PHRASE-LEVEL TRANSDUCTION MODEL WITH REORDERING FOR SPOKEN TO WRITTEN LANGUAGE TRANSFORMATION

Ping Xu, Pascale Fung, Ricky Chan

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, ricky@cse.ust.hk

ABSTRACT

This paper proposes a first-ever phrase-level transduction model with reordering to transform colloquial speech directly to written-style transcription. This model is capable of performing n-m transductions. Our transduction model is trained from a parallel corpus of verbatim transcription and written-style transcription. Deletions, substitutions, insertions are well represented using this model. Inversion transduction cases can also be identified and represented. We implement our transduction model using weighted finite-state transducers (WFSTs), and integrate it into a WFST-based speech recognition search space to give both verbatim speaking-style and written-style transcriptions. Evaluations of our model on Cantonese speech to standard written Chinese show 11.59% relative Word Error Rate (WER) reduction over interpolated language model between Cantonese and standard Chinese speech, 5.72% relative WER reduction and 14.82% relative Bilingual Evaluation Understudy (BLEU) improvement over the word-level transduction model.

Index Terms— spoken to written language transformation, phrase-level transduction, reordering, WFST

1. INTRODUCTION

There is often a large discrepancy between colloquial speaking-style speech and written standard language. In the case of certain language groups, discrepancies can run the gamut from pronunciation, lexical to syntax, as is the case for Cantonese Chinese. Sinitic languages such as Cantonese/Yue, Shanghai/Wu, etc. are officially considered ”dialects” of the standard written Chinese Putonghua (or Mandarin). However, they differ greatly from Mandarin in all aspects and are not mutually comprehensible. In addition to lexical and pronunciation differences, Cantonese differs syntactically from Mandarin as well - we found that there are around 10% cases of syntactic inversion between sentences of the two forms of Chinese. Since Cantonese does not have an official written form, there are very little written texts available for training language models. Manual transcription of Cantonese is also more expensive because transcribers are not familiar with verbatim transcription of Cantonese.

Owing to the high cost of manual transcription, often an interpolation of language models between speaking-style transcription and standard written texts is used in most ASR systems. Others proposed transforming written-style language models to speaking-style language models by word-level transduction, either in a context-independent or context-dependent manner [1, 2, 3]. All such methods attempt to enrich colloquial language model by using a large amount of written-style texts and a small amount of colloquial speech. However, interpolated models do not have enough coverage and word-level transduction assumes 1-1/n-n transduction. Yet, n-m transduction is a common occurrence between spoken and written languages. More importantly, previous work [1, 2, 3] did not consider inversion cases, which frequently occur between Cantonese and Mandarin, as explained above. Inspired by the alignment template model [4] in statistical machine translation (SMT), we propose a phrase-level transduction model with reordering using WFSTs to take into account syntactic discrepancies between speaking-style speech (e.g. Cantonese) and written-style speech (e.g. Mandarin).

We also propose to integrate speech recognition and speaking-to-written style transcription transduction in a globally optimized single system. Previous work decoupled speech recognition and phrase-based translation into a two-step process [5, 6, 7]. We propose to instead incorporate phrase-level transduction into the ASR search network using a WFST-based speech recognition decoder [8] to output both verbatim Cantonese transcriptions and standard written Chinese transcriptions.

2. NOISY-CHANNEL MODEL FOR SPOKEN TO WRITTEN LANGUAGE TRANSFORMATION

In automatic speech recognition, given an observed speech vector X, the decoding process finds the best word sequence $\hat{v}_1^I$ (consists of words $v_1, v_2, ..., v_I$) by maximizing the posterior probability $P(v_1^I|X)$, in which v_1^I is the verbatim

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

transcript representing the faithful transcription of colloquial speech. According to Bayes' law, we can further decompose $P(v_1^I|X)$ into an acoustic model $P(X|v_1^I)$ and a language model (LM) $P(v_1^I)$. However, there is always a lack of verbatim transcript to train the LM $P(v_1^I)$. In order to estimate $P(v_1^I)$, we utilize LM transformation (see Eq. (1)) based on *noisy-channel model* to transform written-style transcript w_1^J (consists of words $w_1, w_2, ..., w_J$) to verbatim transcript v_1^I through maximization of $P(v_1^I|w_1^J)P(w_1^J)$. LM $P(w_1^J)$ can be trained from a large quantity of written-style transcript w_1^J, while *transduction model* $P(v_1^I|w_1^J)$ is estimated from a parallel corpus of aligned v_1^I and w_1^J.

