Lexical Triggers and Latent Semantic Analysis for Cross-Lingual Language Model Adaptation

WOOSUNG KIM and SANJEEV KHUDANPUR
The Johns Hopkins University

In-domain texts for estimating statistical language models are not easily found for most languages of the world. We present two techniques to take advantage of in-domain text resources in other languages. First, we extend the notion of *lexical triggers*, which have been used monolingually for language model adaptation, to the cross-lingual problem, permitting the construction of sharper language models for a target-language document by drawing statistics from related documents in a resource-rich language. Next, we show that *cross-lingual latent semantic analysis* is similarly capable of extracting useful statistics for language modeling. Neither technique requires explicit translation capabilities between the two languages! We demonstrate significant reductions in both perplexity and word error rate on a Mandarin speech recognition task by using these techniques.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language models*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech recognition and synthesis*; I.5.4 [**Pattern Recognition**]: Applications—*Text processing*

General Terms: Experimentation, Human Factors, Languages

Additional Key Words and Phrases: Automatic speech recognition, language model adaptation, latent semantic analysis, lexical trigger, multilingual processing, statistical language modeling

1. INTRODUCTION

Statistical language models are indispensable components of many human language technologies, for example, automatic speech recognition (ASR), machine translation (MT), handwriting and optical character recognition, spelling correction, and information retrieval (IR). The best-known techniques for estimating language models (LMs) require large amounts of text in the domain and language of interest, making this a bottleneck resource for the development of applications in many languages. For a typical example of how performance

Authors' address: Woosung Kim and Sanjeev Khudanpur Center for Language and Speech Processing, The Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218, USA; email: woosung@cs.jhu.edu, khudanpur@jhu.edu.
This research was supported by the National Science Foundation (via Grant Nos. ITR-0225656 and IIS-9982329) and the Office of Naval Research (via Contract No. N00014-01-1-0685).

suffers from the lack of adequate data, the reader may consider the task of automatic transcription of Arabic conversational telephone speech described in Kirchhoff et al. [2002], where it is evident that the word error rates are considerably higher than those for a comparable English language task.

It is often the case, however, that the application in question has been developed in one of the major languages—English, French, German, and so on. It is natural to consider if the vast text resources created in these languages could somehow be put to use in developing models for other languages. This is the problem we set out to investigate.

There have been attempts to overcome this data scarcity problem in other components of speech and language processing systems. In acoustic modeling, Schultz and Waibel [1998] and Byrne et al. [2000] reuse and adapt acoustic models from resource-rich languages to improve acoustic models for resource-deficient languages. Morphological analyzers, noun–phrase chunkers, POS taggers, and so on, have been developed for resource-deficient languages by Yarowsky et al. [2001] by performing linguistic analysis on the resource-rich side of a *parallel* or translation-equivalent corpus, and automatically projecting the annotations to the resource-deficient language via word alignment.

For language modeling, if sufficiently good MT is available between a resource-rich language such as English and a resource-deficient language, say, Chinese, then one may choose English documents related to a target Mandarin story, translate the English documents, and use the resulting Chinese word statistics to adapt a Chinese language model to the target story [cf. Khudanpur and Kim 2002]. The demands placed on the MT component of this technique for cross-lingual language modeling are, arguably, not as severe as they would be if translation were the ultimate end rather than merely the means to an end. A statistical translation lexicon derived using techniques of Brown et al. [1993] may often be adequate. Yet, the assumption of some MT capability presupposes linguistic resources, such as a modest *sentence-aligned* parallel corpus, which may not be available for some languages. We investigate, in this article, whether one could get away with only a *document-aligned* or *comparable* corpus in the resource-rich language.

Two primary means of exploiting cross-lingual information for language modeling are investigated in this paper, neither of which requires any explicit MT capability:

Cross-Lingual Lexical Triggers: If we are constructing a language model for transcribing a Mandarin news story, and we happen to have at hand a number of English newswire articles covering the same event or topic, it is clear that several content-bearing English words will signal the existence of a number of content-bearing Chinese counterparts in the story we are transcribing. If a set of matched English–Chinese stories is provided for training, one can infer which Chinese words an English word would "trigger" by using statistical measures, such as the mutual information between the cooccurrence of the two words in the matched document-pairs. This idea, which is reminiscent of *trigger language models* [Rosenfeld 1996], is investigated in a cross-lingual setting here.

Cross-Lingual Latent Semantic Analysis: Latent semantic analysis (LSA) of a collection of bilingual document-pairs provides a representation of words in

both languages in a common low-dimensional Euclidean space [Dumais et al. 1997]. If the pairing of documents is derived on the basis of content similarity, the Chinese "neighbors" of an English word, for instance, are semantically related words. This provides another means for using English word-frequencies to improve the Chinese language model from English texts for the transcription of a related Mandarin story.

It is shown through empirical evidence that while both techniques yield good statistics for adapting a Chinese language model to a particular story, the goodness of the information varies from story to story. It is through a story-specific adaptation scheme, based on a maximum likelihood criterion, that we derive significant benefits in our task.

The rest of this paper is organized as follows. Section 2 begins, for the sake of completeness, with a review of the cross-lingual story-specific LM. A notion of cross-lingual lexical triggers is proposed in Section 3, which overcomes the need for a sentence-aligned parallel corpus for obtaining translation lexicons. The use of LSA for cross-lingual language modeling is described in Section 4. After a brief detour to describe topic-dependent LMs in Section 5, a description of the database and the experimental results are provided in Section 6 and Section 7, respectively. Finally, Section 8 concludes this paper with future work.

2. CROSS-LINGUAL STORY-SPECIFIC ADAPTATION

Our aim is to sharpen a language model in a resource-deficient language, say, Mandarin, by using data from a resource-rich language, say, English. Of course, any other pair of languages will serve the purpose of this exposition.

Let $d_1^C, \ldots, d_N^C$ denote the text of N *test stories* in a Mandarin news broadcast to be transcribed by an ASR system, and let $d_1^E, \ldots, d_N^E$ denote their corresponding or *aligned* English newswire articles, selected from some contemporaneous text corpus. Correspondence here does not imply that the English document d_i^E needs to be an exact translation of the Mandarin story d_i^C. It is quite adequate, for instance, if the two stories report the same news event. Our approach is expected to be helpful even when the English document is merely on the same general topic as the Mandarin story, although the closer the content of a pair of articles the better the proposed methods are likely to work. Assume for the time being that a sufficiently good Chinese–English story alignment is (somehow) given.

