Another Diversity-Promoting Objective Function for Neural Dialogue Generation

Ryo Nakamura, Katsuhito Sudoh, Koichiro Yoshino, Satoshi Nakamura

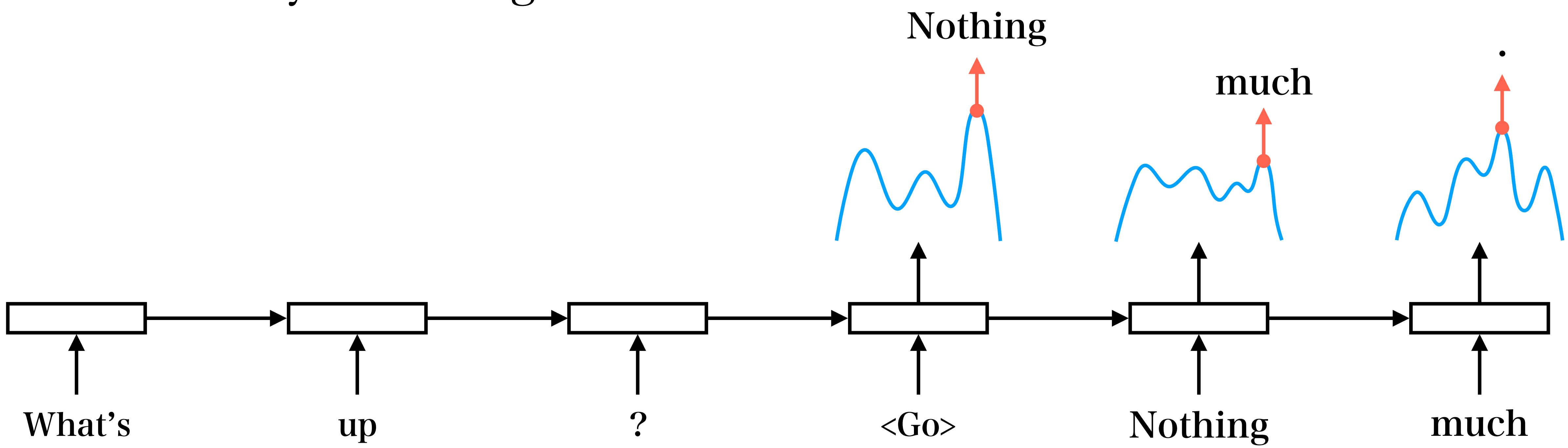
Graduate School of Information Science, Nara Institute of Science and Techonology, Japan

