closer to the observed American speaking style. Filler words such as "uh" and "uhm" were mapped to a unique form. The training texts were reprocessed in order to add a proportion of breath markers (4%), and of filler words (0.5%)[4]. While it would seem more elegant to incorporate these in the LM by interpolating LMs estimated on the clean text (without noises) and on the transcripts (with noises), adding them to the clean texts via a generation model, gave a higher perplexity ($\sim$ 6%) but a lower word error rate ($\sim$ 2% relative). This result can be explained by the observation that breath noise and filler words do not occur at random, but at specific places. Adding them at such places in the clean texts is equivalent to adding a priori information about the distribution of these phenomena in the transcripts.

All of the training texts were processed to include the most common 1000 acronyms found in the training texts[3], and compound words to represent frequent word sequences[4]. This provides an easy way to allow for reduced pronunciations such as /lɛmi/ for "let me" and /gʌnx/ for "going to" or a syllabic-n for the word "and" in "AT&T" in the recognition lexicon.

The recognition vocabulary contains 65,122 words, and includes all words occuring a minimum of 15 times in the BNC (63,954 words) or at least twice in the BNA data (23,234 words). The lexical coverage was 99.14%, 99.53% and 99.73% on the eval96, eval97 and eval98 test sets respectively, for which experimental results are provided.

LANGUAGE MODEL INTERPOLATION

One easy way to combine training material from different sources is to train an n-gram backoff LM per source and to interpolate them. The interpolation weights can be directly estimated on some development data with the EM algorithm. The resulting LM is a mixture of n-gram backoff LMs. This mixture is less practical to use than a single n-gram backoff LM which offers a decoding advantage. In particular, the LIMSI decoder relies on a backoff lexicon tree for word graph generation and even during word graph rescoring it takes advantage of the backoff property to reduce the search space.

An alternative is to simply merge the n-gram counts and train a single n-gram backoff language model on these counts. If some data sources are more representative than others for the task, the n-gram counts can be empirically weighted to minimize the perplexity on a set of development data. While this can be effective, it has to be done by trial and error and cannot easily be optimized. In addition, weighting the n-gram counts can pose problems in properly estimating the backoff coefficients. Using the three available data sources, we compared the two approaches on one hand by generating interpolated 4-gram backoff LMs and on the other hand by merging the n-gram counts with the manually optimized weights. The results obtained with word graph rescoring show that on 3 eval sets the approach which merged the n-gram counts had a slightly higher word error rate (0.2% absolute) 15.73% compared to 15.46%.

Another approach to combining the data sources is to merge the different LM components of the LM mixture obtained as above, thus creating a single n-gram backoff LM as proposed in [9]. The advantage of this approach is that the LM combination can still be properly optimized with the EM algorithm. In the resulting LM there are as many n-grams as there are distincts n-grams in the individual LMs trained on the separate data sets. The backoff coefficients of the merged LM are computed from the interpolated n-gram probabilities, ensuring that the probability mass for each context is equal to 1. Experimental results did not show any difference between the LM mixture and the LM merging approaches. All results given in this paper were obtained by using this last combination strategy.

COMBINING DATA SOURCES

Optimizating Text Selection

Selecting the appropriate LM training material evidently affects the resulting LM accuracies. There is the sometimes conflicting need for sufficient amounts of text data to estimate LM parameters and assuring that the data is representative of the task. For instance, in [5] it was reported that, for the broadcast news transcription task, while the use of all the available newspaper data led to a small decrease in perplexity, it also led to a small increase in the recognition error rate. Therefore, this year all NEWS texts that did not lower the perplexity were eliminated.