Assume further that we have at our disposal a stochastic translation lexicon—a probabilistic model of the form $P_T(c|e)$—which provides the Chinese translation $c \in \mathcal{C}$ of each English word $e \in \mathcal{E}$, where $\mathcal{C}$ and $\mathcal{E}$ respectively denote our Chinese and English vocabularies.

2.1 Computing a Cross-Lingual Unigram Distribution

Let $\hat{P}(e|d_i^E)$ denote the relative frequency of a word e in the document d_i^E, $e \in \mathcal{E}$, $1 \leq i \leq N$. It seems plausible that

$$P_{\text{CL-unigram}}\left(c|d_i^E\right) = \sum_{e \in \mathcal{E}} P_T(c|e)\hat{P}\left(e|d_i^E\right) \quad \forall\, c \in \mathcal{C} \tag{1}$$

would be a good unigram model for the ith Mandarin story d_i^C.

We propose using this cross-lingual unigram statistic to sharpen a statistical Chinese LM used for processing the test story d_i^C. One way to do this is via linear interpolation

$$P_{\text{CL-interpolated}}\left(c_k|c_{k-1}, c_{k-2}, d_i^E\right) = \lambda P_{\text{CL-unigram}}\left(c_k|d_i^E\right) + (1-\lambda)P(c_k|c_{k-1}, c_{k-2}) \tag{2}$$

of the cross-lingual unigram model (1) with a static trigram model for Chinese, where the interpolation weight λ may be chosen offline to maximize the likelihood of some held-out Mandarin stories. The improvement in (2) is expected from the fact that unlike the static text from which the Chinese trigram LM is estimated, d_i^E is semantically close to d_i^C and even the adjustment of unigram statistics, based on a stochastic translation model, may help. Other variations on (2) are easily anticipated, such as

$$\begin{aligned}&\tilde{P}_{\text{CL-interpolated}}\left(c_k|c_{k-1}, c_{k-2}, d_i^E\right)\\ &\quad = \frac{\lambda_{c_k} P_{\text{CL-unigram}}\left(c_k|d_i^E\right) + (1-\lambda_{c_k})P(c_k|c_{k-1}, c_{k-2})}{\sum_{c\in\mathcal{C}} \lambda_c P_{\text{CL-unigram}}\left(c|d_i^E\right) + (1-\lambda_c)P(c|c_{k-1}, c_{k-2})}\end{aligned} \tag{3}$$

where the interpolation weight may be chosen to let content-bearing words be influenced more by the cross-lingual cues than function words by making λ_{c_k} proportional to the *inverse document frequency* of c_k in the Chinese LM training text [cf. e.g., Coccaro and Jurafsky 1998], log-linear interpolation with a global or word-dependent λ's, bucketing the λ's based on Chinese N-gram counts, and so on.

Figure 1 shows the data flow in this cross-lingual LM adaptation approach, where the output of the first pass of an ASR system is used by a cross-lingual information retrieval (CLIR) system to find the English document(s) d_i^E, an MT system computes the statistic of (1), and the ASR system uses the LM (2) in a second pass.

2.2 Obtaining the Matching English Document(s) d_i^E

To illustrate how one may obtain the English document(s) d_i^E to match the Mandarin story d_i^C, let us assume that we also have a stochastic reverse-translation lexicon $P_T(e|c)$. One obtains from the first pass ASR output (cf. Figure 1), the relative frequency estimate $\hat{P}(c|d_i^C)$ of Chinese words c in d_i^C, $c \in \mathcal{C}$, and uses the translation lexicon $P_T(e|c)$ to compute

$$P_{\text{CL-unigram}}\left(e|d_i^C\right) = \sum_{c\in\mathcal{C}} P_T(e|c)\hat{P}\left(c|d_i^C\right) \quad \forall\, e \in \mathcal{E} \tag{4}$$

an English bag-of-words representation of the Mandarin story d_i^C as used in standard vector-based vector-based information retrieval (IR). The document d_i^E with the highest TF-IDF weighted cosine-similarity to d_i^C is then selected

$$d_i^E = \arg\max_{d_j^E} \text{sim}\left(P_{\text{CL-unigram}}\left(e|d_i^C\right), \hat{P}\left(e|d_j^E\right)\right). \tag{5}$$

Readers familiar with IR literature will recognize this to be the standard *query-translation* approach to CLIR.

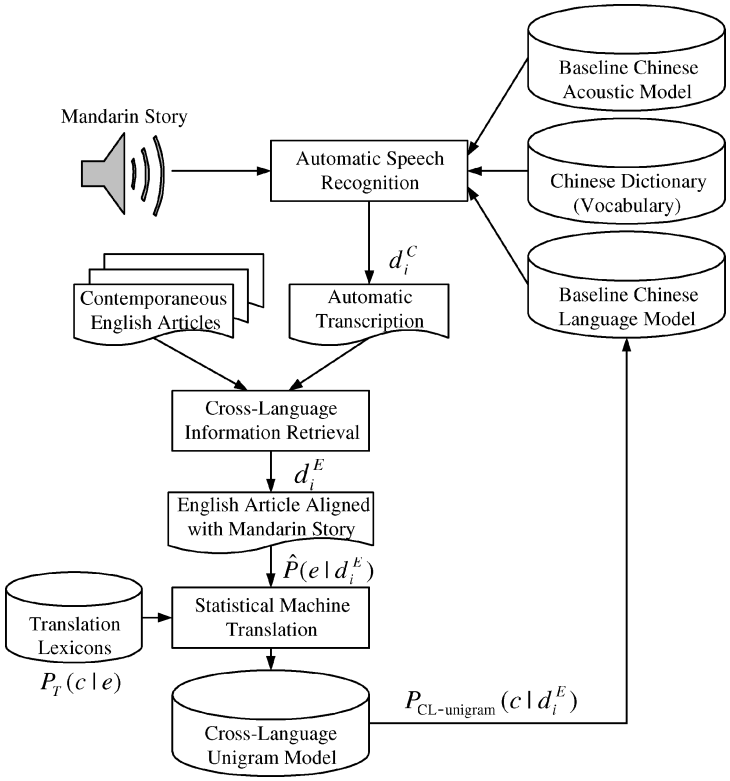

Fig. 1. Story-specific cross-lingual adaptation of an LM.

2.3 Obtaining Stochastic Translation Lexicons

The translation lexicons $P_T(c|e)$ and $P_T(e|c)$ may be created out of an available electronic translation lexicon, with multiple translations of a word being treated as equally likely. Stemming and other morphological analyses may be applied to increase the vocabulary coverage of the translation lexicons.

