ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Speech and Language xxx (2003) xxx–xxx

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Contemporaneous text as side-information in statistical language modeling

Sanjeev Khudanpur [*,1], Woosung Kim [2]

Center for Language and Speech Processing, The Johns Hopkins University, 320 Barton Hall,
3400 North Charles Street, Baltimore, MD 21218, USA

Received 5 February 2003; received in revised form 1 September 2003; accepted 4 September 2003

Abstract

We propose new methods to exploit contemporaneous text, such as on-line news articles, to improve language models for automatic speech recognition and other natural language processing applications. In particular, we investigate the use of text from a resource-rich language to sharpen language models for processing a news story or article in a language with scarce linguistic resources. We demonstrate that even with fairly crude cross-language information retrieval and simple machine translation, one can construct story-specific Chinese language models which exploit cues from a side-corpus of English newswire to significantly improve the performance of language models estimated from a static Chinese corpus. Our investigations cover cases when the amount of available Chinese text is small, and a case when a large Chinese text corpus is available. We examine the effectiveness of our techniques both when the side-corpus contains English documents that are near-translations of the Chinese documents being processed, and when the English side-corpus is merely from contemporaneous and independent news sources. We present experimental results for automatic transcription of speech from the Mandarin Broadcast News corpus.

Keywords: Multi-lingual processing; Statistical language modeling; Automatic speech recognition; Resource-deficient languages; Lexical triggers; Maximum entropy

[*] Corresponding author. Tel.: +1-410-516-4237; fax: +1-410-516-5050.

E-mail addresses: khudanpur@jhu.edu (S. Khudanpur), woosung@cs.jhu.edu (W. Kim).

[1] Sanjeev Khudanpur is with the Department of Electrical & Computer Engineering and the Department of Computer Science.

[2] Woosung Kim is with the Department of Computer Science.

0885-2308/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2003.09.001

1. Exploiting side-information from contemporaneous text

The last decade has seen a dramatic improvement in the capability and performance of automatic systems for processing speech and natural language. This progress may be attributed largely to advances in statistical modeling techniques and procedures for automatic learning from large speech and text corpora. The construction of increasingly accurate and complex stochastic models, in particular, is crucially dependent on the availability of large corpora of transcribed speech and annotated text specific to the language and the application domain. Much of these advances, therefore, have been in languages such as English, French, German and Japanese, and in domains such as air travel information and broadcast news transcription, for which such linguistic resources have been created at considerable cost.

Construction of accurate stochastic models for processing *resource-deficient* languages has recently started receiving attention. A limited amount of linguistic resources can almost always be produced with moderate effort in a language and domain of interest, and we use the term resource-deficiency to imply the lack of tens of hours of orthographically transcribed speech, tens of millions of words of in-domain text, hundreds of thousands of manually parsed sentences, etc.

Methods have been proposed to bootstrap acoustic models for automatic speech recognition (ASR) in resource-deficient languages by reusing acoustic models from resource-rich languages.

- The notion of a universal phone-set has been used, e.g., by Schultz and Waibel (1998), to jointly train acoustic models in multiple languages.
- Acoustic–phonetic models in the target language have been synthesized by Byrne et al. (2000) by matching well-trained models from resource-rich languages to a limited amount of transcribed speech in the target language.

Morphological analyzers, noun–phrase chunkers, part-of-speech taggers etc., have been developed for resource-deficient languages by exploiting translated texts.

- Statistical models have been used by Yarowsky, Ngai, and Wicentowski (2001) to align words in a sentence in the target language with words in, say, the English translation of the sentence; the English side is automatically annotated for the necessary categories (POS tags, NP brackets), and the annotation is projected to the target language via the alignment, producing a "labeled" corpus in the resource-deficient language, from which necessary statistical models are then estimated.

In this paper, we propose techniques for estimating a language model (LM) for ASR and other natural language applications in resource-deficient languages by exploiting related text in resource-rich languages.

When an ASR system needs to be engineered for a specific domain in a new language (e.g., Arabic news broadcasts), a modest amount of domain specific LM training text is usually made available, from which a word-list and a small N-gram LM may be derived. Additional target language text from an unrelated domain (e.g., Arabic web pages) may sometimes be available, and its use to improve performance in the target language and domain has been investigated elsewhere (cf. Berger & Miller, 1998; Scheytt, Geutner, & Waibel, 1998). Abundant domain-specific text in *other* languages (e.g., English news broadcasts) is also often available. Furthermore, for several languages with a sub-par electronic presence, the amount of English text in the domain of interest is likely to far outweigh the amount of in-language

Table 1
The TDT-4 corpus covers news in three languages from the (The TDT4 corpus, 2002)

	Arabic	English	Mandarin
Newswire	An-Nahar Al-Hayat Agence France Press	New York Times Associated Press Wire	Zaobao Xinhua
Radio	VOA Arabic	PRI The World VOA English	VOA Mandarin CNR
Television	Nile TV	CNN Headline News ABC World News Tonight NBC Nightly News MSNBC News with Brian Williams	CCTV CTS CBS-Taiwan

text from all domains. In this paper we investigate methods to use such cross-language in-domain data to improve the LM.

The topic detection and tracking (TDT) task is a concrete example of a large publicly funded technology demonstration program which motivates the research described in this paper. The original TDT corpus contains news broadcasts from four audio sources and two text sources in English as well as one audio source and two text sources in Mandarin (see Graff, Cieri, Martey, & Strassel, 2000). The broadcasts were collected concurrently over a 9 month period in 1998. Arabic language sources have since been added to the TDT collection, as indicated in Table 1. The goal of the TDT program is to demonstrate a system which can automatically track the reporting of a specified event or set of events across all the news sources, to detect new events as soon as they are reported in any one of the sources, etc. The audio sources are transcribed by language-specific ASR systems and the rest of the processing does not explicitly distinguish between speech and text sources. It has been noted in TDT literature that ASR errors, particularly those of named entities and infrequently occurring "content words", degrade fine-grained event detection and information extraction (Allan et al., 1998).

It has been demonstrated on the other hand that even with mediocre ASR, existing cross-language information retrieval (CLIR) techniques can be effectively employed to identify concurrent documents in the English newswire which are on the same topic as an audio story in a target, resource-deficient language. In this paper we propose methods to exploit such contemporaneous English documents, with very limited machine translation (MT), to sharpen the language model for each individual audio news story in the target language, an exercise we call *story-specific language modeling*, as illustrated in Fig. 1. A second-pass transcription of the audio with such sharper models may be employed to improve the ASR accuracy – from being barely adequate for CLIR to possibly being usable for summarization or information extraction.

