Neural Machine Translation with Acoustic Embedding

Takatomo Kano[1], Sakriani Sakti[1,2], Satoshi Nakamura[1,2]

[1]Nara Institute of Science and Technology, Japan [2]RIKEN, Center for Advanced Intelligence Project AIP, Japan

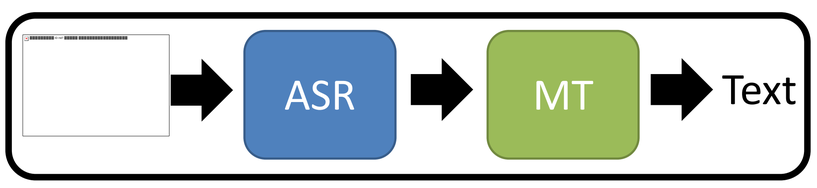
Introduction

■ **Neural machine translation (NMT)**

→ State of the art in MT for several language pairs

→ Often mistranslates words that seem natural in the target context but do not reflect the content of the source sentence

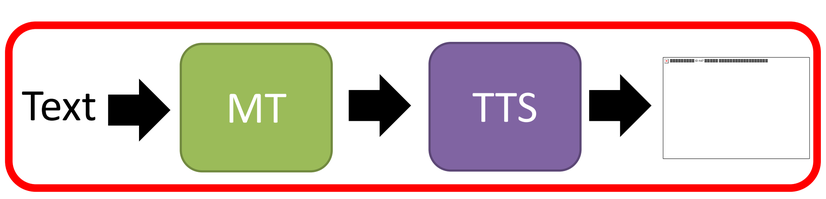
→ To enhance the discriminability, most studies incorporate additional information from the source side (focus on ASR+MT)



■ **We propose:**

→ **Incorporating Acoustic Embedding into NMT (focus on MT+TTS)**

→ **Use TTS embedding to model acoustic information of MT target sequence**



Proposed Approach

■ MT is sensitive to the content of the source & target language sequence; TTS is sensitive to the pronunciation of the target language sequence

■ **Proposed Approach:**

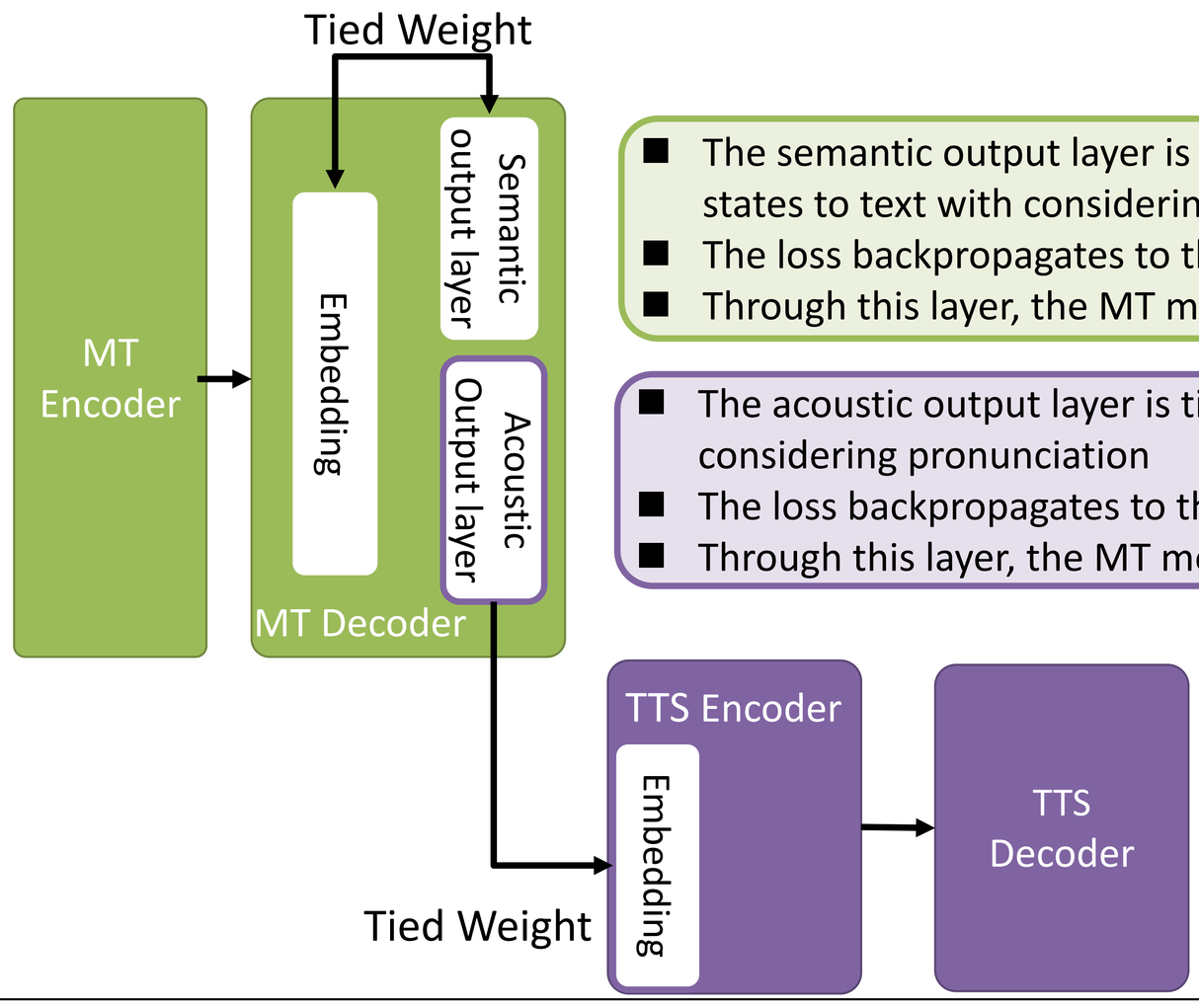
→ Use TTS embedding to restrict MT's target generation