Alternately, they may also be obtained automatically from a parallel corpus of translated and sentence-aligned Chinese–English text using statistical MT techniques, such as the publicly available GIZA++ tools [Och and Ney 2000]. These tools use several iterations of the EM (expectation maximization) algorithm on increasingly complex word-alignment models to infer, among other translation model parameters, the conditional probabilities $P_T(c|e)$ and $P_T(e|c)$ of words c and e being mutual translations.

Unlike standard MT systems, however, we will apply the translation models to entire articles, one word at a time, to get a *bag of translated words*. A sentence-aligned corpus is therefore not necessary for our purposes and a document-aligned corpus ought, in theory, to suffice for obtaining $P_T(c|e)$ and $P_T(e|c)$. However, obtaining translation probabilities in the same manner using very long (document-sized) sentence-pairs has its own issues. We therefore explore two alternative methods for extracting such translation probabilities in Sections 3 and 4.

Finally, for truly resource-deficient languages, one may obtain a translation lexicon via optical character recognition from a printed bilingual dictionary [cf. Doermann et al. 2002]. This task is arguably easier than obtaining a large LM training corpus.

3. CROSS-LINGUAL LEXICAL TRIGGERS

It seems plausible that most of the information one gets from the cross-lingual unigram LM of (1) is in the form of the altered statistics of topic-specific Chinese words conveyed by the statistics of content-bearing English words in the matching story. The translation lexicon used for obtaining the information, however, is an expensive resource. Yet, if one were only interested in the conditional distribution of Chinese words given some English words, there is no reason to require translation as an intermediate step. In a monolingual setting, the mutual information between lexical-pairs cooccurring anywhere within a long "window" of each other has been used to capture statistical dependencies not covered by N-gram LMs [Rosenfeld 1996; Tillmann and Ney 1997]. We use this inspiration to propose the following notion of cross-lingual lexical triggers.

In a monolingual setting, a pair of words (a, b) is considered a trigger-pair if, given a word-position in a sentence, the occurrence of a in any of the preceding word-positions significantly alters the (conditional) probability that the following word in the sentence is b: a is said to *trigger* b. For example, the occurrence of `either` significantly increases the probability of `or` subsequently in the sentence. The set of preceding word-positions is variably defined to include all words from the beginning of the sentence, paragraph, or document, or is limited to a fixed number of preceding words, limited of course by the beginning of the sentence, paragraph, or document.

In the cross-lingual setting, we consider a pair of words (e, c), $e \in \mathcal{E}$ and $c \in \mathcal{C}$, to be a trigger-pair if, given an English–Chinese pair of aligned documents, the occurrence of e in the English document significantly alters the (conditional) probability that the word c appears in the Chinese document: e is said to trigger c. It is plausible that translation-pairs will be natural candidates for trigger-pairs. It is, however, not necessary for a trigger-pair to also be a translation-pair. For example, the occurrence of `Belgrade` in the English document may trigger the Chinese transliterations of `Serbia` and `Kosovo`, and possibly the translations of `China`, `embassy`, and `bomb`! By inferring trigger-pairs from a document-aligned corpus of Chinese–English articles, we expect to be able to discover semantically or topically related pairs in addition to translation equivalences.

3.1 Identification of Cross-Lingual Triggers

Average mutual information, which measures how much knowing the value of one random variable reduces the uncertainty of about another, has been used to identify trigger-pairs. We compute the average mutual information for every English–Chinese word-pair (e, c) as follows.

Let $\{d_i^E, d_i^C\}$, $i = 1, \ldots, N$, now be a document-aligned training corpus of English–Chinese article-pairs. Let $\#d(e, c)$ denote the *document frequency*, that is, the number of aligned article-pairs, in which e occurs in the English article and c in the Chinese. Let $\#d(e, \bar{c})$ denote the number of aligned article-pairs in

which e occurs in the English articles but c *does not* occur in the Chinese article. Let

$$P(e,c) = \frac{\#d(e,c)}{N} \quad \text{and} \quad P(e,\bar{c}) = \frac{\#d(e,\bar{c})}{N}\,. \tag{6}$$

The quantities $P(\bar{e},c)$ and $P(\bar{e},\bar{c})$ are similarly defined. Next let $\#d(e)$ denote the number of English articles in which e occurs, and define

$$P(e) = \frac{\#d(e)}{N} \text{ and } P(c|e) = \frac{P(e,c)}{P(e)}\,. \tag{7}$$

Similarly, define $P(\bar{e})$, $P(c|\bar{e})$ via the document frequency $\#d(\bar{e}) = N - \#d(e)$; define $P(c)$ via the document frequency $\#d(c)$ and so on. Finally, let

$$\begin{aligned} I(e;c) \;=\; & P(e,c)\log\frac{P(c|e)}{P(c)} + P(e,\bar{c})\log\frac{P(\bar{c}|e)}{P(\bar{c})} \\ & + P(\bar{e},c)\log\frac{P(c|\bar{e})}{P(c)} + P(\bar{e},\bar{c})\log\frac{P(\bar{c}|\bar{e})}{P(\bar{c})}\,. \end{aligned} \tag{8}$$

We propose to select word-pairs with high mutual information as cross-lingual lexical triggers.

There are $|\mathcal{E}|\times|\mathcal{C}|$ possible English–Chinese word-pairs, which may be prohibitively large to search for the pairs with the highest mutual information. We filter out infrequent words in each language, say, words appearing less than five times, then measure $I(e;c)$ for all possible pairs from the remaining words, sort them by $I(e;c)$, and select, say, the top one million pairs.

3.2 Estimating Trigger LM Probabilities

Once we have chosen a set of trigger-pairs, the next step is to estimate a probability $P_{\text{Trig}}(c|e)$ in lieu of the translation probability $P_T(c|e)$ in (1), and a probability $P_{\text{Trig}}(e|c)$ in (4).

