A MAXIMUM ENTROPY LANGUAGE MODEL INTEGRATING N-GRAMS AND TOPIC DEPENDENCIES FOR CONVERSATIONAL SPEECH RECOGNITION

Sanjeev Khudanpur and Jun Wu

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218-2686
{sanjeev,junwu}@mail.clsp.jhu.edu

ABSTRACT

A compact language model which incorporates local dependencies in the form of N-grams and long distance dependencies through dynamic topic conditional constraints is presented. These constraints are integrated using the maximum entropy principle. Issues in assigning a topic to a test utterance are investigated. Recognition results on the Switchboard corpus are presented showing that with a very small increase in the number of model parameters, reduction in word error rate and language model perplexity are achieved over trigram models. Some analysis follows, demonstrating that the gains are even larger on content-bearing words. The results are compared with those obtained by interpolating topic-independent and topic-specific N-gram models. The framework presented here extends easily to incorporate other forms of statistical dependencies such as syntactic word-pair relationships or hierarchical topic constraints.

1. INTRODUCTION

Language modeling is a crucial component of systems that convert from various language modalities, including speech and handwriting, to text. Most current algorithms for language modeling, however, tend to suffer from an acute myopia, basing their predictions of the next word on only a few immediately preceding words. When humans are faced with a comparable task they can easily outperform such models using the richer linguistic information available from more complete context. We present a method for exploiting long-distance dependencies through dynamic models of the topic of the conversation. We present a compact model that integrates these dependencies with N-grams in a statistically sound manner in the maximum entropy (ME) framework.

Several models which combine topic related information with N-gram models have been studied, *e.g.*, in [1, 4, 3, 8, 9, 10]. The essential idea comes from the information retrieval (IR) literature where extensive use is made of weighted term-frequencies to discern the topic or genre of a document. Most schemes [4, 8, 10] exploit these differences for language modeling by constructing separate N-gram models for each individual genre or topic to capture these differences. Such a construction however results in fragmentation of the training text, for which the usual remedy is to interpolate such a topic specific N-gram model with a topic-independent model constructed using all the available data. An alternative presented in [3] starts off being similar to this work, but then makes *ad hoc* changes to an exponential model with the limited objective of fast rescoring. The work on read speech in [9] is similar; dynamics there are modeled by cache-like notions rather than a semantic notion of topic. The approach based on latent semantic analysis recently proposed in [1] is a refreshing departure from these methods and presents perplexities on newspaper text.

In the method presented here the term-frequencies are treated as *topic-dependent* salient features of a corpus, just as overall N-gram frequencies are *topic-independent* salient features. An admissible model is then required to satisfy constraints that reflect both the sets of features. The ME principle [5] is used to select a statistical model which meets all these constraints. This method has the advantage that only constraints on those term-frequencies which vary significantly across topics are made topic-dependent while the others are topic-independent. As a result, the model for every topic is trained from all the training data, making it possible to obtain better estimates of the topic-independent components of the model. A model with a small number of free parameters follows as a consequence. Finally, though we do not present experimental results for this, ME features could easily include higher order N-gram frequencies in documents of a particular topic, syntactic word-pair dependencies, frequencies of word-classes, POS constraints *etc.* It is easy to integrate them in a unified manner in this framework, as they are simply other features of the corpus which the model is constrained to match.

Section 2 contains a formulation of this model. Issues in assigning a topic to a test utterance are discussed in Section 3. Section 4 describes experiments on Switchboard, a corpus of spontaneous American English telephone conversations, and provides analysis of the results. Section 4.5 makes some comparisons between our method and the method of combining topic specific N-gram models by linear interpolation.

2. COMBINING N-GRAM AND TOPIC DEPENDENCIES

Let $\mathcal{V}$ denote the vocabulary of a speech recognizer. A language model may be viewed as a family $P(w_k|w_1,\ldots,w_{k-1})$ of conditional probability mass functions (pmfs) over words $w_k \in \mathcal{V}$ which may appear at the k-th position, based on some equivalence classification h_k of the "history" $w_1,\ldots,w_{k-1}$. For a trigram model, $P(w_k|w_1,\ldots,w_{k-1}) = P_3(w_k|w_{k-1},w_{k-2})$ and $h_k = [w_{k-1},w_{k-2}]$.

