Neural Machine Translation Improvement by Acoustic Embedding

Takatomo Kano[1], Sakriani Sakti[1,2], and Satoshi Nakamura[1,2]
[1]Nara Institute of Science and Technology
[2]RIKEN, Center for Advanced Intelligence Project AIP
{kano.takatomo.km0, ssakti, s-nakamura}@is.naist.jp

1 Introduction

Neural machine translation (NMT) has successfully redefined the state of the art in machine translation on several language pairs. One popular framework models the translation process end-to-end using attentional encoder-decoder architecture and treats each word in the vectors of intermediate representation. In such architecture, word embeddings in continuous vector representations are an almost ubiquitous NMT component. These representations are sensitive to the meaning of words and their accurate order and reasonably insensitive to the replacement of the active voice with a passive voice [1]. These representation's behavior assumes that words in similar contexts have similar meanings. Such embeddings represent the semantics of the corresponding words/sequences, allowing semantic similar words to be grouped together in the vector spaces to share statistical power. The embedding layer provides advantages that increase the robustness for rare data and produce more natural outputs than statistical phrase-based translation [2]. Unfortunately, the model maps such similar words too closely, which complicates distinguishing them. Consequently, NMT generates words that seem natural in the target sentence that do not reflect the source sentence ' s original meaning. Many studies also argued that NMT's translations are often fluent but lack accuracy [3, 4, 5]. For example, the system mistakenly translated "may I" for "can I", "dog" for "cat," "Norway" for "Tunisia," and so on. Although it does not destroy the overall naturalness, the sentence ' s entire meaning might be completely different, which makes the error critical.

Osamura et al. proposed a simpler solution to incorporate speech information by an ASR posterior vector. This might resemble word confusion networks (WCNs) [6] that can directly express the ambiguity of word hypotheses at each time point. Osamura et al. [7] reported that acoustic information helps distinguish such semantic similar words as "cut" and "perm" in the encoder part, which helps the decoder find correct attention points and output correct words in the target language. However, these works only focused on the source speech from the source language that was incorporated into the NMT encoder part.

In this research, we learn how to incorporate acoustic information from the target language in collaboration with a text-to-speech (TTS) system. We integrate acoustic information within the NMT decoder by multi-task learning. Our model learns how to embed and translate the word sequences based on their acoustic and semantic differences to help the model choose the correct output word by considering its meaning and pronunciation. To the best of our knowledge, this is the first study that improved NMT in collaboration with TTS. In our proposed method we use a state of the art sequential translation model transformer with tied-embedding as a baseline. Our experiment results show that our proposed approach provides greater improvements than the standard text-based NMT model.

2 Proposed Approach

An encoder-decoder translation model maps an input sequence into a fixed-dimension vector [1]. Such representations are sensitive to the meaning of the sequence and accurate word order, but they are insensitive to the replacement of the active voice with the passive voice. In speech synthesis models, these representations are sensitive to the replacement of the active voice with the passive voice but insensitive to the meaning of the sequence. These attributes create models that are robust to test sets and generate natural sequences. On the other hand, the model sometimes confuses output with similar meaning words and context like a "dog" and "cat" and "may I" and "can I."

In this research we used a pre-trained TTS embedding weight for the NMT output layer. The transformer decoder has two modules that handle the target word: a target word embedding layer and an output layer. We treat these two as inverse mappings and tied their weights [8]. We tied a decoder embedding layer weight and a decoder output layer weight. We added a new output layer where the mapping decoder was hidden using TTS embedding weights that were not updated during training. This model has two types of output layers. The standard decoder output layer weight is tied to the decoder embedding weight. This output layer maps decoder hidden to output for a sensitive sequence meaning. The output layer, is tied with a TTS embedding weight, maps decoder hidden for sensitive sequence pronunciations. We use these to output the results and back-propagate the loss:

$$o_{nmt} = W_{nmt} h_{dec}, \quad (1)$$

$$o_{tts} = W_{tts} h_{dec}, \quad (2)$$

$$loss = (1-\lambda) CE(o_{nmt}, y) + \lambda CE(o_{tts}, y). \quad (3)$$

Here h_{dec} is a decoder hidden sequence, W_{nmt} denotes

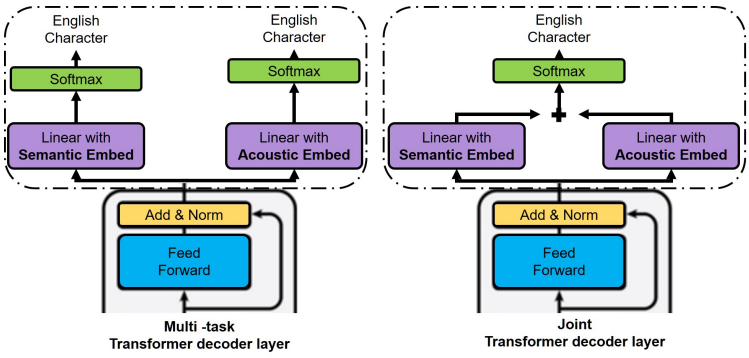

Figure 1: Proposed models architectures

a decoder word embedding weight, and W_{tts} denotes the TTS encoder embedding weight. In this work, W_{tts} is not update during training. We only update W_{nmt} through training. We using soft-max cross entropy (CE) to individually calculate the loss for each output, and λ is the weight for each loss. We call our proposed method multi-task learning:

$$o_y = o_{nmt} + o_{tts}, \quad (4)$$

$$loss = CE(o_y, y). \quad (5)$$

We sum both NMT and TTS weight mapping. Since we do not update the TTS embedding weight during training, the model updates the NMT embedding weight scale based on the degree of each layer's contribution. If the output from the NMT embedding scale greatly exceeds the output from the TTS embedding, then the proposed model resembles a standard NMT. We call our proposed method joint learning. We summary this section in Fig. 1.

