CROSS-LINGUAL LATENT SEMANTIC ANALYSIS FOR LANGUAGE MODELING

Woosung Kim and Sanjeev Khudanpur

Center for Language and Speech Processing
The Johns Hopkins University, Baltimore, MD 21218, USA
woosung@cs.jhu.edu khudanpur@jhu.edu

ABSTRACT

Statistical language model estimation requires large amounts of domain-specific text, which is difficult to obtain in many languages. We propose techniques which exploit domain-specific text in a resource-rich language to adapt a language model in a resource-deficient language. A primary advantage of our technique is that in the process of cross-lingual language model adaptation, we *do not* rely on the availability of any machine translation capability. Instead, we assume that only a modest-sized collection of story-aligned document-pairs in the two languages is available. We use ideas from cross-lingual latent semantic analysis to develop a single low-dimensional *representation* shared by words and documents in both languages, which enables us to (i) find documents in the resource-rich language pertaining to a specific story in the resource-deficient language, and (ii) extract statistics from the pertinent documents to adapt a language model to the story of interest. We demonstrate significant reductions in perplexity and error rates in a Mandarin speech recognition task using this technique.

1. INTRODUCTION

Statistical modeling techniques have been remarkably successful in many speech and natural language processing areas. However, the construction of accurate statistical models requires extensive amounts of training data, and it is extremely difficult to build statistical models for resource-deficient languages such as Arabic due to lack of linguistic resources. Not surprisingly, therefore, the performance of the speech and natural language processing systems on resource-deficient languages is much worse than the preformance on resource-rich languages [1].

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 [2, 3]. Recently, [4] proposed using cross-lingual information retrieval (CLIR) followed by machine translation (MT) to improve a statistical language model (LM) in a resource-deficient language. In spite of considerable success in their experiments, one demand placed by the approach of [4] is that a *sentence-aligned* parallel corpus, which is expensive to obtain, is needed to acquire translation lexicons. For a resource-deficient language, it is easily anticipated that little or no sentence-aligned corpus may be available, and therefore methods which do not require the sentence-aligned parallel corpus must be developed. One possible way out is to extract and use cross-lingual lexical triggers from an easier-to-obtain *document-aligned* corpus as proposed in [5].

In this paper, we propose using latent semantic analysis (LSA) for cross-lingual language modeling, which does not require a sentence-aligned corpus. LSA of a collection of bilingual document-aligned texts provides a representation of words in both languages in a common low-dimensional Euclidean space [6]. This provides another means for using a resource-rich language to improve the LM in a resource-deficient language. We also combine our LSA-based models with trigger-based models [5], and compare their performance with LMs built from a sentence-aligned corpus [4].

Section 2 introduces the basics of cross-lingual story specific language modeling. Section 3 presents cross-lingual latent semantic analysis. Section 4 describes the databases used for our experiments and Section 5 shows the experimental results. Section 6 concludes with future work.

2. CROSS-LINGUAL STORY-SPECIFIC LMS

For the sake of illustration, consider the task of sharpening a Chinese LM for transcribing Mandarin news stories by using an extensive corpus of contemporaneous English newswire text. Of course, any other language pair may serve the purpose of this exposition, and we use Mandarin only to be able to simulate varying levels of data sparseness.

Let $d_1^C, \ldots, d_N^C$ denote the text of N Chinese *test stories* to be transcribed by an ASR system. Since the correct transcriptions d_i^C's are not available in advance, we use the first pass ASR outputs as *pseudo* documents for d_i^C's. We are assuming data sparseness in Chinese (for the sake of ex-

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).