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

3. PHRASE-LEVEL TRANSDUCTION MODEL

Instead of the word-level transduction model, we propose a phrase-level transduction model that not only allows n-m alignments, but also captures the inversion transduction cases.

We define a phrase sequence $\tilde{v}_1^K$ (consists of phrases $\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_K$) segmented from word-level verbatim transcript v_1^I and $\tilde{w}_1^K$ (consists of phrases $\tilde{w}_1, \tilde{w}_2, ..., \tilde{w}_K$) segmented from word-level written-style transcript w_1^J. Furthermore, we define a reordering sequence r_1^K, of which the detail can be found in section 3.2.

The phrase-level transduction model $P(v_1^I|w_1^J)$ is decomposed into four components (see Eq. (2)): *segmentation model* $P(\tilde{w}_1^K|w_1^J)$, *phrase 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 and its WFST implementation, we need to extract two phrase tables for the verbatim transcript and written-style transcript, respectively.

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

The phrase extraction is based on word-to-word alignments of the parallel corpus trained with GIZA++. Fig. 1 shows an example of word-to-word alignment results between the verbatim transcript (Cantonese) and the written-style transcript (standard Chinese), from which phrase-to-phrase alignments are derived by means of identifying deletion, substitution, insertion and inversion.

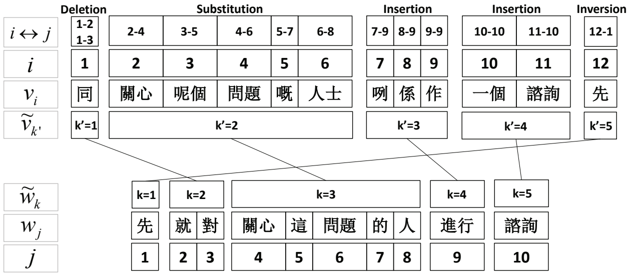


Fig. 1. An example of phrase extraction from word-to-word alignments. i and j are word indexes. k' and k are phrase indexes. $i \leftrightarrow j$ represents the word-to-word alignment.

Deletion can be modeled as alignment (3). According to this alignment, we extract a phrase $\tilde{v}_k = v_i$ for the verbatim transcript and a phrase $\tilde{w}_k = w_j\_w_{j+1}\_\cdots\_w_{j'}$ for the written-style transcript. The "_" symbol is used to indicate the concatenation of consecutive words forming a phrase. Alignment (4) is for substitution with $\tilde{v}_k = v_i\_v_{i+1}\_\cdots\_v_{i'}$, $\tilde{w}_k = w_j\_w_{j+1}\_\cdots\_w_{j'}$, and alignment (5) is for insertion with $\tilde{v}_k = v_i\_v_{i+1}\_\cdots\_v_{i'}$, $\tilde{w}_k = w_j$. An inversion transduction can be identified if it matches any of the three alignment patterns (3)(4)(5) under the condition that j' in current phrase is smaller than j in the previous phrase. With phrase extraction, we can obtain a table of phrases $\{\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_K\}$ for the verbatim transcript and $\{\tilde{w}_1, \tilde{w}_2, ..., \tilde{w}_K\}$ for the written-style transcript.