A natural question to ask, in pursuing the strategy depicted in Fig. 1, is whether sufficient resources for effective CLIR and MT are easier to obtain than additional electronic text for LM training. We argue that, in this setting, the implicit demands on the performance of a CLIR or MT system are far lower than in a setting where CLIR or MT are being evaluated as stand-alone technologies. For instance, if a CLIR system retrieved only a tenth of the relevant documents in response to a particular query, it may rate poorly (e.g., offer modest precision only at unacceptably low recall levels) in a CLIR evaluation. However, if the few documents thus retrieved

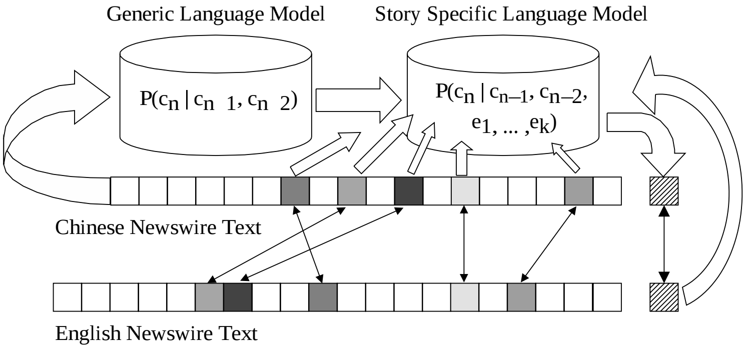
Generic Language Model
Story Specific Language Model
$P(c_n \mid c_n\ _1, c_n\ _2)$
$P(c_n \mid c_{n-1}, c_{n-2}, e_1, \ldots, e_k)$
Chinese Newswire Text
English Newswire Text

Fig. 1. Illustration of cross-lingual story-specific language models.

cover most of the proper names, places and other topic-related entities, which the ASR system would otherwise have difficulties transcribing, then the side-information provided by this CLIR system may be almost as useful as that from a near-perfect CLIR system. Similarly, the word-order of the output of a MT system is of considerable import in its performance evaluation, but may be of little consequence in language modeling: it has been shown by Khudanpur and Wu (1999) that maximum entropy models are as effective at exploiting topic-specific *unigram* frequencies as N-gram models are at exploiting topic-specific *trigram* frequencies. Therefore, the resources needed to put together a CLIR or MT system towards story-specific language modeling may indeed be far lower than for a state-of-the-art CLIR or MT system for stand-alone application. It may, e.g., be possible to obtain a translation lexicon via optical character recognition from a printed bilingual dictionary (cf. Doermann, Ma, Karagol-Ayan, & Oard, 2002), while additional text in electronic form may simply not be available.

The remainder of this paper is organized as follows:

- We begin in Section 2 by describing general techniques for exploiting English text documents for improving a story-specific Chinese language model, assuming that CLIR and MT of a quality sufficiently good for our purposes is available.
  - We also mention some simple methods to obtain such CLIR and MT capabilities.
- We present experimental results in Section 3 for the case when a fairly large Chinese–English parallel text corpus is available for estimating good statistical models for CLIR and MT as well as for language modeling. Our main conclusion is that even in this case, considerable reduction in LM perplexity is obtained by using contemporaneous English text to improve a static Chinese LM.
  - We show further, in Section 3.4, that this improvement is in addition to within-language adaptation of the static Chinese LM to the topic of the test story.

The side-corpus, in this case, happens to contain documents which are close translations of the Chinese documents being processed. Even though neither the knowledge nor even the existence of this correspondence is assumed by our methods, its existence makes this an over-optimistic situation.

- In Section 4, we investigate the story-specific adaptation of a large Mandarin Broadcast News LM using contemporaneous English newswire text.
  - A likelihood-based technique for deciding the number of English articles to use from the

side-corpus, and for dynamically adapting a static Chinese LM to a test story, is presented in Section 4.2.

Our main result is that even when a large amount of text is available for training the Chinese LM, a modest but statistically significant improvement in speech recognition accuracy is provided by the contemporaneous English side-corpus.

- Finally, in Section 5, we study the more compelling situation when little in-domain, in-language text is available for LM training, and only mediocre CLIR and MT systems trained using data from a different domain may be had. We show in this case that significant improvements in speech recognition accuracy are obtained by exploiting the side-information extracted from contemporaneous English stories.

The reader should note that while Chinese plays the role of a resource-deficient language in all our experiments, our techniques are language-independent and applicable to most other languages with little or no modifications. Our primary motivation in choosing Chinese is, indeed, the availability of large Chinese text resources which enable us to make comparative studies. In practice, these techniques will be of benefit for a language in which LM text is truly not plentiful.

A Remark on Terminology: We use the phrase *contemporaneous texts* throughout this paper to mean documents created at about the same period in time; e.g., Chinese and English newspaper articles published in 1998. We specifically avoid the phrase *parallel texts*, which has been used in the literature to connote a collection of exact translations. The phrase *comparable texts*, which has been used to connote multilingual collections in a single genre, usually containing for each document in one language one or more documents on the same topic in the other language(s), may be used here. However, we make no effort to verify whether a Mandarin news story indeed has an English article on the same topic or event in our collection; we try to exploit all English news articles written around that time.

2. Story-specific cross-lingual language modeling

Let $d_1^{\mathrm{C}}, \ldots, d_N^{\mathrm{C}}$ denote the text of N *test stories* in a Mandarin news broadcast to be transcribed by an ASR system, and let $d_1^{\mathrm{E}}, \ldots, d_N^{\mathrm{E}}$ 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 may 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 lexicon – a probabilistic model of the form $P_{\mathrm{T}}(c|e)$ – which provides the Chinese translation $c \in \mathtt{C}$ of each English word $e \in \mathtt{E}$, where $\mathtt{C}$ and $\mathtt{E}$, respectively, denote our Chinese and English vocabularies.

2.1. Cross-lingual unigram distributions

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

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

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

As a first approach, 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

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

of the cross-lingual unigram model (1) with a static trigram model for Chinese, where the interpolation weight λ may be chosen 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. Consequently, the gain from interpolating a trigram LM with (1) are expected to be largest when the d_i^{E}'s are true translations of the respective d_i^{C}'s.