2.1. The ME Framework

We use the long range history $w_1,\ldots,w_{k-1}$ to assign a topic $\mathrm{t}_k = t(w_1,\ldots,w_{k-1})$ to a conversation. The sufficient statistic of the history is thus the triple $h_k = [w_{k-1},w_{k-2},\mathrm{t}_k]$, and

$$P(w_k|w_1,\ldots,w_{k-1}) = P_T(w_k|w_{k-1},w_{k-2},\mathrm{t}_k). \qquad (1)$$

Supported by National Science Foundation Grant No. IRI-9618874

Intuition suggests that not every word in the vocabulary will have strong dependence on the topic of the conversation. Estimating a separate conditional pmf for each $[w_{k-1}, w_{k-2}, \mathrm{t}_k]$ however fragments the training data and may result in poor estimates for such words. Additionally, topic related terms may not be seen in every word-context $[w_{k-1}, w_{k-2}]$. We therefore seek a model which, in addition to topic-independent N-gram constraints, meets topic-dependent *marginal* constraints:

$$\sum_{\mathrm{t}_k} P_T(w_k, \mathrm{t}_k | w_{k-1}, w_{k-2}) = \frac{\#[w_k, w_{k-1}, w_{k-2}]}{\#[w_{k-1}, w_{k-2}]},$$

$$\sum_{w_{k-1}, w_{k-2}} P_T(w_k, w_{k-1}, w_{k-2} | \mathrm{t}_k) = \frac{\#[w_k, \mathrm{t}_k]}{\#\mathrm{t}_k}. \qquad (2)$$

Note that these marginal probabilities are much more reliably estimated than the conditional ones in (1). Unreliable marginal probabilities, *e.g.*, those based on one or two observations, may be completely left out of the model's requirements or replaced with their Good-Turing estimates.

Linear constraints of the form described above define a family of pmfs and we choose the model in this family which has the highest entropy, corresponding qualitatively to the least additional assumptions on (or maximal smoothness of) the model. It is known that the ME model has an exponential form, with one parameter λ corresponding to each linear constraint placed on the model:

$$P_{\underline{\lambda}}(w_k | w_{k-1}, w_{k-2}, \mathrm{t}_k) = \frac{e^{\lambda(w_k)} \cdot e^{\lambda(w_{k-1}, w_k)} \cdot e^{\lambda(w_{k-2}, w_{k-1}, w_k)} \cdot e^{\lambda(\mathrm{t}_k, w_k)}}{Z(\underline{\lambda}, w_{k-1}, w_{k-2}, \mathrm{t}_k)},$$

where Z is a suitable normalization constant. The first three numerator terms correspond to standard N-gram constraints, while the fourth one is a topic-unigram parameter determined by term-frequencies in a particular topic.

2.2. Computational Issues in ME Model Estimation

The generalized iterative scaling (GIS) algorithm [6] is used to compute the ME model parameters $\underline{\lambda}$. Several challenges, predominantly associated with the computational and storage needs of the parameter estimation procedure, must be overcome in order to successfully implement a ME language model (LM) which incorporates topic dependencies with N-gram constraints in a unified manner.

To overcome the increased complexity from the addition of topic-dependent unigram constraints, we partition the training corpus based on the topics of the conversation, perform each iteration of the GIS algorithm for updating $\underline{\lambda}$ separately on each part, and correctly combine the updates. This topic-based division of the corpus reduces the computational complexity by an order of magnitude.

3. TOPIC ASSIGNMENT FOR TEST UTTERANCES

Two issues arise when using a topic-dependent LM for speech recognition. Since the actual spoken words are not available for topic assignment, topic assignment must be based on recognizer hypotheses. We investigate the impact of recognition errors on this process. It is also well known that the topic of a conversation may change as the conversation progresses. We examine whether a topic should be assigned to an entire test conversation, each utterance, or parts of an utterance. We also study if topic assignment for an utterance should be based only on that utterance, include a few preceding utterances, or include a few preceding and succeeding utterances. The results are presented in Section 4.3.

4. EXPERIMENTAL RESULTS ON SWITCHBOARD

The training set for all models described here consists of nearly 1200 conversations containing a total of 2.1 million words. Each conversation is annotated with one of about 70 topics, ranging from Affirmative Action to Woodworking; this is the topic recommended to the callers during data collection, though not every call adheres to its assigned topic. The vocabulary for the task has 22K words and includes all the words in the training and test set. The performance of various LMs is evaluated on a test set of 19 conversations (38 conversation sides) comprising 18000 words in over 2400 utterances.