To optimize the selection of text, the newspaper and commercial transcription sources were split into 5 non-overlapping time periods, based on proximity to the eval98 epoch (15oct96-14nov96). For each of these periods (jan94-sep95, oct95-jun96, jul96-feb97, mar97-aug97, sep97-dec97) separate LMs were constructed for each source. The interpolation coefficient for each component LM was optimized on eval97 data (containing shows recorded in oct96). LMs with very low interpolation coef-

4gram LM	*Word Error rate (and % relative decrease)*			*Perplexity (and % relative decrease)*		
	Eval96	*Eval97*	*Eval98*	*Eval96*	*Eval97*	*Eval98*
NEWS	22.7	15.8	15.3	291.8	246.3	257.4
NEWS+BNA	21.1 (-7)	15.0 (-5)	14.4 (-6)	199.8 (-32)	192.4 (-22)	201.9 (-22)
BNC	20.8 (-1)	14.5 (-3)	14.1 (-2)	209.5 (+5)	196.6 (+2)	198.7 (-2)
BNC+BNA	20.3 (-2)	14.3 (-1)	13.8 (-2)	175.7 (-16)	175.6 (-11)	181.6 (-9)
BNC+NEWS+BNA	20.0 (-1)	14.0 (-2)	13.6 (-1)	167.4 (-5)	163.3 (-7)	168.8 (-7)

Table 1: Word error rate and perplexity for LMs constructed on different sources (NEWS: newspaper & newswire, 340M words; BNA: accurate broadcast news transcripts, 1.5M words; BNC: commercial broadcast news transcripts, 200M words) on 3 evaluation data sets.

ficients were eliminated. Subsets with comparable interpolation coefficients (difference sources or epochs) were merged in order to decrease the size of the resulting LM. Only very small variations in perplexity were observed during this process, and the final optimization resulted in interpolation of four 4-gram LMs, constructed on the following texts: BNC (200M words, interpolation coefficient 0.56); BNA (1.5M words, interpolation coefficient 0.22); NEWS period jan94-sep95 (200M words, interpolation coefficient 0.10); and NEWS period jul96-aug97[1] (141 Mwords, interpolation coefficient 0.12). It can be noted that the weight of BNA LM is equal to the weight of NEWS LMs (0.22) even though the text is much smaller.

Relevance of the Data Sources

The data available in American English may not be available for other languages:

- NEWS data is often available, but usually smaller amounts and with less diversity.
- BNA data corresponds to accurate transcriptions of 200h of speech, which is costly to develop.
- BNC is rare for other languages mainly for commercial reasons.

Several experiments were conducted in order to evaluate the influence of each source on the recognition word error rate. 4-grams LMs were constructed using the following data sets: NEWS only, BNC only, NEWS (0.55) + BNA (0.45), BNC (0.75) + BNA (0.25), BNC (0.56) + NEWS (0.22) + BNA (0.22) (the 4-gram used in the ARPA'98 evaluation). The interpolation coefficients were optimized on the eval97 data.

Recognition results obtained via word graph rescoring using these five LMs are summarized in Table 1 for the three eval data sets. The true differences between models may be slightly larger since all results used the same word graph generated with the BNC+NEWS+BNA LM. These results illustrate the discrepancy between perplexity and word error rate results. Based on perplexity, NEWS+BNA seems to be slightly better than BNC, but the recognition results are the opposite. In general, perplexity alone is not sufficient to measure the relevance of training data sources. Combining the accurate broadcast news transcriptions BNA with other sources is seen to reduce the perplexity, with the largest reduction when the only other source available is NEWS. The commercially produced transcripts BNC are seen to be quite relevant, particularly when combined with more accurate trancripts. Although the LM constructed only on NEWS data has a perplexity 43% higher than the 4-gram LM used in the LIMSI system, the recognition word error rate is only 11% higher.

[1] All data from the same period as the eval98 test set (15/10/96-14/11/96) was excluded.