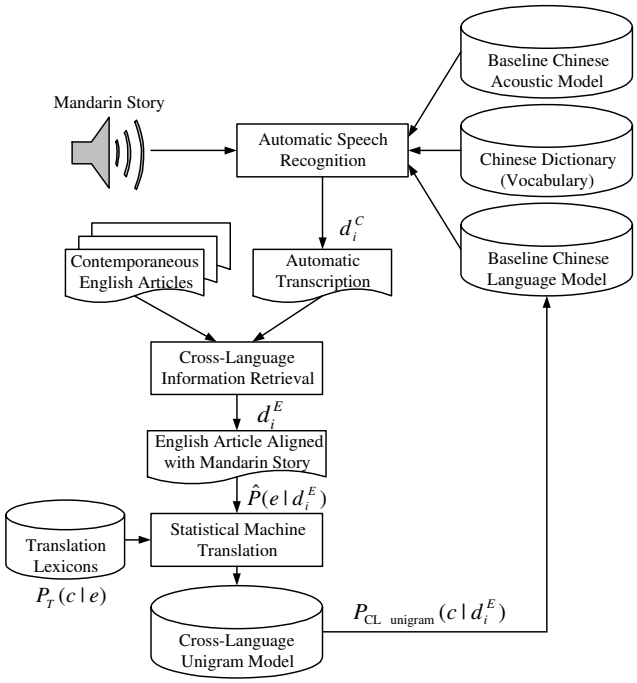

Fig. 1. Data flow diagram for cross-lingual LMs

perimentation) while extensive amounts of English data are assumed available. Therefore, it stands reason to extract some useful information from the *relevant* English documents[1], and use them to improve (sharpen) our Chinese LMs. Obviously, the first step is a conventional CLIR problem: we have to identify English documents relevant to a Chinese test document. Simple vector-based IR with query translation—based on statistical translation tables—has been used for these purposes. Let d_i^E's denote the relevant English documents retrieved by CLIR for d_i^C. Since these are English documents whereas our target language of interest is Chinese, they must be translated into Chinese – which is done in the next step, namely MT. Even with state-of-the-art MT techniques, however, the quality of translated results is far from perfect. Yet, topic-specific or story-specific adaptation information may be enough for our purpose and it can be easily obtained by estimating a Chinese unigram statistic based on translation lexicons (cf. [4]). In short, cross-lingual language modeling is a robust combination of CLIR and MT techniques. Figure 1 shows the data flow in our cross-lingual language modeling approach.

3. CROSS-LINGUAL LATENT SEMANTIC ANALYSIS

LSA is a fully automatic mathematical technique to extract relations between words and/or documents. The basic idea

[1]Here, relevant English documents d_i^E need not be exact translations of d_i^C's; stories about the same topic or event will be adequate.

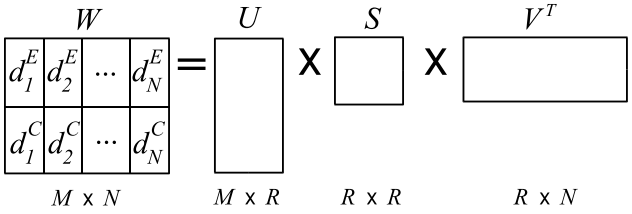

Fig. 2. Cross-lingual Latent Sematic Analysis

is to reduce the dimensionality of the the co-occurrence data, for example a word-document matrix in an IR task, into a smaller but adequate subspace of lower dimension. By applying singular value decomposition (SVD) which is a form of factor analysis, it decomposes the input matrix into the product of three other matrices: one describing the original row entities (words) as vectors of derived orthogonal factor values, another describing the original column entities (documents) in the same way, and the third a diagonal matrix containing singular values such that when the three components are multiplied, the original matrix is reconstructed. The diagonal entries in the third matrix are sorted by the singular values, and by ignoring entries with small singular values, we can approximate the original input matrix. For a detailed exposition of LSA, see [7].

3.1. Latent Semantic Analysis for CLIR

The first use of LSA in Figure 1 is for CLIR. A bilingual collection of story-aligned documents, say in Chinese and English, is assumed to be given. From the document-aligned corpus, we construct a bilingual word-document matrix W by concatenating aligned document-pairs to create bilingual documents. Each *column* in the matrix W corresponds to a document-pair, while each row corresponds to either an English or a Chinese word. The bilingual word-document matrix is then decomposed by SVD, and we obtain three matrices, U, S and V^T, as shown in Figure 2.

Next, suppose we have an additional monolingual collection, say, of English documents with its word-document matrix $\overline{W}$. Since it is an English-only collection, all the rows corresponding to the Chinese words (shown in the lower half in Figure 3) have 0 entries. Nevertheless, we project columns of $\overline{W}$ into the LSA space of the bilingual collection described above by constructing a representation $\overline{V}^T$ for the English-only documents as[2]

$$\overline{V}^T = S^{-1} \times U^T \times \overline{W}. \quad (1)$$

Finally, given a Chinese query (document), we again construct a bilingual document-vector, this time with 0 in all English word-positions, and project it to the LSA space in a manner similar to (1). Since the projected vector has the

[2]Note that matrix U is column orthogonal, which implies $U^T \times U = I$.

