Cross-Lingual Language Modeling with Syntactic Reordering for Low-Resource Speech Recognition

Ping Xu and Pascale Fung
Human Language Technology Center
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
`xuping@ust.hk, pascale@ece.ust.hk`

Abstract

This paper proposes cross-lingual language modeling for transcribing source resource-poor languages and translating them into target resource-rich languages if necessary. Our focus is to improve the speech recognition performance of low-resource languages by leveraging the language model statistics from resource-rich languages. The most challenging work of cross-lingual language modeling is to solve the syntactic discrepancies between the source and target languages. We therefore propose syntactic reordering for cross-lingual language modeling, and present a first result that compares inversion transduction grammar (ITG) reordering constraints to IBM and local constraints in an integrated speech transcription and translation system. Evaluations on resource-poor Cantonese speech transcription and Cantonese to resource-rich Mandarin translation tasks show that our proposed approach improves the system performance significantly, up to 3.4% relative WER reduction in Cantonese transcription and 13.3% relative bilingual evaluation understudy (BLEU) score improvement in Mandarin transcription compared with the system without reordering.

1 Introduction

Statistical language modeling techniques have achieved remarkable success in speech and language processing (Clarkson and Rosenfeld, 1997; Stolcke, 2002). However, this success largely depends on the availability of a large amount of suitable text data in a language. Without sufficient text data for training, it is very difficult to build a practical and usable statistical language model. Therefore, most of the advances have been reported in so called *resource-rich* language such as English, Mandarin and Japanese, after creating linguistic resources of these languages at considerable cost. Today there are more than 6000 living languages spoken in the world (Gordon et al., 2005), and most of them have little transcribed texts and are considered as *resource-poor* languages (Nakov and Ng, 2009). Many of these languages are actually spoken by a huge number of speakers (e.g. some Chinese and Indian languages), and thus there is still a great demand to build speech and language processing systems for these languages.

Owing to data scarcity, most often an interpolation (Bellegarda, 2004) of language models between a resource-poor language and a resource-rich language is used in most low-resource ASR systems. Some researchers have proposed transforming resource-rich language models to resource-poor language models by word-level transduction, either in a context-independent or context-dependent manner (Hori et al., 2003; Akita and Kawahara, 2006; Jensson et al., 2009; Neubig et al., 2010). In (Jensson et al., 2009), a simple dictionary based context-independent transduction from a resource-rich language to a resource-poor language is exploited to improve speech recognition of the resource-poor language. In (Hori et al., 2003; Akita and Kawahara, 2006; Neubig et al., 2010), context-dependent transduction is exploited. In their case, the resource-poor language is a spoken language, and the resource-rich language is a written language. They carried out language model transformation since the input speech

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 766–776, Jeju Island, Korea, 12–14 July 2012.

is in speaking-style and the output text is in written-style.

Others have investigated cross-lingual information between a resource-poor language and a resource-rich language. In (Khudanpur and Kim, 2002), cross-language cues are used to improve a language model of a resource-poor language. They used cross-lingual unigram probabilities trained from a story-specific parallel corpus of the resource-poor and resource-rich languages. They interpolate the language model of the resource-poor language with those unigram probabilities. In (Kim and Khudanpur, 2003), an n-gram language model in a resource-poor language is interpolated with cross-lingual unigram trigger probabilities. These triggers are word pairs of the resource-poor and resource-rich languages with the highest mutual information across these two languages. Another way of estimating those unigram probabilities is using latent semantic analysis by measuring cosine similarities from a document-aligned corpus for any given word pair (Kim and Khudanpur, 2004).

Both interpolation and word-level transduction approaches fail to meet the challenge of syntactic discrepancies between the resource-poor and resource-rich languages. This syntactic discrepancies exist, for example, even between the Sinitic languages and Indian languages[1] of the same family. Sinitic languages such as Cantonese/Yue, Shanghai/Wu, etc. are officially considered as "dialects" of the standard Chinese Mandarin (or Putonghua)[2]. However, they differ greatly from Mandarin in all aspects and are not mutually comprehensible. For instance, in addition to lexical and pronunciation differences, Cantonese Chinese (Lee, 2011) differs syntactically from Mandarin as well - we found that there are approximately 10% syntactic inversions between sentences of the two forms of Chinese.

We suggest that a better approach than interpolation and word-level transduction is to use *cross-lingual language modeling* with syntactic reordering. A reordering model with reordering constraints, such as ITG constraints (Wu, 1997), IBM constraints (Berger et al., 1996), and local constraints (Kumar and Byrne, 2005) can account for the syntactic differences. It has been shown in (Zens and Ney, 2003; Kanthak et al., 2005; Dreyer et al., 2007) that ITG constraints perform better than other constraints when tackling the reordering between many language pairs. Previous work on weighted finite-state transducer (WFST) based speech translation such as (Casacuberta et al., 2004; Zhou et al., 2005; Zhou et al., 2006; Mathias and Byrne, 2006; Matusov et al., 2006; Saon and Picheny, 2007) only train the reordering model using IBM constraints, local constraints or ad hoc rules. We will use ITG constraints, which have only been applied to text translation tasks before, to model the syntactic differences in cross-lingual language modeling for speech recognition.

We will implement a *cross-lingual language model* using WFSTs, and integrate it into a WFST-based speech recognition search space to give both resource-poor language and resource-rich language transcriptions. This creates an integrated speech transcription and translation framework.