→ Embedding functions as a feature extraction and word reconstruction module

→ **Type 1: Multi-task approach:** Handles output from MT and TTS embedding output layers individually

→ **Type 2: Joint NMT+TTS approach:** Summation output from NMT and TTS embedding output layer

→ **Framework:** Transformer NMT & TTS (input: Japanese subwords; output: English characters)



■ The semantic output layer is tied to MT decoder embedding that maps decoder hidden states to text with considering content

■ The loss backpropagates to the decoder through the semantic output layer

■ Through this layer, the MT models the target content

■ The acoustic output layer is tied to TTS embedding that maps hidden states to text with considering pronunciation

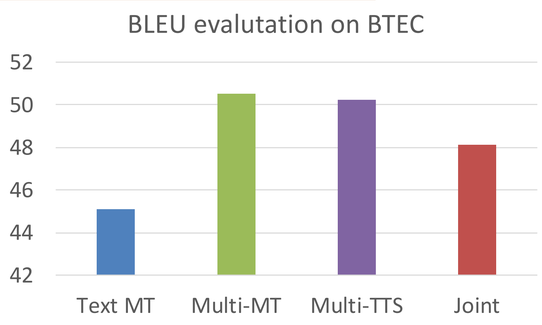
■ The loss backpropagates to the decoder through the acoustic output layer

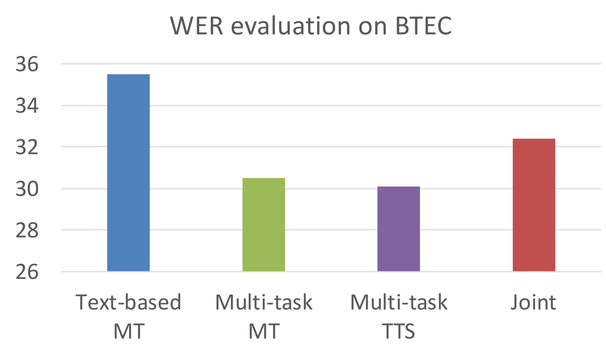
■ Through this layer, the MT models the target pronunciation