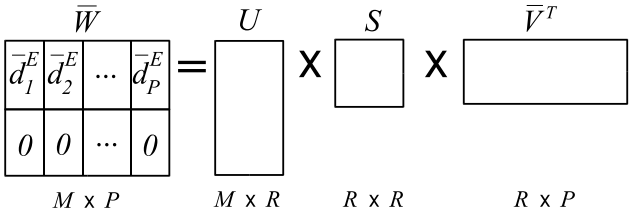

Fig. 3. Folding-in a monolingual corpus into LSA

same dimension as columns of $\overline{V}^T$, we compare a Chinese query with an English document by simply measuring the *cosine similarity* (dot-product) of their vectors. See [6] for details.

3.2. Latent Semantic Analysis for MT

We have shown how LSA can be used for CLIR; this is accomplished by first projecting each query and document into the low dimensional space, then measuring similarities in the projected space. A similar idea is used to compute similarity between an English and a Chinese word. Each word in either vocabulary, represented as a row in W, may also be represented by the corresponding row of U. Again, since all the words are in the same dimension, we can compare words regardless of which language they belong to. Remember that we are not interested in exact translations; all we need is a probability, $P_{\text{LSA}}(c|e)$, that a Chinese word c occurs in a document given an English word e is in its aligned counterpart. For each English word e, we first select similar Chinese words based on cosine similarity, and then estimate the translation probability as

$$P_{\text{LSA}}(c|e) = \frac{Sim(c,e)^{\gamma}}{\sum_{c' \in \mathcal{C}} Sim(c,e)^{\gamma}} \tag{2}$$

where $\gamma \gg 1$ as suggested in [8]. Having estimated (2), we build our LSA-based LM analogous to [4].

$$P_{\text{LSA-unigram}}(c|d_i^E) = \sum_{e \in \mathcal{E}} P_{\text{LSA}}(c|e)\hat{P}(e|d_i^E) \tag{3}$$

$$P_{\text{LSA-interpolated}}(c_k|c_{k-1},c_{k-2},d_i^E) = \tag{4}$$
$$\lambda P_{\text{LSA-unigram}}(c_k|d_i^E) + (1-\lambda)P(c_k|c_{k-1},c_{k-2}).$$

4. LM TRAINING AND ASR TEST CORPORA

We have chosen the experimental ASR setup created in the 2000 Johns Hopkins Summer Workshop to study Mandarin pronunciation modeling[9]. The approximately 10 hours of acoustic training data 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, not phone(me)s, as subword units.

We use the Hong Kong News (**HKNews**) text corpus as our parallel text for the training stochastic translation lexicons using the GIZA++ toolkit, and for the SVD decomposition of Figure 2.

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 Chinese training text into words. This vocabulary gives an OOV rate of 5% on the test data, described below.

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, 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.

Our test set, a subset [9] of the NIST 1997 and 1998 HUB-4NE benchmark tests, contains 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 LMs are evaluated by rescoring 300-best lists extracted from these two sets of lattices. 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.

5. EXPERIMENTAL RESULTS

We perform the SVD of Figure 2 using document-pairs from the HKNews corpus, and retain about $R = 700$ singular values in S and the corresponding singular vectors U and V. This provides the basis for the word-to-word similarity computation of (2). We project the NAB-TDT corpus on to this space via (1), and use the resulting representations $\overline{V}$ for the English documents. This provides the basis for the CLIR step used to match a Mandarin story being transcribed by the ASR system with English documents in NAB-TDT.

We begin by rescoring the 300-best lists from the bigram lattices with trigram models. The results for the XINHUA and the HUB-4NE corpora are reported in Table 1. For each test story d_i^C, we then perform CLIR using the first pass

Language Model	Perp	WER	CER	p-value
XINHUA Trigram	426	49.9%	28.8%	–
LSA-interpolated	364	49.3%	28.9%	0.043
Trig+LSA-intpl	351	49.0%	28.7%	0.002
CL-interpolated	346	48.8%	28.4%	< 0.001
HUB4-NE Trigram	1195	60.1%	44.1%	–
LSA-interpolated	695	58.6%	43.1%	<0.001
Trig+LSA-intpl	686	58.7%	43.2%	<0.001
CL-interpolated	630	58.8%	43.1%	< 0.001

Table 1. Word-Perplexity and ASR WER comparisons

ASR output to choose the most similar English documents d_i^E's from NAB-TDT. Then we create the cross-lingual unigram of (3). We next find the interpolation weight λ in (4) that maximizes the likelihood of the 1-best hypotheses of all test utterances in a story obtained from the first ASR pass. We finally rescore the 300-best lists using the *LSA-interpolated* LM, and report results[3] in Table 1.

