TOWARDS END-TO-END SPEECH-TO-TEXT TRANSLATION WITH TWO-PASS DECODING

Tzu-Wei Sung, Jun-You Liu, Hung-yi Lee, Lin-shan Lee

National Taiwan University
{b03902042, b03502040, hungyilee}@ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

Speech-to-text translation (ST) refers to transforming the audio in source language to the text in target language. Mainstream solutions for such tasks are to cascade automatic speech recognition with machine translation, for which the transcriptions of the source language are needed in training. End-to-end approaches for ST tasks have been investigated because of not only technical interests such as to achieve globally optimized solution, but the need for ST tasks for the many source languages worldwide which do not have written form. In this paper, we propose a new end-to-end ST framework with two decoders to handle the relatively deeper relationships between the source language audio and target language text. The first-pass decoder generates some useful latent representations, and the second-pass decoder then integrates the output of both the encoder and the first-pass decoder to generate the text translation in target language. Only paired source language audio and target language text are used in training. Preliminary experiments on several language pairs showed improved performance, and offered some initial analysis.

Index Terms— Speech-to-Text Translation, End-to-End Model, Unwritten Language

1. INTRODUCTION

Speech-to-text translation (ST) refers to transforming the audio in source language to the text in target language. Conventional approaches for ST perform machine translation (MT) on the top of the automatic speech recognition (ASR) output. In these approaches, source language transcriptions are always needed regardless of whether ASR and MT are trained separately or jointly [1, 2]. However, among the thousands of languages worldwide, most of them do not have an acknowledged written form nor even been well described in documents [3]. These include not only many indigenous as well as aboriginal languages without a literary tradition, but many dialects used only for daily conversations instead of written communication. In order to eliminate the need for source language transcriptions, in addition to other technical consideration such as achieving globally optimized solutions, directly trained on source language audio paired with target language text translations was considered end-to-end ST [4, 5, 6, 7], for example, using the sequence-to-sequence model with attention mechanism [4].

In this paper, we propose to insert an extra decoder into the typical encoder-decoder architecture [8, 9] to better handle the relatively deeper relationships between source language audio and target language text. In this way, the encoder is followed by two decoders, but only paired source language audio and target language text are given during training. The first-pass decoder may generate some useful latent representations referred to as intermedia here (may be close to ASR transcriptions mixed with subword units of the source language), based on which the second-pass decoder generates the target language text. Similar two-pass decoder architecture was used in text-based MT [10, 11] before, but has not yet reported on ST. Initial experiments on several language pairs showed improved performance.

2. PROPOSED APPROACH

As in Fig. 1, the proposed approach includes three components: an encoder $\mathcal{E}$ with parameter set θ_e, a first-pass decoder $\mathcal{D}_1$, and a second-pass decoder $\mathcal{D}_2$ respectively with parameter sets θ_1 and θ_2. The input is an acoustic feature sequence in source language of length T, $x = \{x_1, x_2, ..., x_T\}$, while the output is a word sequence in target language of length M, $y = \{y_1, y_2, ..., y_M\}$.

2.1. The encoder

The encoder at the left lower part of Fig. 1 is the same as those used in prior works [1, 2, 4, 5, 7], based on bidirectional Long Short-Term Memory (LSTM). A input sequence x of length T is encoded into T hidden states $h = \{h_1, h_2, \cdots, h_T\}$, where $h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i]$ for the bidirectional parts. More precisely, the forward encoder generates $\overrightarrow{h}_i = LSTM(x_i, \overrightarrow{h}_{i-1})$ while the backward encoder generates $\overleftarrow{h}_i = LSTM(x_i, \overleftarrow{h}_{i+1})$, where $\overrightarrow{h}_0$ and $\overleftarrow{h}_{T+1}$ are zero vectors.