CLASS-BASED LANGUAGE MODELS

Word class based language models can be combined with word n-gram language models to better estimate n-grams for unseen word sequences. In this work, word classes are automatically obtained by means of a clustering algorithm, where each word can belong to only one class. Words are clustered into a fixed number of classes according to their word contexts in the training texts. We assume that the 2-gram frequencies are sufficient statistics and use the conditional probabilities $p(v_k|v_i)$ (of observing v_k after v_i in the text), estimated as: $p(v_k|v_i) = N(v_i v_k)/N(v_i)$, where $N(v_i, v_k)$ is the frequency of the 2-gram (v_i, v_k) and $N(v_i)$ the frequency of v_i in the text. When the vocabulary words are clustered into classes, the initial distribution of the bigram probabilities $p(v_k|v_i)$ is replaced by the class based distribution $q(v_k|v_i) = \Pr(v_k|C(v_k)) * p(C(v_k)|C(v_i))$ where $C(\cdot)$ is the word class mapping function and $p(C(v_k)|C(v_i))$ is the probability of observing $C(v_k)$ after $C(v_i)$ in the text. The mapping criterion corresponds to the minimization of the Kullback-Leibler divergence between the word-based 2-gram distribution and the class-based 2-gram distribution (depending of the clustering). This criterion is equivalent to minimizing the perplexity of the training texts[7]. The classification procedure uses a Monte-Carlo algorithm to explore the search space by randomly selecting a word and associating it with a randomly selected class. If the Kullback-Leibler divergence is reduced, the new classification is kept. The number of classes is fixed *a priori*. Initially all words are gathered in one class, corresponding to the 1-gram distribution. The optimal number of classes is obtained *a posteriori* by measuring the perplexities of a held-out text with the class-based language models.

Class-based 3-gram models were built using a Witten-Bell discounting strategy and interpolated with word-based 3-gram and 4-gram models. Figure 1 shows the perplexity of the eval96 data set as a function of the interpolation coefficient for three models: a word-based 3-gram model interpolated with a class-based 3-gram model with 270 classes (tg+tg270cl); a word-based 4-gram model interpolated with the same class-based 3-gram model (fg+tg270cl); and the same word-based 4-gram model interpolated with a class-based 3-gram model with 1000 classes (fg+tg1000cl). The

	Word Error rate (and % relative decrease)			*Perplexity (and % relative decrease)*		
System Step	*Eval96*	*Eval97*	*Eval98*	*Eval96*	*Eval97*	*Eval98*
3-gram	21.0	14.6	14.2	181.4	176.7	182.6
4-gram	20.2 (-4)	14.3 (-2)	13.7 (-4)	167.4 (-8)	163.3 (-8)	168.8 (-8)
4-gram class	19.8 (-2)	13.9 (-3)	13.6 (-1)	163.6 (-2)	159.9 (-2)	165.6 (-2)

Table 2: Word error rates and perplexity after each decoding step with the Nov'98 system on 3 evaluation data sets.

optimal value of the interpolation coefficient is seen to be in the range $[0.1, 0.2]$, in which there is no real difference in perplexity for the 4-gram model with 270 or 1000 classes.

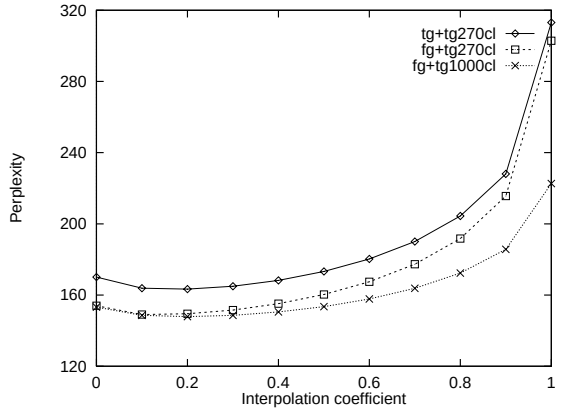

Figure 1: Perplexity of 3-gram and 4-gram word LM interpolated with a class-based 3-gram LM (270 and 1000 classes) on the eval96 test data.