This paper is organized as follows: Section 2 presents our proposed cross-lingual language modeling with syntactic reordering. In Section 3, we discuss speech recognition with cross-lingual language models. Section 4 and 5 give the experimental setup and results. We conclude our work at the end of this paper.

2 Cross-lingual Language Modeling with Syntactic Reordering

In automatic speech recognition (ASR), given an observed source speech vector X, the decoding process searches the best word sequence $\hat{v}_1^I$ (consists of words $v_1, v_2, ..., v_I$) by maximizing the posterior probability $P(v_1^I|X)$, where v_1^I is the source transcript representing the transcription of the source speech (see Eq. (1)). According to Bayes' law, we can decompose $P(v_1^I|X)$ into an acoustic model $P(X|v_1^I)$ and a language model $P(v_1^I)$. If a source language L_v is a resource-rich language, then the language model $P(v_1^I)$ can be well estimated from sufficient training texts. However, if the source lan-

[1] For example, Hindi and Malayalam (Geethakumary, 2002).

[2] Since Cantonese does not have an official written form, there are very few written texts available for training language models. In this paper, we treat Cantonese as a typical resource-poor language and Mandarin as a typical resource-rich language. This language pair will be used for illustration purposes throughout this paper.

guage L_v is a resource-poor language, then the language model $P(v_1^I)$ cannot be reliably or robustly estimated due to lack of training texts.

$$\begin{aligned}
\hat{v}_1^I &= \arg\max_{v_1^I} P(v_1^I|X) && (1)\\
&= \arg\max_{v_1^I} P(X|v_1^I)P(v_1^I)\\
&= \arg\max_{v_1^I} P(X|v_1^I)\sum_{w_1^J} P(v_1^I|w_1^J)P(w_1^J)\\
&\approx \arg\max_{v_1^I} P(X|v_1^I)\max_{w_1^J} P(v_1^I|w_1^J)P(w_1^J)
\end{aligned}$$

Since this paper tackles the language modeling challenge for low-resource speech recognition, here we just assume that the source language L_v is a resource-poor language. We further assume that there is a target language L_w, which is a resource-rich language closely related to the language L_v. In order to improve the language model $P(v_1^I)$ of the resource-poor language L_v, we introduce *cross-lingual language modeling* by decomposing the language model $P(v_1^I)$ into a translation model $P(v_1^I|w_1^J)$ and a language model $P(w_1^J)$ of the resource-rich language L_w (see Eq. (1)). w_1^J is the target resource-rich language transcript that consists of words $w_1, w_2, ..., w_J$. $P(v_1^I|w_1^J)P(w_1^J)$ is defined as a *cross-lingual language model*. It leverages the abundant statistics from the language model $P(w_1^J)$ to improve the language model $P(v_1^I)$ of the resource-poor language.

The translation model $P(v_1^I|w_1^J)$ can be estimated by addressing the discrepancies between the resource-poor language L_v and the resource-rich language L_w, which can be modeled from a parallel corpus of the L_v transcript v_1^I and the L_w transcript w_1^J. For the syntactic inversions, we reorder the word or phrase positions of the L_w language model into those of the L_v language model. We have observed that most of the words are aligned monotonically between L_v and L_w within a phrase. This paper, therefore only considers phrase-level reordering, which effectively preserves the monotonic word sequences within phrases, and significantly reduces the number of reordering paths compared with word-level reordering.

2.1 Preprocessing: Phrase Extraction and Segmentation

Our discussion starts with phrase extraction from the parallel corpus. We define a phrase sequence $\tilde{v}_1^K$ (consists of phrases $\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_K$) segmented from the word-level L_v transcript v_1^I and $\tilde{w}_1^K$ (consists of phrases $\tilde{w}_1, \tilde{w}_2, ..., \tilde{w}_K$) segmented from the word-level L_w transcript w_1^J. Furthermore, we define a reordering sequence r_1^K, of which the detail can be found in Section 2.2.

The phrase-level translation model $P(v_1^I|w_1^J)$ is decomposed into four components (see Eq. (2)): *segmentation model* $P(\tilde{w}_1^K|w_1^J)$, *phrasal reordering model* $P(r_1^K|\tilde{w}_1^K, w_1^J)$, *phrase-to-phrase transduction model* $P(\tilde{v}_1^K|r_1^K, \tilde{w}_1^K, w_1^J)$ and *reconstruction model* $P(v_1^I|\tilde{v}_1^K, r_1^K, \tilde{w}_1^K, w_1^J)$. Before presenting each component model, we need to extract two phrase tables for the L_v transcript and the L_w transcript, respectively.

$$\begin{aligned}
P(v_1^I|w_1^J) \approx \max_{\tilde{v}_1^K, r_1^K, \tilde{w}_1^K} & P(\tilde{w}_1^K|w_1^J)\cdot\\
& P(r_1^K|\tilde{w}_1^K, w_1^J)\cdot\\
& P(\tilde{v}_1^K|r_1^K, \tilde{w}_1^K, w_1^J)\cdot\\
& P(v_1^I|\tilde{v}_1^K, r_1^K, \tilde{w}_1^K, w_1^J) \quad (2)
\end{aligned}$$

The phrase extraction is based on word-to-word alignments of the parallel corpus. We train word alignments in both directions with GIZA++, and then symmetrize the two alignments using the *refined method* (Och and Ney, 2003). Figure 1 shows an example of word-to-word alignment results between an L_v transcript (Cantonese) and an L_w transcript (Mandarin), from which phrase-to-phrase alignments are derived by identifying deletion, substitution, insertion and inversion.