2.2. The first-pass decoder

The first-pass decoder $\mathcal{D}_1$ at the right lower part of Fig. 1 is a conventional decoder with attention. It generates a series of

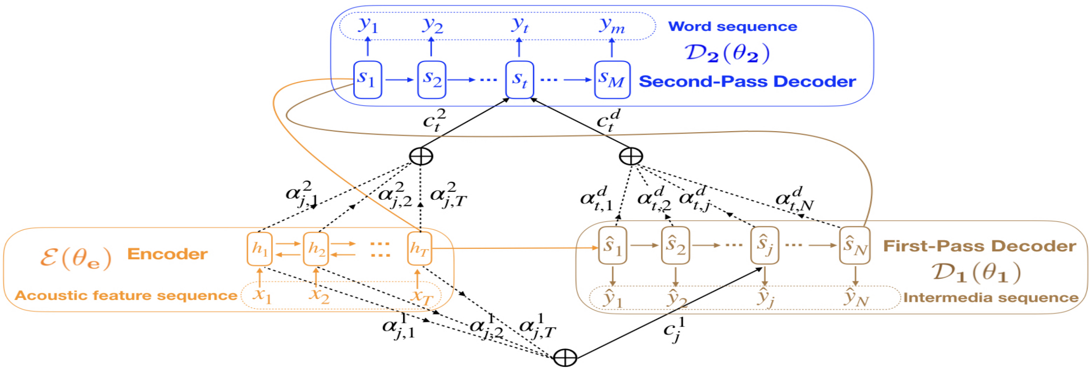

Fig. 1. The proposed framework for speech-to-text translation (ST). The orange, brown, and blue parts are encoder, first-pass and second-pass decoder respectively. The input is the acoustic feature sequence in source language, $x = \{x_1, x_2, ..., x_T\}$, while the output is the word sequence in target language, $y = \{y_1, y_2, ..., y_M\}$.

N hidden states $\hat{s} = \{\hat{s}_1, \cdots, \hat{s}_j, \cdots, \hat{s}_N\}$, and a first-pass output sequence of length N, $\hat{y} = \{\hat{y}_1, \cdots, \hat{y}_j, \cdots, \hat{y}_N\}$. At time index j within $\hat{s}$ or $\hat{y}$, context information c_j^1 is obtained via the attention over the hidden state of the encoder $\mathcal{E}$, $h = \{h_1, h_2, \cdots, h_T\}$, as shown at the bottom of Fig. 1:

$$\begin{aligned} c_j^1 &= \sum_{i=1}^{T} \alpha^1_{j,i} h_i, \\ \alpha^1_{j,i} &= \frac{exp(e^1_{j,i})}{\sum_{i'=1}^{T} exp(e^1_{j,i'})}, \\ e^1_{j,i} &= (v^1_\alpha)^T \tanh(W^1_{\hat{s}} \hat{s}_{j-1} + W^1_h h_i), \end{aligned} \tag{1}$$

where v^1_α, $W^1_{\hat{s}}$ and W^1_h are the parameters of $\mathcal{D}_1$, denoted by θ_1. The hidden state is $\hat{s}_j = LSTM([y_{t-1}; c^1_j], \hat{s}_{j-1})$. The first hidden state $\hat{s}_1$ is initialized by the last hidden state h_T of the encoder $\mathcal{E}$, as shown by the arrow connecting them. After obtaining $\hat{s}_j$, an affine transformation is applied on the concatenated vector $[\hat{s}_j; c^1_j; \hat{y}_{j-1}]$. The results of the transformation are then fed into a softmax layer, and the output $\hat{y}_j$ is sampled from the multinomial distribution obtained.

2.3. The second-pass decoder

Here the hidden states h and $\hat{s}$ of the encoder $\mathcal{E}$ and the first-pass decoder are used by the second-pass decoder $\mathcal{D}_2$ to generate a hidden state sequence $s = \{s_1, \cdots, s_t, \cdots, s_M\}$, and a output sequence $y = \{y_1, \cdots, y_t, \cdots, y_M\}$, where M is the length of the output sequence.