Variations on (2) are easily anticipated, such as

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

where the interpolation weight may be chosen to let content-bearing words be influenced more by the cross-lingual cues than function words, e.g., by making λ_{c_k} proportional to the *inverse document frequency* of c_k (cf. Coccaro & Jurafsky, 1998; Iyer & Ostendorf, 1999) in the Chinese LM training text. Other variations include log-linear interpolation with a global or word-dependent λ's, and bucketing the λ's based on Chinese N-gram counts.

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

Remark: It is not surprising that $P_{\text{CL-unigram}}(c|d_i^{\text{E}})$, by itself, is often a poor language model, and is easily outperformed by its trigram counterpart $P(c|c_{k-1}, c_{k-2})$ in (2), even if the latter were estimated from a small corpus. For instance, one of our Chinese trigram models, trained only only the ASR acoustic training transcripts (to be discussed in Table 8), has a perplexity of 1195, while $P_{\text{CL-unigram}}(c|d_i^{\text{E}})$ has a perplexity of 2342. We will therefore not report any results for the case of $\lambda = 1$ in (2).

2.2. Obtaining the matching English documents 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 $P_{\text{T}}(e|c)$. One obtains from the first pass ASR output, cf. Fig. 2, a relative frequency estimate $\hat{P}(c|d_i^{\text{C}})$ of Chinese words $c \in \mathtt{C}$ in d_i^{C}, and uses $P_{\text{T}}(e|c)$ to get

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

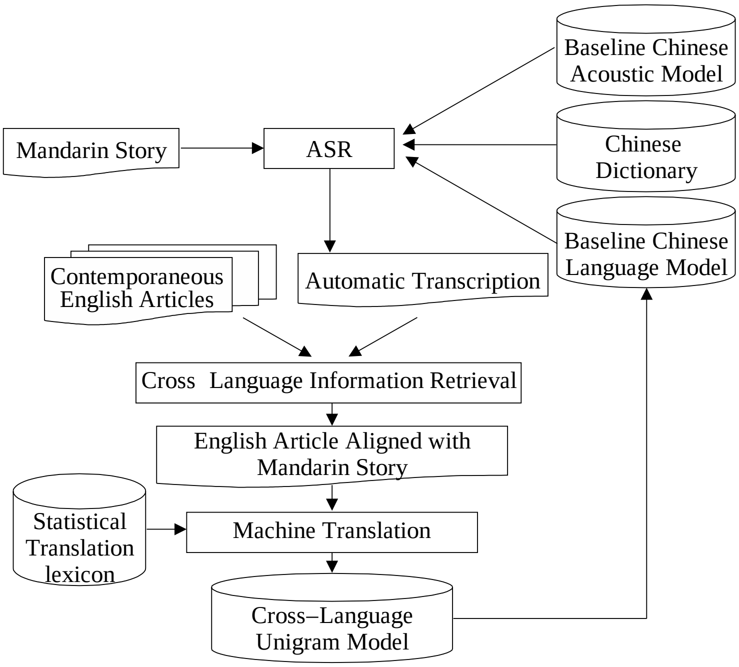
Baseline Chinese Acoustic Model
Mandarin Story
ASR
Chinese Dictionary
Baseline Chinese Language Model
Contemporaneous English Articles
Automatic Transcription
Cross Language Information Retrieval
English Article Aligned with Mandarin Story
Statistical Translation lexicon
Machine Translation
Cross–Language Unigram Model

Fig. 2. Story-specific cross-lingual priming of a language model.

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

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

Knowledgeable readers will recognize this as the *query-translation* approach to CLIR. Readers interested in details are referred to standard texts (cf. Baeza-Yates & Ribero-Neto, 1999).

2.3. Obtaining stochastic translation lexicons $P_T(c|e)$ and $P_T(e|c)$

The models $P_{\mathrm{T}}(c|e)$ and $P_{\mathrm{T}}(e|c)$ 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 developed by Och (2000). These tools use several iterations of the EM algorithm on increasingly complex word-alignment models to infer, among other translation model parameters, the conditional probabilities $P_{\mathrm{T}}(c|e)$ and $P_{\mathrm{T}}(e|c)$ of words c and e being mutual translations. Unlike standard MT systems, however, we will apply these 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 $P_T(e|c)$.

3. Language modeling experiments on the Hong Kong News Corpus

We use the Hong Kong News parallel text corpus (2000) for all the preliminary experiments reported in this Section. The corpus contains 18,147 aligned translation-equivalent 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. After removing a few articles containing nonstandard Chinese characters, we divide the corpus, by random selection, into article-pairs for training, cross-validation and other development, evaluation. All the Chinese articles, training, development and evaluation sets included, have been automatically segmented into words (by Radev et al., 2001), and the resulting corpus statistics are noted in Table 2.

Note that the statistics for the English portion of the corpus are in harmony with those for the Chinese portion due to the fact that the article-pairs are indeed translations of each other. We use the obvious notation C-train, C-dev and C-eval, and E-train, E-dev and E-eval to denote the six corpus partitions described in Table 2.

Only perplexity and out-of-vocabulary rate measurements are performed on the evaluation portion of the corpus; no parameters are tuned on it, nor any iterative diagnostics performed.

Results reported in this section expand upon some preliminary experiments reported earlier by Khudanpur and Kim (2002).

3.1. Baseline Chinese language model estimation

We estimate a standard trigram LM, using Good-Turing discounting and Katz back-off, from C-train. Its perplexity on C-dev and C-eval is reported in Section 3.3. We considered basing all our Chinese LMs on character N-grams instead of words, but went with a word-based LM primarily because we believe that the cross-language cues will be directly beneficial to a word-based model.

Table 2
Partition of the Hong Kong News corpus into training (Train), cross-validation and development (Dev) and evaluation (Eval) sets

Language	Chinese			English		
Corpus partition	Train	Dev	Eval	Train	Dev	Eval
Number of documents	16,010	750	682	16,010	750	682
Number of word tokens	4.2M	255K	177K	4.3M	263K	182K
Number of characters	6.2M	376K	260K	–	–	–
Word-vocabulary size	41K		–	39K		–
Out-of-vocabulary rate	–	–	0.4%	–	–	0.4%

Table 3
Performance of story-specific language models with cross-lingual cues

Language model (# words)	C-dev perplexity		C-eval perplexity	
	Word	Character	Word	Character
Baseline trigram (4.2M)	106	23.7	62.5	16.7
CL-interpolated with d_i^{E} from CLIR	90.1	21.2	51.3	14.6
CL-interpolated with true d_i^{E}	89.7	21.1	51.2	14.6

Chinese LM discussions, particularly for ASR, frequently report *character perplexity* (instead of word perplexity) and character error rates, mainly to facilitate comparison across approaches that use different word-segmentations. We too report character perplexity: while calculating the average perplexity of a set of sentences, we simply divide the total log-probability of a sentence by the number of characters in the sentence rather than the number of segmented words. All our models, however, assign probability to entire words.