RIKEN, Center for Advanced Intelligence Project AIP, Japan

{nakamura.ryo.nm8, sudoh, koichiro, s-nakamura}@is.naist.jp

Neural Dialogue Generation

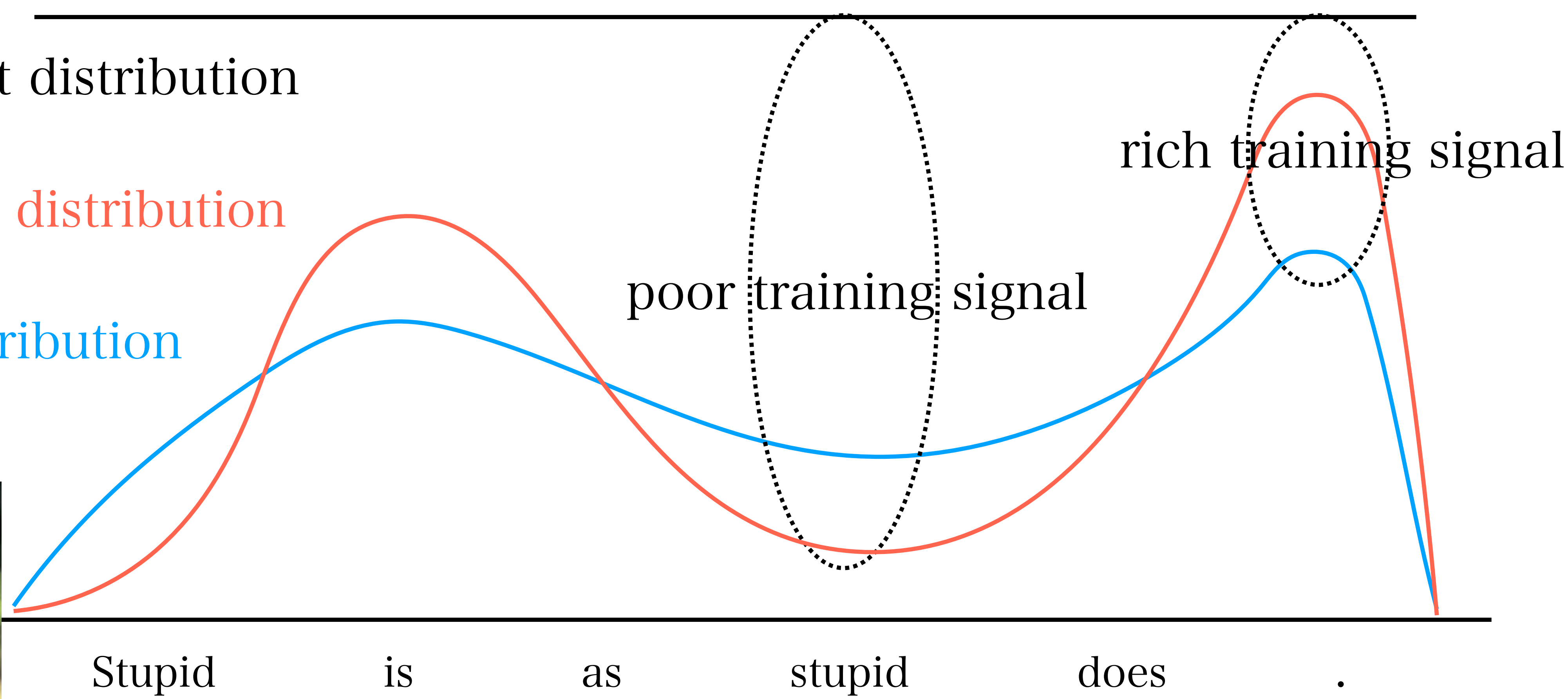
- An open-domain dialogue system that generates a response word by word using a trained neural network (e.g., seq2seq)
- Generation-base is more flexible than retrieval-base, but fluency, consistency are not good

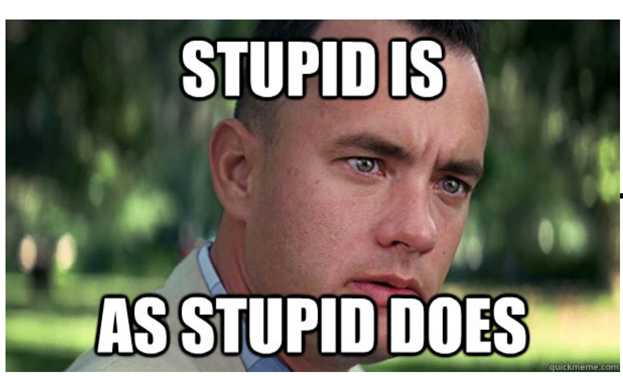


- In particular, the generated response has **low diversity** and tends to be a generic response like "I don't know." **Why?**

During training

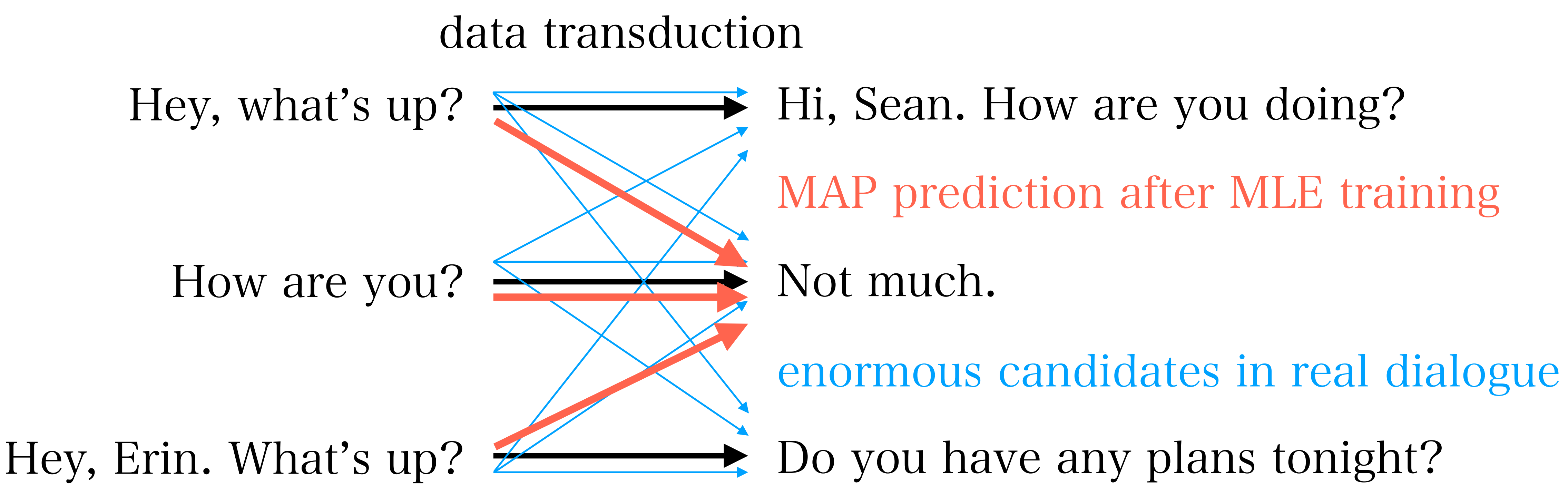
- Frequent words in training set supply more training penalties than rare words
- Therefore, large occurrence probabilities is assigned to frequent words





