Cross-Lingual Lexical Triggers in Statistical Language Modeling*

Woosung Kim
The Johns Hopkins University
3400 N. Charles St., Baltimore, MD
woosung@cs.jhu.edu

Sanjeev Khudanpur
The Johns Hopkins University
3400 N. Charles St., Baltimore, MD
khudanpur@jhu.edu

Abstract

We propose new methods to take advantage of text in resource-rich languages to sharpen statistical language models in resource-deficient languages. We achieve this through an extension of the method of lexical triggers to the cross-language problem, and by developing a likelihood-based adaptation scheme for combining a trigger model with an N-gram model. We describe the application of such language models for automatic speech recognition. By exploiting a side-corpus of contemporaneous English news articles for adapting a static Chinese language model to transcribe Mandarin news stories, we demonstrate significant reductions in both perplexity and recognition errors. We also compare our cross-lingual adaptation scheme to monolingual language model adaptation, and to an alternate method for exploiting cross-lingual cues, via cross-lingual information retrieval and machine translation, proposed elsewhere.

1 Data Sparseness in Language Modeling

Statistical techniques have been remarkably successful in automatic speech recognition (ASR) and natural language processing (NLP) over the last two decades. This success, however, depends crucially on the availability of accurate and large amounts of suitably annotated training data and it is difficult to build a usable statistical model in their absence. Most of the success, therefore, has been witnessed in the so called *resource-rich* languages. More recently, there has been an increasing interest in languages such as Mandarin and Arabic for ASR and NLP, and data resources are being created for them at considerable cost. The data-resource bottleneck, however, is likely to remain for a majority of the world's languages in the foreseeable future.

Methods have been proposed to bootstrap acoustic models for ASR in resource deficient languages by reusing acoustic models from resource-rich languages (Schultz and Waibel, 1998; Byrne et al., 2000). Morphological analyzers, noun-phrase chunkers, POS taggers, etc., have also been developed for resource deficient languages by exploiting translated or *parallel* text (Yarowsky et al., 2001). Khudanpur and Kim (2002) recently proposed using cross-lingual information retrieval (CLIR) and machine translation (MT) to improve a statistical language model (LM) in a resource-deficient language by exploiting copious amounts of text available in resource-rich languages. When transcribing a news story in a resource-deficient language, their core idea is to use the first pass output of a rudimentary ASR system as a query for CLIR, identify a contemporaneous English document on that news topic, followed by MT to provide a rough translation which, even if not fluent, is adequate to update estimates of word frequencies and the LM vocabulary. They report up to a 28% reduction in perplexity on Chinese text from the Hong Kong News corpus.

*This research was supported by the National Science Foundation (via Grant No ITR-0225656 and IIS-9982329) and the Office of Naval Research (via Contract No N00014-01-1-0685).

In spite of their considerable success, some shortcomings remain in the method used by Khudanpur and Kim (2002). Specifically, stochastic translation lexicons estimated using the IBM method (Brown et al., 1993) from a fairly large *sentence-aligned* Chinese-English parallel corpus are used in their approach — a considerable demand for a resource-deficient language. It is suggested that an easier-to-obtain *document-aligned* comparable corpus may suffice, but no results are reported. Furthermore, for each Mandarin news story, the single best matching English article obtained via CLIR is translated and used for priming the Chinese LM, no matter how good the CLIR similarity, nor are other well-matching English articles considered. This issue clearly deserves further attention. Finally, ASR results are not reported in their work, though their proposed solution is clearly motivated by an ASR task. We address these three issues in this paper.

Section 2 begins, for the sake of completeness, with a review of the cross-lingual story-specific LM proposed by Khudanpur and Kim (2002). 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. After a brief detour to describe topic-dependent LMs in Section 4, a description of the ASR task is provided in Section 5, and ASR results on Mandarin Broadcast News are presented in Section 6. The issue of how many English articles to retrieve and translate into Chinese is resolved by a likelihood-based scheme proposed in Section 6.1.