3.2. Estimating statistical translation lexicons

The Hong Kong News corpus has been automatically aligned at the sentence level (cf. Radev et al., 2001). We use GIZA++, a statistical machine translation training tool (cf. Och, 2000), to train an IBM-Model-3 translation system from the 16,000 article pairs from our training set. We extract the translation tables from GIZA++ and use them as translation lexicons $P_{\mathrm{T}}(c|e)$ and $P_{\mathrm{T}}(e|c)$. Note that since we apply these translation models word-by-word to entire English documents to get the statistic of (1), or to entire Chinese documents in (4) for CLIR, a sentence aligned corpus is not crucial.

3.3. Language model perplexities on test data

We first assume that the *true* alignment of a Chinese test document d_i^{C} with its English counterpart d_i^{E} is given, and compute the language model of (2), henceforth called the *CL-interpolated LM*. The interpolation weight λ is chosen to minimize the perplexity of the C-dev data, and then reused blindly on the C-eval data. We report the average perplexity [3] for C-dev and C-eval in Table 3; both word- and character-perplexity are reported for completeness.

Next, we relax the assumption that the true story-alignment is given. For each Chinese article d_i^{C}, we use the reverse translation model $P_{\mathrm{T}}(e|c)$ described earlier to create an English bag-of-words representation of (4), and use it to find the English document with the highest cosine similarity as described in (5) – this document then plays the role of d_i^{E}. Again, the interpolation

[3] Note that in results reported here, the entire article d_i^{C} is used to determine d_i^{E}, which in turn conditions the probability assigned to words in d_i^{C}. Strictly speaking, this is inappropriate conditioning of the probabilistic model. However, the theoretically correct version of conditioning the LM for each word c_k only on $c_1, \ldots, c_{k-1}$, is known from many other cases to produce nearly identical results – due to the robust determination of d_i^{E} even with small values of k – so we proceed with this somewhat tainted but defensible investigation.

weight λ is chosen to minimize the perplexity of the C-dev data, and reused blindly on the C-eval. The results in Table 3 indicate that the CL-interpolated LM is quite robust to CLIR errors.

As an aside, the correct English document is retrieved from E-dev for 92% of the articles in C-dev, and from E-eval for 89% of the articles in C-eval. The E-dev and E-eval sets are small in size relative to document collections used for benchmarking information retrieval systems. If one were looking at English newswire feed on a given day for an article to match a Chinese story, however, it ought to be feasible to narrow the search down to a few hundred candidate articles.

A remark on the low perplexity of C-eval: We were surprised by the significantly lower perplexity of the C-eval data relative to the C-dev data for all models, as seen in Table 3. We did not initially perform a detailed diagnosis, due to the fear that we may inadvertently learn more about our evaluation set than would be considered proper in a blind evaluation. Upon request from one of the referees, we did look into this further. Unfortunately, other than the higher N-gram coverage of the C-eval data (53% of the word-positions covered by trigrams, 36% by back-off to bigrams and 11% to unigrams) as opposed to the C-dev data (45% covered by trigrams, 42% by back-off to bigrams and 13% to unigrams), we were unable to find any "explanation" for the difference. Recall that the corpus partition was made by random selection, as noted in Table 2. We somewhat reluctantly write off the difference between the two data sets to chance, and instead focus on the relative differences between different LMs on each data set.

3.4. Contrast with topic-specific language models

The linear interpolation of the story-dependent unigram model (1) with a story-independent trigram model, as described in (2), is very reminiscent of monolingual topic-dependent language models (cf. Seymore & Rosenfeld, 1997; Clarkson & Robinson, 1997; Iyer & Ostendorf, 1999). This motivates us to construct topic-dependent LMs and contrast their performance with the models in Table 3. We proceed as follows.

The 16,000 articles in C-train are each represented by a bag-of-words vector $\hat{P}(c|d_i^{\mathrm{C}})$. These 16,000 vectors are then clustered into 100 classes using a standard K-means clustering algorithm. Random initialization is used to seed the algorithm, and standard TF-IDF weighted cosine-similarity is used as the "metric" for clustering. Five iterations of the K-means algorithm are performed, and the resulting 100 clusters are deemed to represent different *topics*. A bag-of-words *centroid* created from all the articles in a cluster is used to represent each topic. Topic-dependent unigram and trigram LMs, denoted $P_t(c)$ and $P_t(c_k|c_{k-1}, c_{k-2})$ respectively, are also computed for each topic t exclusively from the articles in its cluster.

For each article d_i^{C} in C-dev or C-eval, a bag-of-words vector is generated in the same manner as was done for C-train, and the topic-centroid having the highest cosine-similarity to it is chosen as the topic t_i of d_i^{C}. Topic-dependent LMs are then constructed for each article d_i^{C} as

$$P_{\text{Topic-unigram}}(c_k|c_{k-1}, c_{k-2}, t_i) = \lambda P_{t_i}(c_k) + (1-\lambda)P(c_k|c_{k-1}, c_{k-2}) \tag{6}$$

and

$$P_{\text{Topic-trigram}}(c_k|c_{k-1}, c_{k-2}, t_i) = \lambda P_{t_i}(c_k|c_{k-1}, c_{k-2}) + (1-\lambda)P(c_k|c_{k-1}, c_{k-2}). \tag{7}$$

The development set C-dev is used to estimate the global interpolation weight λ. The perplexity of the resulting topic-dependent models is reported in Table 4.

Table 4
Performance of topic-specific language models, and their interpolation with story-specific models that use cross-lingual cues

Language model (# words)	C-dev perplexity		C-eval perplexity	
	Word	Char	Word	Char
Baseline trigram (4.2M)	106	23.7	62.5	16.7
Topic-unigram	94.6	21.9	57.4	15.8
Topic-trigram	84.4	20.3	49.3	14.2
Topic-trigram + CL-interpolated	80.1	19.6	44.6	13.3

We conclude from a comparison of Tables 3 and 4 that contemporaneous side-information, even if cross-lingual, is more accurate than the static topic-unigram statistics of (6), but (2) lacks the contextual awareness of the topic-trigram statistics of (7).