During evaluation

- Dialogue generation is a **many-to-many** transduction task in which contents vary depending on the context
- Frequent words are applicable in any context, so they tend to be candidates for generation
- As a result, only the most likely response is generated



Break down low diversity problem

During training

No suggestion. We challenged it!

- **Frequent words supply more penalties than rare words**
- Due to lack of data and data imbalance (Serban et al. 2016)
- Softmax Cross-Entropy (SCE) loss is not good because all words are handled equally regardless of lack and imbalance

During evaluation

Measures already suggested

- **Maximum A-Posteriori (MAP) predicts the most likely response only**
- A way to generate unlikely response using Maximum Mutual Information (MMI) is reported in (Li et al. 2016)

Previous research

- Maximum Mutual Information (Li et al. 2016)

$$\hat{T} = \operatorname{argmax} \{\log p(T|S) - \lambda \log p(T)\},$$

- MMI-antiLM suppresses language model-like generation by subtracting a language model term $\log p(T)$ from transduction model term $\log p(T|S)$.

- They used MLE during training and used MMI-antiLM during evaluation.

- In practice, MMI-antiLM generates token y:

$$y = \operatorname{argmax} \{\log \operatorname{softmax}(x - \lambda u)\},$$

Proposed method

Softmax Cross-Entropy loss

- SCE loss treats each token class equally

$$L_{\mathrm{sce}} = -\log\left(\frac{\exp(x_c)}{\sum_k^{|V|}\exp(x_k)}\right)$$

You
do
not
talk
about
Fight
Club
.

Inverse Token Frequency loss

- ITF loss scales smaller loss for frequent token classes

$$\begin{aligned} L_{\mathrm{tfd}} &= w_c L_{\mathrm{sce}} \\ w_c &= \frac{1}{\mathrm{freq}(\mathrm{token}_c)^{\lambda}} \end{aligned}$$

You
do
not
talk
about
Fight
Club
.

Advantages compared to previous works

- No special inference method. You can use common greedy search
- ITF loss can be easily incorporated. **Just replace loss function!**
- Training with ITF loss is as stable as training with SCE loss.
- ITF models yield **state-of-the-art diversity** and maintains quality.

Code examples with PyTorch

Very Easy!!

```
sce_loss = nn.NLLLoss(weight=None)
```

SCE loss

```
def get_weights(_lambda):
    weights = torch.zeros(vocab_size)
    for token, index in token2index.items():
        weight = 1 / (token2freq[token]**_lambda)
        weights[index] = weight
    return weights
```

Inverse Token Frequency

```
weights = get_weights(_lambda=0.4)
itf_loss = nn.NLLLoss(weight=weights)
```

ITF loss

Experiment setups

Datasets

- OpenSubtitles (En) — 5M turns and 0.4M episodes
- Twitter (En/Ja) — 5M/4.5M turns and 2.5M/0.7M episodes