2 Cross-Lingual Story-Specific LMs

For the sake of illustration, consider the task of sharpening a Chinese language model for transcribing Mandarin news stories by using a large corpus of contemporaneous English newswire text. Mandarin Chinese is, of course, not resource-deficient for language modeling — 100s of millions of words are available on-line. However, we choose it for our experiments partly because it is sufficiently different from English to pose a real challenge, and because the availability of large text corpora in fact permits us to simulate controlled resource deficiency.

Let $d_1^C, \ldots, d_N^C$ denote the text of N *test stories* to be transcribed by an ASR system, and let $d_1^E, \ldots, d_N^E$ denote their corresponding or *aligned* English newswire articles. 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. This 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 given.

Assume further that we have at our disposal a stochastic translation dictionary — 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 LM

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, $\forall\, c \in \mathcal{C}$,

$$P_{\mathrm{CL-unigram}}(c|d_i^E) = \sum_{e \in \mathcal{E}} P_T(c|e)\hat{P}(e|d_i^E)\,, \quad (1)$$

would be a good unigram model for the i-th Mandarin story d_i^C. We use 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

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

of the cross-lingual unigram model (1) with a static trigram model for Chinese, where the interpolation weight λ may be chosen off-line 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.

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 CLIR system to find an English document d_i^E, an MT system

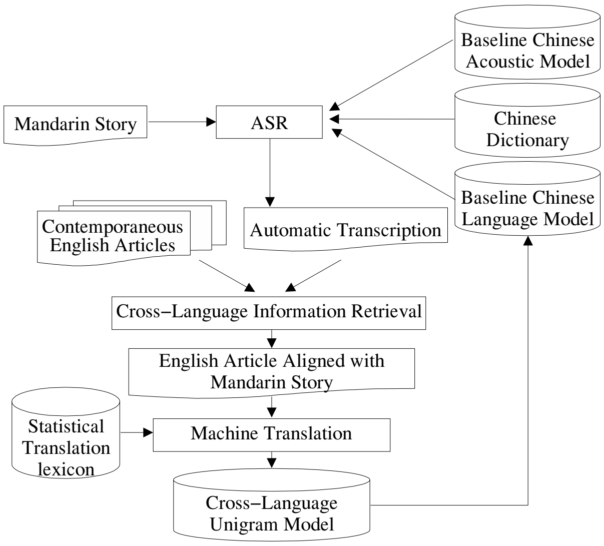

Figure 1: Story-Specific Cross-Lingual Adaptation of a Chinese Language Model using English Text.

computes the statistic of (1), and the ASR system uses the LM of (2) in a second pass.

2.2 Obtaining Matching English Documents

To illustrate how one may obtain the English document d_i^E to match a 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, $\forall\, e \in \mathcal{E}$,

$$P_{\text{CL}-\text{unigram}}(e|d_i^C) = \sum_{c\in\mathcal{C}} P_T(e|c)\hat{P}(c|d_i^C)\,, \quad (3)$$

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

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

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

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 machine translation techniques, such as the publicly available GIZA++ tools (Och and Ney, 2000), as done by Khudanpur and Kim (2002). Unlike standard MT systems, however, we apply the translation models to entire articles, one word at a time, to get a *bag of translated words* — cf. (1) and (3).

Finally, for truly resource deficient languages, one may obtain a translation lexicon via optical character recognition from a printed bilingual dictionary (cf. Doerman 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 co-occurring 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. E.g. 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. E.g., 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 infering 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*, i.e., 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

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

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

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

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 $\#d(c)$, etc. Finally, let

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

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 5 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 1 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 (3).

