INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Structured-based Curriculum Learning for End-to-end English-Japanese Speech Translation

Takatomo Kano, Sakriani Sakti, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{kano.takatomo.km0,ssakti,s-nakamura}@is.naist.jp

Abstract

Sequence-to-sequence attentional-based neural network architectures have been shown to provide a powerful model for machine translation and speech recognition. Recently, several works have attempted to extend the models for end-to-end speech translation task. However, the usefulness of these models were only investigated on language pairs with similar syntax and word order (e.g., English-French or English-Spanish). In this work, we focus on end-to-end speech translation tasks on syntactically distant language pairs (e.g., English-Japanese) that require distant word reordering. To guide the encoder-decoder attentional model to learn this difficult problem, we propose a structured-based curriculum learning strategy. Unlike conventional curriculum learning that gradually emphasizes difficult data examples, we formalize learning strategies from easier network structures to more difficult network structures. Here, we start the training with end-to-end encoder-decoder for speech recognition or text-based machine translation task then gradually move to end-to-end speech translation task. The experiment results show that the proposed approach could provide significant improvements in comparison with the one without curriculum learning.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Translating a spoken language, in other words recognizing speech and automatically having one's words translated into another language, is extremely complex. One traditional approach in speech-to-text translation systems must construct automatic speech recognition (ASR) and machine translation (MT) system, both of which are independently trained and tuned. Given a speech input, the ASR system processes and transforms the speech into the text in the source language, and then MT transforms the text in the source language to corresponding text in the target language [1]. The basic unit for information sharing between these components is only words at the text level. Even though significant progress has been made and various commercial speech translation systems have been introduced, this approach continues to suffer from several major limitations.

One of the drawbacks is that speech acoustics might involve both linguistic and paralinguistic information (i.e., prosody, intonation, accent, rhythm, emphasis, or emotion), but such paralinguistic information is not a factor in written communication, and much cannot even be expressed in words. Consequently, the words output by ASR have lost all of their paralinguistic information, and only the linguistic parts are translated by the MT system. Some studies have proposed including additional component to just handle paralinguistic translation, but this step introduces more complexity and delay [2, 3, 4]. Another noted problem is that over half of the world's languages actually have no written form and are only spoken. Another solution is to translate directly from phoneme-based transcription. However, the performance of a phoneme-based ASR is usually low, and errors in the ASR stage can propagate throughout the translation process [5]. Therefore, it would be useful to find ways beyond the current conventional approach to directly translate from the speech of the source language to the text of the target language.

Recently, deep learning has shown much promise in many tasks. A sequence-to-sequence attention-based neural network is one architecture that provides a powerful model for machine translation and speech recognition [6, 7]. Recently, several works have extended models for end-to-end speech translation (ST) tasks. Duong et al. [8]. directly trained attentional models on parallel speech data. But their work is only applicable for Spanish-English language pairs with similar syntax and word order (SVO-SVO). Furthermore, it focused on alignment performance. The only attempt to build a full-fledged end-to-end attentional-based speech-to-text translation system is Bérard et al. [9]. But, that work was only done on a small French-English synthetic corpus, because these language share similar word order (SVO-SVO). For such languages, only local movements are sufficient for translation.

This paper proposes a first attempt to build an end-to-end attention-based ST system on syntactically distant language pairs that suffers from long-distance reordering phenomena. We train the attentional model on English-Japanese language pairs with SVO versus SOV word order. To guide the encoder-decoder attentional model to learn this difficult problem, we proposed a structured-based curriculum learning strategy. Unlike the conventional curriculum learning that gradually emphasize difficult data examples, we formalize CL strategies that start the training with an end-to-end encoder-decoder for speech recognition or text-based machine translation tasks and gradually train the network for end-to-end speech translation tasks by adapting the decoder or encoder parts. Here we start the training with an end-to-end encoder-decoder for speech recognition or a text-based machine translation task and gradually move to an end-to-end speech translation task.

2. Related Works

Curriculum learning, which is one learning paradigm, is inspired by the learning processes of humans and animals that learn from easier aspects and gradually increase to more difficult ones. Although the application of such training strategies to machine learning has been discussed between machine learning and cognitive science researchers going back to Elman (1993) [10], CL's first formulation in the context of machine learning was introduced by Bengio et al. (2009) [11].