At time index t within s or y, $\mathcal{D}_2$ takes the previous hidden state s_{t-1} generated by itself, the contextual information c^2_t from the encoder $\mathcal{E}$, and the contextual information c^d_t from the first-pass decoder $\mathcal{D}_1$ as inputs. c^2_t is generated in exactly same way in Eq. (1), except $\hat{s}_{j-1}$ in Eq. (1) is replaced by s_{t-1} here and the different set of model parameters, v^2_α, W^2_s and W^2_h, as shown at the middle part of Fig. 1. c^d_t is the context vector used in $\mathcal{D}_2$ at time index t based on the attention over the hidden states $\hat{s}$ of $\mathcal{D}_1$, as shown at the midden right of Fig. 1:

$$\begin{aligned} c^d_t &= \sum_{i=1}^{N} \alpha^d_{t,i} \hat{s}_i, \\ \alpha^d_{t,i} &= \frac{exp(e^d_{t,i})}{\sum_{i'=1}^{N} exp(e^d_{t,i'})}, \\ e^d_{t,i} &= (v^d_\alpha)^T \tanh(W^d_s s_{t-1} + W^d_{\hat{s}} \hat{s}_i), \end{aligned} \tag{2}$$

where v^d_α, W^d_s, and $W^d_{\hat{s}}$ are the parameters of $\mathcal{D}_2$, denoted by θ_2. Then, we calculate the hidden state s_t in $\mathcal{D}_2$ as $s_t = LSTM([y_{t-1}; c^2_t; c^d_t], s_{t-1})$. The first hidden state s_1 is initialized based on the last hidden h_T and $\hat{s}_N$ of $\mathcal{E}$ and $\mathcal{D}_1$, as shown by the arrows i Fig. 1. The concatenated vector $[s_t; c^2_t; c^d_t; y_{t-1}]$ is finally transformed to generate y_t, the output sequence at time index t. So $\mathcal{D}_2$ actually takes complete information represented by hidden states $\hat{s}$ in $\mathcal{D}_1$ and h in $\mathcal{E}$ through the initialization of s_1 and the contextual vectors c^d_t and c^2_t aggregating the information extracted in $\mathcal{D}_1$ and $\mathcal{E}$.

2.4. Training

$D = \{(x, y^*)\}$ denotes the training corpus, where x is an acoustic feature sequence in source language, and y^* is the reference text transcription of x in target language or the learning target for the output of $\mathcal{D}_2$. We never know learning target for $\mathcal{D}_1$ since it is not given, but in the experiments we found also using y^* as the learning target of $\mathcal{D}_1$ is helpful. The objective is therefore below:

$$\begin{aligned} &\mathcal{J}(\theta_e, \theta_1, \theta_2) = \\ &\frac{1}{|D|} \sum_{(x,y^*) \in D} \{\lambda \log P(y^*|\theta_e, \theta_1, \theta_2) + (1-\lambda) \log P(y^*|\theta_e, \theta_1)\}, \end{aligned} \tag{3}$$

where $|D|$ is the size of D. The first term is to jointly train the encoder and the two decoders to maximize the log likelihood of the reference transcription y^* at the output of $\mathcal{D}_2$. The second term is to maximize the log likelihood of the reference transcription y^* at the output of $\mathcal{D}_1$. λ is a hyper-parameter close to 1.

	Taiwanese	Mboshi	Fisher/Callhome
Emitting type	Character	Subword	Subword
Encoder layers	1	1	3
Encoder hidden size	64	128	256
Decoder hidden size	128	256	256

Table 1. Individual settings for each corpus.

3. EXPERIMENTS

Four different datasets were used in the initial experiments.

3.1. Corpus

Taiwanese, often called Min-Nan, is one out of the hundreds of dialect for Chinese [12], for which there does not exist a unified orthographic system. The *Taiwanese-Chinese* dataset[1] was collected from a local television program company, including approximately 120 hours of Taiwanese audio spoken by a single speaker, Dharma Master Cheng Yen, and the corresponding 120K sentences of Chinese text translations. We split this corpus into training, development, and testing sets with 80%, 10%, and 10% respectively.