Following the maximum likelihood approach proposed by Tillmann and Ney [1997], one could choose the trigger probability $P_{\text{Trig}}(c|e)$ to be based on the unigram frequency of c among Chinese word tokens in that subset of aligned documents d_i^C which have e in d_i^E, namely

$$P_{\text{Trig}}(c|e) = \frac{\sum_{i\,:\,d_i^E \ni e} N_{d_i^C}(c)}{\sum_{c'\in\mathcal{C}}\sum_{i\,:\,d_i^E \ni e} N_{d_i^C}(c')}\,. \tag{9}$$

As an ad hoc alternative to (9), we also use

$$P_{\text{Trig}}(c|e) = \frac{I(e;c)}{\sum_{c'\in\mathcal{C}} I(e;c')} \tag{10}$$

where we set $I(e;c) = 0$ whenever (e,c) is not a trigger-pair, and find it to be somewhat more effective (cf. Section 7.4). Thus (10) is used henceforth in this paper. Analogous to (1), we set

$$P_{\text{Trig-unigram}}\big(c|d_i^E\big) = \sum_{e\in\mathcal{E}} P_{\text{Trig}}(c|e)\hat{P}\big(e|d_i^E\big) \tag{11}$$

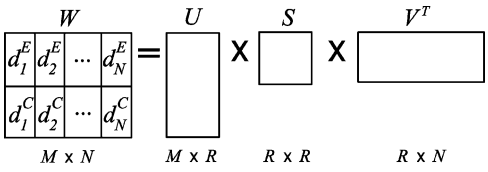

Fig. 2. SVD of a word-document matrix for CL-LSA.

and, again, we build the interpolated model

$$P_{\text{Trig-interpolated}}\left(c_k|c_{k-1}, c_{k-2}, d_i^E\right) = \lambda P_{\text{Trig-unigram}}\left(c_k|d_i^E\right) + (1-\lambda)P(c_k|c_{k-1}, c_{k-2}). \quad (12)$$

4. CROSS-LINGUAL LATENT SEMANTIC ANALYSIS

CL-LSA is a standard automatic technique to extract corpus-based relations between words or documents. It has been successfully used to retrieve, for example, English documents via Chinese queries. We describe it briefly here for completeness.

Assume that a document-aligned Chinese–English bilingual corpus is provided. The first step in CL-LSA is to represent the corpus as a word-document *cooccurrence frequency matrix* W in which each row represents a word in one of the two languages, and each column a document-pair. If the size of the Chinese plus English vocabularies $|\mathcal{C} \cup \mathcal{E}|$ is M, and the corpus has N document-pairs, then W is a $M \times N$ matrix. To begin with, each element w_{ij} of W contains the frequency (count) of the ith word in the jth document-pair. Next, each row of W is weighted by some function, which deemphasizes frequent (function) words in either language, such as the inverse of the number of documents in which the word appears. Next, singular value decomposition (SVD) is performed on W and, for some $R \ll \min\{M, N\}$, we approximate W by its largest R singular values and the corresponding singular vectors as

$$W \approx U \times S \times V^T \quad (13)$$

where columns of U and V are orthogonal left- and right-singular vectors of W, and S is a diagonal matrix whose entries $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_R$ are the corresponding singular values [Berry et al. 1995]. If R is less than the true rank of W, then (13) is the least-squares approximation of W among all rank-R matrices (Figure 2).

In view of this rank-R approximation, the jth column W_{*j} of W, or the document-pair d_j^E and d_j^C, is a linear combination of the columns of $U \times S$, the *weights* for the linear combination being provided by the jth column of V^T. Equivalently the *projection*

$$V_{*j}^T \approx S^{-1} \times U^T \times W_{*j} \quad (14)$$

of the jth column of W on to the *basis* formed by the column-vectors of $U \times S$ provides an R-dimensional representation for d_j^E and d_j^C. Similarly, projecting the ith row of W, which represents the distribution of the ith word, onto the basis formed by row-vectors of $S \times V^T$ provides an R-dimensional representation of words.

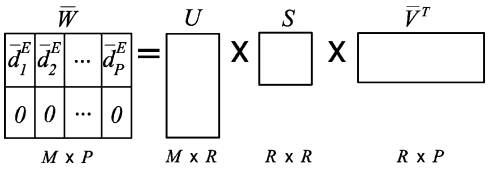

Fig. 3. Folding-in a monolingual corpus into LSA.

It is clear that a Chinese–English translation-pair (e, c), used consistently in the paired-documents, will yield two similar rows in W, and hence their R-dimensional representations will be very close to each other. An elegant consequence of CL-LSA is that other topically or semantically related words also end up having very similar R-dimensional representations, as do documents on those topics. This common R-dimensional representation of words and documents has therefore come to be called *semantic space*.

4.1 Cross-Language IR

CL-LSA provides a way to measure the similarity between a Chinese query and an English document without using a translation lexicon $P_T(e|c)$ as required by (4) in Section 2.2.

Assume that a modest-sized document-aligned Chinese–English corpus has been used to construct the matrices U, S, and V of (13). Next, an additional corpus of English documents is given, a Chinese query is provided, and our task is to find a document from the English corpus, which is topically related to the Chinese query.

We construct a word-document matrix $\overline{W}$ using the English corpus, much as we did earlier with the document-aligned corpus. All rows corresponding to the Chinese vocabulary items have zeros in this matrix, as illustrated in Figure 3. Yet, we proceed to project these documents d_j^E into the semantic space as suggested by (14), and obtain the R-dimensional representations $\overline{V}^T$ for these documents. We then project the Chinese query d_i^C also into the R-dimensional space. Finally, we use the cosine-similarity (5) between the query and document representations to find the English document d_i^E, which is most similar to the Chinese document d_i^C.

4.2 LSA-Derived Translation Probabilities

We use the CL-LSA framework to construct the translation model $P_T(c|e)$ of (1). In the matrix W of (13), each word is represented as a row no matter whether it is English or Chinese. As discussed earlier, projecting these words into R-dimensional space yields rows of U, and the semantic similarity of an English word e and a Chinese word c may be measured by the cosine-similarity of their R-dimensional representation. We extend this notion and construct a word–word translation model, $\forall\, c \in \mathcal{C}$, $\forall\, e \in \mathcal{E}$, as

$$P_{\mathrm{T}}(c|e) = \frac{\mathrm{sim}(c, e)^{\gamma}}{\sum_{c' \in \mathcal{C}} \mathrm{sim}(c', e)^{\gamma}} \tag{15}$$

where $\gamma \gg 1$, as suggested by Coccaro and Jurafsky [1998]. This leads to an LSA-based alternative to (2), as

$$P_{\text{LSA-interpolated}}\left(c_k|c_{k-1}, c_{k-2}, d_i^E\right) = \lambda P_{\text{LSA-unigram}}\left(c_k|d_i^E\right) + (1-\lambda)P(c_k|c_{k-1}, c_{k-2}). \quad (16)$$

Remark 1. We note that *our technique* for exploiting a large English corpus to improve Chinese LMs, as well as the use of a document-aligned Chinese–English corpus to overcome the need for a translation lexicon, *stands in direct competition to a recent method* proposed by Kim and Khudanpur [2003]. We make direct comparisons with their results by choosing identical training and test conditions, as described below.

