Computer Speech and Language (2000) **14**, 355–372
Article No. 10.1006/csla.2000.0149
Available online at http://www.idealibrary.com on IDEAL

Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling

Sanjeev Khudanpur[†] **and Jun Wu**

Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, U.S.A.

Abstract

A new statistical language model is presented which combines collocational dependencies with two important sources of long-range statistical dependence: the syntactic structure and the topic of a sentence. These dependencies or constraints are integrated using the maximum entropy technique. Substantial improvements are demonstrated over a trigram model in both perplexity and speech recognition accuracy on the Switchboard task. A detailed analysis of the performance of this language model is provided in order to characterize the manner in which it performs better than a standard N-gram model. It is shown that topic dependencies are most useful in predicting words which are semantically related by the subject matter of the conversation. Syntactic dependencies on the other hand are found to be most helpful in positions where the best predictors of the following word are not within N-gram range due to an intervening phrase or clause. It is also shown that these two methods individually enhance an N-gram model in complementary ways and the overall improvement from their combination is nearly additive.

1. Introduction

It is generally acknowledged that while N-gram models are the mainstay of statistical language modeling for speech recognition, machine translation, information retrieval and other natural language-processing applications, they make little or no use of either the rich syntactic structure or the semantic content in natural language. Several attempts to alleviate this drawback have been made in the recent past and we begin with a brief discussion of a few representative examples of such work, along with the observation that these efforts have been in one of the two directions mentioned above—either incorporation of topic dependencies or of syntactic structure—but not both.

Several models which combine topic-related information with N-gram models have been studied, e.g. by Bellegarda (1998), Clarkson and Robinson (1997), Chen and Rosenfeld (1998), Iyer and Ostendorf (1996), Kneser, Peters and Klakow (1997) and Martin, Liermann and Ney (1997). The essential idea in all these papers comes from the information

[†] E-mail: [sanjeev,junwu]@clsp.jhu.edu

0885–2308/00/040355 + 18 $35.00/0

retrieval (IR) literature where extensive use is made of weighted word frequencies to discern the topic or genre of a document. Most schemes (Iyer & Ostendorf, 1996; Clarkson & Robinson, 1997; Martin *et al.*, 1997) exploit differences in word frequencies across topics by constructing separate N-gram models for each individual genre or topic. Such a construction however results in fragmentation of the training text by topic, for which the usual remedy is to interpolate each topic-specific N-gram model with a topic-independent model constructed using all the available data. An alternative presented in Chen and Rosenfeld (1998) starts off being similar to the technique which will be presented in this paper, but then makes *ad hoc* changes to an exponential model with the limited objective of fast rescoring. The work on read speech in Kneser *et al.* (1997) is also somewhat similar to our work. However, dynamic shifts in topic there are modeled by a unigram cache rather than a semantic notion of topic. Lafferty and Suhm (1996) described a topic-dependent model very similar to the topic-dependent model presented here and demonstrated improvement in perplexity over a bigram model with a relatively small 5000-word vocabulary (Lafferty, 1996). The approach based on latent semantic analysis recently proposed in Bellegarda (1998) is a refreshing departure from these methods and is most similar to our approach in philosophy. It remains of interest to us to compare our method with Bellegarda's under identical experimental conditions.

Relatively fewer models which utilize syntactic structure for improving over N-gram models have been proposed (cf. Lafferty, Sleator and Temperly (1992) and Chelba *et al.* (1997)). Notable among them are Chelba and Jelinek (1998, 1999), who have used a left-to-right parser to extract syntactic heads and used them to enhance trigram models. To illustrate the kinds of dependencies they attempt to capture, consider the following sentence which illustrates the mechanism used by these models.

> Financial **officials** *in the former British colony* **consider** the contract essential to the revival of the Hong Kong futures exchange.

The word *consider* is clearly more at home as part of the bigram *officials consider* as opposed to *colony consider*. This intuitive dependence is incorporated into the above-mentioned models by identifying that the prepositional phrase "in the former British colony" modifies the noun phrase "Financial officials" and by suppressing its role in predicting the following word, *consider*. This is achieved, e.g. by a predictor which uses the preceding head-words instead of preceding words. Furthermore, noun phrases such as "Bank managers", "Securities traders" or "Stock brokers" play exactly the same syntactic role as "Financial officials". Therefore, if the bigram *officials consider* happens to be unseen, the nonterminal label NP$'$ of the plural noun phrase provides a coarser classification of the history from which to predict the word *consider*. This is achieved via deleted interpolation in Chelba and Jelinek (1998, 1999). Thus the syntactic heads bring both syntactic structure and an improved *back-off* scheme, so to speak, to bear on the prediction problem.

We present a model which combines dependencies of both kinds: we incorporate semantic dependencies through a notion of topic inferred by IR techniques applied to the long-range history at any given word position, and we incorporate syntactic structure by parsing in a left-to-right manner the sentence fragment up to the word position of interest. We use maximum entropy techniques to combine these two kinds of statistical dependencies into a single unified model. This leads to a fairly parsimonious model or representation.

The remainder of this paper is organized as follows. Section 2 discusses the way in which we use the maximum entropy framework to combine *N-gram and topic dependencies*. A brief theoretical exposition is followed by model parameterization, parameter estimation issues, issues of assigning a topic to test utterances, experimental results and analysis. A performance comparison with some other topic-dependent language models is also shown in this section. Section 3 discusses our maximum entropy framework to combine *syntactic constraints with N-grams*. Here too, the theoretical exposition, model parameterization, experimental results and analysis are provided in one place. Finally, in Section 4, we present a language model which combines *N-gram, syntactic and topic dependencies*. We demonstrate in this section that the improvements in performance from using syntactic and topic dependencies are nearly additive and argue why this may be the case. We conclude in Section 5 with some remarks and mention our ongoing research.