For comparison, we also report the *CL-interpolated* results of [4] which use a superior translation lexicon derived from a sentence-aligned corpus, both for CLIR to find d_i^E and instead of $P_{\text{LSA}}(c|e)$ in (3). Finally, we note that the technique of cross-lingual lexical triggers reported in [5] also assumes only a document-aligned corpus as done here, and an interpolation of their model with ours does not require any additional resources. We perform this interpolation and report the results as *Trig+LSA-intpl* in Table 1.

As Table 1 shows, the LSA-interpolated model shows significant reduction in both perplexity (15-42%) and WER (0.6-1.5% absolute) over the baseline trigram model both when a moderate amount of LM training text is available (XINHUA) and when it is really scarce (HUB4-NE). It also performs only slightly worse than the CL-interpolated model of [4], which requires the more expensive sentence-aligned corpus. Finally, the interpolation of our LSA-based model and the trigger-based model of [5] brings further gains, removing the remaining gap from the CL-interpolated model: the p-values of the differences between CL-interpolated and Trig+LSA-intpl models are 0.58 for XINHUA and 0.79 for HUB4-NE. Since a large document-aligned corpus is much easier to obtain than a large sentence-aligned one, our technique has the potential for further gains from larger bilingual training sets.

6. CONCLUSIONS

We have demonstrated cross-lingual language modeling techniques that require the bilingual corpus only to be document-aligned, which is a realistic reflection of the situation in a resource-deficient language. Effectively, we have proposed methods to build cross-lingual language models which do not require machine translation. By using latent semantic analysis of a *document-aligned* corpus, we we have demonstrated a significant reduction in perplexity (18-42%) and WER (0.9-1.4%) over a trigram model. Performance statistically indistinguishable from a previously published method predicated on good MT capabilities can be achieved by our methods.

We are developing ways to extend the LSA-based model beyond the cross-lingual unigram statistics (2) to higher order N-grams.

[3] All p-values are based on the standard NIST MAPSSWE test, indicating statistical significance of WER improvement over the trigram baseline.

7. REFERENCES

[1] K. Kirchhoff, J. Bilmes, J. Henderson, R. Schwartz, P. Schone, M. Noamany, S. Das, G. Ji, M. Egan, and F. He, "Novel speech recognition models for arabic," *Johns Hopkins Summer Workshop*, 2002.

[2] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *Proc. ICSLP*, Sydney, Australia, 1998, vol. 5, pp. 1819–1822.

[3] W. Byrne, P. Beyerlein, J. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and W. Wang., "Towards language independent acoustic modeling," in *Proc. ICASSP*, 2000, vol. 2, pp. 1029–1032.

[4] S. Khudanpur and W. Kim, "Using cross-language cues for story-specific language modeling," in *Proc. ICSLP*, Denver, CO, 2002, vol. 1, pp. 513–516.

[5] W. Kim and S. Khudanpur, "Cross-lingual lexical triggers in statistical language modeling," in *Proc. of EMNLP*, Sapporo, Japan, 2003, pp. 17–24.

[6] S. Dumais, T. Letsche, L. Littman, and T. Landauer, "Automatic cross-language retrieval using latent semantic indexing," in *AAAI Spring Symposuim on Cross-Language Text and Speech Retrieval*, 1997.

[7] M. Berry, S. Dumais, and G. O'brien, "Using linear algebra for intelligent information retrieval," *SIAM review*, vol. 37(4), pp. 573–595, 1995.

[8] N. Coccaro and D. Jurafsky, "Towards better integration of semantic predictors in statistical language modeling," in *Proc. ICSLP*, Sydney, Australia, 1998, vol. 6, pp. 2403–2406.

[9] P. Fung, W. Byrne, Z. Fang, T. Kamm, L. Yi, S. Zhangjian, V. Venkataramani, and U. Ruhi, "Pronunciation modeling of mandarin casual speech," *Johns Hopkins Summer Workshop*, 2000.