Prior to phrasal reordering, the segmentation model $P(\tilde{w}_1^K|w_1^J)$ implemented by a segmentation WFST S_w is applied to segment a word sequence w_1^J in the L_w language model into a phrase sequence $\{\tilde{w}_1, \tilde{w}_2, ..., \tilde{w}_K\}$. The maximum number of words that can be segmented into one phrase is controlled by a *segmentation order* s. An example of S_w is shown in Figure 3(a1). It segments a word sequence $\{w_1, w_2, w_3\}$ into a phrase sequence $\{w_1, w_2\_w_3\}$ after performing *composition* (Mohri, 2009) with the target L_w language model (see Figure 3(b1 & b2))[3].

[3]The "_" symbol is used to indicate the concatenation of con-

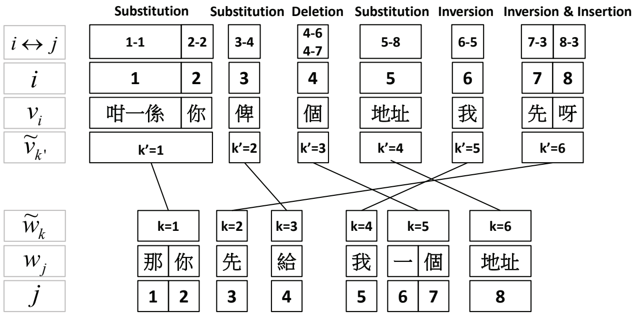


Figure 1: An example (in English: Please give me an address first) of phrase extraction from word-to-word alignments. i and j are word indexes. k' and k are phrase indexes. $i \leftrightarrow j$ represents the word-to-word alignment. $k \leftrightarrow k'$ represents the indentified phrase-to-phrase alignment.

2.2 Phrasal Reordering Model

Given a phrase sequence $\{\tilde{w}_1, \tilde{w}_2, ..., \tilde{w}_K\}$ of the L_w transcript, the role of the reordering model $P(r_1^K | \tilde{w}_1^K, w_1^J)$ is to reorder phrase positions of the L_w transcript into those of the L_v transcript by permutation of $\tilde{w}_1^K$ according to a reordering sequence $\{r_1^K : r_k \in \{1, 2, ..., K\}, r_k \neq r_{k' \neq k}\}$. The phrase sequence $\{\tilde{w}_1, \tilde{w}_2, ..., \tilde{w}_K\}$ is therefore reordered into $\{\tilde{w}_{r_1}, \tilde{w}_{r_2}, ..., \tilde{w}_{r_K}\}$ consequently (see Figure 2 where $K = 3$). Since arbitrary permutations of K phrases are NP-hard (Knight, 1999), reordering constraints have to be set over r_1^K to reduce the number of permutations.

There are three reordering constraints widely used in statistical machine translation, namely local constraints, IBM constraints and ITG constraints. Here we would like to point out that this is the first time that reordering constraints have been incorporated into a cross-lingual language model for speech recognition.

Reordering Constraints

Local constraints make the restriction that one phrase can jump at most $L-1$ phrases either forward or backward, where L is the reordering distance (or window size of permutation)[4]. The generation of r_1^K under local constraints can be viewed as solving of the following problem (Kløve, 2009):

How many permutations of $\{1, 2, \ldots k \ldots, K\}$ satisfy $|r_k - k| < L$ for all k?

IBM constraints, a superset of local constraints (Dreyer et al., 2007), generate permutations r_1^K deviate from the monotonic phrase order $\{r_1^K : r_k = k\}$. More specifically, any phrase position r_k can be selected from the positions of the first m yet uncovered phrases (see Eq. (3)). A typical value of m is 4 (Zens and Ney, 2003), and we write IBM constraints with $m = 4$ as IBM(4).

$$r_k \in \begin{cases} \{1, 2, ..., k-1+m; r_k \neq r_{k' \neq k}\} \\ \text{if } k \leq K+1-m, \\ \{1, 2, ..., K; r_k \neq r_{k' \neq k}\} \\ \text{if } K+1-m < k \leq K. \end{cases} \tag{3}$$

ITG constraints provide a more faithful coverage of syntactic reordering in the parallel data than local constraints and IBM constraints. Our presentation of ITG constraints starts with defining of some permutation sets. Let S_K be the set of permutations on {1,2,…,K}. A permutation $r_1^K \in S_K$, where $r_1^K = r_1 r_2 \ldots r_K$, contains a subsequence of type $\tau \in S_M$ if and only if a sequence of indices $1 \leq i_1 < i_2 < \ldots < i_M \leq K$ exists such that $r_{i_1} r_{i_2} \ldots r_{i_M}$ has all the same pairwise comparisons as τ. We denote the set of permutations of S_K not containing subsequences of type τ by $S_K(\tau)$. If we have sets $S_K(\tau_1), \ldots, S_K(\tau_p)$, we denote the set $S_K(\tau_1) \cap \ldots \cap S_K(\tau_p)$ by $S_K(\tau_1, \ldots, \tau_p)$ (Barcucci et al., 2000). ITG constraints allow the permutation set $S_K(3142, 2413)$, which forbids subsequence of type $(3, 1, 4, 2)$ and its dual $(2, 4, 1, 3)$. Explicitly, ITG constraints avoid any permutation r_1^K satisfying either $r_{i_2} < r_{i_4} < r_{i_1} < r_{i_3}$ or $r_{i_3} < r_{i_1} < r_{i_4} < r_{i_2}$, where $1 \leq i_1 < i_2 < i_3 < i_4 \leq K$. In (Wu, 1997), these forbidden subsequences are called "inside-out" transpositions. They are fairly distorted matchings, and hardly observed in real parallel data.