Following the maximum likelihood approach proposed by Tillman 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

$$P_{\text{Trig}}(c|e) = \frac{\sum_{i\,:\,d_i^E \ni e} N_{d_i^C}(c)}{\sum_{c' \in \mathcal{C}} \sum_{i\,:\,d_i^E \ni e} N_{d_i^C}(c')}. \quad (4)$$

As an ad hoc alternative to (4), we also use

$$P_{\text{Trig}}(c|e) = \frac{I(e; c)}{\sum_{c' \in \mathcal{C}} I(e; c')}, \quad (5)$$

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 6.2). Thus (5) is used henceforth in this paper. Analogous to (1), we set

$$P_{\text{Trig-unigram}}(c|d_i^E) = \sum_{e \in \mathcal{E}} P_{\text{Trig}}(c|e) \hat{P}(e|d_i^E), \quad (6)$$

and, again, we build the interpolated model

$$\begin{aligned} & P_{\text{Trig-interpolated}}(c_k|c_{k-1}, c_{k-2}, d_i^E) = \quad (7) \\ & \lambda P_{\text{Trig-unigram}}(c_k|d_i^E) + (1 - \lambda) P(c_k|c_{k-1}, c_{k-2}). \end{aligned}$$

4 Topic-Dependent Language Models

The linear interpolation of the story-dependent unigram models (1) and (6) with a story-independent trigram model, as described above, is very reminiscent of monolingual topic-dependent language models (cf. e.g. (Iyer and Ostendorf, 1999)). This motivates us to construct topic-dependent LMs and contrast their performance with these models.

To this end, we represent each Chinese article in the training corpus by a bag-of-words vector, and cluster the vectors using a standard K-means algorithm. 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 j-th cluster, $1 \leq j \leq K$.

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

$$P_{\text{Topic}-\text{trigram}}(c_k|c_{k-1}, c_{k-2}, t_i) = \qquad (8)$$

$$\lambda P_{t_i}(c_k|c_{k-1}, c_{k-2}) + (1-\lambda) P(c_k|c_{k-1}, c_{k-2})$$

and used in a second pass of recognition.

Alternatives to topic-dependent LMs for exploiting long-range dependencies include cache LMs and monolingual lexical triggers; both unlikely to be as effective in the presence of significant ASR errors.

5 ASR 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 hours) 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. 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 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 amounts 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 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 (WER) and character error rates (CER), the latter being independent of any difference in segmentation of the ASR output and reference transcriptions.

6 ASR Performance of Cross-Lingual LMs

We begin 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. Table 1 shows the perplexity and WER for XINHUA and HUB-4NE.

Language model	Perp	WER	p-value
XINHUA trigram	426	49.9%	–
CL-interpolated	375	49.5%	0.208
HUB-4NE trigram	1195	60.1%	–
CL-interpolated	750	59.3%	< 0.001

Table 1: Word-Perplexity and ASR WER of LMs based on single English document and global λ.

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 text is scarce, the CL-interpolated LM delivers considerable benefits via the large English corpus.

6.1 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 2 shows the perplexity of the reference transcriptions of one typical test story under the LM (2) as a function of the number of English documents chosen for creating (1). For each choice of the number of English documents, 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 $\frac{1}{10}$-th of the *range of similarity scores*, not necessarily the top-100 documents, compute $P_{\mathrm{CL-unigram}}(c|d_i^E)$, 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 $\frac{2}{10}$-th of the range of similarity scores, the top $\frac{3}{10}$-th, etc.,

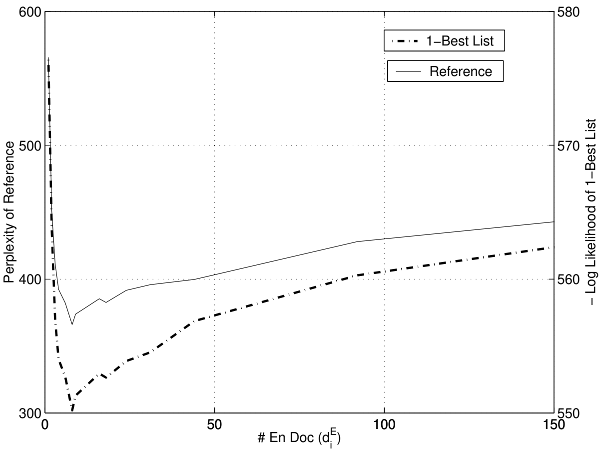

Figure 2: Perplexity of the Reference Transcription and the Likelihood of the ASR Output v/s Number of d_i^E for a Typical Test Story.

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 2 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.

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.