2. Combining topic and N-gram dependencies

In the method presented in this section, unigram frequencies collected from all documents on a specific topic t in the language model (LM) training corpus are treated as *topic-dependent* salient features of the corpus, just as overall N-gram frequencies are *topic-independent* salient features. A class of models is constructed in which the probability of a word w_i depends on the preceding $N - 1$ words, $w_{i-1}, \dots, w_{i-N+1}$, as well as the topic t of the current document. In our case, $N = 3$. An admissible model is defined to be any joint distribution $P(w_i, w_{i-1}, w_{i-2}, t_i)$ which satisfies constraints that reflect both the sets of salient features: within every topic, P assigns a conditional probability to N-grams in such a way that the marginal (unigram) probability of a word coincides with the topic-specific relative frequency of the word, and the total probability of each N-gram equals its relative frequency in the corpus. The maximum entropy (ME) principle (Jaynes, 1982; Csiszár, 1991) is used to select the “smoothest” statistical model from this class of models.

This method of imposing constraints has the advantage that only those word 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.

Section 2.1 contains a formulation of this model, which was proposed earlier in Khudanpur and Wu (1999). Issues in assigning a topic to a test utterance are discussed in Section 2.2. Section 2.3 describes experiments on Switchboard, a corpus of spontaneous American English telephone conversations (see Godfrey, Holliman and McDaniel (1992) for a description of the corpus), and provides analysis of the results. Section 2.3.5 makes some comparisons between our technique and the technique of combining topic-specific N-gram models by linear interpolation.

2.1. Parameterization of the maximum entropy model

Let $\mathcal{V}$ denote the vocabulary of a speech recognizer. A language model may be viewed as a family $P(w_i|w_1, \dots, w_{i-1})$ of conditional probability mass functions (pmfs) over words $w_i \in \mathcal{V}$ which may appear at the ith position, based on some equivalence classification h_i of the “history” $w_1, \dots, w_{i-1}$. For a trigram model, $P(w_i|w_1, \dots, w_{i-1}) \approx P(w_i|w_{i-1}, w_{i-2})$ and $h_i = [w_{i-1}, w_{i-2}]$.

2.1.1. The maximum entropy framework

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

$$P(w_i|w_1,\dots,w_{i-1}) \approx P(w_i|w_{i-1},w_{i-2},t_i). \tag{1}$$

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 $[w_{i-1}, w_{i-2}, t_i]$ however fragments the training data and may result in poor estimates for such words. Additionally, topic-related words may not be seen in every word context $[w_{i-1}, w_{i-2}]$. We therefore seek a model which, in addition to topic-independent N-gram constraints,

$$\sum_{t_i} P(w_i|w_{i-1},w_{i-2},t_i)P(t_i) = \frac{\#[w_{i-2},w_{i-1},w_i]}{\#[w_{i-2},w_{i-1}]}, \tag{2}$$

$$\sum_{t_i,w_{i-2}} P(w_i|w_{i-1},w_{i-2},t_i)P(w_{i-2},t_i) = \frac{\#[w_{i-1},w_i]}{\#[w_{i-1}]}, \tag{3}$$

meets topic-dependent *marginal* constraints

$$\sum_{w_{i-1},w_{i-2}} P(w_i|w_{i-1},w_{i-2},t_i)P(w_{i-1},w_{i-2}) = \frac{\#[t_i,w_i]}{\#t_i}. \tag{4}$$

Note that these marginal probabilities are much more reliably estimated from counts $\#[\cdot]$ of the corresponding events observed in the corpus than the conditional probabilities in (1). Unreliable marginal probabilities, e.g. those based on one or two observations, may be completely left out of the model's requirements. Alternately, or in addition to leaving out some constraints, the relative frequency estimates of the marginal probabilities on the right-hand sides of Equations (2)–(4) may be replaced by their corresponding Good–Turing estimates. Finally, since our primary interest is in the conditional model $P(w_i|w_{i-1}, w_{i-2}, t_i)$, empirically observed relative frequencies $\hat{P}(t_i)$, $\hat{P}(w_{i-2}, t_i)$ and $\hat{P}(w_{i-1}, w_{i-2})$ may be substituted for the model-based probabilities $P(t_i)$, $P(w_{i-2}, t_i)$ and $P(w_{i-1}, w_{i-2})$, respectively, in Equations (2)–(4).

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:

$$P_{\underline{\lambda}}(w_i|w_{i-1},w_{i-2},t_i) = \frac{e^{\lambda(w_i)}\cdot e^{\lambda(w_{i-1},w_i)}\cdot e^{\lambda(w_{i-2},w_{i-1},w_i)}\cdot e^{\lambda(t_i,w_i)}}{Z(\underline{\lambda},w_{i-1},w_{i-2},t_i)}, \tag{5}$$

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 word frequencies in particular topics.

2.1.2. Computational issues in model estimation

An improved version of the generalized iterative scaling or GIS algorithm of Darroch and Ratcliff (1972) presented by S. Della Pietra, V. Della Pietra and Lafferty (1995) is used to compute the ME model parameters $\underline{\lambda}$. However, a straightforward use of this algorithm turns out to be quite tedious and some algorithmic challenges must be overcome in order to incorporate topic dependencies with N-gram constraints. These are predominantly associated with the computational and storage needs of the parameter estimation procedure.

The GIS algorithm, stated very loosely, proceeds as follows. We start with an exponential model as given by (5) with arbitrary initial values for the parameters, say $\underline{\lambda}^0$. At the kth iteration, $k = 1, 2, \ldots,$ the marginal probabilities of interest, namely the left-hand sides of Equations (2)–(4), are computed using $P_{\underline{\lambda}^{k-1}}$ and compared to the right-hand sides. If the two do not coincide for any particular constraint, only the model parameter $\lambda^{k-1}(\cdot)$ corresponding to the specific constraint is adjusted to $\lambda^k(\cdot)$ to achieve equality. It can be shown that under suitable conditions, this procedure converges to the model with the maximum entropy among all models that satisfy the constraints. An interesting alternate interpretation of this procedure is that it yields, from the class of exponential models of the form (5), the model that assigns the *maximum likelihood* to the observed LM training data.