In order to get an intuitive sense of the reordering capability of those three constraints, we list the number of permutations under local constraints, IBM constraints as well as ITG constraints[5] in Table 1.

secutive words forming a phrase.

[4] The concept of reordering distance also applies to other constraints.

[5] Interestingly, when $K = L$, the number of permutations under ITG constraints $N_{ITG} = |S_K(3142, 2413)|$, and $|S_K(3142, 2413)|$ equals the $K-1$-th Schröder numbers s_{K-1} (Ehrenfeucht et al., 1998)

Table 1: Comparison of permutation number under local constraints (N_{Local}), IBM constraints ($N_{IBM(4)}$) and ITG constraints (N_{ITG}). The comparison is constrained by the phrase number K and the reordering distance L.

		K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
	N_{Local}	2	3	5	8	13	21	34	55	89
L=2	$N_{IBM(4)}$	2	3	5	8	13	21	34	55	89
	N_{ITG}	2	3	5	8	13	21	34	55	89
	N_{Local}	2	6	14	31	73	172	400	932	2177
L=3	$N_{IBM(4)}$	2	6	14	31	73	172	400	932	2177
	N_{ITG}	2	6	12	25	57	124	268	588	1285
	N_{Local}	2	6	24	78	230	675	2069	6404	19708
L=4	$N_{IBM(4)}$	2	6	24	78	230	675	2069	6404	19708
	N_{ITG}	2	6	22	52	122	321	885	2304	5880
	N_{Local}	2	6	24	120	504	1902	6902	25231	95401
L=5	$N_{IBM(4)}$	2	6	24	96	330	1066	3451	11581	39264
	N_{ITG}	2	6	22	90	236	602	1714	5269	16385
	N_{Local}	2	6	24	120	720	3720	17304	76110	329462
L=6	$N_{IBM(4)}$	2	6	24	96	384	1374	4718	16275	57749
	N_{ITG}	2	6	22	90	394	1108	3014	9038	29618

We can see that given the same K ($K \leq 10$) and L ($L \leq 6$), IBM constraints have less permutations than local constraints, and ITG constraints have less permutations than IBM constraints in general (only one exception when $K = L = 6$). These observations indicate that ITG constraints can filter out more unlikely permutations for a fixed reordering distance, resulting in longer distance reordering capability.

Table 1 also tells us that the phrase number K and the reordering distance L for any of the constraints cannot be too large for practical implementation. For instance, if $L = 6$ and K goes from 6 to 7, the order of magnitude of N_{Local}, $N_{IBM(4)}$ and N_{ITG} increases from 2 to 3. Hence, phrases for permutation should be selective to cover the most possible re-orderings. If long reordering distances are allowed, unlikely permutations should be pruned so that the memory consumption becomes manageable.

Reordering Sequence Distribution

So far we have discussed the issue that how to generate permutations for the reordering model using reordering constraints. Another issue is how to parameterize the reordering sequence distribution. Both ITG constraints and other constraints assume that all permutations are equally probable. However, it makes sense to restrict those non-monotonic reorderings when performing the translation. This not only helps the search of the most likely permutation, but also guides the pruning of unlikely permutations.

$$
\begin{aligned}
P(r_1^K|\tilde{w}_1^K, w_1^J) &= P(r_1)\prod_{k=2}^{K} P(r_k|r_{k-1}, \tilde{w}_1^K) \\
&= P(r_1)\prod_{k=2}^{K} P(r_k|r_{k-1}) \qquad (4)
\end{aligned}
$$

We make a first order Markov assumption over the phrasal reordering model $P(r_1^K|\tilde{w}_1^K, w_1^J)$ (see Eq. (4)). The reordering sequence distribution is parameterized to assign decreasing likelihood to phrase reorderings $\{\tilde{w}_{r_1}, \tilde{w}_{r_2}, \ldots, \tilde{w}_{r_K}\}$ that diverge from the original word order (Och et al., 1999; Kumar et al., 2005). Suppose $\tilde{w}_{r_k} = w_l^{l'}$ and $\tilde{w}_{r_{k-1}} = w_q^{q'}$, the reordering sequence distribution is set as Eq. (5), where p_0 is a tuning factor. We normalize the probabilities $P(r_k|r_{k-1})$ such that $\sum_{k'=1, k'\neq r_{k-1}}^{K} P(r_k = k'|r_{k-1}) = 1$.

$$
\begin{aligned}
P(r_k|r_{k-1}) &= p_0^{|l-q'-1|} \\
P(r_1 = k) &= \frac{1}{K}; k \in \{1, 2, ..., K\}
\end{aligned} \qquad (5)
$$

Assume that we have a phrase sequence $\{\tilde{w}_1, \tilde{w}_2, \tilde{w}_3\}$, Figure 2 shows the phrasal reordering model implemented by a reordering WFST Ω_r under the first order Markov assumption for this phrase sequence.