Mboshi is a Bantu language spoken in the Republic of Congo. It is endangered and lacks a stable orthographic system. The *Mboshi-French* dataset was collected during a language documentation process [13], including 5517 utterances (about 4.4 hours) in Mboshi audio aligned to French text translations. The Mboshi speech was produced by three speakers in Congo-Brazzaville. Following the prior work [1, 2], we randomly sampled 100 utterances from training data taken as development set, and used the original development set as testing set.

Fisher and Callhome Spanish-English Speech Translation corpus (LDC2014T23) [14] contained English reference translations and Spanish audio in the Fisher-Spanish corpus (LDC2010S01, LDC2010T04) and Callhome-Spanish corpus (LDC96S35, LDC96T17), for telephone conversations between mostly native Spanish speakers in a variety of dialects. The Fisher-Spanish dataset consisted of 819 transcribed conversations (roughly 160 hours), while the Callhome-Spanish corpus comprised 120 spontaneous conversations (roughly 20 hours); both aligned at the utterance level. Following [5, 14], we trained our models on Fisher *train* set, and evaluated on Fisher *test* set and Callhome *evltest* set. While evaluating the BLEU-4 scores, we used the four references for Fisher and one reference for Callhome.

3.2. Implementation

The models were implemented with Tensor2tensor [15]. We used a dropout rate of 0.2, and trained the models using Adam with an initial learning rate of 0.001. L2 weight decay was used with a weight of 10^{-6}. Batch size for each step varied depending on audio length [16], but never exceeding larger than 128 due to the memory constraints. Beam search with beam size 10 was performed during inference. We also found length normalization [17] did not lead to better performance for all corpora, so we set it to 0.0 for *Taiwanese-Chinese* and 0.6 for others. The other settings are in Table 1.

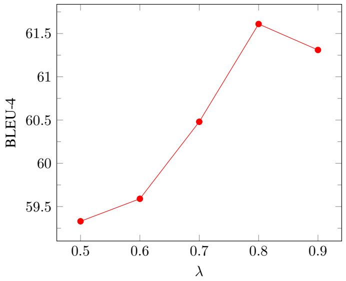

Fig. 2. Character-level BLEU-4 scores for different λ on *Taiwanese-Chinese* testing set.

Following the prior work [5], we used 80 channel log mel filterbank features extracted from 25ms windows with a hop size of 10ms. We also concatenated them with delta and delta-delta features organized with a shape of $T \times 80 \times 3$. To reduce computational complexity in the following layers, they further went through two consecutive convolutional layers, each comprising 128 kernels with shape 3×3 and a stride of 2×2, followed by layer normalization and ReLU activation. Hence, the time scale was downsampled by a factor of 4. For *Fisher and Callhome* corpus, following the prior work [5], the downsampled features additionally went through the bidirectional convolutional LSTM [18], convolving across frequency axis with kernel size 3. This new set of features was then fed into the bidirectional LSTM encoder described in Section 2.1.

For the *Taiwanese-Chinese* corpus, all punctuations were removed, and we did not segment the character sequences into words. So the models are character-based, with a total of 2956 distinct characters. For the *Mboshi-French* dataset, the French transcriptions have been tokenized and cleaned up. We detokenized them, and adopted subword-based models, with a total of 2356 distinct subword units. For *Fisher and Callhome* corpus, following the prior work [5], we lowered all letters and removed all punctuations excluding the apostrophes. We also used subword-based models with a total of 9841 distinct subword units.

3.3. Results

We first tuned the trade-off factor λ in Eq. (3) on the *Taiwanese-Chinese* testing set. We increased λ from 0.5 to 1.0 with increment of 0.1. The model did not work when $\lambda = 1.0$, or the first-pass decoder played an important role here. The results in Fig. 2 showed the best performance was achieved when $\lambda = 0.8$. Therefore, we used $\lambda = 0.8$ for later experiments

[1] The *Taiwanese-Chinese* dataset was provided by Da Ai Television, Taiwan and was consolidated by Taiwan AI Labs.

