

# Recent Advances to Sequence-To-Sequence Learning

Tsz Kin Lam

Department of Computational Linguistics  
lam@cl.uni-heidelberg.de

24th Oct, 2019

## **Lecturer: Tsz Kin Lam**

- 1st-year PhD student of Prof. Stefan Riezler
- Concentrations: Neural and Speech Translations, RL and interactive techniques
- Personal research interests: (Multi-agent) machine learning, Generative models, Bayesian related techniques.
- I do not know anything about linguistics ...
- Office Hour: by appointment or try to talk to me when you meet me.

## Proseminar / Hauptseminar

- ① Read all selected papers
- ② Actively participate to the discussions, in particular,
  - prepare  $\sim 2$  questions for each presentation (Questions shown on moodle)
- ③ Present a paper + summary of questions discussed
- ④ +implementation project (HS)

### Presentation and discussion:

- Depends on the number of participants
  - **35  $\leq$  duration (mins)  $\leq$  45 and  $\sim$ 15 minutes of Q&A**, or
  - **20  $\leq$  duration (mins)  $\leq$  25 and  $\sim$ 15 minutes of Q&A** if more participants than the number of weeks.
- **A 5 to 10 minute summary** of the previous discussion at the beginning of next lesson
  - Key points of the papers
  - Questions we discussed and their solutions
- The format of Q&A is decided by the presenters:
  - ① Allow interruptions and questions at anytime during the talk
  - ② Q&A only after your presentation.
  - ③ or a mix of it.
- Please send me your slides 1-week before your presentations.

Papers  $\rightarrow$  [Course webpage](#)

# Papers - Module 1: Neural network architecture

Gehring, Jonas, et al. "Convolutional Sequence to Sequence Learning"

- Convolution layers vs Recurrent Layers, e.g. parallelizability
- The use of positional embedding and residual connections in Seq2Seq learning.

Chen, Mia Xu, et al. "Quasi-Recurrent Neural Networks"

- How to speed up LSTM/RNN without using convolutions, especially in small batch size and long sequence cases.

Vaswani, Ashish, et al. "Attention is all you need"

- Self-attention vs convolution and recurrent layers
- Multi-head attention?
- Layer normalisation

Pham, Ngoc-Quan, et al. "Very Deep self-attention networks for end2end speech recognition"

- Applying Transformer to the setting of speech translation - straightforward?
- Stochastic layers

## Papers - Module 1: Neural network architecture Cont

Chen, Mia Xu, et al. "The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation"

- RNMT+ and hybrid architecture
- label smoothing

Xia, Yingce, et al. "Deliberation Networks: Sequence Generation Beyond One-Pass Decoding"

- What makes Microsoft Research Translate sounds no. 1 (?)
- Two-pass decoding

Oord, Aaron van den, et al. "WaveNet: A generative model for raw audio"

- Dilated convolution
- Gated Activation units
- Conv layer of kernel size of 1x1

# Summary for module 1

After module 1, we expect to know:

- Pros and Cons for convolution, recurrent and self-attention layers
- Application in text-to-text, speech-to-text or text-to-speech cases.
- Small but important training tricks, e.g., label smoothing.

### **Something about data instead of model - low-resources scenarios**

He, Di, et al. "Dual Learning for Machine Translation"

- What makes Microsoft Research Translate sounds no. 2
- Data-level unsupervised dual learning  $\implies$  ... ?

Baskar, Murali Karthick, et al. "Semi-supervised Sequence-to-sequence ASR using Unpaired Speech and Text"

- Fine-tuning ASR and TTS using unpaired speech and text data
- Auto-encoder ...

Liu, Alexander H., et. al. "Adversarial training of end-to-end speech recognition using a criticizing language model"

- GAN setting
- Fine-tuning the ASR with discriminator based on paired and unpaired text.

Chorowski, Jan, and Navdeep Jaitly "Towards better decoding and language model integration in sequence to sequence models"

- Using unpaired data to enhance decoding
- should be read together with the paper: Gulcehre, Caglar, et al. "On Using Monolingual Corpora in Neural Machine Translation"

Sriram, Anuroop, et al. "Training Seq2Seq Models Together with Language Models"

- Cold Fusion vs shallow and deep fusion
- Unsupervised pre-training

## Summary for module 2

After module 2, we expect to know:

- How to leverage unpaired data to improve the performance of seq2seq both in terms of training and decoding.

Gu, Jiatao, et al. "Non-Auto regressive NMT"

- Basics

Gu, Jiatao, Qi Liu, and Kyunghyun Cho "Insertion-based Decoding with Automatically Inferred Generation Order"

- Should read together with: Stern, Mitchell, et al. "Insertion Transformer: Flexible sequence generation via insertion operations"
- Decoding in a binary-tree like format

Ghazvininejad, Marjan, et al. "Mask-Predict: Parallel Decoding of Conditional Masked Language Models"

- Ask Julia

## Summary for module 3

After module 3, we expect to know:

- Is non-auto regressive idea sensible in seq2seq?
- How fancy research in machine translation can be?

Please sign up by next class.  
I will give a tutorial about Seq2Seq next week  
We start our discussions on 7th Nov 2019