Figure 3(a2) gives one more example of Ω_r, which reorders the phrase sequence $\{w_1, w_2\_w_3\}$ into $\{w_2\_w_3, w_1\}$[6]. Within the WFST paradigm, reordering models under any of those constraints can be integrated into the cross-lingual language model.

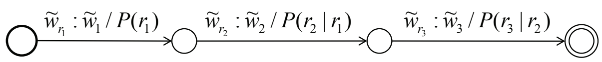


Figure 2: An example of reordering WFST Ω_r implementing the phrasal reordering model under the first order Markov assumption.

2.3 Phrase-to-Phrase Transduction Model

Once the phrase sequence of the L_w transcript is reordered into the L_v transcript order, we use the phrase-to-phrase transduction model specified in Eq. (6) to perform the cross-language transduction. Given sufficient parallel training data, the context-dependent phrase-to-phrase transduction model can be estimated using the GIATI method (Casacuberta and Vidal, 2004). However, for the translation task with scarce training data, the context-dependent transduction probabilities may not be reliably estimated. Therefore, we assume that a phrase $\tilde{v}_k$ is generated independently by each phrase $\tilde{w}_{r_k}$. $C(\tilde{v}_k, \tilde{w}_{r_k})$ is the number of times that phrase $\tilde{v}_k$ is aligned to $\tilde{w}_{r_k}$ in the parallel corpus. This model can be implemented by a WFST T_{vw} which transduces $\tilde{v}_k$ to $\tilde{w}_{r_k}$. Figure 3(a3) shows an example of T_{vw} transducing $v_2\_v_3$ to $w_2\_w_3$.

$$
\begin{aligned}
P(\tilde{v}_1^K | r_1^K, \tilde{w}_1^K, w_1^J) &= P(\tilde{v}_1^K | r_1^K, \tilde{w}_1^K) \\
&= \prod_{k=1}^{K} P_k(\tilde{v}_k | \tilde{w}_{r_k}) \\
&= \prod_{k=1}^{K} \frac{C(\tilde{v}_k, \tilde{w}_{r_k})}{\sum_{\tilde{v}_k} C(\tilde{v}_k, \tilde{w}_{r_k})} \qquad (6)
\end{aligned}
$$

2.4 Reconstruction Model

Reconstruction model $P(v_1^I | \tilde{v}_1^K, r_1^K, \tilde{w}_1^K, w_1^J)$ operates in the opposite direction as the segmentation model. It generates a word sequence v_1^I from a phrase sequence $\tilde{v}_1^K$. The reconstruction model can be implemented by a WFST R_v. An example of R_v is shown in Figure 3(a4), which reconstructs a phrase $v_2\_v_3$ into a word sequence $\{v_2, v_3\}$.

[6]For simplicity, reordering sequence distributions are not shown there.

3 Speech Recognition with Cross-Lingual Language Models

The translation model $P(v_1^I | w_1^J)$ can be constructed via WFST *composition* (denoted by $\circ$) (Mohri, 2009) of all the component models as shown in Eq. (7) and Figure 3, where $\mathcal{T}$ is the final composed WFST that transduces v_1^I to w_1^J.

$$\mathcal{T} = R_v \circ T_{vw} \circ \Omega_r \circ S_w \qquad (7)$$

The cross-lingual language model G_{cl} is constructed through composition (see Eq. (8)) of the translation model and a resource-rich language model G.

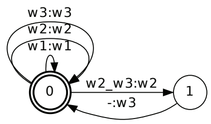
$$G_{cl} = \mathcal{T} \circ G = R_v \circ T_{vw} \circ \Omega_r \circ S_w \circ G \qquad (8)$$

As the way of integrating a resource-rich language model G into ASR search space (Mohri et al., 2008), we can integrate the cross-lingual language model G_{cl} into ASR search space in a globally optimized way as well. The search space can be implemented using a transducer ASR, which is formulated with a unified WFST approach as shown in Eq. (9). Here H transduces HMM states to context-dependent phones. C represents a transduction from context-dependent phones to context-independent phones. L is a lexicon transducer which maps context-independent phone sequences to word strings restricted to the input symbols of the cross-lingual language model transducer G_{cl}.

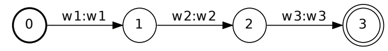
$$ASR = H \circ C \circ L \circ G_{cl} \qquad (9)$$

Eq. (9) outputs the recognition result in a resource-rich language. If recognition system requires recognition outputs in a resource-poor language, then the search space should be constructed as Eq. (10), where π is a *projection* (Mohri, 2009) operator which projects the input label to the output label. Before decoding, the recognition transducer ASR can be optimized by a determinization operation right after each composition.

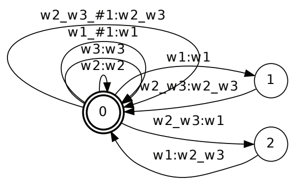
$$ASR = H \circ C \circ L \circ \pi(G_{cl}) \qquad (10)$$