A very large part of the computational cost in the GIS algorithm is in determining the normalization term

$$Z(\underline{\lambda}, w_{i-1}, w_{i-2}, t_i) = Z = \sum_{w \in \mathcal{V}} e^{\lambda(w)} \cdot e^{\lambda(w_{i-1},w)} \cdot e^{\lambda(w_{i-2},w_{i-1},w)} \cdot e^{\lambda(t_i,w)}$$

for each history $[w_{i-1}, w_{i-2}, t_i]$ in the training corpus. As explained by Rosenfeld (1994), this sum may be simplified by dividing the vocabulary into words which *do not* have bigram, trigram or topic-dependent constraints, and words which *do*.

$$Z = \sum_{w:\lambda(*,w)=0} e^{\lambda(w)} + \sum_{w:\lambda(*,w)\neq 0} e^{\lambda(w)} \cdot e^{\lambda(w_{i-1},w)} \cdot e^{\lambda(w_{i-2},w_{i-1},w)} \cdot e^{\lambda(t_i,w)}. \tag{6}$$

Note that the first sum goes over the same set of words for every history $[w_{i-1}, w_{i-2}, t_i]$ and therefore needs to be computed only once in each iteration of the GIS algorithm. After this simplification, previously published methods require computation, in each seen history, of a sum of as many factors as there are *conditional constraints*, i.e. the number of constraints in which a vocabulary word w appears along with some component of the conditioning event $[w_{i-1}, w_{i-2}, t_i]$.

In case of a standard trigram model, the number of seen histories $[w_{i-1}, w_{i-2}]$ is the number of distinct bigrams in the LM training text. The average number of conditional constraints in each seen history is a very small fraction of the vocabulary. For the Switchboard corpus, this average number is about 200, compared to a vocabulary size of about 20 000 words. In the case of the topic-dependent model, the number of seen histories $[w_{i-1}, w_{i-2}, t_i]$ is considerably larger. More significantly, the number of topic-dependent unigram constraints is a much larger fraction of the vocabulary because the topic-unigram constraints $[t_i, w]$ are active in every bigram history $[w_{i-1}, w_{i-2}]$. This results in considerably more computation if parameters are estimated in the usual 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.

(1) We partition the set of seen histories $\mathcal{H}$ based on topic t as:

$$\mathcal{H}_t = \{\text{“history}_i\text{”} = (w_{i-1}, w_{i-2}, t_i = t)\}.$$

(2) Separately for each training partition $\mathcal{H}_t$, we compute

$$\underline{\lambda}_t^k = \{\lambda_t^k(\text{“history”}, w)\},$$

the kth update in the iterative procedure for computing model parameters $\underline{\lambda}$. The result of this partitioning is that within each topic-cluster $\mathcal{H}_t$ of the training corpus, $t_i = t$ for each position i and the topic-conditional constraints, corresponding to $\lambda(t, w)$, may be treated just like regular unigram constraints corresponding to $\lambda(w)$. Their contribution to the normalization constant Z for all seen histories $[w_{i-1}, w_{i-2}, t]$ in $\mathcal{H}_t$ need therefore be computed only once, as described for unigrams in Equation (6).

(3) We then obtain $\underline{\lambda}^k$, the overall kth update of model parameters, as

$$\lambda^k(\text{“history”}, w) = -\log\left\{\sum_{t=1}^{T} \hat{P}(t) \cdot e^{-\lambda_t^k(\text{“history”}, w)}\right\},$$

where $\hat{P}(t)$ is the fraction of the histories in the LM training corpus that belong to partition $\mathcal{H}_t$. It is relatively straightforward to verify that this is the correct aggregation of the partial updates.

This topic-based division of the corpus reduces the computational complexity by an order of magnitude.

2.2. *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 following utterances. The results are presented in Section 2.3.3.

2.3. *Experimental results on Switchboard*

The vocabulary for this speech recognition task has 22 K words. The language model training set for all models described here consists of nearly 1200 Switchboard 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 suggested to the callers during data collection, though not every call adheres to its assigned topic. The acoustic models used for recognition are state-clustered cross-word triphone HMMs with 12 Gaussian mixture output densities, trained with MF-PLP acoustic features from about 60 hours of speech data. The performance of various LMs is evaluated on a test set of 19 conversations (38 conversation sides) comprising 18 000 words in over 2400 utterances.

For every test utterance, a list of the 100-best hypotheses is generated by an HTK-based recognizer (Young, Jansen, Odell, Ollasen & Woodland, 1995) using 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.

2.3.1. *Baseline experiments*

Table I 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. Note that singleton trigrams have no constraints on their probability in the ME model. The smaller back-off N-gram model is constructed to be comparable with the ME model and will be treated as the baseline model throughout this paper.

TABLE I. Perplexity and WER of back-off trigram models and a maximum entropy model with trigram constraints

Model (N-gram cutoffs)	Perplexity	WER
Back-off (no cutoffs)	79.2	38.7%
Back-off ($B \geq 1$, $T \geq 2$)	78.8	38.5%
ME ($B \geq 1$, $T \geq 2$)	78.9	38.3%

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 technique.

2.3.2. *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 ($K \sim 70$), 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 relative frequency f_t in a cluster t differs significantly from its relative frequency f in the whole corpus are designated as topic-related words. We designate all words w for which

$$f_t(w) \log \frac{f_t(w)}{f(w)} \geq 3$$

to be words related to topic t. There are roughly 300 such words for every topic cluster, about 16 K such words in the 22 K vocabulary, and they constitute about 8% of the 2.1 million training tokens. ME models are trained with the constraints of the kind (4) on these words in addition to the N-gram constraints.

2.3.3. *Topic assignment during testing*

To use a topic-dependent model for rescoring, a topic must be assigned to test utterances. 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 (Bellegarda, 1998; Florian & Yarowsky, 1999) to assign a topic to test sentences. We investigate four options for this assignment: (i) manual assignment of topics to the conversation, automatic topic assignment[1] 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 II, clearly indicate that even with a WER of over 38%, 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.