6.2 Comparison of Cross-Lingual Triggers 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 (5), and $P_T(e|c)$ in (3) 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.

As Table 2 shows, the trigger-based model (Trig-interpolated) performs only slightly worse than the CL-interpolated model. One explanation for this degradation is that the CL-interpolated model is trained from the sentence-aligned corpus while the trigger-based model is from the document-aligned corpus. There are two steps which 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.

We note with some satisfaction that even simple trigger-pairs selected on the basis of mutual information are able to achieve perplexity and WER reductions comparable to a stochastic translation lexicon: the smallest p-value at which the difference between the WERs of the CL-interpolated LM and the Trig-interpolated LM in Table 2 would be significant is 0.4 for XINHUA and 0.7 for HUB-4NE.

Triggers (4) vs (5): We compare the alternative $P_{\text{Trig}}(\cdot|\cdot)$ definitions (4) and (5) 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 (4), compared to 367 using (5). Similarly, on the HUB-4NE test set, using (4) yields 736, while (5) yields 727. Therefore, (5) has been used throughout.

XINHUA					HUB-4NE			
Perp	WER	CER	p-value	Language model	Perp	WER	CER	p-value
426	49.9%	28.8%	–	Baseline Trigram	1195	60.1%	44.1%	–
381	49.1%	28.4%	0.003	Topic-trigram	1122	60.0%	44.1%	0.660
367	49.1%	28.6%	0.004	Trig-interpolated	727	58.8%	43.3%	< 0.001
346	48.8%	28.4%	< 0.001	CL-interpolated	630	58.8%	43.1%	< 0.001
340	48.7%	28.4%	< 0.001	Topic + Trig-interpolated	730	59.2%	43.5%	0.002
326	48.5%	28.2%	< 0.001	Topic + CL-interpolated	631	59.0%	43.3%	< 0.001
320	48.3%	28.1%	< 0.001	Topic + Trig- + CL-interp.	627	59.0%	43.3%	< 0.001

Table 2: Perplexity and ASR Performance 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.

7 Conclusions and Future Work

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 nontrivial amount of training data is available, as seen on the 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 HUB-4NE corpus. Most of these gains come from the optimal choice of adaptation parameters. The ASR test data we used in our experiments is derived from a different source than the corpus on which the translation and trigger models are trained, and the techniques work even when the bilingual corpus is only document-aligned, which is a realistic reflection of the situation in a resource-deficient language.

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

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):269 – 311.

W. Byrne, P. Beyerlein, J. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and W. Wang. 2000. Towards language independent acoustic modeling. In *Proc. ICASSP*, volume 2, pages 1029 – 1032.

P. Fung et al. 2000. Pronunciation modeling of mandarin casual speech. *2000 Johns Hopkins Summer Workshop*.

D. Doermann et al. 2002. Lexicon acquisition from bilingual dictionaries. In *Proc. SPIE Photonic West Article Imaging Conference*, pages 37–48, San Jose, CA.

R. Iyer and M. Ostendorf. 1999. Modeling long-distance dependence in language: topic-mixtures vs dynamic cache models. *IEEE Transactions on Speech and Audio Processing*, 7:30–39.

S. Khudanpur and W. Kim. 2002. Using cross-language cues for story-specific language modeling. In *Proc. ICSLP*, volume 1, pages 513–516, Denver, CO.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *ACL00*, pages 440–447, Hongkong, China, October.

D. Pallett, W. Fisher, and J. Fiscus. 1990. Tools for the analysis of benchmark speech recognition tests. In *Proc. ICASSP*, volume 1, pages 97–100, Alburquerque, NM.

R. Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, 10:187–228.

T. Schultz and A. Waibel. 1998. Language independent and language adaptive large vocabulary speech recognition. In *Proc. ICSLP*, volume 5, pages 1819–1822, Sydney, Australia.

C. Tillmann and H. Ney. 1997. Word trigger and the em algorithm. In *Proceedings of the Workshop Computational Natural Language Learning (CoNLL 97)*, pages 117–124, Madrid, Spain.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proc. HLT 2001*, pages 109 – 116, San Francisco CA, USA.