(A) Ground truth		(B) Proposed	(C) Proposed (first-pass)	(D) $\mathcal{M}_{\text{seq2seq}(1)}$
(i)	沒有漏失的正法	沒有漏失的正法	沒老息的正法	沒有老實的正法
	(no missing orthodox dharma)	(no missing orthodox dharma)	(no old breath dharma)	(no honest orthodox dharma)
(ii)	bô lāu sit ê tsìng huat	bô lāu sit ê tsìng huat	bô lāu sit ê tsìng huat	bô láu sit ê tsìng huat
(i)	自然增長大愛	自然增長大愛	自然前長大愛	自然清重大愛
	(naturally growing philanthropy)	(naturally growing philanthropy)	(naturally front growing philanthropy)	(naturally clear and heavy philanthropy)
(ii)	tsū jiân tsing tióng tuā ài	tsū jiân tsing tióng tuā ài	tsū jiân tsîng tiông tuā ài	tsū jiân tshing tiông tuā ài

Table 2. Example output sequences for the different approaches: (i) the output Chinese character sequence (with English translation in the parentheses), (ii) the phoneme sequence when the character sequence in (i) was produced in Taiwanese.

	Taiwanese	Mboshi	Fisher *test*	Callhome *evltest*
$\mathcal{M}_{\text{seq2seq}(1)}$	59.61	4.09	36.74	13.51
$\mathcal{M}_{\text{seq2seq}(2)}$	59.62	3.96	37.50	13.99
Proposed	**61.54**	**6.55**	**38.30**	**14.46**

Table 3. BLEU-4 scores for each corpus, character-level for *Taiwanese-Chinese*, and word-level for others.

	CER (Taiwanese)	n-gram precisions (Mboshi)
$\mathcal{M}_{\text{seq2seq}(1)}$	0.2701	17.3 / 5.6 / 2.7 / 1.2
$\mathcal{M}_{\text{seq2seq}(2)}$	0.2672	16.3 / 5.3 / 2.4 / 1.2
Proposed	**0.2531**	**21.4 / 8.6 / 4.8 / 3.1**

Table 4. CER for *Taiwanese-Chinese* and n-gram precisions for *Mboshi-French*.

Ground truth	$\mathcal{M}_{\text{seq2seq}(1)}$	$\mathcal{M}_{\text{seq2seq}(2)}$	Proposed (first-pass)	Proposed
29.73	38.11	40.34	59.15	**34.43**

Table 5. Perplexity of output sequences for *Taiwanese-Chinese*.

on all corpora.

We present the results on the four corpora considered, compared with two baselines both using attentional sequence-to-sequence model but with one decoder only instead of two here: (i) this decoder with one layer of LSTMs, denoted by $\mathcal{M}_{\text{seq2seq}(1)}$; and (ii) this decoder with two stacked LSTM layers, denoted by $\mathcal{M}_{\text{seq2seq}(2)}$. We considered $\mathcal{M}_{\text{seq2seq}(2)}$ because our model has two decoders. We used `t2t-bleu` to evaluate BLEU-4 score [19]. Character-level BLEU-4 score was used for *Taiwanese-Chinese* corpus (since each Chinese character has its meaning), and word-level BLEU-4 score (subword units grouped into words for evaluation) for the others.

The results are in Table 3, from which we can see that the proposed approach clearly outperformed the baselines. The BLEU-4 scores varied significantly across the different corpora, obviously because these corpora are very different. For example, the *Taiwanese-Chinese* corpus was large but produced by a single speaker with a relatively higher quality recording, so the score is much higher.

Because Taiwanese can be considered as a dialect of Chinese, and therefore the *Taiwanese-Chinese* task may be considered as a speech recognition task and evaluated in character error rate (CER). This is listed in the first column of Table 4, where a lower CER was achieved by the proposed approach. Also, because BLEU-4 score is integrated from n-gram precision rates ($n = 1, 2, 3, 4$), we noted for the low-resourced *Mboshi-French* corpus, the very low BLEU-4 score was due to the very poor 3-gram and 4-gram match, although the 1-gram match was not too bad. This is reported in the second column of Table 4, in which we see even for the low-resourced *Mboshi-French* corpus, the proposed approach also offered better 1,2,3,4-gram precisions than the baselines.