■ **Expectation:**

MT becomes sensitive to content and pronunciation

Experiments



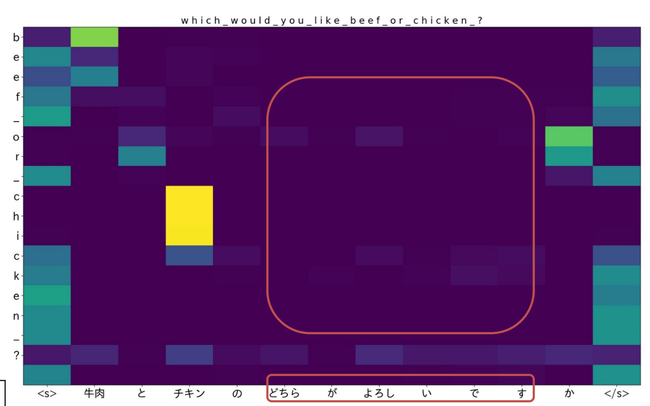


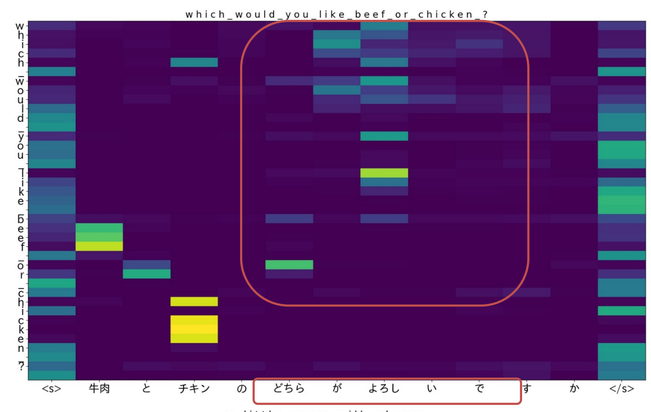
→ Proposed methods outperform text-based transformer NMT

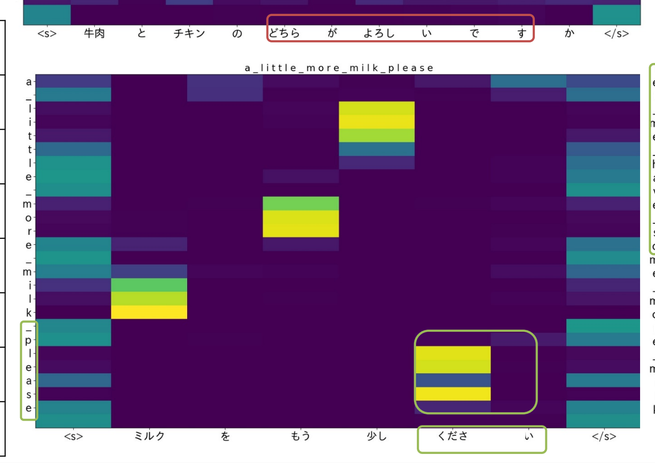
→ New output layer (TTS embedding weight) improves WER

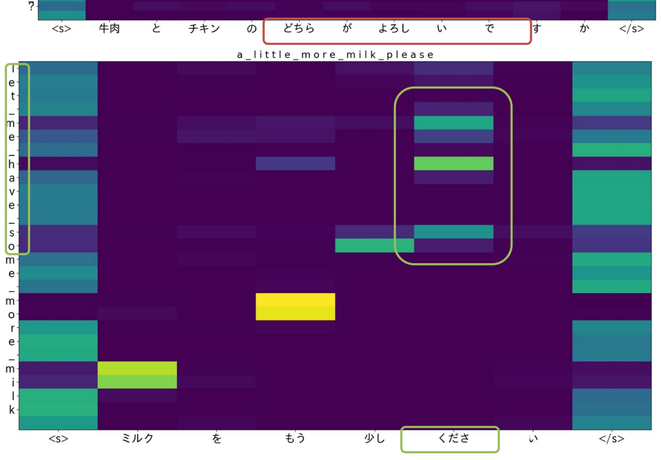
Source	Gyuniku to Chikin no dochira ga yoroshi i de su ka
Reference	which would you like beef or chicken ?
Text MT	** beef or chicken ?
Multi-MT	which would you like beef or chicken ?
Source	Miruk wo mou sukoshi **ku dasa** i
Reference	a little more milk please
Text MT	let me have some more milk
Multi-MT	a little more milk please

■Dataset: BTEC parallel text corpus (train: 480k; dev: 1k; test: 500)









Conclusion

■ **Conclusion**

→ Used TTS embedding weight to map translation results

→ Sensitive to sequence meaning and pronunciation

→ Outperformed text-based transformer NMT

→ Can learn multi-modal information from text

■ **Future works**

→ Consider ASR, NMT and TTS fully joint optimization

→ Utilize other kinds of information, e.g. images

→ Investigate other translation data such as TED Talks