5. TOPIC-DEPENDENT LANGUAGE MODELS

The combination of the story-dependent unigram models (1) and (11) with a story-independent trigram model using linear interpolation, as described above, seems to be a good choice as they are complementary. There are several references showing effectiveness of monolingual topic-dependent language models [cf. e.g., Iyer and Ostendorf 1999], and our approach may be regarded as similar to the monolingual topic-dependent language model. This motivates us to construct topic-dependent LMs and contrast their performance with our models.

To this end, we use a well-known K-means clustering algorithm. First, we represent each Chinese article in the training corpus by a bag-of-words vector. Then, we use random initialization to seed the algorithm, and a standard TF-IDF weighted cosine-similarity as the "metric" for clustering. We perform a few iterations of the K-means algorithm, and deem the resulting clusters as representing different *topics*. We then use a bag-of-words *centroid* created from all the articles in a cluster to represent each topic. Topic-dependent trigram LMs, denoted $P_j(c_k|c_{k-1}, c_{k-2})$, are also computed for each topic exclusively from the articles in the jth cluster, $1 \le j \le K$.

Each Mandarin test story is represented by a bag-of-words vector $\hat{P}(c|d_i^C)$ generated from the first-pass ASR output, and the topic-centroid t_i having the highest TF-IDF weighted cosine-similarity to it is chosen as the topic of d_i^C. Topic-dependent LMs are then constructed for each story d_i^C as

$$P_{\text{Topic-trigram}}(c_k|c_{k-1}, c_{k-2}, t_i) = \lambda P_{t_i}(c_k|c_{k-1}, c_{k-2}) + (1-\lambda)P(c_k|c_{k-1}, c_{k-2}) \quad (17)$$

and used in a second pass of recognition.

One may wonder, as one of the anonymous referees did, whether a topic-unigram $P_{t_i}(c_k)$ would make for a fairer comparison with the models of (12) and (16). While this may be partially true, we do not make a comparison with such a model. We believe based on results reported in Khudanpur and Kim [2002] that the topic-trigram of (17) is a better model, making for an informative, even if unfair comparison.

6. TRAINING AND TEST CORPORA

We investigate the use of the techniques described above for improving ASR performance on Mandarin news broadcasts using English newswire texts. We have

chosen the experimental ASR setup created in the 2000 Johns Hopkins Summer Workshop to study Mandarin pronunciation modeling, extensive details about which are available in Fung et al. [2000]. The acoustic training data ($\sim$10 h) for their ASR system was obtained from the 1997 Mandarin Broadcast News distribution, and context-dependent state-clustered models were estimated using initials and finals as subword units.

6.1 Parallel Corpus

We use the Hong Kong News (*HKNews*) text corpus as our parallel text—this is used for training of GIZA++, construction of trigger-pairs and the cross-lingual mate retrieval experiment. The corpus contains 18,147 *document-aligned* documents.[1] Chinese–English article-pairs, dating from July 1997 to April 2000, released by the Information Services Department of Hong Kong Special Administrative Region of the People's Republic of China; through the Linguistic Data Consortium [LDC 2000]. After removing a few articles containing non-standard Chinese characters, we divide the corpus, by random selection, into 16,010 article-pairs for training, 750 pairs for testing. This results in a 4.2M-word Chinese training set, and a 177K-word Chinese test set. By comparison, the English counterpart of our Chinese training corpus contains 4.3M-words in the training set and 182K-words in the test set.

6.2 Monolingual Corpora

Two Chinese text corpora and an English corpus are used to estimate LMs in our experiments. A vocabulary $\mathcal{C}$ of 51K Chinese words, used in the ASR system, is also used to segment the training text. This vocabulary gives an out-of-vocabulary (OOV) rate of 5% on the test data.

XINHUA: We use the Xinhua News corpus of about 13 million words to represent the scenario when the amount of available LM training text borders on adequate, and estimate a baseline trigram LM for one set of experiments.

HUB-4NE: We also estimate a trigram model from *only* the 96K words in the transcriptions used for training acoustic models in our ASR system. This corpus represents the scenario when little or no additional text is available to train LMs.

NAB-TDT: English text contemporaneous with the test data is often easily available. For our test set, described below, we select (from the North American News Text corpus) articles published in 1997 in The Los Angeles Times and The Washington Post, and articles from 1998 in the New York Times and the Associated Press news service (from TDT-2 corpus). This amount to a collection of roughly 45,000 articles containing about 30 million words of English text; a modest collection by CLIR standards.

Our ASR test set is a subset [Fung et al. 2000] of the NIST 1997 and 1998 HUB-4NE benchmark tests, containing Mandarin news broadcasts from three sources for a total of about 9800 words. We generate two sets of lattices using

[1]This is actually a *sentence-aligned* corpus, which satisfies the requirement to train GIZA++, however, we use *document-level* alignments for trigger and LSA experiments ignoring *sentence-level* alignments.

the baseline acoustic models and *bigram* LMs estimated from XINHUA and HUB-4NE. All our LMs are evaluated by rescoring 300-best lists extracted from these two sets of lattices. The 300-best lists from the XINHUA bigram LM are used in all XINHUA experiments, and those from the HUB-4NE bigram LM in all HUB-4NE experiments. We report both word error rates (WERs) and character error rates (CERs), the latter being independent of any difference in segmentation of the ASR output and reference transcriptions. The oracle best (worst) WERs for 300-best list are 34.4% (94.4%) for the XINHUA LM and 39.7% (95.5%) for the HUB-4NE LM.

7. EXPERIMENTAL RESULTS

7.1 Cross-Lingual Mate Retrieval

In order to see how CL-LSA performs compared to vector-based IR, Dumais et al. [1997] conducted a cross-language mate retrieval experiment. Starting from a parallel corpus, they divide it into training and test data, use training data to build SVD matrices, U, S, and V, and measure the performance of how well each IR system finds corresponding mate (aligned document) in the other language for a given test document in one language. According to their French–English experiments, CL-LSA shows superior results to vector-based IR (98.4% vs. 48.5% on average in terms of accuracy). However, since there is almost no common word between French and English documents, we cannot expect to have a good result from the vector-based IR system. Nevertheless, the accuracy of 48.5% reflects the fact that French and English are quite similar to each other.

We follow their steps for CL-LSA experiments for the Chinese–English parallel collection (HKNews). We use the training set for SVD decomposition and project the English test set documents into low R-dimensional space. For a given Chinese test document, again we project it into R-dimensional space, find the most similar one among English test document by comparing it in R-dimensional space. We use cosine-similarity score to measure the similarity between each Chinese and English document. Finally, we measure how accurately CL-LSA selects the corresponding English mate.