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

TABLE II. Topic assignment based on erroneous recognizer hypotheses causes little degradation in performance

Source of text for topic classification	Perplexity	WER
None (baseline)	78.8	38.5%
Manual assignment	73.1	37.8%
Ref. transcriptions	73.8	37.8%
10-Best hypotheses	74.4	37.9%
Oracle (optimal)	72.5	37.7%

TABLE III. Dynamic topic assignment for individual utterances based on the current and three preceding utterances

Source of text for topic classification	Perplexity	WER
None (baseline)	78.8	38.5%
Ref. transcriptions	73.3	37.8%
10-Best hypotheses	73.5	37.8%

TABLE IV. Topic dynamics viewed through (dis)agreement of utterance-level and conversation-level topic assignment

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%

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 III. Note that utterance-level topic assignment of Table III is more effective than the conversation-level assignment (Table II). Adding topic-dependent constraints *reduces absolute WER by* 0.7% *and relative perplexity by 7%*.

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 IV, 8 out of 10 utterances prefer the topic-independent model and are filler utterances, probably serving vital discourse functions (e.g. acknowledgements, 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, e.g. *social security* and *health care for the elderly*.

2.3.4. 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

TABLE V. Analysis of performance gains from the topic-dependent model

Language model	Topic words		Nontopic words	
(N-gram cutoffs)	Perplexity	WER	Perplexity	WER
ME ($B \geq 1$, $T \geq 2$)	3795	37.7%	62.2	38.5%
ME-Topic ($B \geq 1$, $T \geq 2$)	356	35.5%	66.6	37.9%

TABLE VI. Performance improvement on content-bearing words

Language model	Content words		Stop words	
(N-gram cutoffs)	Perplexity	WER	Perplexity	WER
ME ($B \geq 1$, $T \geq 2$)	8941	42.2%	36.4	37.6%
ME-Topic ($B \geq 1$, $T \geq 2$)	3923	40.8%	37.1	36.9%

topics, and the others. About 7% of the tokens in the test set have topic-dependent constraints. The WER of each set of tokens is divided in the same way, i.e. insertions or deletions of topic-dependent words as well as a substitution that involves a topic-dependent word count towards an error in the first category. Table V 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, about 25% of tokens in the test set are content bearing and the remainder are stop words. Table VI presents the performance gains analyzed for this partition. As expected, Tables V and VI show that the gain in perplexity comes predominantly from content-bearing words, and the 1.4% improvement in WER on these words is greater than the overall WER improvement. This may be a more important performance measure than the overall WER for tasks such as spoken document retrieval and automatic content extraction.

2.3.5. Maximum entropy vs. interpolated topic N-grams

Compared to the back-off trigram model which has about 250 K parameters, the topic-conditional ME models introduce only about 16 K 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 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 Clarkson and Robinson (1997), Iyer and Ostendorf (1996) and Martin *et al.* (1997).

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 2.3.2. 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 VII 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.

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 tri-

TABLE VII. Comparison with 70 interpolated topic N-gram models

Model (N-gram cutoff)	#Params	Perplexity	WER
Back-Off ($B \geq 1$, $T \geq 2$)	499 K	79.2	38.5%
Back-Off + Topic 1-gram	+70 × 11 K	78.4	38.5%
Back-Off + Topic 2-gram	+70 × 26 K	77.3	38.3%
Back-Off + Topic 3-gram	+70 × 55 K	76.1	38.1%
ME-Topic ($B \geq 1$, $T \geq 2$)	+16 K	73.5	37.8%

TABLE VIII. Perplexity and WER of a cache-based model and a maximum entropy model with topic constraints

Model (N-gram cutoffs)	Perplexity	WER
Back-off ($B \geq 1$, $T \geq 2$)	78.8	38.5%
Back-off + 1-gram Cache	75.2	38.9%
ME-Topic ($B \geq 1$, $T \geq 2$)	73.5	37.8%

gram models. This, we argue, is due to the systematic integration of topic-dependent and topic-independent constraints in our model.

2.3.6. Maximum entropy vs. cache-based models

We have also implemented a cache language model for comparison with the topic-dependent maximum entropy model. Specifically, at each position i in the test corpus, we estimate a unigram model based on all the words seen so far in that particular conversation side. This unigram model is then interpolated with the back-off trigram model and used in rescoring.

Since the correct (spoken) words are not available to the recognizer, all the words in the N-best hypotheses for the preceding utterances in a conversation side are considered in estimating the cache model. This has the beneficial effect that words appearing in multiple hypotheses are likely to be correct and hence get counted multiple times, while words which do not appear in many hypotheses are potentially due to recognition errors and are naturally discounted.

Usually, one would tune this value of N as well as the interpolation coefficient on some held-out set. However, in this case as well, we (cheat and) choose the value of N (100) and the interpolation weight (0.1) of the cache language model to minimize the perplexity of the test set. The results in Table VIII show that though interpolation with the cache model results in reduced perplexity, there is no reduction in WER from using a cache model. Cache models have, however, been shown to reduce WER on other tasks.

While it is a bit surprising to see the WER increase with a cache model, it is perhaps explained by the relatively high error rates in the hypotheses from which the model is estimated. To support this argument, we compare the recognition output of the baseline trigram model and the cache-interpolated model of Table VIII in the following manner. Each recognition error is first marked as an *insertion*, *substitution* or *deletion* in the usual way. Next, each error token is marked as a *repeated* error if the word in question has undergone the same kind of error in the portion of the history used by the cache model, e.g. if a word w in the reference transcription was deleted from the hypothesis for the first time in a conversation side at some position i and deleted again at position $j > i$, then j is a case of a repeated deletion of w, though i is not. We find that compared to the baseline trigram model, the cache-based model has about a 0.6% higher rate of repeated errors. Comparing this to the overall increase in error rate of only 0.4%, it seems reasonable to conclude that a small reduction of nonrepeated errors is offset by an increase in the number of repeated errors, pointing to the detrimental effect of using a cache model at high error rates.