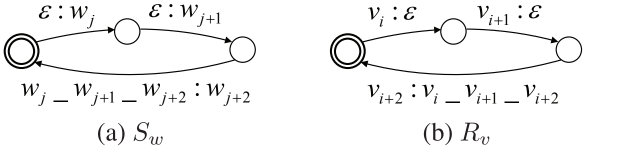
$$v_i \leftrightarrow w_j, v_i \leftrightarrow w_{j+1}, \cdots, v_i \leftrightarrow w_{j'} \qquad (3)$$

$$v_i \leftrightarrow w_j, v_{i+1} \leftrightarrow w_{j+1}, \cdots, v_{i'} \leftrightarrow w_{j'} \qquad (4)$$

$$v_i \leftrightarrow w_j, v_{i+1} \leftrightarrow w_j, \cdots, v_{i'} \leftrightarrow w_j \qquad (5)$$

3.1. Segmentation and Reconstruction Models

Segmentation model $P(\tilde{w}_1^K|w_1^J)$ segments word sequence w_1^J into K phrases. We define *segmentation order* s, where $s = j' - j + 1$, to represent the maximum number of words that can be segmented into one phrase. The WFST implementation of the segmentation model is described in Fig. 2(a). It shows a portion of segmentation transducer S_w for the written-style transcript when segmentation order $s = 3$. 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. Fig. 2(b) shows a portion of the reconstruction transducer R_v for the verbatim transcript consistent with S_w.



(a) S_w

(b) R_v

Fig. 2. WFST implementation of the segmentation model (a) and the reconstruction model (b).

3.2. Phrase Reordering Model

Fig. 1 shows that the phrase order of the verbatim transcript may differ from the written-style transcript. We define a phrase reordering model $P(r_1^K|\tilde{w}_1^K, w_1^J)$, which reorders phrase positions of the written-style transcript into those of the verbatim transcript 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 $\{\tilde{w}_{r_1}, \tilde{w}_{r_2}, ..., \tilde{w}_{r_K}\}$.

$$\begin{aligned} P(r_1^K|\tilde{w}_1^K, w_1^J) &= P(r_1^K|\tilde{w}_1^K) \\ &= P(r_1)\prod_{k=2}^{K} P(r_k|r_{k-1}, \tilde{w}_1^K) \end{aligned} \quad (6)$$

We make a first order Markov assumption over the phrase reordering model as shown in Eq. (6). The reordering sequence distribution is parameterized to assign decreasing likelihood to phrase re-orderings that diverge from the original word order [9]. 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. (7), where p_0 is a tuning factor. We normalize the probabilities $P(r_k|r_{k-1})$ such that $\sum_{k'=1, k' \neq r_{k-1}}^{K} P(r_k = k'|r_{k-1}) = 1$.

$$\begin{aligned} P(r_k|r_{k-1}) &= p_0^{|l-q'-1|} \\ P(r_1 = k) &= \frac{1}{K}; k \in \{1, 2, ..., K\} \end{aligned} \quad (7)$$

Assume that we have a phrase sequence $\{\tilde{w}_1, \tilde{w}_2, \tilde{w}_3\}$, Fig. 3 shows the WFST implementation of phrase reordering model for this phrase sequence.

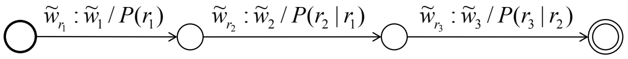

$\tilde{w}_{r_1} : \tilde{w}_1 / P(r_1)$
$\tilde{w}_{r_2} : \tilde{w}_2 / P(r_2 | r_1)$
$\tilde{w}_{r_3} : \tilde{w}_3 / P(r_3 | r_2)$

Fig. 3. Phrase reordering transducer Ω_r for phrase sequence $\{\tilde{w}_1, \tilde{w}_2, \tilde{w}_3\}$.