An obvious experiment is to interpolate the cross-language unigram and the topic trigram models with the baseline trigram model, which we do and report in Table 4. The further reduction in perplexity suggests that the topic-wide and article-specific cues obtained from the two models are considerably complementary.

4. Speech recognition experiments on Mandarin Broadcast News

We next apply the techniques described above for improving ASR performance on Mandarin news broadcasts using English newswire texts. We have chosen the experimental ASR setup created to study Mandarin pronunciation modeling in the 2000 Johns Hopkins Summer Workshop, extensive details about which are available from Fung et al. (2000). The acoustic training data (~10 h) for their ASR system was obtained from the 1997 Mandarin Broadcast News distribution (see Hub-4 Mandarin Broadcast News speech corpus, 1998), and context-dependent state-clustered models were estimated using initials and finals as subword units.

ASR evaluation data, containing Mandarin news broadcasts from three sources, were selected from the 1997 and 1998 NIST HUB-4NE benchmark tests (NIST Hub-4 Evaluation, 1997–98). About 1250 studio quality utterances (F0 condition), amounting to about 9800 words, were selected from the two test sets, and lattices were generated using the baseline acoustic models and a bigram language model by Fung et al. (2000). All experiments in this section are based on rescoring a 300-best list extracted from these bigram lattices. We report both word error rates (WER) and character error rates (CER), the latter being independent of any differences in segmentation of the ASR output and reference transcriptions. The bigram LM used to generate the lattices has a 51,000-word vocabulary, and yields a WER of 50.7% and a CER of 29.6%. Choosing the most erroneous hypothesis in each of the the 300-best lists yields a WER of 92.3% (CER 59.4%) and the least erroneous, a WER of 32.2% (CER 15.2%), leaving ample room for reasonable conclusions from rescoring.

We use two large text corpora, which we name PDXR and NAB-TDT for easy reference, to estimate language models for this ASR task.

PDXR: We use a large Chinese text corpus collected from five sources, The People's Daily, China Radio News International, Xinhua News and transcriptions of Mandarin news broadcasts from Voice of America and CCTV International. The original text is segmented to produce (only) words which match the pronunciation lexicon used by our ASR system, resulting in about 291 million words, from which bigram and trigram LMs are estimated, using the SRI LM toolkit, with Good-Turing discounting and Katz back-off.

NAB-TDT: English text contemporaneous with the test data is available in the form of the North American News Text corpus, from which we select articles published in 1997 in The Los Angeles Times and The Washington Post. We also use the TDT-2 text corpus to extract articles from the New York Times and the Associated Press news service published in 1998. This amounts to a collection of roughly 45,000 articles containing about 30 million words of English text; a modest collection by today's CLIR standards.

It should be clear that with over 290 million words of Chinese LM training text, this is far from the scenario of a resource-deficient language we initially set out to address. However, we use experiments in this section to benchmark our techniques in the large-data regime, and conduct experiments with a simulated shortage of Chinese LM training text in the following section. The amount of acoustic training data used here ($\sim$10 h) is quite representative of typical resource-deficient languages.

What is quite realistic about these experiments is that the English newswire text is not a translation of the Mandarin news we are attempting to transcribe. Indeed it is guaranteed by design only to be contemporaneous, and we hope that the CLIR will be sufficiently good to find articles covering the same events as the Mandarin story being processed. We do not use the timestamp on the newswire articles or the test story to assist the CLIR system; an option which could potentially further improve system performance.

4.1. Recognition performance of cross-lingual language models

We begin with the first pass recognition output, which includes the 1-best transcription of each test utterance, and a 300-best list as described above. We rescore the 300-best lists using a trigram model estimated from the large PDXR corpus, and report the error rates in Table 5. Next, for each test story d_i^{C}, we perform CLIR to choose the single best-matching English document d_i^{E} from NAB-TDT and create the CL-interpolated LM of (2). The translation models $P_{\mathrm{T}}(c|e)$ and $P_{\mathrm{T}}(e|c)$ are reused from Section 2.2.

Since we did not provide for separate development and evaluation partitions of the test set, we choose the *global* interpolation weight λ for the story-specific CL-interpolated LM of (2) via a

Table 5
ASR performance for the large Chinese trigram LM, a CL-interpolated LM, a topic-dependent trigram LM and its interpolation with the CL-interpolated LM

Language model (# words)	Perplexity	WER (%)	CER (%)	p-value
Baseline PDXR trigram (291M)	283	47.6	26.9	–
CL-interpolated	260	47.3	26.8	0.226
Topic-trigram	247	47.3	26.7	0.174
Topic-trigram + CL-interpolated	236	46.9	26.4	0.005

simple EM procedure that maximizes the total likelihood of the 1-best hypotheses of all the test utterances from the first ASR pass. The performance of this model is also reported in Table 5.

One way to gauge whether an observed improvement in WER from one ASR system to another on a test corpus is a reliable result is to view the performance of the two ASR systems on each utterance in the test corpus as an independent trial, and to *count* how often one system outperforms the other. Under the null hypothesis – that the two ASR systems have the same *underlying* WER and the observed WER improvement is a fluke – one then computes the probability of this observed count. The lesser the probability of the observed count under the null hypothesis, the more confident one feels accepting the alternate hypothesis – that the improvement is reliable. A specific version of this statistical significance test, implemented by Pallett, Fisher, and Fiscus (1990) and dubbed the matched pairs sentence-segment test, is widely accepted and we use it throughout this paper to compute the probability of an observed WER improvement being by chance. If this probability is less than an application- or user-specified *p*-value, then the WER difference is considered significant. Sometimes, the improvement is said to be “significant at a *p*-value of” this probability.

Table 5 indicates the *p*-value at which a WER improvement over the baseline trigram LM is statistically significant under the matched pairs sentence-segment test.

The CL-interpolated LM, not too surprisingly, provides an insignificant reduction in WER over the Chinese trigram LM based on the 291 million-word PDXR corpus, even though there is some reduction in LM perplexity.

Finally, analogous to Section 3.4, we create 100 topic-clusters from the PDXR corpus and rescore the 300-best list using the topic-dependent LM of (7). Note, that the improvement in ASR performance obtained by the topic-dependent LM is very comparable to the CL-interpolated LM. This is remarkable, since unigram statistics from a *single* English document are used by the CL-interpolated LM, while each topic-cluster used by the topic-trigram LM has *hundreds* of Chinese documents. We believe that the contemporaneous 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 WER improvement obtained by the interpolation of the two LMs, as shown on the last line.