We further trained a recurrent neural network based language model [20] with hidden size 256 using RNNLM toolkit [21] to evaluate the perplexity of the *Taiwanese-Chinese* testing set output sequences. The results are listed in Table 5, where we see not only the proposed approach gave the lowest perplexity, but the output sequences from the first-pass decoder had much highest perplexity, but significantly reduced by the second-pass decoder.

3.4. Examples

We further analyze the results with two typical examples in the *Taiwanese-Chinese* corpus in Table 2, where part (i) is the output Chinese character sequence (with English translation in the parentheses), while row (ii) is the phoneme sequence when the character sequence in (i) was produced in Taiwanese. The results of the proposed approach but using the first-pass decoder only and the baseline $\mathcal{M}_{\text{seq2seq}(1)}$ are also listed in columns (C)(D). We can see that the first-pass decoder of the proposed approach and $\mathcal{M}_{\text{seq2seq}(1)}$ (columns (C)(D)) both generated partly incorrect character sequence in part (i), but these character sequences sounded close to the audio it produced in Taiwanese. So the first-pass decoder of the proposed approach and $\mathcal{M}_{\text{seq2seq}(1)}$ are likely to "transcribe" the audio by "Chinese characters with Taiwanese pronunciations". With the second-pass decoder applied in addition, these "transcriptions" were corrected to some degree.

4. CONCLUSION

In this work, we consider the end-to-end speech-to-text translation task and extend the conventional sequence-to-sequence model by adding an additional second-pass decoder, which considers both the encoder output and the first-pass decoder output simultaneously. Improved performance was obtained and some analysis of the results are reported in this paper.

5. REFERENCES

[1] Antonios Anastasopoulos and David Chiang, "Tied multitask learning for neural speech translation," in *NAACL-HLT*, 2018.

[2] Antonios Anastasopoulos and David Chiang, "Leveraging translations for speech transcription in low-resource settings," in *Interspeech*, 2018.

[3] Laurent Besacier, Bowen Zhou, and Yuqing Gao, "Towards speech translation of non written languages," in *SLT*, 2006.

[4] Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *CoRR*, vol. abs/1612.01744, 2016.

[5] Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Interspeech*, 2017.

[6] Alexandre Berard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin, "End-to-end automatic speech translation of audiobooks," in *ICASSP*, 2018.

[7] Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater, "Low-resource speech-to-text translation," in *Interspeech*, 2018.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[9] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *NIPS*, 2015.

[10] Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu, "Deliberation networks: Sequence generation beyond one-pass decoding," in *NIPS*, 2017.

[11] Xiangwen Zhang, Jinsong Su, Yue Qin, Yang Liu, Rongrong Ji, and Hongji Wang, "Asynchronous bidirectional decoding for neural machine translation," *CoRR*, vol. abs/1801.05122, 2018.

[12] Ethnologue: Languages of the world, "List of Taiwan's dialects," `https://www.ethnologue.com/country/tw/languages`, [Online Catologue].

[13] Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noël Kouarata, Lori Lamel, Hélène Bonneau-Maynard, Markus Müller, Annie Rialland, Sebastian Stüker, François Yvon, and Marcely Zanon Boito, "A very low resource language speech corpus for computational language documentation experiments," *CoRR*, vol. abs/1710.03501, 2018.

[14] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur, "Improved speech-to-text translation with the fisher and callhome spanish–english speech translation corpus," in *IWSLT*, 2013.

[15] Ashish Vaswani, Samy Bengio, Eugene Brevdo, François Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Lukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit, "Tensor2tensor for neural machine translation," in *AMTA*, 2018.

[16] Martin Popel and Ondrej Bojar, "Training tips for the transformer model," *CoRR*, vol. abs/1804.00247, 2018.

[17] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016.

[18] Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NIPS*. 2015.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[20] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur, "Recurrent neural network based language model," in *Interspeech*, 2010.

[21] Tomas Mikolov, Stefan Kombrink, Anoop Deoras, and Lukás Burget, "Rnnlm-recurrent neural network language modeling toolkit," 2011.