Using CL might help avoid bad local minima and speed up the training convergence and improve generalization. These advantages have been empirically demonstrated in various tasks,

http://dx.doi.org/10.21437/Interspeech.2017-944

including shape recognition [11], object classification [12], and language modeling tasks [13]. However, most studies focused on how to organize the sequence of the learning data examples in the context of single task learning. Bengio at al. [11] proposed curriculum learning for multiple tasks. But again, all of the tasks still belonged to the same type of problem, which is object classification, where those tasks shared the same input and output spaces.

In contrast with most previous CL studies, (1) we utilize CL strategy not for simple recognition/classification problems, but for sequence-to-sequence based neural network learning problems in speech translation tasks; (2) the attentional-based neural network is not trained directly for the speech translation task using similar but more and more difficult speech translation data. Instead we formalize CL strategies that start the training with an end-to-end encoder-decoder for speech recognition or text-based machine translation tasks and gradually train the network for end-to-end speech translation tasks by adapting the decoder or encoder parts respectively; (3) those different tasks of speech recognition, text-based machine translation, and speech translation used in structured-based CL do not share the same input and output spaces, as in the CL of multiple tasks.

3. Basic Attention-based Speech Translation

We built our end-to-end speech translation system upon the standard attention-based encoder-decoder neural networks architecture [7, 14] that consists of encoder, decoder, and attention modules. Given input sequence $\mathbf{x} = [x_1, x_2, ..., x_N]$ with length N, the encoder produces a sequence of vector representation $h^{enc} = (h_1^{enc}, h_2^{enc}, ..., h_N^{enc})$. Here we used a bidirectional recurrent neural network with long short-term memory (bi-LSTM) units [15], which consist of forward and backward LSTMs. The forward LSTM reads the input sequence from x_1 to x_N and estimates forward $\overrightarrow{h^{enc}}$, while the backward LSTM reads the input sequence in reverse order from x_N to x_1 and estimates backward $\overleftarrow{h^{enc}}$. Thus, for each input x_n, we obtain h_n^{enc} by concatenating forward $\overrightarrow{h^{enc}}$ and backward $\overleftarrow{h^{enc}}$.

The decoder, on the other hand, predicts target sequence $\mathbf{y} = [y_1, y_2, ..., y_T]$ with length T by estimating conditional probability $p(\mathbf{y}|\mathbf{x})$. Here, we use uni-directional LSTM (forward only). Conditional probability $p(\mathbf{y}|\mathbf{x})$ is estimated based on the whole sequence of the previous output:

$$p(y_t|y_1, y_2, ..., y_{t-1}, x) = softmax(W_y \tilde{h}_t^{dec}). \quad (1)$$

Decoder hidden activation vector $\tilde{h}_t^{dec}$ is computed by applying linear layer W_c over context information c_t and current hidden state h_t^{dec}:

$$\tilde{h}_t^{dec} = \tanh(W_c[c_t; h_t^{dec}]). \quad (2)$$

Here c_t is the context information of the input sequence when generating current output at time t, estimated by the attention module over encoder hidden states h_n^{enc}

$$c_t = \sum_{n=1}^{N} a_t(n) * h_n^{enc}, \quad (3)$$

where variable-length alignment vector a_t, whose size equals the length of input sequence x, is computed by

$$\begin{aligned} a_t(n) &= \text{align}(h_n^{enc}, h_t^{dec}) \quad (4) \\ &= \text{softmax}(\text{dot}(h_n^{enc}, h_t^{dec}). \end{aligned}$$

This step is done to assist the decoder to find relevant information on the encoder side based on the current decoder hidden states. There are several variations to calculate align(h_n^{enc}, h_t^{dec}). Here we simply use the dot product between the encoder and decoder hidden states [16].

In this study, we apply this basic architecture for various tasks:

- ASR system
  Input sequence $\mathbf{x} = [x_1, ..., x_N]$ is the input speech sequence of the source language, and target sequence $\mathbf{y} = [y_1, ..., y_T]$ is the predicted corresponding transcription in the source language.
- MT system
  Input sequence $\mathbf{x} = [x_1, ..., x_N]$ is the word sequence of the source language, and target sequence $\mathbf{y} = [y_1, ..., y_T]$ is the predicted corresponding word sequence in the target language.
- ST system
  Input sequence $\mathbf{x} = [x_1, ..., x_N]$ is the input speech sequence of the source language, and target sequence $\mathbf{y} = [y_1, ..., y_T]$ is the predicted corresponding word sequence in the target language.