4.2. Maximum likelihood based interpolation

The experiments so far naïvely used the single most similar English document, and a common λ in (2), for each Mandarin test story, no matter how similar its best-matching English document is to a given Mandarin story. This was acceptable when working with the Hong Kong News corpus, because we were guaranteed that there indeed was one English document which was a translation of the Chinese document being processed. It stands to reason in this case, however, 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 when the match is good, and a smaller λ otherwise. For experiments in this subsection, we select a different λ in CL-interpolated model for each test story, again based on maximizing the likelihood of the 1-best output. The other remaining issue then is the choice and the number of English documents to translate.

4.2.1. *M-best English documents*

In one experiment, we choose a predetermined number M of the best-matching English documents for each Mandarin news story. We experimented with values of 1, 10, 30, 50, 80 and 100, and found that $M = 30$ gave us the best LM performance, but it was only slightly better than the case of $M = 1$ in Table 5. The details are not interesting and therefore omitted.

4.2.2. *All English documents above a similarity threshold*

The argument against a predetermined number M of the best-matching documents may be that it ignores the goodness of the match. An alternative is to take, for each Mandarin test story, all English documents whose similarity to the test story exceeds a certain threshold. As this *common* threshold is lowered, starting from a high value, the rank-order in which English documents get selected for a test story is thus the same as for choosing the M-best documents. However, the number of English documents selected now varies from story to story even for a common threshold; it is possible now that for some test stories, even the best-matching English document falls below the threshold at which other stories have found more than one good match. The solid curve in the middle of Fig. 3 shows the perplexity of the reference transcriptions of *all* the test stories as a function of this common similarity threshold. Note that a threshold of 0.12 gives the lowest perplexity, but the reduction is not very large.

Fig. 3 also shows, via the two broken curves, the effect of varying the similarity threshold on the perplexities of two *individual* stories. It is clear that different thresholds are optimal for different stories. We therefore turn our attention to a *story-specific* strategy for choosing the similarity threshold.

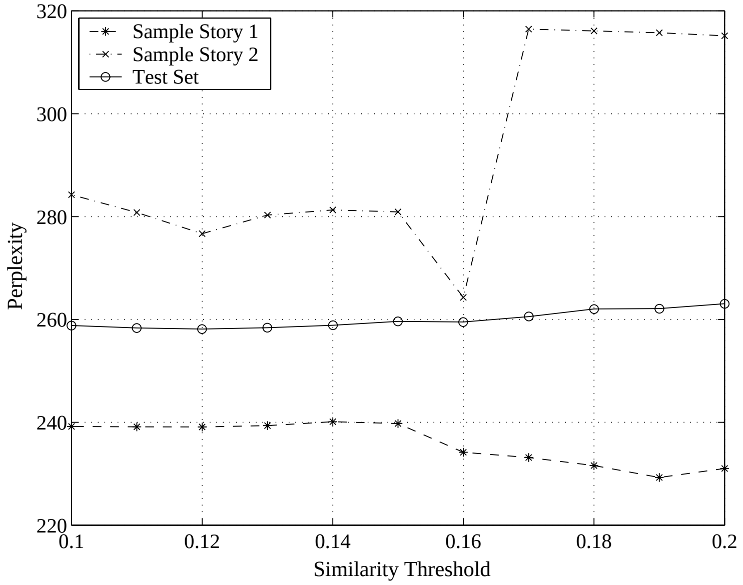

Fig. 3. Perplexity v/s similarity threshold for selecting a story-specific number of d_i^E's. Perplexity averaged over the entire test set (solid curve) and over two individual stories in the test set (broken curves) is plotted for each similarity threshold.

Table 6
ASR results with a likelihood-based story-specific number of d_i^{E}'s and λ

Language model (# words)	Perplexity	WER (%)	CER (%)	p-value
Baseline PDXR trigram (291M)	283	47.6	26.9	–
CL-interpolated	248	47.1	26.8	0.028
Topic-trigram	247	47.3	26.7	0.174
Topic-trigram + CL-interpolated	225	46.8	26.5	0.003

4.2.3. Likelihood-based selection of English documents and interpolation weight

For each Mandarin test story, we choose the 1000-best-matching English documents and divide the *dynamic range* of their similarity scores evenly into ten intervals. Next, we choose documents in the top one-tenth of the range of similarity scores, compute $P_{\text{CL-unigram}}(c|d_i^{\mathrm{E}})$, determine the λ in (2) that maximizes the likelihood of the first pass 1-best output of only the utterances in that story, and record this likelihood. We repeat this with documents in the top two-tenths of the range of similarity scores, the top three-tenths, etc., and obtain the likelihood as a function of the similarity threshold. We choose the threshold that maximizes the likelihood of the first pass 1-best 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.

Modest but statistically significant WER improvements are obtained from the CL-interpolated LM using this likelihood-based story-specific adaptation scheme, as seen in Table 6, where the p-value at which a WER improvement over the trigram LM is significant is also shown.

Note, also from Table 6, that this story-specific CL-interpolated LM outperforms the topic-trigram LM, and the small additional gain from their interpolation points to their continued complementarity.

5. Speech recognition experiments in a resource-deficient setting

The amount of transcribed speech used for estimating the acoustic models for the Mandarin ASR task described in the previous section, $\sim$10 h, is fairly representative of a resource-deficient. If we reduce the amount of language modeling text also down to a modest amount, it creates a realistic setting to explore the use of the cross-lingual side-information for resource-deficient languages.

We therefore investigate the use of two smaller text data sets to estimate language models for this ASR task.

XINHUA: We use the Xinhua News portion of about 13 million words from the PDXR corpus described above to represent the more frequent scenario when a modest amount of LM training text is available, and again estimate a trigram LM, using the same 51,000-word vocabulary as before.

HUB-4NE: We also estimate language models from *only* the 96,000 words in the transcriptions used for training acoustic models in our ASR system. This represents the extreme case when no additional LM training text is available.

We continue to use the full NAB-TDT data set for finding contemporaneous English documents for the Mandarin stories being transcribed, as well as the translation models $P_{\mathrm{T}}(c|e)$ and $P_{\mathrm{T}}(e|c)$

estimated earlier from an unrelated corpus, namely the Hong Kong News corpus. The ASR evaluation set is also the same as before, and we again rescore 300-best lists produced by a bigram LM.

There is some concern in rescoring with, say, the HUB-4NE trigram LM trained on 96,000 words, an 300-best list generated by the PDXR bigram LM trained on 290 million words of text. We therefore regenerate the lattices, once using the baseline acoustic models with a bigram LM trained from the XINHUA corpus described above and, once again, with a bigram LM trained from only the HUB-4NE transcriptions.