3. Combining syntactic and N-gram dependencies

Although it is widely acknowledged that the syntactic structure of a sentence should be helpful in predicting words, many challenges must be met to integrate syntactic structure into a language model. An outline of our initial efforts in this direction appears in Wu and Khudanpur (1999).

The structure of statistical language models and that of syntax are substantially different. Statistical language models are represented as multi-dimensional tables of conditional probabilities, while the syntactic structure of a sentence is usually embedded in trees. Therefore, a probabilistic model for syntactic information is required before it can be employed in statistical language modeling. Furthermore, parsers that generate syntactic structures are usually designed to work on correct sentences, and they must be modified to parse erroneous partial hypotheses in speech recognition.

We parse all sentences in the training data by the left-to-right parser presented by Chelba and Jelinek (1998). This parser generates a stack $\mathcal{S}_i$ of candidate parse trees T_{ij} for a sentence prefix $W_1^{i-1} = w_1, w_2, \ldots, w_{i-1}$ at position i. The parser also assigns a probability $P(w_i|W_1^{i-1}, T_{ij})$ to each possible following word w_i given the jth partial parse T_{ij}, and a likelihood function $\rho(W_1^{i-1}, T_{ij})$ for the jth partial parse, according to

$$P(w_i|W_1^{i-1}) = \sum_{T_{ij}\in\mathcal{S}_i} P(w_i|W_1^{i-1}, T_{ij}) \cdot \rho(W_1^{i-1}, T_{ij}), \tag{7}$$

where

$$\rho(W_1^{i-1}, T_{ij}) = \frac{P(W_1^{i-1}, T_{ij})}{\sum_{T_{ij}\in\mathcal{S}_i} P(W_1^{i-1}, T_{ij})}. \tag{8}$$

Each nonterminal node in the jth parse tree T_{ij} has a head-word. This head-word dominates the subtree rooted at this node and thus can be regarded as the representative of the syntactic constituent spanned by the subtree. In a parse tree spanning a complete sentence, the head-word at the root node is typically the main verb of the sentence. For a sentence prefix, the syntactic analysis is partial, and more than one constituent may await higher level parsing decisions as shown in Figure 1. The lexical heads of such constituents are called *exposed* head-words. The exposed head-words preceding the word “after” in Figure 1, for instance, are “contract” and “ended”. For a word w_i at position i, we label the exposed head-words from right to left as h_{i-1}, h_{i-2}, etc., dropping the subscript j for simplicity of notation. The reader should bear in mind that the preceding head-word exposed at position i will depend on the choice of the candidate partial parse T_{ij}. Details of parse generation and identification of head-words may be found in Chelba and Jelinek (1998).

We assume that the immediate history (w_{i-2}, w_{i-1}) and last two exposed head-words h_{i-2}, h_{i-1} of the partial parse T_{ij} carry most of the useful information. Of course some syntactic information will be lost under this assumption. However, it leads to a very important simplification:

$$P(w_i|W_1^{i-1}, T_{ij}) \approx P(w_i|w_{i-2}, w_{i-1}, h_{i-2}, h_{i-1}). \tag{9}$$

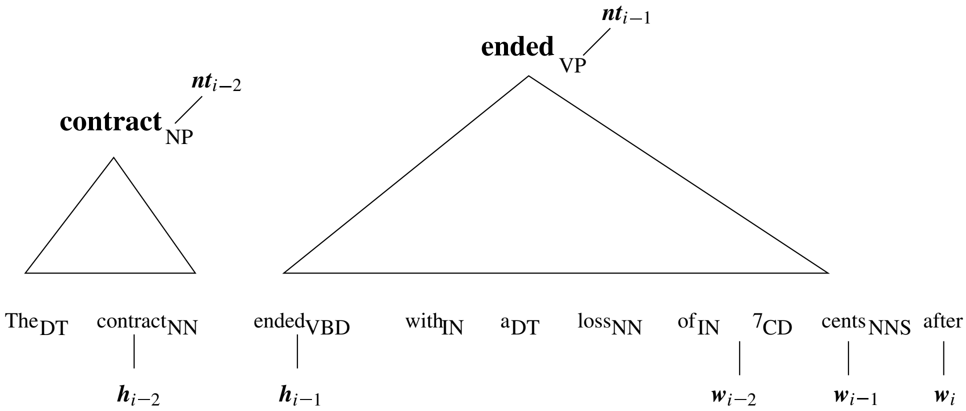

Figure 1. An example of a partial parse (reproduced from Chelba and Jelinek (1998)).

Note that the syntactic information is represented in the same form as N-grams by this representation! The main difference is that unlike w_{i-2}, w_{i-1}, the head-words h_{i-2}, h_{i-1} are *latent* variables, i.e. unlike w_{i-2} and w_{i-1} which are directly observable at any position i, the identities of h_{i-2} and h_{i-1}, which would be known only from the correct syntactic analysis of the utterance, must be treated as missing data.

The "correct" exposed head-words are not known even for the training corpus. Therefore we use the top N candidate partial parses as potentially correct parses with the appropriate probability. We accumulate counts for each seen 5-tuple $(w_1, w_2, h_1, h_2, w_3)$ in the training data as follows. If a particular position i has the trigram (w_{i-2}, w_{i-1}, w_i) and the jth partial parse T_{ij} exposes head-words (h_{i-2}, h_{i-1}) with probability ρ, then the 5-tuple $(w_{i-2}, w_{i-1}, h_{i-2}, h_{i-1}w_i)$ gets a fractional count of ρ. The counts obtained in this manner for all frequent tuples are set as constraints in the syntactic ME model.

During testing, each hypothesis is parsed by the parser and the ME model is invoked for each partial parse with its appropriate exposed head-words. The word probability is calculated by Equations (7)–(9), where the probability ρ of a partial parse is taken directly from the parsing model (8).