4. Attention-based Speech Translation with Curriculum Learning

The training process of the attention-based encoder-decoder model is basically more difficult than the standard neural network model [17] because an attention-based model needs to jointly optimize three different (encoder, decoder, and attention) modules simultaneously. Utilizing the attention-based encoder-decoder architecture for constructing a direct ST task is obviously difficult because the model needs to solve two complex problems: (1) learning how to process a long speech sequence and map it to the corresponding words, similar to the issues focused on in the field of ASR [6]; (2) learning how to make good alignment rules between source and target languages, similar to the issues discussed in the field of MT [7, 18]. Furthermore, we utilize attention-based encoder-decoder architecture to construct a ST system on syntactically distance language pairs that suffer from long-distance reordering phenomena and train the attentional model on English-Japanese language pairs with SVO versus SOV word order. Therefore, to assist the encoder-decoder model to learn this difficult problem, we proposed a structured-based curriculum learning strategy.

In our CL strategy, the attentional-based neural network is not trained directly for speech translation tasks using similar but more and more difficult speech translation data, instead we formalize structured-based CL strategies that start the training with an end-to-end encoder-decoder for ASR or text-based MT tasks and gradually train the network for end-to-end ST tasks. In other words, we train the attentional encoder-decoder architecture by starting from a simpler task, switch a certain part of the structure (encoder or decoder) in each training phase, and set it to a more difficult target task. In this way, the difficulty of the problems increases gradually in each training phase, as in CL strategies.

Figure 1 illustrates the attention-based speech translation training phases, and the details are described below.

1. **CL type 1: Start from an attention-based ASR system**
   Here the curriculum learning for each phases is designed as follows:

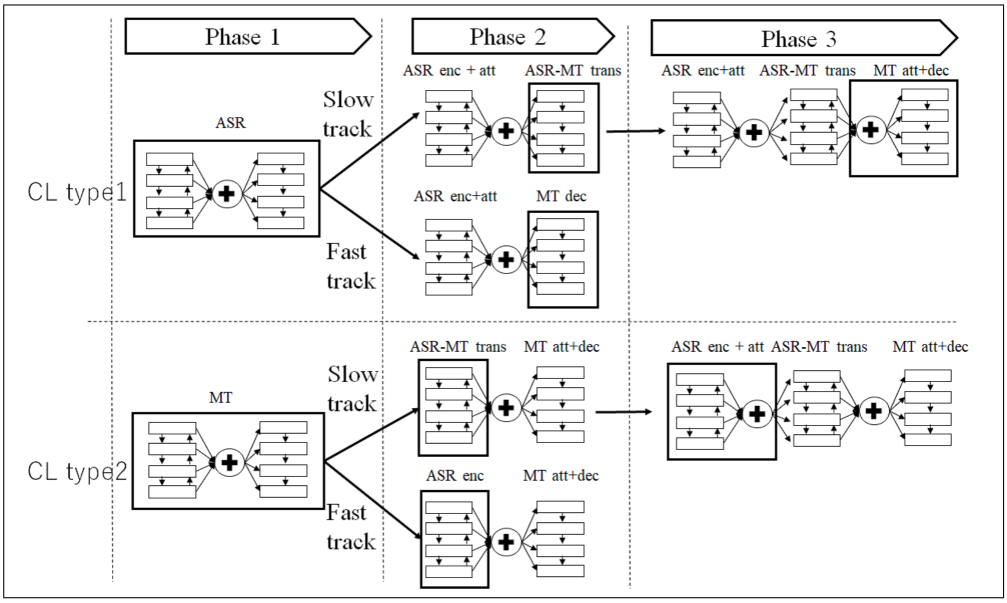

Figure 1: *Attention-based speech translation training phases with CL-based concept.*

(a) **Fast track**

Phase 1 We train an attentional-based encoder-decoder neural network for a standard ASR task, which predicts the corresponding transcription of the input speech sequence in the source language.

Phase 2 Next we replace the ASR decoder with a new decoder and retrain it to match the MT decoder's output. The model now predicts the corresponding word sequence in the target language given the input speech sequence of the source language.

(b) **Slow track**

Phase 1 As before, we train the attentional-based encoder-decoder neural network for a standard ASR task, which predicts the corresponding transcription of the input speech sequence in the source language.

Phase 2 Then we replace the ASR decoder with a new decoder and retrain it to match the MT encoder's output this work as ASR-MT transcoder. The model's objective now is to predict the word representation (like the MT encoder's output) that is good for the corresponding word sequence in the source language given the input speech sequence of the source language. Here, as a loss function, we calculate the mean squared error between the output of the new decoder with the ouput of the MT encoder.