XINHUA: The top scoring hypotheses in the XINHUA bigram lattices yield a WER of 53.2% and a CER of 31.4%. Choosing the most erroneous hypothesis in each of the the 300-best lists yields a WER of 94.4% (CER 60.5%) and the least erroneous, a WER of 34.4% (CER 16.2%).

HUB-4NE: The top scoring hypotheses in the HUB-4NE bigram lattices yield a WER of 60.9% and a CER of 44.4%. Choosing the most erroneous hypothesis in each of the the 300-best lists yields a WER of 95.5% (CER 68.2%) and the least erroneous of 39.7% (CER 27.5%).

All experiments in this section are based on rescoring a 300-best list extracted from one of these lattices.

For the sake of comparison with the results of Section 4, we limit the rescoring to the 300-best list, though we note with some concern the high WER of even the best hypotheses for HUB-4NE. Such high error rates, however, are not unusual for ASR systems trained, e.g., with 10 h of speech and 100,000 words of LM text.

5.1. Recognition performance of cross-lingual language models

We rescore the 300-best lists using trigram LMs estimated from the XINHUA and HUB-4NE corpora, and report the results in Tables 7 and 8. For each test story d_i^{C}, we again perform CLIR to choose the best-matching English document d_i^{E} from NAB-TDT and create the cross-lingual unigram model of (1), choose a *global* λ via an EM procedure that maximizes the likelihood of the 1-best hypotheses of the first ASR pass, and construct story-specific CL-interpolated LMs, one each for the XINHUA and HUB-4NE trigrams. The performance of the CL-interpolated models are also reported in Tables 7 and 8.

Compared to the 290 million-word PDXR corpus of Table 5, a slightly greater reduction in WER over the baseline trigram LM is obtained using the CL-interpolated LM on the 13 million-

Table 7
ASR performance for the small Chinese LM, CL-interpolated LM, and its interpolation with a topic-dependent trigram LM for the XINHUA corpus

Language model (# words)	Perplexity	WER (%)	CER (%)	*p*-value
Baseline XINHUA trigram (13M)	426	49.9	28.8	–
CL-interpolated	375	49.5	28.7	0.208
Topic-trigram	381	49.1	28.4	0.003
Topic-trigram + CL-interpolated	352	49.1	28.3	0.004

Table 8
ASR performance for the tiny Chinese LM, CL-interpolated LM, and its interpolation with a topic-dependent trigram LM for the HUB-4NE corpus

Language model (# words)	Perplexity	WER (%)	CER (%)	p-value
Baseline HUB-4NE trigram (96K)	1195	60.1	44.1	–
CL-interpolated	750	59.3	43.7	<0.001
Topic-trigram	1122	60.0	44.1	0.660
Topic-trigram + CL-interpolated	752	59.4	43.7	0.015

word XINHUA corpus, as seen from Table 7. But the statistical significance of the improvement is still suspect. Furthermore, it seems that the XINHUA corpus, due to reasons we do not fully understand, is able to provide a significantly better topic-dependent LM.

However, when only the 96,000-word HUB-4NE corpus is available, significant reduction in both perplexity and error rates over the baseline trigram model are evident in Table 8 due to the CL-interpolated LM. The HUB-4NE corpus, furthermore, is too small to produce good topic-dependent models, and our efforts to produce even 4 topic-centroids meets will little success. This clearly establishes the utility of the CL-interpolated model in a resource-deficient setting.

5.1.1. Likelihood-based selection of English documents and interpolation weight

The CL-interpolated models of Tables 7 and 8 use a single common interpolation weight λ for all test stories. Analogous to Section 4.2.3, we next perform a likelihood-based selection of the number of English documents and interpolation weight for each test story, and adaptively estimate the XINHUA or HUB-4NE CL-interpolated models in a story-specific manner. Tables 9 and 10 show ASR results for this likelihood-based story-specific adaptation scheme.

Table 9
ASR results with a likelihood-based story-specific number of d_i^{E}'s and λ for XINHUA

Language model (# words)	Perplexity	WER (%)	CER (%)	p-value
Baseline XINHUA trigram (13M)	426	49.9	28.8	–
CL-interpolated	346	48.8	28.4	<0.001
Topic-trigram	381	49.1	28.4	0.003
Topic-trigram + CL-interpolated	326	48.5	28.2	<0.001

Table 10
ASR results with a likelihood-based story-specific number of d_i^{E}'s and λ for HUB-4NE

Language model (# words)	Perplexity	WER (%)	CER (%)	p-value
Baseline HUB-4NE trigram (96K)	1195	60.1	44.1	–
CL-interpolated	630	58.8	43.1	<0.001
Topic-trigram	1122	60.0	44.1	0.660
Topic-trigram + CL-interpolated	631	59.0	43.3	<0.001

Significant improvements are apparent, for both the corpora, in perplexity and error rates, due to the CL-interpolated LM. We hope that the reader is also convinced of the increasing utility, with diminishing training data size, of the proposed methods.

6. Summary of the main results and conclusions

The results of Table 3 amply demonstrate the reduction in language model perplexity that is attainable by exploiting side-information from contemporaneous articles, even if in another language, and those of Table 4 suggest that these reductions in perplexity via cross-lingual texts are over and above comparable semantic information provided by within-language topic-dependent models. An overall perplexity reduction of 15–28% over a trigram model was demonstrated on the Hong Kong News corpus using some of the simpler ideas presented here.

The ASR results reported in Table 6 demonstrate that even when a large amount of LM training text is available in the language of interest, statistically significant reductions in recognition error rates are yielded by simple interpolated models derived from cross-lingual side-information. This advocates the use of contemporaneous side-information in whatever language it may be found. An overall perplexity reduction of 12% and a modest WER reduction of 0.5% (absolute) over a Chinese trigram LM was demonstrated, on a subset of the Mandarin Broadcast News test set, using English newswire. The gains went up to 20% and 0.8% (absolute) respectively when Chinese topic-dependencies were also exploited.

Finally, the ASR results reported in Tables 9 and 10 clearly demonstrate that when LM training text is hard to come by in the language or domain of interest, significant improvements may be obtained by exploiting domain-specific text from other languages, even when only very simple CLIR and MT techniques are available. This represents the most compelling scenario for using these techniques. Using English newswire, perplexity reductions of 23–47% and WER reductions of 1.3–1.4% (absolute) over Chinese trigram LMs estimated from limited amounts of training text were demonstrated for Mandarin Broadcast News transcription.