3.3. Phrase-to-Phrase Transduction Model

Once the phrase sequence of the written-style transcript is reordered into the verbatim transcript order, we use the phrase-to-phrase transduction model specified in Eq. (8) to perform the transduction. It assumes that a phrase $\tilde{v}_k$ is generated independently by each phrase $\tilde{w}_{r_k}$. This model can be easily implemented by a transducer T_{vw} which transduces $\tilde{v}_k$ to $\tilde{w}_{r_k}$.

$$P(\tilde{v}_1^K|r_1^K, \tilde{w}_1^K, w_1^J) = \prod_{k=1}^{K} P_k(\tilde{v}_k|\tilde{w}_{r_k}) \quad (8)$$

3.4. Phrase-Level Transduction Via WFSTs

Our phrase-level transduction model $P(v_1^I|w_1^J)$ can be constructed via WFST *composition* [10] (denoted by $\circ$) of all the component models as shown in Eq. (9), where $\mathcal{T}$ is the final composed WFST that transduces v_1^I to w_1^J.

$$\mathcal{T} = R_v \circ T_{vw} \circ \Omega_r \circ S_w \quad (9)$$

Now the recognition model for colloquial speech in Eq. (1) can be implemented using a transducer ASR, which is formulated with a unified WFST approach as shown in Eq. (10).

$$ASR = H \circ C \circ L \circ \pi(\mathcal{T} \circ G) \quad (10)$$

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 a LM [8]. The LM can either be $\mathcal{T} \circ G$ or $\pi(\mathcal{T} \circ G)$, where G is a LM to be transformed. π is a *projection* [10] operator which projects the input label to output label. $\mathcal{T} \circ G$ outputs the written-style recognition result, and $\pi(\mathcal{T} \circ G)$ outputs the speaking-style result. Before decoding, the recognition transducer ASR is optimized by *determinization* [8] operation right after each composition.

4. EXPERIMENTAL SETUP

We evaluate our phrase-level transduction model on Cantonese parliamentary speech from the Hong Kong Legislative Council. Currently we only have 3364 parallel transcribed sentences containing 15.7 hours of speech. It is separated into two sets, Set A (3.8 hours, 664 sentences) and Set B (11.9 hours, 2700 sentences). Set A is only used for evaluation of WER and BLEU score. The WER evaluation is on the speaking-style output against the verbatim transcription (manual transcription). The BLEU score evaluation is on the written-style output against the written-style transcription (Hansard transcription). The parallel transcriptions of Set B constitute the parallel corpus, which includes verbatim transcription of 106k words and written-style transcription of 80k words. Besides the parallel corpus, we have a set of additional Hansard transcription, which has 31M words.

The acoustic model is a tied-state cross-word triphone model with 39-dimensional MFCC feature trained from Set B. It comprises 70 Cantonese phoneme HMMs as well as silence, short pause and noise. The interpolated LM using words as modeling units is trained from the additional Hansard transcription together with the parallel corpus.

All transduction models are trained from the parallel corpus. Our reordering model permutes K phrases. Empirically, we find that $K \leq 5$ is capable of capturing most of the inversion transduction cases.

Decoding is performed by T^3 Decoder [11], which is a state-of-the-art WFST-based LVCSR speech decoder.

Table 1. WER for various recognition models. G is the interpolated LM. $\mathcal{T}_w^n$ is the word-level n-to-n TM. $\mathcal{T}_p^s$ is the phrase-level TM, where s is the segmentation order.

Models	WER (%)		
$H \circ C \circ L \circ G$	29.85		
$H \circ C \circ L \circ \pi(\mathcal{T}_w^n \circ G)$	$n = 1$	$n = 2$	$n = 3$
	29.09	**27.99**	28.37
$H \circ C \circ L \circ \pi(\mathcal{T}_p^s \circ G)$ (without/with Reordering)	s=2	27.66/27.09	
	s=3	**27.05/26.39**	
	s=4	27.52/27.01	
	s=5	28.11/27.63	

Table 2. BLEU score for the best word-level TM and phrase-level TM.