Phase 3 Finally, we combine the MT attention and decoder modules to perform the speech translation task from the source speech sequence to the target word sequence and train the whole architecture using a softmax cross-entropy function.

2. **CL type 2: Start from attention-based MT system** Similar to CL type 1, we construct an attentional-based ST system for both fast and slow tracks, but instead of starting with an ASR system, we start with the MT system. In this case, the model gradually adapts the encoder part from the MT encoder to more closely resemble the ASR encoder.

5. Experimental Set-Up and Results

5.1. Experimental Set-Up

We conducted our experiments using a basic travel expression corpus (BTEC) [19, 20]. The BTEC English-Japanese parallel corpus consists of 4.5-k training sentences and 500 sentences in the test set. Since corresponding speech utterances for this text corpus are unavailable, we used the Google text-to-speech synthesis[1] to generate a speech corpus of the source language.

The speech utterances were segmented into multiple frames with a 25-ms window size and a 10-ms step size. Then we extracted 23-dimension filter bank features using Kaldi's feature extractor [21] and normalized them to have zero mean and unit variance. As for the text corpus, using one-hot vectors results in large sparse vectors due to a large vocabulary. In this study, we incorporated word embedding that learns the dense representation of words in a low-dimensional vector space.

We further used this data to build an attention-based ASR and MT system, a direct ST system, and a CL-based ST-system. Table 1 summarizes the network parameters. For all the systems, we used the same learning rate and adopted Adam[22] to all of the models.

5.2. Results and Discussion

We applied the attentional encoder-decoder architecture described in Section 3 to train the ASR, MT, and direct ST systems. We also constructed an ASR+MT cascade system. For our proposed models, we also applied the CL-based attentional

[1] Google TTS: https://pypi.python.org/pypi/gTTS

Table 1: *Model settings for each system*

ASR system	
Input units	23
Hidden units	512
Output units	27293
LSTM layer depth	2
MT system	
Source vocabulary	27293
Target vocabulary	33155
Embed size	128
Input units	128
Hidden units	512
Output units	33155
LSTM layer depth	2
Optimization	
Initial learning rate	0.001000
Learning descend rate	1.800000
Optimizing method	Adam [22]

encoder-decoder architecture described in Section 4 to train CL type 1 and CL type 2 (fast and slow tracks). Unfortunately, CL type 2 failed to converge. This might be due to the large divergence between the MT encoder in the text input space to the ASR encoder in the speech input space. The successfully trained systems are listed below.

Baseline ASR: speech-to-text model of source language.

Baseline MT: text-to-text translation model from source language to target language.

Baseline ASR+MT: speech-to-text translation model by cascading speech-to-text in source language with a text-to-text translation model.

Direct ST Enc-Dec: direct end-to-end speech translation model using a single attention-based neural network.

Proposed ST Enc-Dec (CL type 1 - Fast Track): end-to-end speech translation model trained using CL type 1 (fast track).

Proposed ST Enc-Dec (CL type 1 - Slow Track): end-to-end speech translation model trained using CL type 1 (slow track).

The performance of our ASR system achieved a 9.4% word error rate (WER). The remaining systems were evaluated based on translation quality using a standard automatic evaluation metric BLEU+1 [23].

First, we show how our proposed methods work during the training steps. Fig.2 illustrates the softmax cross-entropy until 15 epochs. The MT system has easiest task, which is translating the text in the source language to the corresponding target language. The loss decreased quite fast. On the other hand, direct ST training is hard, and therefore it gave the worst performance (its loss only decreased 0.04 from epochs 1 to 15). By using our CL-based proposed method, we can further decrease the loss. Specifically, the one that trained with CL type 1 - Slow Track successfully outperformed the text-based MT system.

Next, we investigated the translation quality of the models summarized in Fig.3. The results also reveal that the direct attentional ST system is difficult. Direct ST Enc-Dec model seems to be over-fitting the language model and could not handle the input speech utterances. The results also demonstrated that our proposed ST Enc-Dec (CL type 1 - Fast Track) model significantly improved the baseline. The best performance was achieved by the proposed ST Enc-Dec (CL type 1 - Slow Track) model, which even surpassed the text-based MT and cascade ASR+MT systems. This system is constructed with [ASRenc+att]+[ASRdec-MTenc]+[MTatt+dec] (Fig.1). The combination of the second and third parts actually resembles a conventional MT system. Therefore, from the MT system viewpoint, the additional components in the first part, which introduced more noise to the input of the MT system, might function as a denoising encoder-decoder that prevents over-fitting.