7. Discussion

It stands to reason that while the crude nature of the translation model $P_{\mathrm{T}}(c|e)$ used here is unlikely to result in extraction of any detailed knowledge from an English document d_i^{E} about *word sequences* in a Chinese test story d_i^{C}, it is probably through some mechanism of “topic focussedness” that the story-specific LMs obtain a reduced perplexity. It is well known that, in case of monolingual topic-dependent models, interpolating with a story- or topic-dependent trigram model is more effective. Story-specific trigram statistics cannot, unfortunately, be easily obtained in our case due to the limited MT capabilities one may reasonably assume. However, it was demonstrated by Khudanpur and Wu (1999) in a monolingual setting that using maximum entropy techniques to impose constraints on topic-conditional unigram probabilities in a trigram model is as effective as interpolation of a topic-independent trigram model with a topic-dependent trigram model. We are therefore investigating maximum entropy models for incorporating the unigram statistics of (1) into a Chinese trigram model.

In a monolingual setting, (Rosenfeld, 1996) has proposed using *lexical triggers* from the word-history and a maximum entropy framework for adaptive language modeling. We are extending this idea of lexical triggers to our cross-lingual setting (Kim & Khudanpur, 2003).

An important advantage of using lexical triggers is that, unlike (2), there is *no explicit need for a translation lexicon* $P_T(c|e)$ for language modeling. Some form of translation may, of course, still be implicitly needed in the CLIR component. It suffices during LM training, however, to have a few contemporaneous English documents matched to each document in the Chinese LM training corpus, and when a test story is being processed, trigger words in contemporaneous English documents may be used directly. This relaxation of the need for a translation lexicon may indeed be necessary in some resource-deficient settings.

Acknowledgements

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

References

Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., Lavrenko, V., Hoberman, R., Caputo, D., 1998. Topic-based novelty detection – final report. Proceedings of the Johns Hopkins Summer Workshop, Baltimore, MD. Available from <http://www.clsp.jhu.edu/ws99>.

Baeza-Yates, R., Ribero-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley, Reading, MA.

Berger, A., Miller, R., 1998. Just-in-time language modeling. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, vol. 2, Seattle, WA, pp. 705–708.

Byrne, W., Beyerlein, P., Huerta, J., Khudanpur, S., Marathi, B., Morgan, J., Peterek, N., Picone, J., Vergyri, D., Wang, W., 2000. Towards language independent acoustic modeling. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, vol. 2, Istanbul, Turkey, pp. 1029–1032.

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, vol. 2, Munich, Germany, pp. 799–802.

Coccaro, D., Jurafsky, D., 1998. Towards better integration of semantic predictors in statistical language modeling. Proceedings of the International Conference on Spoken Language Processing, vol. 6, Sydney, Australia, pp. 2403–2406.

Doermann, D., Ma, H., Karagol-Ayan, B., Oard, D., 2002. Lexicon acquisition from bilingual dictionaries. Proceedings of the SPIE Photonic West Electronic Imaging Conference, San Jose, CA, pp. 37–48.

Fung, P., Byrne, W., Zheng, F., Kamm, T., Liu, Y., Song, Z., Venkataramani, V., Ruhi, U., 2000. Pronunciation modeling of mandarin casual speech – final report. Proceedings of the Johns Hopkins Summer Workshop, Baltimore, MD. Available from <http://www.clsp.jhu.edu/ws2000/groups/mcs>.

Graff, D., Cieri, C., Martey, N., Strassel, S., 2000. The tdt-3 text and speech corpus. Proceedings of the Topic Detection and Tracking Workshop, Vienna, VA.

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

Khudanpur, S., Kim, W., 2002. Using cross-language cues for story-specific language modeling. Proceedings of the International Conference on Spoken Language Processing, vol. 1, Denver, CO, pp. 513–516.

Khudanpur, S., Wu, J., 1999. A maximum entropy language model to integrate *n*-grams and topic dependencies for conversational speech recognition. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, vol. 1, Phoenix, AZ, pp. 553–556.

Kim, W., Khudanpur, S., 2003. Cross-lingual lexical triggers in statistical language modeling. Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing, Sapporo, Japan, pp. 17–24.

Och, F., 2000. Giza++ tools for training statistical translation models. Available from <http://www-i6.informatik.rwth-aachen.de/~och/>.

Pallett, D., Fisher, W., Fiscus, J., 1990. Tools for the analysis of benchmark speech recognition tests. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, vol. 1, Alburquerque, NM, pp. 97–100.

Radev, D., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Çelebi, A., Qi, H., Drabek, E., Liu, D., 2001. Automatic summarization of multiple (multilingual) documents – final report. Proceedings of the Johns Hopkins Summer Workshop, Baltimore, MD. Available from <http://www.clsp.jhu.edu/ws2001>.

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

Scheytt, P., Geutner, P., Waibel, A., 1998. Serbo-croatian lvcsr on the dictation and broadcast news domain. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, vol. 2, Seattle, WA, pp. 897–900.

Schultz, T., Waibel, A., 1998. Language independent and language adaptive large vocabulary speech recognition. Proceedings of the International Conference on Spoken Language Processing, vol. 5, Sydney, Australia, pp. 1819–1822.

Seymore, K., Rosenfeld, R., 1997. Using story topics for language model adaptation. Proceedings of the Eurpoean Conference on Speech Communication and Technology, vol. 4, Rhodes, Greece, pp. 1987–1990.

Yarowsky, D., Ngai, G., Wicentowski, R., 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. Proceedings of the Human Language Technologies Workshop, Santa Monica, CA, pp. 109–116.

Hong Kong News parallel text corpus, 2000. Available through the Linguistic Data Consortium. Available from <http://www.ldc.upenn.edu/Catalog/LDC2000T46.html>.

Hub-4 Mandarin Broadcast News speech corpus, 1998. Available through the Linguistic Data Consortium. Available from <http://www.ldc.upenn.edu/Catalog/LDC98S73.html>.

The 1997 Hub-4NE Broadcast News Evaluation – Mandarin, and the 1998 Hub-4 Broadcast News Evaluation, 1997–98. Conducted by the National Institute of Standards and Technology. Available from <http://www.nist.gov/speech/tests/index.htm>.

The TDT4 corpus, 2002. Available through the Linguistic Data Consortium. Available from <http://www.ldc.upenn.edu/Projects/TDT4/Annotation/label_instructions.html>.