For the vector-based IR system, rather than doing CLIR where there are almost no common words, we use statistical MT before applying vector-based IR. First, we get the *translated* bag-of-words English document by using the GIZA++ translation dictionary, $P_T(e|c)$, from a given Chinese test document. Once we have a *translated* bag-of-words English document (from a Chinese query document), we can apply standard vector-based IR from then on. In other words, we find the most similar document from the English test set again by measuring cosine-similarity.

Table I shows the accuracy of the cross-lingual mate retrieval experiment. Contrary to the results in Dumais et al. [1997], but not surprisingly, vector-based IR performs quite well giving 92.4% of accuracy even though we are testing on much harder language pairs (Chinese–English). This is purely due to the use of fairly well-tuned translation dictionary $P_T(e|c)$ before applying vector-based IR.

Table I. Cross-Lingual Mate Retrieval Results

Model	Accuracy
Vector-based IR	92.4%
CL-LSA (R = 76)	70.9%
CL-LSA (R = 207)	80.4%
CL-LSA (R = 447)	87.6%
CL-LSA (R = 693)	90.2%

Table II. Word-Perplexity and ASR WER of LMs Based on Single English Document and Global λ

Language Model	Perp	WER	CER	p-Value
XINHUA trigram	426	49.9%	28.8%	–
CL-interpolated	375	49.5%	28.7%	0.208
HUB-4NE trigram	1195	60.1%	44.1%	–
CL-interpolated	750	59.3%	43.7%	$<$0.001

Our first CL-LSA result with small low rank ($R = 76$) performs far below vector-based IR. Assuming that the low rank of 76 is too small for our database, we test the accuracy by changing the low rank of SVD. As we increase the low rank, the accuracy increases and we get the similar performance to vector-based IR when R is 693.[2] From this experiment, we can conclude that even without using the translation dictionary, CL-LSA performs close to cross-lingual vector-based IR after MT.

7.2 Baseline ASR Performance of Cross-Lingual LMs

Since our eventual goal is to improve ASR performance, we proceed by rescoring the 300-best lists from the bigram lattices with trigram models. For each test story d_i^C, we perform CLIR using the first pass ASR output to choose the most similar English document d_i^E from NAB-TDT. Then we create the cross-lingual unigram model of (1). We also find the interpolation weight λ, which maximizes the likelihood of the 1-best hypotheses of all test utterances from the first ASR pass. All translation tables, $P_T(e|c)$ for CLIR and $P_T(c|e)$ for back translation into the target language, Chinese, used here come from GIZA++ translation tables. Table II shows the perplexity and WER for XINHUA and HUB-4NE.

All p-values reported in this paper are based on the standard NIST MAPSSWE test [Pallett et al. 1990], and indicate the statistical significance of a WER improvement over the corresponding trigram baseline, unless otherwise specified.

Evidently, the improvement brought by CL-interpolated LM is not statistically significant on XINHUA. On HUB-4NE, however, where Chinese LM text is scarce, the CL-interpolated LM delivers considerable benefits via the large English corpus.

[2]Due to the memory limitation, 693 was the maximum of our system and we were not able to go further.

7.3 Likelihood-Based Story-Specific Selection of Interpolation Weights and the Number of English Documents per Mandarin Story

The experiments above naïvely used the one most similar English document for each Mandarin story, and a global λ in (2), no matter how similar the best matching English document is to a given Mandarin news story. Rather than choosing one most similar English document from *NAB-TDT*, it stands to reason that choosing more than one English document may be helpful if many have a high similarity score, and perhaps not using even the best matching document may be fruitful if the match is sufficiently poor. It may also help to have a greater interpolation weight λ for stories with good matches, and a smaller λ for others. For experiments in this subsection, we select a different λ for each test story, again based on maximizing the likelihood of the 1-best output given a CL-unigram model. The other issue then is the choice and the number of English documents to translate.

N-best documents: One could choose a predetermined number N of the best matching English documents for each Mandarin story. We experimented with values of 1, 10, 30, 50, 80, and 100, and found that $N = 30$ gave us the best LM performance, but only marginally better than $N = 1$ as described above. Details are omitted, as they are uninteresting.

All documents above a similarity threshold: The argument against always taking a predetermined number of the best matching documents may be that it ignores the goodness of the match. An alternative is to take all English documents whose similarity to a Mandarin story exceeds a certain predetermined threshold. As this threshold is lowered, starting from a high value, the *order* in which English documents are selected for a particular Mandarin story is the same as the order when choosing the N-best documents, but the number of documents selected now varies from story to story. It is possible that for some stories, even the best matching English document falls below the threshold at which other stories have found more than one good match. We experimented with various thresholds, and found that while a threshold of 0.12 gives us the lowest perplexity on the test set, the reduction is insignificant. This points to the need for a story-specific strategy for choosing the number of English documents, instead of a global threshold.

Likelihood-based selection of the number of English documents: Figure 4 shows the perplexity of the reference transcriptions of two typical test stories under the LM (2) as a function of the number of English documents chosen for creating (1). As the curve shows, the perplexity varies according to the number of English documents (d_i^E) and obviously, the best performance is achieved at different points for each story. For each choice of the number of English documents, also, the interpolation weight λ in (2) is chosen to maximize the likelihood (also shown) of the first pass output. This suggests that choosing the number of English documents to maximize the likelihood of the first pass ASR output is a good strategy.

For each Mandarin test story, we choose the 1000-best-matching English documents and divide the *dynamic range* of their similarity scores evenly into 10 intervals. Next, we choose the documents in the top one-tenth of the *range of similarity scores*, not necessarily the top 100 documents, compute $P_{\text{CL-unigram}}(c|d_i^E)$,

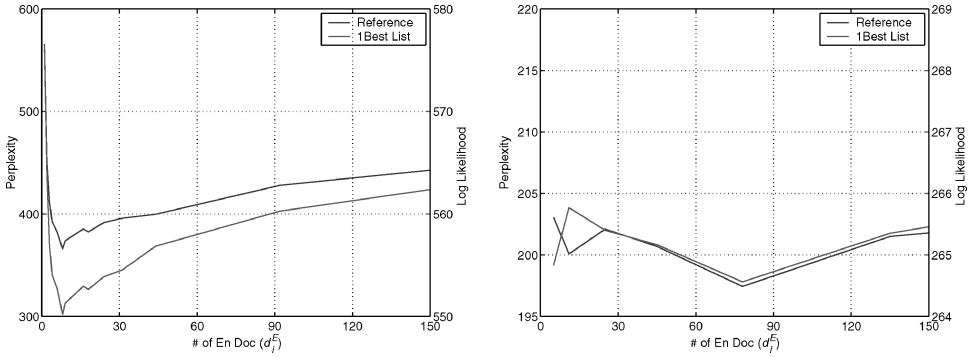

Fig. 4. Perplexity of the reference transcription and the likelihood of the ASR output versus number of d_i^E for typical test stories.