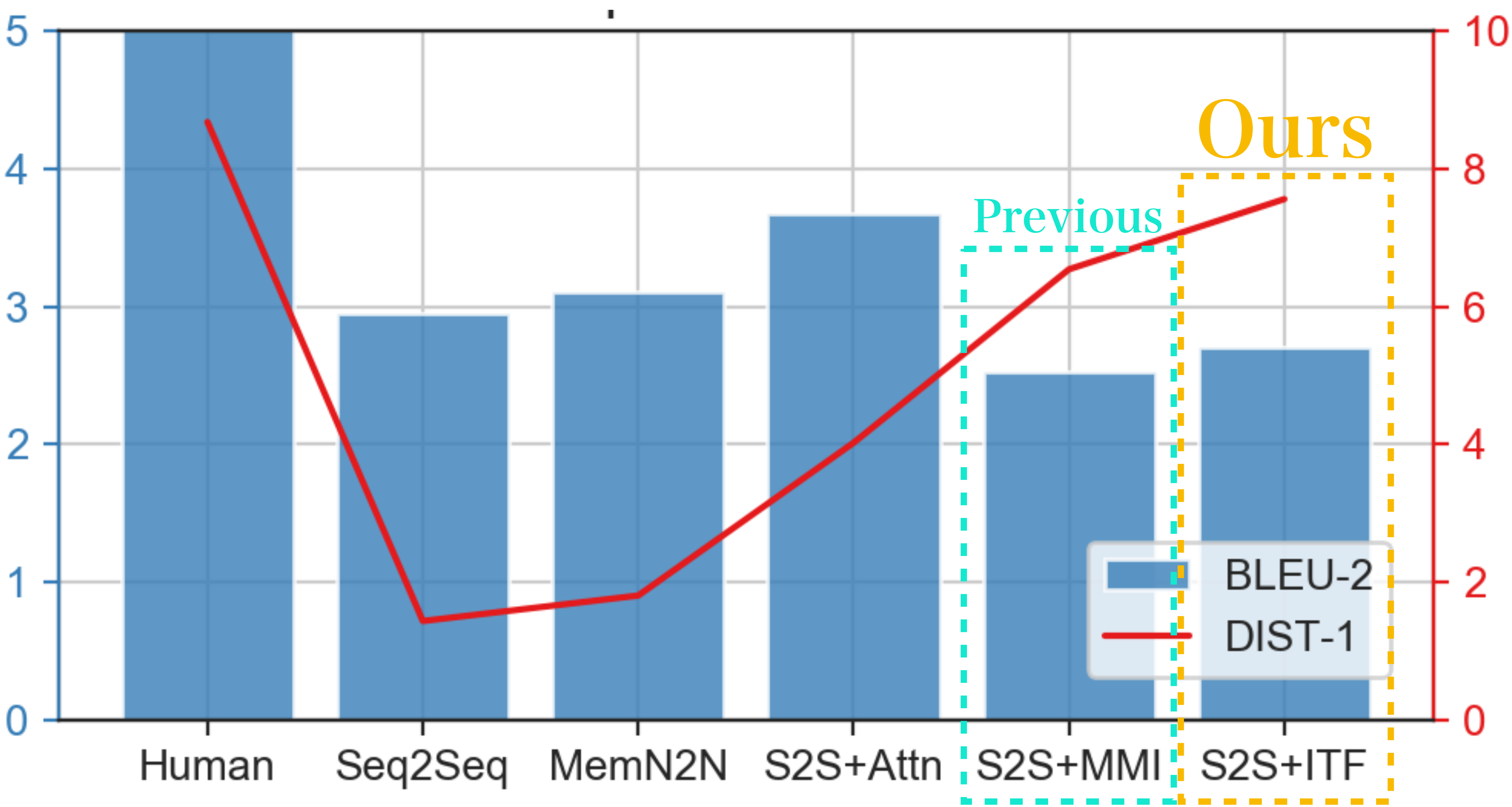
Baselines

- Seq2Seq — 4 layers Bi-LSTM w/ residual connections
- Seq2Seq + Attention
- Seq2Seq + MMI
- MemN2N — considering dialogue history using memory

