Language Model Integration Based on Memory Control for Sequence to Sequence Speech Recognition

Jaejin Cho[1], Shinji Watanabe[1], Takaaki Hori[2], Murali Karthick Baskar[3], Hirofumi Inaguma[4], Jesus Villalba[1], Najim Dehak[1]

[1]Center for Language and Speech Processing, Johns Hopkins University
[2]Mitsubishi Electric Research Laboratories (MERL)
[3]Brno University of Technology
[4]Kyoto University

Introduction

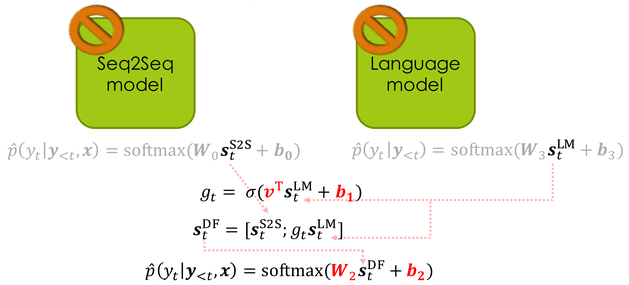
- Task: **Language model (LM) integration to help sequence-to-sequence (S2S) ASR training**
- Proposal:
  - Update of the hidden/cell states in S2S LSTM decoder using LM information
  - Use of the LM information for both character inference and states update in decoder
  - 3 variants with the idea
- System:
  - S2S attention model with CTC-loss as a regularizer
  - LM trained ahead before the S2S model training

Background: LM integration in S2S model

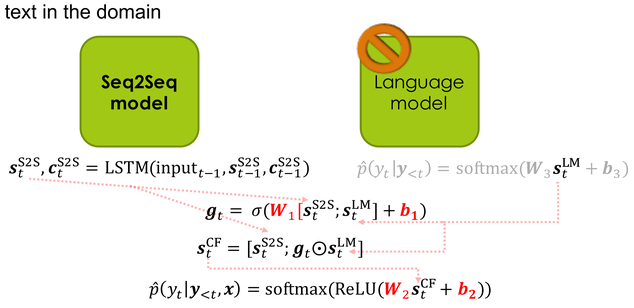
- **LM integration in decoding**:
  - **Shallow fusion (SF)**: Linear interpolation between two scores with a hyper parameter

$$\hat{y} = \underset{y}{\operatorname{argmax}} \left(\log p(y|x) + \gamma \log p(y) \right)$$

  - **Deep fusion (DF)**: Parameter learning to connect LM and S2S model



- **LM Integration in training**:
  - **Cold fusion (CF)**: Training of S2S model in the help of trained LM, ideally with extra unpaired text in the domain