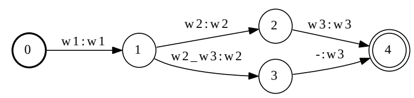
(a1) Segmentation WFST S_w



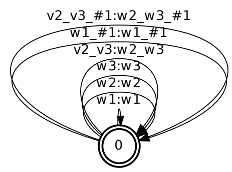
(b1) Written-style language model G



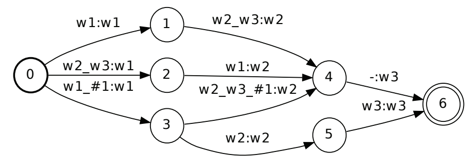
(a2) Reordering WFST Ω_r



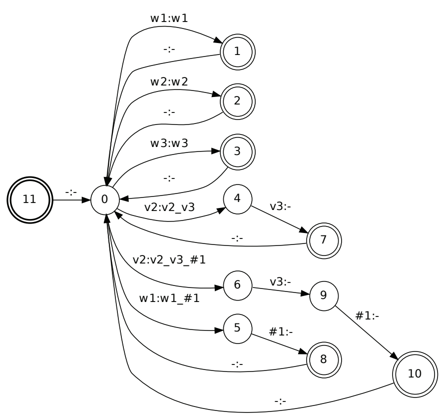
(b2) $S_w \circ G$



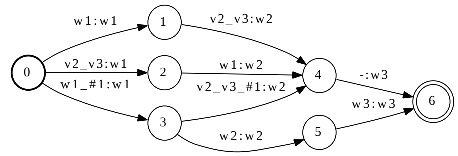
(a3) Phrase-to-phrase transduction WFST T_{vw}



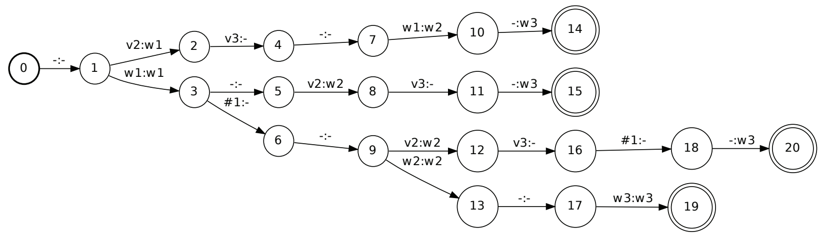
(b3) $\Omega_r \circ S_w \circ G$



(a4) Reconstruction WFST R_v



(b4) $T_{vw} \circ \Omega_r \circ S_w \circ G$



(b5) $R_v \circ T_{vw} \circ \Omega_r \circ S_w \circ G$

Figure 3: Illustration of constructing a cross-lingual language model via WFSTs: a word sequence $\{w_1, w_2, w_3\}$ represented by the L_w language model G (b1) is segmented into a phrase sequence $\{w_1, w_2\_w_3\}$ (b2); $\{w_1, w_2\_w_3\}$ is reordered into $\{w_2\_w_3, w_1\}$ (b3); phrase $w_2\_w_3$ is transduced to $v_2\_v_3$ (b4); phrase $v_2\_v_3$ is reconstructed into a word sequence $\{v_2, v_3\}$ (b5). wk and vk represent w_k and v_k, respectively. "-" refers to ϵ or null symbol. Auxiliary symbols $\#1, \#2, \cdots$ are used to make the WFST *determinizable* (Mohri, 2009) such that the transducer can be optimized by a *determinization* (Mohri, 2009) operation which significantly reduces the search network size.

4 Experimental Setup

4.1 Corpus and Model Training

To investigate the performance of our proposed cross-lingual language models, we have chosen Cantonese as a resource-poor language and Mandarin as a resource-rich language. We have collected Cantonese parliamentary speech from the Hong Kong Legislative Council. Currently we only have 4152 parallel transcribed sentences containing 19.4 hours of speech. It is separated into three sets, a training set (11.9 hours, 2700 sentences), a development set (3.7 hours, 788 sentences), and an evaluation set (3.8 hours, 664 sentences). The sentences in the evaluation set are a bit longer than those in the development set. The parallel transcriptions of the training set constitute a parallel corpus, which includes Cantonese transcription (manual transcription) of 106k words and Mandarin transcription (Hansard[7] transcription) of 80k words. The statistics of substitutions, insertions, deletions and inversions identified in the parallel corpus are shown in Table 2. Besides the parallel corpus, we have a set of additional Mandarin transcriptions, which has 31M words.

Table 2: No. of substitutions, insertions, deletions and inversions identified in the parallel corpus with different segmentation order s.

Segmentation Order	$s = 2$	$s = 3$	$s = 4$	$s = 5$
Substitutions	30921	22723	19011	17106
Insertions	4657	3820	3641	3295
Deletions	1365	1158	1066	1030
Inversions	3000	2876	2814	2779
Total	39943	30577	26532	24210

The training set is used for training an acoustic model (including H and C) using a Maximum Likelihood criterion. It adopts 13 MFCC coefficients, together with 13 delta coefficients and 13 acceleration coefficients as the acoustic features. The acoustic model comprises 73 Hidden Markov Models (HMMs) to represent 70 Cantonese phonemes as well as silence, short pause, and noise. During the acoustic model training, tied-state cross-word triphones are constructed by decision tree clustering.

[7]Hansard is a name of the printed transcripts of parliamentary debates.

The parallel corpus is used for training the translation model $\mathcal{T}$. Together with the parallel corpus, the additional Mandarin transcriptions are used for training an interpolated word-level trigram language model G, where the lexicon size is about 28K. A modified scheme of Kneser-Ney discounting is applied for the language model G with a back-off threshold of 1 for unigram and 2 for bigram. The cross-lingual language model G_{cl} can be obtained by composition of $\mathcal{T}$ and G.