Models	BLEU Score
The Best Word-Level TM	27.94
The Best Phrase-Level TM	32.08

5. EXPERIMENTAL RESULTS

Table 1 shows WER results for the interpolated LM, word-level and phrase-level transduction model (TM). The best word-level TM $\mathcal{T}_w^2$ gives 6.23% relative WER reduction over the interpolated LM. The phrase-level TMs $\mathcal{T}_p^2, \mathcal{T}_p^3, \mathcal{T}_p^4$ consistently outperform word-level TM even without reordering. It is noteworthy that $\mathcal{T}_p^3$ without reordering can reduce the WER by 3.36% relative to the best word-level TM. When increasing the segmentation order s, $\mathcal{T}_p^3$ outperforms $\mathcal{T}_p^2$, suggesting that grouping more words into one phrase does improve the effectiveness of transduction. However, further increasing s decreases the performance. The reason is probably that Chinese phrases tend to be 2 to 3 words long.

Table 1 also shows the effectiveness of the proposed reordering model, which gives 0.4%∼0.7% absolute WER reduction over those without reordering. For example, $\mathcal{T}_p^3$ with reordering can reduce the WER by 0.66%. In Table 2, the phrase-level TM shows 14.82% relative BLEU score improvement over the word-level TM, which further demonstrates the effectiveness of phrase-level transduction. All the improvements are statistically significant according to the two-proportion z-test at 99% confidence.

6. CONCLUSION AND DISCUSSION

In this paper, we propose a first-ever integrated model of speech recognition with phrase-based transduction to decode Cantonese speech into both verbatim transcriptions and standard written transcriptions. We use a large amount of Mandarin data and a small amount of Cantonese data, as well as some Cantonese-to-Mandarin parallel data, with focus on solving the language modeling challenge with limited verbatim training data.

We propose to transform the language model of the standard written Chinese to Cantonese language model. However, instead of word-to-word transformation, this paper proposes a phrase-level transduction model with reordering to achieve better transduction that allows n-m transduction. Our proposed model gives 5.72% relative improvement on WER for verbatim Cantonese transcription and 14.82% relative improvement on BLEU score for Cantonese-to-Mandarin transduction. Our proposed model can be applied to many other low-resource languages with insufficient verbatim transcription for language model training. Possible future research direction is to generalize our model to speech translation, though more advanced reordering models may be explored.

7. ACKNOWLEDGMENTS

We 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. This work is partially supported by ITS/189/09 and CERG#612211.

8. REFERENCES

[1] T. Hori, D. Willett, and Y. Minami, “Language model adaptation using wfst-based speaking-style translation,” in *Proc. of ICASSP*, 2003, vol. 1, pp. 228–231.

[2] Y. Akita and T. Kawahara, “Efficient estimation of language model statistics of spontaneous speech via statistical transformation model,” in *Proc. of ICASSP*, 2006, vol. 1, pp. 1049–1052.

[3] G. Neubig, Y. Akita, S. Mori, and T. Kawahara, “Improved statistical models for smt-based speaking style transformation,” in *Proc. of ICASSP*, 2010, pp. 5206–5209.

[4] F.J. Och and H. Ney, “The alignment template approach to statistical machine translation,” *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, 2004.

[5] B. Zhou, S. Chen, and Y. Gao, “Constrained phrase-based translation using weighted finite-state transducers,” in *Proc. of ICASSP*, 2005, vol. 1, pp. 1017–1020.

[6] L. Mathias and W. Byrne, “Statistical phrase-based speech translation,” in *Proc. of ICASSP*, 2006, vol. 1, pp. 561–564.

[7] F. Casacuberta, H. Ney, F.J. Och, et al., “Some approaches to statistical and finite-state speech-to-speech translation,” *Computer Speech & Language*, vol. 18, no. 1, pp. 25–47, 2004.

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

[9] F.J. Och, C. Tillmann, and H. Ney, “Improved alignment models for statistical machine translation,” in *Proc. of the Joint SIGDAT Conf. on EMNLP and VLC*, College Park, MD, USA, June 1999, pp. 20–28.

[10] M. Mohri, “Weighted automata algorithms,” *Handbook of Weighted Automata*, pp. 213–254, 2009.

[11] P.R. Dixon, T. Oonishi, K. Iwano, and S. Furui, “Recent development of wfst-based speech recognition decoder,” in *Proc. of 2009 APSIPA Annual Summit and Conference*, Sapporo, Japan, October 2009, pp. 138–147.