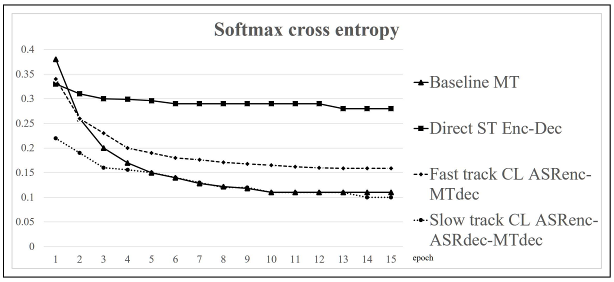

Figure 2: *Softmax cross-entropy of each epoch*

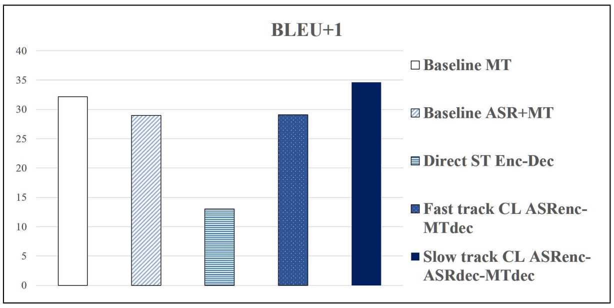

Figure 3: *Translation accuracy of each model*

6. Conclusions

In this paper, we achieved English-Japanese end-to-end speech to text translation without being affected by ASR error. Our proposals utilized structured-based CL strategies for training attentional-based ST systems in which we start with the training of attentional ASR and gradually train the network for end-to-end ST tasks by adapting the decoder part. Experimental results demonstrated that the learning model is stable and its translation quality outperformed the standard MT system. The best performance was achieved by our proposed model. Our current results, however, still rely on synthetic data. In the future, we will investigate the effectiveness of our proposed method using natural speech data, investigate various possible language pairs, paralinguistic information, and expand the speech-to-text translation task to a speech-to-speech translation task.

7. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101,JP17H00747 and JP17K00237.

8. References

[1] S. Nakamura, K. Markov, H. Nakaiwa, G. K. andHisashi Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, “The ATR multilingual speech-to-speech translation system,” *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. Vol 14, NO.2, 2006.

[2] T. Kano, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Generalizing continuous-space translation of paralinguistic information," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 2614–2618.

[3] Q. T. Do, S. Sakti, G. Neubig, and S. Nakamura, "Transferring emphasis in speech translation using hard-attentional neural network models," in *17th Annual Conference of the International Speech Communication Association (InterSpeech 2016)*, September 2016.

[4] P. D. Aguero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, May 2006, pp. I–I.

[5] L. Deng, A. Acero, and X. He, "Why word error rate is not a good metric for speech recognizer training for the speech translation task?" in *Proc. ICASSP*. IEEE, May 2011.

[6] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *CoRR*, vol. abs/1506.07503, 2015. [Online]. Available: http://arxiv.org/abs/1506.07503

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[8] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *HLT-NAACL*, 2016.

[9] A. Brard and O. Pietquin, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *30th Conference on Neural Information Processing Systems*, 2016.

[10] J. L. Elman, "Sequence to sequence learning with neural networks," *Cognition*, vol. Volume 48, Issue 1, 1993.

[11] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," 2009.

[12] C. Gong, D. Tao, S. Maybank, and J. Yang, "Multi-modal curriculum learning for semi-supervised image classification," *IEEE Transactions on Image Processing*, vol. Vol 24, Issue 7, 2016.

[13] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *CoRR*, vol. abs/1509.00685, 2015. [Online]. Available: http://arxiv.org/abs/1509.00685

[14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *CoRR*, vol. abs/1409.3215, 2014.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[16] M. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *CoRR*, vol. abs/1508.04025, 2015.

[17] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4960–4964.

[18] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, ser. NAACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 48–54. [Online]. Available: http://dx.doi.org/10.3115/1073445.1073462

[19] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-tospeech translation," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2003, pp. 381–384.

[20] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1674–1682, Sept 2006.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011, iEEE Catalog No.: CFP11SRW-USB.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[23] C.-Y. Lin and F. J. Och, "Orange: a method for evaluating automatic evaluation metrics for machine translation," in *Proceedings of Coling 2004*. Geneva, Switzerland: COLING, Aug 23–Aug 27 2004, pp. 501–507. [Online]. Available: http://www.aclweb.org/anthology/C04-1072