4.2 Decoding and Evaluation Method

Decoding of the speech recognition search space ASR is performed by T^3 Decoder (Dixon et al., 2009), which is a state-of-the-art WFST-based LVCSR speech decoder. Decoding of ASR in Eq. (9) gives Mandarin outputs. Decoding of ASR in Eq. (10) gives Cantonese outputs.

In our experiments, we use the following evaluation criteria:

WER (word error rate). The WER is computed as the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated sentence into the reference sentence (Zens et al., 2004). The WER relates the speech recognition accuracy. The lower WER, the better.

BLEU (bilingual evaluation understudy) score. The BLEU score measures the precision of n-grams (unigrams, bigrams, trigrams and fourgrams) with respect to a reference translation with a penalty for too short sentences (Papineni et al., 2002). The BLEU score reflects the translation accuracy. The larger BLEU score, the better.

We perform WER evaluation of decoding outputs of Eq. (10) and BLEU score evaluation of decoding outputs of Eq. (9) using the evaluation set. The WER evaluation is on the Cantonese output against the Cantonese reference transcription (manual transcription). The BLEU score evaluation is on the Mandarin output against the Mandarin reference transcription (Hansard transcription).

4.3 Parameter Settings

The performance of our proposed cross-lingual language models is sensitive to many parameters. Firstly, segmentation order s affects phrase extraction. The optimal value depends on the language

Table 3: WER and BLEU score for decoding results of $H \circ C \circ L \circ G$, $H \circ C \circ L \circ \pi(G_{cl})$ without reordering, and $H \circ C \circ L \circ \pi(G_{cl})$ with reordering under various constraints.

Models	$H \circ C \circ L \circ G$	$H \circ C \circ L \circ \pi(G_{cl})$ $G_{cl} = \mathcal{T}_3 \circ G$ $\mathcal{T}_3 = R_v \circ T_{vw} \circ S_w$	$H \circ C \circ L \circ \pi(G_{cl})$ $G_{cl} = \mathcal{T}_3 \circ G, \mathcal{T}_3 = R_v \circ T_{vw} \circ \Omega_r \circ S_w$		
			Local Constraints	IBM Constraints	ITG Constraints
WER(%)	29.85	27.05	26.35	26.20	26.13
BLEU	N/A	29.23	32.29	32.81	33.12

pair and the size of corpus. Secondly, p_0 in the first order Markov assumption affects the decoding results. Thirdly, the number of reordering permutations or paths are formidable when the reordering distance L is long as suggested by Table 1. Therefore, we apply histogram pruning to reordering paths, which only maintains top N most likely ones. The development set is used for tuning parameters p_0 and N.

5 Experimental Results

The evaluation results of the proposed cross-lingual language models G_{cl} with reordering under various constraints are presented in Table 3, where $G_{cl} = \mathcal{T}_s \circ G = \mathcal{T}_3 \circ G$.[8] In general, reordering has a significant effect on enhancing the performance of recognition and translation in the sense of WER reduction and BLEU improvement. Compared with the cross-lingual language model without reordering, the cross-lingual language model with reordering under local constraints gives 0.70% absolute WER reduction and 3.06 absolute BLEU improvement. The cross-lingual language model with reordering under IBM constraints gives 0.85% absolute WER reduction and 3.58 absolute BLEU improvement. The cross-lingual language model with reordering under ITG constraints yields the best performance, with 0.92% absolute WER reduction and 3.89 absolute BLEU improvement. All WER improvements pointed out here are statistically significant at 99% confidence according to a two-proportional z-test, and all BLEU improvements are statistically significant at 95% confidence according to a paired student t-test using bootstrap resampling.

[8]We have chosen segmentation order $s = 3$ because it works the best in our system.

6 Conclusions

We have proposed cross-lingual language modeling with phrase-level syntactic reordering for low-resource speech recognition. The cross-lingual language modeling enriches a resource-poor language model by leveraging the language model from a closely related resource-rich language. It provides an effective method to solve the low-resource language modeling challenge by using a large amount of resource-rich language (e.g. Mandarin) data and a small amount of resource-poor language (e.g. Cantonese) data, as well as some parallel data of resource-poor and resource-rich languages. With a cross-lingual language model, our ASR system can decode speech into transcriptions, either in a resource-poor language or a resource-rich language, using a single WFST-based speech decoder.

We have presented a first end-to-end WFST source to target language transcription and translation system with syntactic reordering and global optimization. Our work is the first to use ITG constraints for the syntactic reordering in such an integrated system. We also did comparative study of ITG constraints, IBM constraints and local constraints in the reordering model, for completeness. We have also presented the determinizable design of each transducer for composing a cross-lingual language model such that we can optimize the search network by determinization. This is crucially important to successfully build a practical integrated system, and, of course, the work is extremely challenging.

Experiments on Cantonese recognition and Cantonese to Mandarin translation tasks have shown that our proposed cross-lingual language model substantially improves the performance of the recognition and translation. The best system gives 12.5% relative WER reduction in Cantonese (resource-poor language) transcriptions over the system using interpolation. The best reordering model gives 3.4% relative WER reduction and 13.3% relative BLEU score improvement in Mandarin (resource-rich language) transcriptions over the system without reordering. The improvements have been found to be statistically significant.

Even though the objective of our work is for speech recognition, our proposed cross-lingual language modeling can be easily applied to speech translation of other language pairs for efficient direct decoding from source speech to target text.