Table III. Word-Perplexity and ASR WER of LMs with a Likelihood-Based Story-Specific Selection of the Number of English Documents d_i^E's and Interpolation Weight λ for Each Mandarin Story

Language Model	Perp	WER	CER	p-Value
XINHUA trigram	426	49.9%	28.8%	–
CL-interpolated	346	48.8%	28.4%	<0.001
Topic-trigram	381	49.1%	28.4%	0.003
Topic + CL-interpolated	326	48.5%	28.2%	<0.001
HUB-4NE trigram	1195	60.1%	44.1%	–
CL-interpolated	630	58.8%	43.1%	<0.001
Topic-trigram	1122	60.0%	44.1%	0.660
Topic + CL-interpolated	631	59.0%	43.3%	<0.001

determine the λ in (2) that maximizes the likelihood of the first pass output of only the utterances in that story, and record this likelihood. We repeat this with documents in the top two-tenth of the range of similarity scores, the top three-tenth, and so on, and obtain the likelihood as a function of the similarity threshold. We choose the threshold that maximizes the likelihood of the first pass output. Thus the number of English documents d_i^E in (1), as well as the interpolation weight λ in (2), are chosen dynamically for each Mandarin story to maximize the likelihood of the ASR output. Table III shows ASR results for this *likelihood-based story-specific adaptation* scheme.

Note that significant WER improvements are obtained from the CL-interpolated LM using likelihood-based story-specific adaptation even for the case of the XINHUA LM. Furthermore, the performance of the CL-interpolated LM is even better than the topic-dependent LM. This is remarkable, since the CL-interpolated LM is based on unigram statistics from English documents, while the topic-trigram LM is based on trigram statistics. We believe that the contemporaneous and story-specific nature of the English document leads to its relatively higher effectiveness. Our conjecture, that the *contemporaneous* cross-lingual statistics and *static* topic-trigram statistics are complementary, is supported by the significant further improvement in WER obtained by the interpolation of the two LMs, as shown on the last line for XINHUA.

Table IV. Word-Perplexity and ASR WER Comparisons of CL-Interpolated, Trigger-Interpolated, and LSA-Interpolated Models

Language Model	Perp	WER	CER	p-Value
XINHUA Trigram	426	49.9%	28.8%	–
CL-interpolated	346	48.8%	28.4%	<0.001
Trig-interpolated	367	49.1%	28.6%	0.004
LSA-interpolated	364	49.3%	28.9%	0.043
Trig+LSA-interpolated	351	49.0%	28.7%	0.002
HUB-4NE Trigram	1195	60.1%	44.1%	–
CL-interpolated	630	58.8%	43.1%	<0.001
Trig-interpolated	727	58.8%	43.3%	<0.001
LSA-interpolated	695	58.6%	43.1%	<0.001
Trig+LSA-interpolated	686	58.7%	43.2%	<0.001

The significant gain in ASR performance in the resource-deficient HUB-4NE case are obvious. The small size of the HUB-4NE corpus makes topic-models ineffective.

7.4 Comparison of Cross-Lingual Triggers and CL-LSA with Stochastic Translation Dictionaries

Once we select cross-lingual trigger-pairs as described in Section 3, $P_T(c|e)$ in (1) is replaced by $P_{\text{Trig}}(c|e)$ of (10), and $P_T(e|c)$ in (4) by $P_{\text{Trig}}(e|c)$. Therefore, given a set of cross-lingual trigger-pairs, the trigger-based models are free from requiring a translation lexicon. Furthermore, a document-aligned comparable corpus is all that is required to construct the set of trigger-pairs. We, otherwise, follow the same experimental procedure as above.

For CL-LSA test, LSA is applied in two parts. First, we do CLIR with LSA without even using $P_T(e|c)$ as described in Section 4.1. Once CLIR is done, again LSA is applied to build $P_{\text{LSA}-\text{unigram}}(c|e)$ as a replacement for $P_T(c|e)$ as explained in Section 4.2. While the other models require translation tables in both direction, $P_T(e|c)$ and $P_T(c|e)$, LSA requires only one directional translation table, $P_T(c|e)$.

As Table IV shows, the trigger-based model (Trig-interpolated) performs only slightly worse than the CL-interpolated model. Also, the LSA-based model shows similar results to the trigger-based model. One explanation for this degradation is that the CL-interpolated model is trained from the sentence-aligned corpus while the trigger-based model and the LSA-based model are from the document-aligned corpus. There are two steps that could be affected by this difference, one being CLIR and the other being the translation of the d_i^E's into Chinese. Some errors in CLIR may however be masked by our *likelihood-based story-specific adaptation* scheme, since it finds optimal retrieval settings, dynamically adjusting the number of English documents as well as the interpolation weight, even if CLIR performs somewhat suboptimally. Furthermore, a document-aligned corpus is much easier to build. Thus, a much bigger and more reliable comparable corpus may be used, and eventually more accurate trigger-pairs will be acquired.

Triggers (9) versus (10): We compare the alternative $P_{\text{Trig}}(\cdot|\cdot)$ definitions (9) and (10) for replacing $P_T(\cdot|\cdot)$ in (1). The resulting CL-interpolated LM (2) yields a perplexity of 370 on the XINHUA test set using (9), compared to 367 using (10).

Table V. Word-Perplexity and ASR WER Comparisons Using a MRD

Language Model	Perp	WER	CER
XINHUA Trigram	426	49.9%	28.8%
MRD-interpolated	387	49.9%	29.0%
HUB-4NE Trigram	1195	60.1%	44.1%
MRD-interpolated	770	60.1%	44.1%

Similarly, on the HUB-4NE test set, using (9) yields 736, while (10) yields 727. Therefore, (10) has been used throughout.