For every test utterance, a list of the 2500-best hypotheses is generated by an HTK-based recognizer [11] using state-clustered crossword triphone HMMs with Gaussian mixture output densities and a back-off bigram LM. The recognition word error rate (WER) for *rescoring* these hypotheses and the average perplexity of the transcriptions of the test set are reported here.

4.1. Baseline Experiments

Table 1 shows the performance of standard back-off trigram models and an ME model with only N-gram constraints. The minimum count for a bigram to be included in a model is indicated by B, that for a trigram is by T. We leave out infrequent bigrams trigrams from the ME model for rapid training; models with all N-grams are presented later. The smaller back-off N-gram model is constructed to calibrate the corresponding ME model with N-gram constraints. It can be seen that when only N-gram constraints are used, the ME model essentially replicates the performance of the corresponding back-off N-gram model. Any improvements (or degradations) which adding topic-dependent constraints may yield is thus attributable to those features rather than the ME method.

Model (N-gram cutoffs)	Perplexity	WER
Back-off (no cutoffs)	79.2	43.2%
Back-off (B≥4, T≥2)	83.4	43.8%
ME (B≥4, T≥2)	84.7	43.7%

Table 1: Perplexity and WERs of Back-Off Trigram Models and Maximum Entropy Models with Trigram Constraints.

4.2. Estimation of Topic-Conditional Models

Each conversation side in the training corpus is processed to obtain a representative vector of weighted frequencies of vocabulary terms excluding stop words, where a stop word is any of a list of about 700 words with low semantic content which are ignored by the topic classifier. These vectors are then clustered using a K-means procedure (K∼70), with the initial cluster assignments being derived from the 70 manually assigned topics of the conversations. The resulting cluster assignment is then fixed for each conversation side for the remainder of the training process.

Words whose unigram frequency f_t in a cluster t differs significantly from its frequency f in the whole corpus are designated as topic-related words. We choose all words w for which $f_t(w) \log \frac{f_t(w)}{f(w)} \geq 3$ to be a word related to topic t. There are roughly 300 such words for every topic cluster, about 16K such words in the 22K vocabulary, and they constitute about 8% of the 2.1 million training tokens. ME models are trained with the constraints of the kind (2) on these words in addition to the N-gram constraints.

4.3. Topic Assignment During Testing

To use a topic-dependent model for rescoring, a topic must be assigned[1] to test utterances. We investigate four options for this assignment: (i) manual assignment of topics to the conversation, automatic topic assignment[2] based on (ii) the reference transcriptions or (iii) the 10-best hypotheses generated by the first recognition pass, and (iv) assignment by an *oracle* to minimize perplexity (or WER). The results, presented in Table 2, clearly indicate that even with a WER of over 40%, there is only a small loss in perplexity and a negligible loss in WER when the topic assignment is based on recognizer hypotheses instead of the correct transcriptions. Comparisons with the oracle indicate that there is little room for further improvement.

Source of Text for Topic Classification	Perplexity	WER
None (Baseline)	84.7	43.7%
Manual Assignment	76.5	42.9%
Ref. Transcriptions	77.1	43.0%
10-Best Hypotheses	77.4	43.1%
Oracle (optimal)	75.8	42.7%

Table 2: Topic Assignment Based on Erroneous Recognizer Hypotheses Causes Little Degradation in Performance.

We have also investigated topic assignment at several granularities and found that the best recognition performance is achieved by assigning a topic to each utterance based on the 10-best hypotheses of the current and the three preceding utterances. These results are presented in Table 3. Note that utterance-level topic assignment of Table 3 is more effective than the conversation-level assignment (Table 2). Adding topic-dependent constraints *reduces absolute WER by 0.8% and relative perplexity by 9.7%*.

Source of Text for Topic Classification	Perplexity	WER
None (Baseline)	84.7	43.7%
Ref. Transcriptions	75.5	42.8%
10-Best Hypotheses	76.5	42.9%
Oracle (optimal)	71.2	40.1%

Table 3: Dynamic Topic Assignment for Individual Utterances Based on the Current and Three Preceding Utterances.

To gain insight into improved performance from utterance-level topic assignment, we examine agreement between topics assigned at the two levels. As seen in Table 4, 8 out of 10 utterances prefer the topic-independent model and are filler utterances, probably serving vital discourse functions (e.g. acknowledgments, back-channel responses). Of the remaining utterances, a majority are closest to the topic which was assigned at the conversation-level. While a large fraction are closer to a topic other than the one preferred at the conversation-level, this is not an equally remarkable result as, in many of these cases, the topic assigned at the conversation-level is a close second or the two topics are similar.