7 Acknowledgments

This work is partially supported by ITS/189/09 and CERG#612211. The authors would like to thank Dr. Tasuku Oonishi for providing access to the T^3 decoder, and thank Prof. Sadaoki Furui and his team for useful discussions. Thanks should go to Yue Yu and Percy Cheung for collecting the Cantonese and Mandarin parallel data. Thanks also go to Ricky Chan for training the Cantonese acoustic model and Dr. Markus Saers for helping on training the GIZA word-to-word alignment models.

References

Y. Akita and T. Kawahara. 2006. Efficient estimation of language model statistics of spontaneous speech via statistical transformation model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1049–1052.

E. Barcucci, A. Del Lungo, E. Pergola, and R. Pinzani. 2000. Permutations avoiding an increasing number of length-increasing forbidden subsequences. *Discrete Mathematics and Theoretical Computer Science*, 4(1):31–44.

J.R. Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech communication*, 42(1):93–108.

A.L. Berger, P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, A.S. Kehler, and R.L. Mercer. 1996. Language translation apparatus and method using context-based translation models. US Patent 5,510,981.

F. Casacuberta and E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225.

F. Casacuberta, H. Ney, F.J. Och, et al. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech & Language*, 18(1):25–47.

P. Clarkson and R. Rosenfeld. 1997. Statistical language modeling using the cmu-cambridge toolkit. In *5th European Conference on Speech Communication and Technology*.

P.R. Dixon, T. Oonishi, K. Iwano, and S. Furui. 2009. Recent development of wfst-based speech recognition decoder. In *Proceedings of 2009 APSIPA Annual Summit and Conference*, pages 138–147, Sapporo, Japan.

M. Dreyer, K. Hall, and S. Khudanpur. 2007. Comparing reordering constraints for smt using efficient bleu oracle computation. In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 103–110, Rochester, New York.

A. Ehrenfeucht, T. Harju, P. Ten Pas, and G. Rozenberg. 1998. Permutations, parenthesis words, and schröder numbers. *Discrete mathematics*, 190(1):259–264.

V. Geethakumary. 2002. A contrastive analysis of hindi and malayalam. *Language in India*.

R.G. Gordon, B.F. Grimes, and Summer Institute of Linguistics. 2005. *Ethnologue: Languages of the world*, volume 15. SIL International, Dallas TX, USA.

T. Hori, D. Willett, and Y. Minami. 2003. Language model adaptation using wfst-based speaking-style translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 228–231.

A.T. Jensson, T. Oonishi, K. Iwano, and S. Furui. 2009. Development of a wfst based speech recognition system for a resource deficient language using machine translation. In *Proceedings of APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, pages 50–56.

S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 167–174. Association for Computational Linguistics.

S. Khudanpur and W. Kim. 2002. Using cross-language cues for story-specific language modeling. In *7th International Conference on Spoken Language Processing*.

W. Kim and S. Khudanpur. 2003. Cross-lingual lexical triggers in statistical language modeling. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 17–24. Association for Computational Linguistics.

W. Kim and S. Khudanpur. 2004. Cross-lingual latent semantic analysis for language modeling. In *Proceed-*

ings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume 1, pages I257–I260. IEEE.

T. Kløve. 2009. Generating functions for the number of permutations with limited displacement. *The Electronic Journal of Combinatorics*, 16(R104).

K. Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607–615.

S. Kumar and W. Byrne. 2005. Local phrase reordering models for statistical machine translation. In *Proceedings of Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 161–168, Vancouver, Canada.

S. Kumar, Y. Deng, and W. Byrne. 2005. A weighted finite state transducer translation template model for statistical machine translation. *Natural Language Engineering*, 12(1):35–75.

J. Lee. 2011. Toward a parallel corpus of spoken cantonese and written chinese. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1462–1466, Chiang Mai, Thailand.

L. Mathias and W. Byrne. 2006. Statistical phrase-based speech translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 561–564.

E. Matusov, S. Kanthak, and H. Ney. 2006. Integrating speech recognition and machine translation: Where do we stand? In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages V1217–V1220. IEEE.

M. Mohri, F. C. N. Pereira, and M. Riley. 2008. Speech recognition with weighted finite-state transducers. *Handbook on Speech Processing and Speech Communication, Part E: Speech Recognition*.

M. Mohri. 2009. Weighted automata algorithms. *Handbook of Weighted Automata*, pages 213–254.

P. Nakov and H.T. Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 3, pages 1358–1367. Association for Computational Linguistics.

G. Neubig, Y. Akita, S. Mori, and T. Kawahara. 2010. Improved statistical models for smt-based speaking style transformation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5209.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

F.J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Joint SIGDAT Conf. on EMNLP and VLC*, pages 20–28, College Park, MD, USA.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

G. Saon and M. Picheny. 2007. Lattice-based viterbi decoding techniques for speech translation. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 386–389. IEEE.

A. Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing*.

D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

R. Zens and H. Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 144–151, Sapporo, Japan. Association for Computational Linguistics.

R. Zens, H. Ney, T. Watanabe, and E. Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 205–211, Geneva, Switzerland. Association for Computational Linguistics.

B. Zhou, S.F. Chen, and Y. Gao. 2005. Constrained phrase-based translation using weighted finite-state transducers. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1017–1020.

B. Zhou, S. F. Chen, and Y. Gao. 2006. Folsom: A fast and memory-efficient phrase-based approach to statistical machine translation. In *Spoken Language Technology Workshop*, pages 226–229. IEEE.