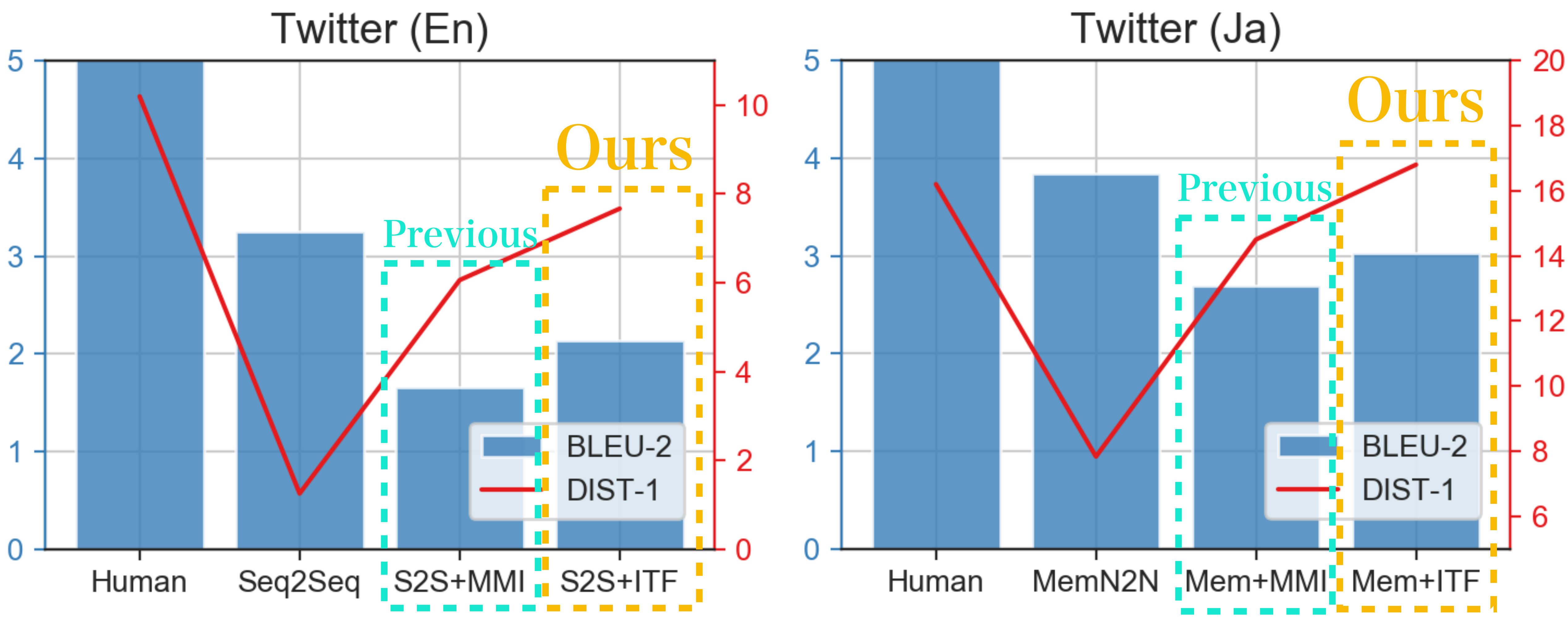
Evaluation metrics

- BLEU-1/2 — n-grams matching between all hypo. and all ref.
- DIST-1/2 — distinct n-grams in all generated responses

Result on OpenSubtitles



Result on Twitter



- ITF models outperform the MMI on both of BLEU-2 and DIST-1
- ITF model achieves a **ground truth-level DIST-1** score of 16.8 on Japanese Twitter Dataset

A generated sample on OpenSubtitles

SRC	**Does he know what's going on?**
TGT	**He knows he's leaving.**
MLE	No.
MMI	No.
ITF	He's got a lot of trouble.

A generated sample on Twitter

SRC	**12 gb ram at 384 gb/sec (gddr5x).. if this is true than damnn!.. you want**
TGT	**eurogamer also said that it speculates 384 and 12 gb of ram by placement of ram in original video**
MLE	i'm not sure if it is worth it.
MMI	mwr gpu is the best.
ITF	rambo is a newer one with chromebook7 connector, laptop router, hdmi cables.

Summary

- SCE loss + MAP prediction => Low diversity => Dull Response
- SCE loss + MMI inference => High diversity and good quality

A Diversity-Promoting Objective Function for Neural Conversation Models

Jiwei Li[1*] **Michel Galley**[2] **Chris Brockett**[2] **Jianfeng Gao**[2] **Bill Dolan**[2]

[1]Stanford University, Stanford, CA, USA

- ITF loss + MAP prediction => Very high diversity and good quality

Another Diversity-Promoting Objective Function for Neural Dialogue Generation

Ryo Nakamura, Katsuhito Sudoh, Koichiro Yoshino, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Techonology, Japan

RIKEN, Center for Advanced Intelligence Project AIP, Japan