Source of Text for Topic Classification	Agreement of Conv. & Utt. Level Topics	Utt. Level Topic When Disagreeing With Conv.	
		Other Topic	No Topic
Ref. Trans.	12.7%	7.1%	80.3%
10-Best Hyps.	9.9%	7.0%	83.1%

Table 4: Topic Dynamics Viewed Through (Dis)agreement of Utterance- and Conversation-Level Topic Assignment.

4.4. Analysis of Recognition Performance

To see if we indeed improve the model where we aim to improve it, the vocabulary is divided into two sets: all those words which have topic-conditional unigram constraints for any of the topics, and the others. Each word token in the reference transcription is then marked as belonging to one of the two sets and their perplexity is calculated separately. The words in the recognizer's output are also similarly marked, and each recognition error is assigned to one of the two sets, separating the WER over the two sets of words. About 7% of the tokens in the test set have topic-dependent constraints. Table 5 shows a breakdown of the results over the set of topic-dependent and -independent words for ME models with and without topic-dependent constraints. We also divide the vocabulary simply into content-bearing words and stop words (as defined earlier). Under this partition, nearly 25% of tokens in the test set are content-bearing and the remainder are stop words. Table 6 presents the performance gains analyzed for this partition.

Language Model (N-gram cutoffs)	Topic Words		Nontopic Words	
	Ppl	WER	Ppl	WER
ME (B$\geq$4, T$\geq$ 2)	3936	42.8%	63.9	43.8%
ME-Topic (B$\geq$4, T$\geq$ 2)	354	40.5%	68.2	43.1%

Table 5: Analysis of Performance Gains From Topic-Dep. LM.

Language Model (N-gram cutoffs)	Content Words		Stop Words	
	Ppl	WER	Ppl	WER
ME (B$\geq$4, T$\geq$ 2)	225	43.4%	58.8	43.8%
ME-Topic (B$\geq$4, T$\geq$ 2)	177	41.9%	57.2	43.2%

Table 6: Performance Improvement on Content-Bearing Words.

It is clear from Tables 5 and 6 that the gain in perplexity comes predominantly from content-bearing words, and the *1.5% improvement in WER on content-bearing words* is greater than the overall WER improvement; an important consideration for end users.

4.5. ME v/s Interpolated Topic N-Grams

Compared to the back-off trigram model which has about 250K parameters, the topic-conditional ME models introduce only about 16K additional parameters which modify probabilities of a few hundred words in the context of each topic. An alternative to this modeling approach is to partition the training data, build separate

[1] A hard decision is made by assigning the closest matching topic in the results presented here, though the formalism extends easily to soft topic decisions. We employ a standard cosine similarity measure commonly used in the IR community [1, 7] to assign a topic to test sentences.

[2] The null topic, which defaults to a topic-independent baseline model, is available as one of the choices to the topic classifier.

N-gram models for each topic and, since each topic N-gram is trained on a much smaller dataset, interpolate this topic specific model with a topic-independent model trained on all the data to obtain a smooth topic-dependent model. This is comparable to the approach described, e.g., in [4, 8, 10].

We construct back-off unigram, bigram and trigram models specific to each topic using the partitioning of the 2.1 million word corpus used for the ME models as described in Section 4.2. We interpolate each topic-specific N-gram with the topic-independent trigram model to obtain smooth topic-dependent N-gram models. Usually, one would tune the interpolation coefficient on some held out set. In this case, however, we (cheat and) choose the interpolation weight to minimize the perplexity of the test set under each interpolated model. Table 7 shows the recognition performance of the interpolated models. The topic for each test utterance for the interpolated model is the same as the one used for the ME topic model.

Model (N-gram cutoff)	#Params	Perplexity	WER
Back-Off (B≥4, T≥2)	253K	83.4	43.8%
Back-Off + Topic 1-gram	+70×11K	83.0	43.8%
Back-Off + Topic 2-gram	+70×26K	78.8	43.2%
Back-Off + Topic 3-gram	+70×55K	77.6	43.0%
ME-Topic (B≥4, T≥ 2)	+16K	76.5	42.9%

Table 7: Comparison with 70 Interpolated Topic N-Gram Models.

It may thus be argued that the ME approach permits us to combine via unigram constraints as much effective information as one would get by interpolating topic specific trigram models. This, we argue, is due to the systematic integration of topic-dependent and topic-independent constraints in our model.