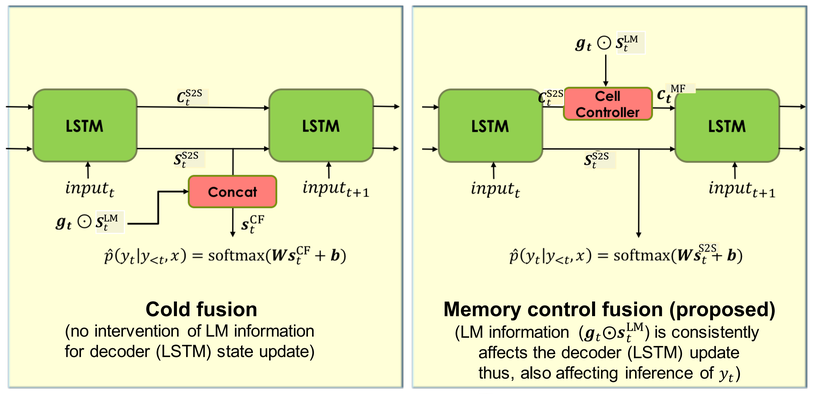
References

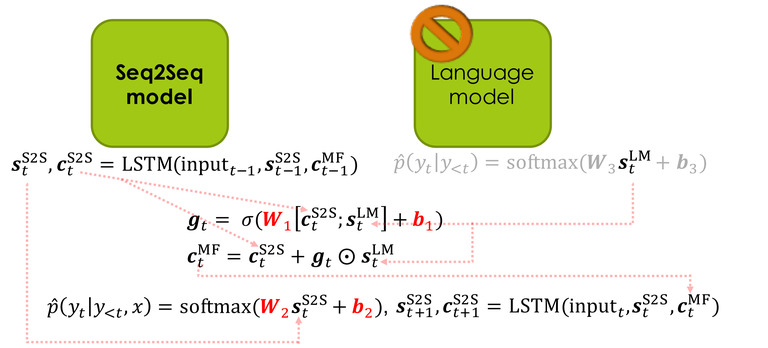
[1] Anuroop Sriram, et al. "Cold fusion: Training seq2seq models together with language models," *arXiv preprint arXiv:1708.06426*, 2017.

[2] Shubham Toshniwal, et al. "A comparison of techniques for language model integration in encoder-decoder speech recognition," *arXiv preprint arXiv:1807.10857*, 2018.

Proposed method: Memory control fusion (MF)

- Belongs to the second category, **LM integration in training**
- **Controls the hidden/cell (memory) states in S2S decoder using LM information**
- Affects **both inference and the states update in the decoder** over time

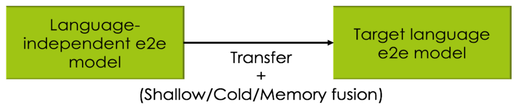




- 3 variants:
  1) Cell update (MF1), 2) CF+ MF1 (MF2) , 3) Cell & State Update (MF3)

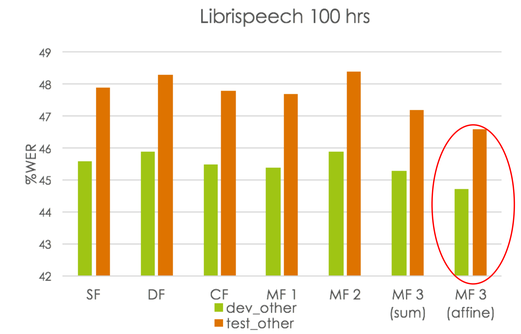
Experimental setup

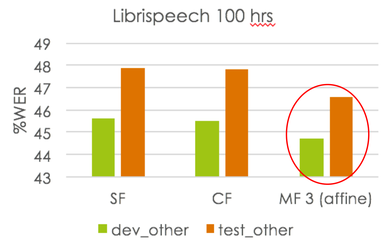
- **Mono-lingual ASR**
  - Paired data (Speech and its transcript): **Librispeech** 100/960 hrs
  - **External text** (not paired with speech): **10 times** of whole paired text
  - S2S: 8-layer BLSTM encoder + 1-layer LSTM decoder, LM: 2-layer LSTM
- **Transfer learning** from a language-independent model to a target model
  - **Language-independent model**: Trained with **10 Babel languages** (~643 hrs) not including the target language
  - **Target model**: Trained on a target language data, **Swahili** (~50 hrs) with initialized parameters from the language-independent model

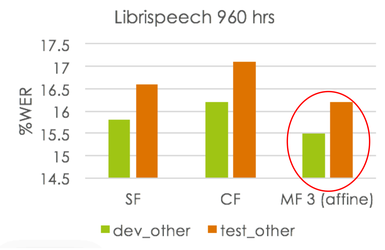


Results

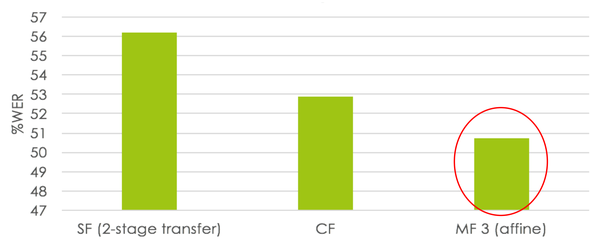
- Mono-lingual setup







- Transfer learning setup



Conclusion

- Consistent improvements compared to the previous methods
- Third variant updating both hidden/cell states worked the best
- ~2 to 4% relative improvement in WER in mono-lingual setup
- ~9 to 10% relative improvement in WER in multi-lingual transfer learning setup

Acknowledgements

The work reported here was started during JSALT 2018, and supported by JHU with gifts from Amazon, Facebook, Google, Microsoft and Mitsubishi Electric.