3 Experiment

We conducted our experiments using a basic travel expression corpus (BTEC) [9, 10]. The BTEC Japanese-English parallel corpus consists of 480-k utterances. We removed the sentences that have more than 100 characters and used this dataset to build a baseline and proposed sub-words for the characters for the Transformer NMT. Next we demonstrate our proposed translation performance and compare it with the baseline text-based NMT model. We used OpenNMT[1] to make a baseline and implemented our proposed model on it. Here is a summary of baseline and our proposed models:

Baseline Text-based NMT
This is a baseline transformer-based text-to-text translation model.

Proposed Multi-task$_{NMT}$
Proposed model with multi-task learning and a semantic weight output layer in test (Fig. 1).

Proposed Multi-task$_{TTS}$
Proposed model with multi-task learning and an acoustic weight output layer in test (Fig. 1).

[1]OpenNMT: http://opennmt.net/

Proposed Joint
Proposed model with joint learning and both output layer in test decoding (Fig. 1).

All models performed a beam search (beam size is 5) algorithm for character sequence auto-regressive decoding.The baseline text-based NMT and our proposed model used the same settings, and the trainable number parameters are the same between the proposed model and the baseline.

Table 1: Translation quality of Japanese-to-English

Model	BLEU score	WER
Text-based NMT	45.10	35.5%
Multi-task$_{NMT}$	**50.51**	30.5%
Multi-task$_{TTS}$	50.23	**30.1%**
Joint	48.12	32.4%

Table 2: Translation results of Japanese-to-English

Source	ミルク を もう 少し く だ さ い
Target	a little more milk please
Text-based NMT	**let me have some** more milk
Multi-task$_{NMT}$	a little more milk please
Source	牛肉 と チキン の どちら が よろし い で す か
Target	which would you like beef or chicken ?
Text-based NMT	** beef or chicken ?
Multi-task$_{NMT}$	which would you like beef or chicken ?

Table 1 shows that our proposed method successfully improved the BLEU scores by 5-points from the text-based NMT. For further discussion of the model behaviors, Table 2 lists the translation results from each model. Each proposed model output a sentence whose meaning was very similar to the meaning of the target sentence. This means that each proposed model extracted the meaning of the source sentence and mapped it to the decoder state. But in contrast in the text-based NMT baseline, the text-based NMT model failed to choose the correct word from the decoder state. The output layer is usually one simple linear regression layer that maps a vector from a continuous narrow space to a large discrete space. If the model maps a similar word too closely, then the output layer cannot separate it again. On the other hand, our proposed model output a correct word for each sentence. This reveals that by incorporating acoustic embedding and constructing a model in a multi-task fashion with two output layers, each layer can map the decoder state to different output with different weights. The hidden representation might be sensitive for both semantic

and pronunciation similarities. Therefore, our proposed model can choose the correct word that not only depends on its meaning but also on its pronunciation.

4 Conclusion

We used TTS embedding weight to map translation results. This approach created an NMT model that is sensitive to sequence meaning and pronunciation. Our proposed method outperformed a standard transformer with BLEU scores. We first considered NMT and TTS collaboration. Our proposed method made an NMT that can learn such multi-modal information as text meaning and pronunciation from a text. Future work will consider ASR, NMT, and TTS deep joint optimization to improve the translation performance and improve NMT so that it can handle other kinds of information, such as images. Furthermore, we will apply our proposed approach to more difficult translation data such as TED Talks.

5 Acknowledgements

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

References

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014.

[2] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*, 2007.

[3] Philip Arthur, Graham Neubig, and Satoshi Nakamura. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1557–1567, 2016.

[4] Xing Wang, Zhengdong Lu, Zhaopeng Tu, Hang Li, Deyi Xiong, and Min Zhang. Neural machine translation advised by statistical machine translation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3330–3336, 2017.

[5] 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, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[6] Lidia Mangu, Eric Brill, and Andreas Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400, 2000.

[7] Kaho Osamura, Takatomo Kano, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. Using spoken word posterior features in neural machine translation. 21:22, 2018.

[8] Nikolaos Pappas, Lesly Miculicich Werlen, and James Henderson. Beyond weight tying: Learning joint input-output embeddings for neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 73–83, 2018.

[9] Gen-ichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. Creating corpora for speech-to-speech translation. In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*, 2003.

[10] Gen-ichiro Kikui, Seiichi Yamamoto, Toshiyuki Takezawa, and Eiichiro Sumita. Comparative study on corpora for speech translation. *IEEE Trans. Audio, Speech & Language Processing*, 14(5):1674–1682, 2006.