5. TOPIC-DEPENDENT ME MODELS INCLUDING N-GRAMS WITH LOWER COUNTS

For rapid turnaround in the experiments described in the preceding section we compare ME and ME-Topic models which do not impose constraints on low-count bigrams (B<4) and trigrams (T<2). We also implement topic-dependent ME models with constraints on less frequent N-grams so as to compare them with the best back-off models. The performance of these models is shown in Table 8. It is clear that the topic conditioning reduces (absolute) WER by about the same amount in each case over the corresponding back-off model without topic constraints, and topic-dependent ME models *reduce absolute WER by 0.6% and relative perplexity by 7%* over the best trigram model. The improvements on content-bearing words is about twice as much (c.f. Section 4.4).

Model (N-gram cutoffs)	Perplexity	WER
Back-off (no cutoffs)	79.2	43.2%
Back-off (B≥1, T≥2)	78.8	43.2%
ME (B≥1, T≥2)	78.9	43.1%
ME-Topic (B≥1, T≥ 2)	73.5	42.7%
Back-off (B≥2, T≥2)	80.1	43.3%
ME (B≥2, T≥2)	80.3	43.1%
ME-Topic (B≥2, T≥ 2)	74.3	42.6%

Table 8: Final Comparison of N-Gram and Topic-Dependent LMs.

6. CONCLUDING REMARKS

We have described a ME language model which combines topic dependencies and N-gram constraints in a unified fashion and provides small but significant performance gains at the cost of few additional parameters. The performance improvement on content-bearing words is even more significant. A small number of topic-dependent unigrams are able to provide this improvement because the information they provide is *complementary* but well integrated with the modeling ability of N-grams.

Since the framework itself extends easily to combining other dependencies, our current efforts are in the direction of exploiting syntactic structure obtained from a left to right partial parse of the utterance as described in [2]. The syntactic constraints will provide information which complements both N-grams and topic dependencies. Additional constraints such as word class frequencies based on parts of speech, hierarchical topic dependencies *etc.* are also under consideration in order to further extend the model and derive benefits from the flexibility offered by the ME framework.

7. ACKNOWLEDGMENTS

Substantial assistance with the topic classification was provided by Radu Florian and David Yarowsky. In particular, Radu Florian provided the list of stop words and the algorithm for cosine-similarity used here for automatically assigning topics to test utterances.

8. REFERENCES

[1] J. R. Bellegarda, “Exploiting Both Local and Global Constraints for Multispan Statistical Language Modeling,” in *Proc. ICASSP’98*, Vol. 2, pp. 677-680, May 12-15, 1998.

[2] C. Chelba and F. Jelinek, “Exploiting Syntactic Structure for Language Modeling,” in *Proc. COLING-ACL’98*, Vol. 1, pp. 225-231, Aug. 10-14, 1998.

[3] S. F. Chen *et al*, “Topic Adaptation for Language Modeling Using Unnormalized Exponential Models,” in *Proc. ICASSP’98*, Vol. 2, pp. 681-684, May 12-15, 1998.

[4] P. Clarkson and A. Robinson, “Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache,” in *Proc. ICASSP’97*, Vol. 2, pp. 799-802, Apr. 21-25, 1997.

[5] I. Csiszár, “Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems,” *The Annals of Statistics*, Vol. 19, No 4, Dec. 1991.

[6] J. N. Darroch and D. Ratcliff, “Generalized Iterative Scaling for Log-Linear Models,” *Annals Math. Stats.*, Vol. 43, 1972.

[7] R. Florian, “Exploiting Nonlocal and Syntactic Word Relationships in Language Models,” Ph.D. Qualifying Project Report, CS Dept, Johns Hopkins University, Aug., 1998.

[8] R. Iyer and M. Ostendorf, “Modeling Long Range Dependencies in Language,” in *Proc. ICSLP’96*, Vol. 1, pp. 236-239, Oct. 3-6, 1996.

[9] R. Kneser *et al*, “Language Model Adaptation Using Dynamic Marginals,” in *Proc. EUROSPEECH’97*, Vol. 4, pp. 1971-1974, Sept. 22-25, 1997.

[10] S. C. Martin *et al*, “Adaptive Topic-Dep. Language Modeling Using Word-Based Varigrams,” *Proc. EUROSPEECH’97*, Vol. 3, pp. 1447-1450, Sept. 22-25, 1997.

[11] S. Young, J. Jansen, J. Odell, D. Ollasen, P. Woodland, *The HTK Book (Version 2.0)*, ECRL, Cambridge, 1995.