Finally, we build an interpolated model (Trig+LSA-interpolated) from the trigger-based model and the LSA-based model hoping to achieve similar results to CL-interpolated model. In terms of perplexities, we achieve much closer results (351 vs. 346 in XINHUA and 630 vs. 686 in HUB-4NE). Even though there are slight differences in terms of WER (28.4% vs. 28.7% in XINHUA and 43.1% vs. 43.2% in HUB-4NE), the differences are not statistically significant.[3]

We note with some satisfaction that we extract comparable information from the *document-aligned* corpus to a stochastic translation lexicon, which cannot be acquired unless a *sentence-aligned* corpus is available. This is achieved by using simple trigger-pairs selected on the basis of mutual information and the LSA-based model.

7.5 Comparison of Stochastic Translation with Manually Created Dictionaries

While the notion of inducing stochastic translation lexicons from aligned bilingual text is appealing, it is also worth investigating the scenario in which a modest-sized, manually created, machine-readable dictionary (MRD) is available. We used a widely available Chinese–English translation lexicon[4] for this purpose. In order to make meaningful comparisons, we used the very same procedure for CLIR and CL-unigram construction as done for the models reported in Table IV, but we used the MRD in place of a stochastic translation lexicon. In other words, instead of using the translation probabilities derived using GIZA++ for the CL-interpolated LM, cross-lingual triggers for the Trig-interpolated LM and CL-LSA for the LSA-interpolated LM respectively, we used 18K English-to-Chinese entries and 24K Chinese-to-English entries from the LDC translation lexicon. It is clear from the results reported in Table V that while the MRD leads to a reduction in perplexity, no reduction in WER is obtained.

This *should not* lead the reader to conclude that a MRD can be completely dispensed-with for cross-lingual applications. A MRD is often crucial for obtaining bilingual text alignment of acceptable quality. Recall that our techniques are predicated on the capability to obtain such bilingual text.

We instead conclude that the best application of a MRD in a resource-deficient language is in the process of obtaining either sentence- or document-aligned bilingual text from available text repositories. A stochastic

[3]The p-values measured between CL-interpolated and Trig+LSA-interpolated are 0.58 and 0.79 for XINHUA and HUB-4NE, respectively.

[4]See `http://www.ldc.upenn.edu/Projects/Chinese/LDC_ch.htm`.

translation lexicon derived from such text should then be used in the actual cross-lingual application.

8. CONCLUSIONS

We have demonstrated a statistically significant improvement in ASR WER (1.4% absolute) and in perplexity (23%) by exploiting cross-lingual side-information even when a nontrivial amount of training data *is* available, as seen on the 13M-word XINHUA corpus. Our methods are even more effective when LM training text is hard to come by in the language of interest: 47% reduction in perplexity and 1.3% absolute in WER as seen on the 96K-word HUB-4NE corpus. Most of these gains come from the optimal choice of adaptation parameters. The ASR test data we used in our experiments are derived from a different news source than the text corpus on which the translation and trigger models are trained, which points to the robustness of the inferred statistics.

The techniques work even when the bilingual corpus is merely document-aligned, which is a realistic reflection of the situation in a resource-deficient language. Effectively, we have proposed methods to build cross-lingual language models, which do not require MT. By using mutual information statistics and latent semantic analysis from a *document-aligned* corpus, we can extract a significant amount of information for language modeling. Experimental results show that performance statistically equal to previously published methods predicated on MT capabilities can be achieved by our methods.

We are developing maximum entropy models to more effectively combine the multiple information sources we have used in our experiments, and expect to report the results in the near future.

REFERENCES

BERRY, M. ET AL. 1995. Using linear algebra for intelligent information retrieval. *SIAM Rev. 37,* 4, 573–595.

BROWN, P., PIETRA, S. D., PIETRA, V. D., AND MERCER, R. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist. 19,* 2, 269–311.

BYRNE, W. ET AL. 2000. Towards language independent acoustic modeling. In *Proceedings of the ICASSP*, vol. 2. 1029–1032.

COCCARO, N. AND JURAFSKY, D. 1998. Towards better integration of semantic predictors in statistical language modeling. In *Proceedings of the ICSLP*, Sydney, Australia, vol. 6. 2403–2406.

DOERMANN, D. ET AL. 2002. Lexicon acquisition from bilingual dictionaries. In *Proceedings of the SPIE Photonic West Article Imaging Conference*, San Jose, CA. 37–48.

DUMAIS, S. ET AL. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Spring Symposuim on Cross-Language Text and Speech Retrieval*.

FUNG, P. ET AL. 2000. Pronunciation modeling of Mandarin casual speech. *2000 Johns Hopkins Summer Workshop*. Available at http://www.clsp.jhu.edu/ws2000/groups/mcs.

IYER, R. AND OSTENDORF, M. 1999. Modeling long-distance dependence in language: topic-mixtures vs dynamic cache models. *IEEE Trans. Speech Audio Process. 7*, 30–39.

KHUDANPUR, S. AND KIM, W. 2002. Using cross-language cues for story-specific language modeling. In *Proceedings of the ICSLP*, Denver, CO, vol. 1. 513–516.

KIM, W. AND KHUDANPUR, S. 2003. Cross-lingual lexical triggers in statistical language modeling. In *Proceedings of the EMNLP*, Sapporo, Japan. 17–24.

KIRCHHOFF, K. ET AL. 2002. Novel speech recognition models for Arabic. *2002 Johns Hopkins Summer Workshop*. Available at http://www.clsp.jhu.edu/ws2002/groups/arabic.

LDC. 2000. Hong Kong news parallel text corpus. Available through the Linguistic Data Consortium. http://www.ldc.upenn.edu/Catalog/LDC2000T46.html.

OCH, F. AND NEY, H. 2000. Improved statistical alignment models. In *ACL*, Hongkong, China. 440–447.

PALLETT, D., FISHER, W., AND FISCUS, J. 1990. Tools for the analysis of benchmark speech recognition tests. In *Proceedings of the ICASSP*, Alburquerque, NM, vol. 1. 97–100.

ROSENFELD, R. 1996. A maximum entropy approach to adaptive statistical language modeling. *Comput. Speech Lang. 10*, 187–228.

SCHULTZ, T. AND WAIBEL, A. 1998. Language independent and language adaptive large vocabulary speech recognition. In *Proceedings of the ICSLP*, Sydney, Australia, vol. 5. 1819–1822.

TILLMANN, C. AND NEY, H. 1997. Word trigger and the EM algorithm. In *Proceedings of the Workshop Computational Natural Language Learning (CoNLL 97)*, Madrid, Spain. 117–124.

YAROWSKY, D., NGAI, G., AND WICENTOWSKI, R. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT*, Santa Monica, CA. 109–116.

Received August 2003; revised June 2004; accepted July 2004