DISCUSSION OF RESULTS

Word error rates and perplexities for the Nov'98 system in the last decoding steps are given in Table 2, on the evaluation data used in last 3 ARPA benchmark tests. The results on eval96 data, which are the closest to unrestricted broadcast news shows, are 30% higher than the results on the other eval datasets. The relative decrease in word error rate in going from a 3-gram LM to a 4-gram class is about 9.5%. The class-based model was build with 270 classes. This gain is partly due to the use of a 4-gram instead of a 3-gram word model and partly due to the interpolation of the 4-gram LM with the class model. A relative reduction in perplexity of about 10% is obtained, with most (8%) contributed by the 4-gram LM. The word error reductions observed between steps are due to the combined effect of a more accurate LM and an additional adaption of the acoustic models based on hypothesis of the previous pass.

From the perplexity variation measured during the text selection optimization process, it can be concluded that a certain amount of older NEWS texts can help to better estimate the core n-grams, whereas more temporally related NEWS texts help to better estimate n-grams involving words specific to the eval epoch. We found that using only about half of the available NEWS data led to smaller and lower perplexity LMs. Even relatively small amounts of accurate transcriptions help substantially in reducing the perplexity, but the overall impact is smaller than might be expected, if a large corpus of commercial transcriptions is available. The optimized interpolation coefficients of the 3 sources (NEWS, BNC and BNA), resulted in the same importance for NEWS and BNA, and comparable relative word error reductions were observed in Table 1. Given sufficient quantities of broadcast new transcriptions, newspaper and newswire texts appear to not be necessary.

Overall the combined improvements in text selection, interpolation, 4-gram and class-based LMs led to a 20% reduction in the perplexity of the LM of the final pass (3-gram class interpolated with a word 4-gram) compared with the 3-gram LM used in the the LIMSI Nov'97 BN system.

CONCLUSION

In this paper we have discussed the problem of language modeling for the transcription of broadcast news data. Different approaches for training the language model were explored and tested in the context of a complete transcription system. We have addressed normalization of the training texts and selective combination of the different materials, demonstrating the relative influence of each text source. Language model interpolation is a good tool to combine the estimates from different sources with no additional cost relative to a classical backoff LM. The use of 4-gram instead of 3-gram, as well as the use of interpolated 3-gram class model brings a small but significant gain. In part due to language modeling improvements (perplexity reduction of 20%), the LIMSI Nov'98 system has a relative word error rate 20% under that of the Nov'97 system.

REFERENCES

[1] J.L. Gauvain et al., "Recent Advances in Transcribing Television and Radio Broadcasts", Eurospeech'99, Budapest, Sept.1999.

[2] M. Adda et al., "Large Vocabulary Speech Recognition in French," *ICASSP'99*, Phoenix, pp. 45-48, March 1999.

[3] J.L. Gauvain, L. Lamel, G. Adda, D. Matrouf, "The LIMSI 1995 Hub3 System," *Proc. DARPA Speech Recognition Workshop*, Arden House, Feb. 1996.

[4] J.L. Gauvain et al., "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *ARPA Speech Recognition Workshop*, Chantilly, pp. 56-63, Feb. 1997.

[5] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI 1997 Hub-4E Transcription System", *DARPA Broadcast News Transcription & Understanding Workshop*, pp. 75-79, Landsdowne, Feb. 1998.

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

[7] M. Jardino "Multilingual stochastic n-gram class language models," *ICASSP-96*, Atlanta, May 1996.

[8] K. Seymore, S. Chen, M. Eskenazi, R. Rosenfeld, "Language and Pronunciation Modeling in the CMU 1996 Hub4 Evaluation. *Proc. DARPA Speech Recognition Workshop*, Chantilly, Virginia, pp. 141-146, Feb. 1997.

[9] P.C. Woodland, T. Neieler, E. Whittaker, "Language Modeling in the HTK Hub5 LVCSR", presented at the 1998 Hub5E Workshop, Sept. 1998.