3.1. Parameterization of the maximum entropy model

It is clearly unreasonable to expect to obtain enough statistics to estimate a separate conditional pmf for each history $(w_{i-2}, w_{i-1}, h_{i-2}, h_{i-1})$ in (9). Therefore, the maximum entropy model for the dependency structure of (9) is done in much the same manner as was done for the topic-conditional model in Section 2.1. In particular, we impose the following marginal constraints on the model,

$$\sum_{h_{i-1},h_{i-2}} P(w_i|w_{i-1}, w_{i-2}, h_{i-1}, h_{i-2})\hat{P}(h_{i-1}, h_{i-2}) = \frac{\#[w_{i-2}, w_{i-1}, w_i]}{\#[w_{i-2}, w_{i-1}]},$$

$$\sum_{w_{i-2},h_{i-1},h_{i-2}} P(w_i|w_{i-1}, w_{i-2}, h_{i-1}, h_{i-2})\hat{P}(w_{i-2}, h_{i-1}, h_{i-2}) = \frac{\#[w_{i-1}, w_i]}{\#[w_{i-1}]},$$

$$\sum_{w_{i-1},w_{i-2}} P(w_i|w_{i-1}, w_{i-2}, h_{i-1}, h_{i-2})\hat{P}(w_{i-1}, w_{i-2}) = \frac{\#[h_{i-2}, h_{i-1}, w_i]}{\#[h_{i-2}, h_{i-1}]},$$

$$\sum_{w_{i-1},w_{i-2},h_{i-2}} P(w_i|w_{i-1}, w_{i-2}, h_{i-1}, h_{i-2})\hat{P}(w_{i-1}, w_{i-2}, h_{i-2}) = \frac{\#[h_{i-1}, w_i]}{\#[h_{i-1}]},$$

where $\#[\cdot]$ and $\hat{P}$ represent observed counts and relative frequencies, respectively, of the corresponding events. Note that the first two kinds of constraints involve regular N-gram counts, and the last two involve what we call *head-word N-gram* counts. To obtain reliable estimates of the marginal probabilities for constraining the model, Good–Turing discounting is applied to the low counts before using them on the right-hand sides of the equations above. Furthermore, trigram constraints based on very low counts (<2) are completely dropped. We then seek a model which satisfies the remaining constraints and has the maximum entropy.

It is well known that the maximum entropy model has an exponential form

$$\begin{aligned} &P_{\underline{\lambda}}(w_i|w_{i-1}, w_{i-2}, h_{i-1}, h_{i-2}) \\ &\quad = \frac{e^{\lambda(w_i)} \cdot e^{\lambda(w_{i-1},w_i)} \cdot e^{\lambda(w_{i-2},w_{i-1},w_i)} \cdot e^{\lambda(h_{i-1},w_i)} \cdot e^{\lambda(h_{i-2},h_{i-1},w_i)}}{Z(\underline{\lambda}, w_{i-2}, w_{i-1}, h_{i-2}, h_{i-1})}, \end{aligned}$$

where Z is a normalization constant. Note, again, that the first three terms in the numerator correspond to standard N-gram constraints. The last two terms represent head-word N-gram constraints.

The parameters $\underline{\lambda}$ of the model are again trained by the improved iterative scaling algorithm (Della Pietra *et al.*, 1995). The heavy computational load in the training procedure is distributed to many computers by the parallel training method described in Section 2.1.2. The computational cost however is much higher than that for a trigram model in this case due to the head-word N-gram constraints.

3.2. *Experimental results on Switchboard*

The experimental set-up for the language models with syntactic constraints is identical to that described in Section 2.3. Hence the baseline results of Table I are valid for the following experiments as well.

3.2.1. *Recognition performance of the syntactic model*

We estimate an ME model with syntactic and N-gram constraints as described above. The cutoffs for N-grams are the same as the baseline models (singleton trigrams are dropped), and the cutoffs for head-word bigrams and trigrams are 2 and 1, respectively. Note that since headword N-grams may have fractional counts ρ, even a cutoff of 1 may result in excluding potential constraints. The results of this model with syntactic and N-gram constraints are shown in Table IX, where the performance of the model with topic-dependent constraints is reproduced from Table III for comparison. As noted earlier, topic-dependent constraints alone reduce perplexity by about 7% and WER by 0.7% (absolute). Independently of these, simple syntactic constraints (head-word N-grams) result in a 7% reduction in perplexity and 0.8% (absolute) in WER.

3.2.2. *Analysis of recognition performance*

Our choice of topic-dependent constraints is clearly aimed at manipulating the probabilities of words whose frequency within conversations on a particular topic are vastly different from their frequency in the corpus. These are typically content-bearing words which are interrelated by the topic of the conversation. Thus we expect improvements from the topic model to

TABLE IX. Performance of a language model with syntactic constraints

Model constraints	Perplexity	WER
Baseline trigram	79.0	38.5%
N-gram + Topic	73.5	37.8%
N-gram + Syntactic	73.5	37.7%

TABLE X. Performance improvement on different history categories

Language model	$h_{i-1} = w_{i-1}$ and $h_{i-2} = w_{i-2}$		$h_{i-1} \neq w_{i-1}$ or $h_{i-2} \neq w_{i-2}$	
	Perplexity	WER	Perplexity	WER
Trigram	78.8	37.8%	79.7	40.3%
Syntactic	73.1	37.4%	74.2	38.8%

be concentrated on such words. Syntactic dependencies on the other hand are expected to be most helpful in positions where the best predictors of the following word are not within N-gram range due to an intervening phrase or clause. We perform some analysis in this section to see if this is true.

To see whether we indeed improve the model when syntactic head-words are different from N-grams, all histories in the test sentences are divided into two categories: those histories whose head-words (h_{i-2}, h_{i-1}) *coincide* with the two immediately preceding words (w_{i-2}, w_{i-1}) and those whose head-words do not. About 75% of the histories in the test set belong to the former category, perhaps explaining the considerable power of the trigram model. The perplexity and WER for these two sets are counted separately. Table X gives the analysis.

When $h_{i-1} \neq w_{i-1}$ or $h_{i-2} \neq w_{i-2}$, the reduction in perplexity and WER resulting from the syntactic model is consistent with the intuition exemplified by Figure 1, and supports our claim that the syntactic head-words carry meaningful information that is not captured by N-grams. Two other contrasts seen in Table X and not explained by such intuition, are worth pointing out.

Firstly, for either model, the perplexity when $h_{i-1} \neq w_{i-1}$ or $h_{i-2} \neq w_{i-2}$ is *not* dramatically higher than when $h_{i-1} = w_{i-1}$ and $h_{i-2} = w_{i-2}$. We do not have a satisfactory explanation for this fact, though we speculate that this is a corpus- or genre-specific result and more of a contrast may be expected between the two cases in written texts such as the Wall Street Journal corpus.

Secondly, note that the syntactic model reduces perplexity by about the same amount both for cases when $(h_{i-2}, h_{i-1}) = (w_{i-2}, w_{i-1})$ and when $(h_{i-2}, h_{i-1}) \neq (w_{i-2}, w_{i-1})$. But there is a significant reduction in WER in one case and not in the other. For this we offer a somewhat more convincing explanation. When $h_{i-1} = w_{i-1}$ and $h_{i-2} = w_{i-2}$, one may expect to be inside a common phrase in the utterance. This expectation is reinforced by the fact that in 63% of cases when $h_{i-1} = w_{i-1}$ and $h_{i-2} = w_{i-2}$, the trigram (w_{i-2}, w_{i-1}, w_i) turns out to have been seen in the LM training data. When $h_{i-1} \neq w_{i-1}$ or $h_{i-2} \neq w_{i-2}$, the trigram (w_{i-2}, w_{i-1}, w_i) turns out to have been seen in training in only 45% of the cases. But this is not all. The LM training corpus is just a superset of the acoustic training corpus for Switchboard. Consequently, a frequently seen trigram (w_{i-2}, w_{i-1}, w_i) implies that the context-dependent acoustic model for w_i is likely to have been better estimated than the case

TABLE XI. ME vs. interpolated syntactic models (cf. Chelba and Jelinek (1999))

Language model	Perplexity	WER
Baseline trigram	79.0	38.5%
Interpolated syntactic	75.5	37.9%
Maximum entropy	73.5	37.7%

when (w_{i-2}, w_{i-1}, w_i) is an unseen trigram. There is some evidence pointing towards this in Table X: the baseline trigram, for roughly the same perplexity, has a lower WER when $h_{i-1} = w_{i-1}$ and $h_{i-2} = w_{i-2}$. Based on this inference about the relative strengths of the acoustic model in the two history categories, an improvement in the language model may be expected to have a larger impact on WER when the acoustic model is relatively deficient. This is supported by the greater reduction in WER by the syntactic model when $h_{i-1} \neq w_{i-1}$ or $h_{i-2} \neq w_{i-2}$ as compared to $(h_{i-2}, h_{i-1}) = (w_{i-2}, w_{i-1})$.

3.2.3. Maximum entropy vs. interpolated syntactic models

Analogous to Section 2.3.5, it is of interest to compare our maximum entropy technique of combining syntactic and N-gram dependencies with more traditional interpolation techniques. Due to more than sheer coincidence, the experimental results reported so far are on the same language model training corpus, baseline recognizer and test set as used by Chelba and Jelinek (1999), who use deleted interpolation instead of maximum entropy. Thus such a comparison is easily made.

In their work, the standard trigram model and trigram models based on the syntactic heads, i.e. the 5-tuple $(h_{i-2}, nt_{i-2}, h_{i-1}, nt_{i-1}, w_i)$, where the nonterminal labels are used for a coarser classification of the history, are combined via deleted interpolation. It may be noted that we do not take advantage of the nonterminal labels in the work presented here. Another minor difference between the experimental conditions is that they rescore a first-pass lattice using an A^*-search algorithm, while we rescore the 100-best hypotheses from the first recognition pass. However, the results are still very comparable. We therefore simply reproduce their experimental results in Table XI for comparison. Note that the maximum entropy technique slightly but consistently outperforms interpolation, as also reported in other work on maximum entropy models. However, the main argument for maximum entropy remains its ability to combine many diverse sources of statistical dependence in an elegant unified manner.

4. Combining topic, syntactic and *N*-gram dependencies

Finally, we construct a model with both topic and syntactic constraints. Due to the flexible manner in which constraints can be introduced in the maximum entropy framework, the combination of these constraints results in another exponential model of the form

$$P(w_i|w_{i-1}, w_{i-2}, h_{i-1}, h_{i-2}, t_i) \\ = \frac{e^{\lambda(w_i)} \cdot e^{\lambda(w_{i-1},w_i)} \cdot e^{\lambda(w_{i-2},w_{i-1},w_i)} \cdot e^{\lambda(h_{i-1},w_i)} \cdot e^{\lambda(h_{i-2},h_{i-1},w_i)} \cdot e^{\lambda(t_i,w_i)}}{Z(\underline{\lambda}, w_{i-2}, w_{i-1}, h_{i-2}, h_{i-1}, t_i)},$$

where Z is a normalization constant, and the parameters $\underline{\lambda}$ are computed to satisfy constraints on the marginal probability of N-grams, head-word N-grams and topic-conditional unigrams.

TABLE XII. Performance of a language model with topic and syntactic constraints

Model constraints	Perplexity	WER
Baseline trigram	79.0	38.5%
N-gram + Topic	73.5	37.8%
N-gram + Syntactic	73.5	37.7%
N-gram + Topic + Syntax	69.9	37.2%

TABLE XIII. Perplexity and WER for content words and stop words

Language	Content Wds		Stop Wds	
model	Perplexity	WER	Perplexity	WER
Trigram	8941	42.2%	36.4	37.6%
Topic	3923	40.8%	37.1	36.9%
Syntactic	4825	41.6%	36.6	36.8%
Composite	3699	40.4%	35.6	36.4%

TABLE XIV. Performance improvement on different history categories

Language model	$h_{i-1} = w_{i-1}$ $h_{i-2} = w_{i-2}$		$h_{i-1} \neq w_{i-1}$ or $h_{i-2} \neq w_{i-2}$	
	Perplexity	WER	Perplexity	WER
Trigram	78.8	37.8%	79.7	40.3%
Topic	73.0	37.3%	74.4	39.1%
Syntactic	73.1	37.4%	74.2	38.8%
Composite	69.8	37.0%	69.9	38.0%

The recognition performance of this model, shown in row 4 of Table XII, surpasses that of the other three models. The perplexity on the test set reduces by 12% (relative) and WER by 1.3% (absolute) compared to the baseline trigram model. The results also show that the gains from these two sources of long-range dependence are nearly additive on both perplexity and WER.

4.1. *Analysis of recognition performance*

As seen in Table VI, the gain from the topic-dependent language model was largely due to improved prediction of content-bearing words, and Table X showed that the WER reduction in the case of the syntactic constraints came largely from improved prediction when the two immediately preceding exposed heads were not within trigram range. In Tables XIII and XIV, we perform a similar *post hoc* analysis of our recognition results for the composite language model described above.

It is clear from these two tables that in each category, the model with semantic and syntactic features is able to perform better than the models with either syntactic or semantic features alone. This further reinforces the notion that the information provided by the two types of cues in the language model "context" are to a considerable extent independent.

5. Concluding remarks

A composite language model that incorporates two diverse sources of long-range dependence with N-grams has been described. A perplexity reduction of 12% and a WER reduction of 1.3% (absolute) are achieved on the Switchboard task. The performance improvement on content words is even more significant when topic information is employed. Both the topic constraints and syntactic constraints are helpful when useful predictors of the following word are beyond N-gram range, and the syntactic model is more powerful in this case. These two sources of nonlocal dependencies are complementary and their gains over a trigram model are almost additive. The results show the benefits of integrating various sources of information under the ME framework in improving language modeling.

More gains from syntactic structure is expected by using the nonterminal labels in the partial parse as shown in Chelba and Jelinek (1998). Work is in progress to incorporate this information in the language model. All experiments in this paper are based on N-best rescoring, but we will also use this method in lattice rescoring in the future. Finally, we are working to extend this model to other corpora, e.g. Broadcast News.

The authors thank Radu Florian and David Yarowsky for their help on topic detection and data clustering and Frederick Jelinek and Ciprian Chelba for providing the syntactic model (parser) for the experimental results reported here. William Byrne provided crucial expertise for the generation of the first-pass recognition lattices, from which the 100-best lists used in these experiments were extracted. The anonymous referees provided several useful suggestions to improve this paper. This research was supported by the U.S. National Science Foundation via Grant No IRI 9618874 under the STIMULATE program.

References

Bellegarda, J. R. (1998). Exploiting both local and global constraints for multispan statistical language modeling. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Seattle, WA, pp. 677–680.

Chelba, C. *et al.* (1997). Structure and performance of a dependency language model. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 2775–2778.

Chelba, C. & Jelinek, F. (1998). Exploiting syntactic structure for language modeling. *Proceedings of the COLING-ACL Meeting,* Montreal, Canada, pp. 225–231.

Chelba, C. & Jelinek, F. (1999). Recognition performance of a structured language model. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1567–1570.

Chen, S. & Rosenfeld, R. (1998). Topic adaptation for language modeling using unnormalized exponential models. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Seattle, WA, pp. 681–684.

Clarkson, P. & Robinson, A. (1997). Language model adaptation using mixtures and an exponentially decaying cache. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Munich, Germany, pp. 799–802.

Csiszár, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, **19**, 2032–2066.

Darroch, J. N. & Ratcliff, D. (1972). Generalized iterative scaling for Log-Linear models. *The Annals of Mathematical Statistics*, **43**, 1470–1478.

Della Pietra, S., Della Pietra, V. & Lafferty, J. (1995). Inducing features of Random fields. *Computer Science Technical Report No CMU-CS-95-144*, Carnegie Mellon University.

Florian, R. & Yarowsky, D. (1999). Dynamic nonlocal language modeling via hierarchical topic-based adaptation. *Proceedings of ACL99*.

Godfrey, J., Holliman, E. & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 517–520. Available online at `http://www.ldc.upenn.edu/`.

Iyer, R. & Ostendorf, M. (1996). Modeling long range dependencies in language. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 236–239.

Jaynes, E. T. (1982). On the rationale of maximum entropy methods. *Proceedings of the IEEE*, **70**, 939–952.

Khudanpur, S. & Wu, J. (1999). A maximum entropy language model integrating N-grams and topic dependencies for conversational speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Phoenix, AZ.

Kneser, R., Peters, H. & Klakow, D. (1997). Language model adaptation using dynamic marginals. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1971–1974.

Lafferty, J. (1996). Gibbs–Markov models. *Computing Science and Statistics*, **27**, 370–377.

Lafferty, J., Sleator, D. & Temperly, D. (1992). Grammatical trigrams: A Probabilistic model of link grammar. *Proceedings of the 1992 AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, Cambridge, MA.

Lafferty, J. & Suhm, B. (1996). Cluster expansions and iterative scaling for maximum entropy language models. In *Maximum Entropy and Bayesian Methods*, (Hanson, K. and Silver, R., eds), Kluwer Academic Publishers.

Martin, S. C., Liermann, J. & Ney, H. (1997). Adaptive Topic-dep. Language modeling using Word-based varigrams. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1447–1450.

Rosenfeld, R. (1994). Adaptive statistical language modeling: A maximum entropy approach. *Computer Science Technical Report No. CMU-CS-94-138*, Carnegie Mellon University.

Wu, J. & Khudanpur, S. (1999). Combining nonlocal, syntactic and N-gram dependencies in language modeling. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Budapest, Hungary.

Young, S., Jansen, J., Odell, J., Ollasen, D. & Woodland, P. (1995). The HTK Book (Version 2.0), Entropic Cambridge Research Laboratory.

(Received 20 April 2000, and accepted for publication 21 July 2000)