UndNet - A Mutual Understanding Maximization Framework

Ruochen Zhang | ruochen_zhang@brown.edu,
Zejiang Shen | zejiang_shen@brown.edu

BROWN Computer Science

「Motivation」

- Popularity and importance regarding conversational AIs rises in all lines of business, yet how to generate the most appropriate response stays challenging.
- Traditional generative models frame dialog generation as machine translation problem [1], neglecting that similar sentences could not ensure identical understanding in different perspectives.
- UndNet is proposed and implemented with the aim of maximizing the mutual understandings of the conversation participants.

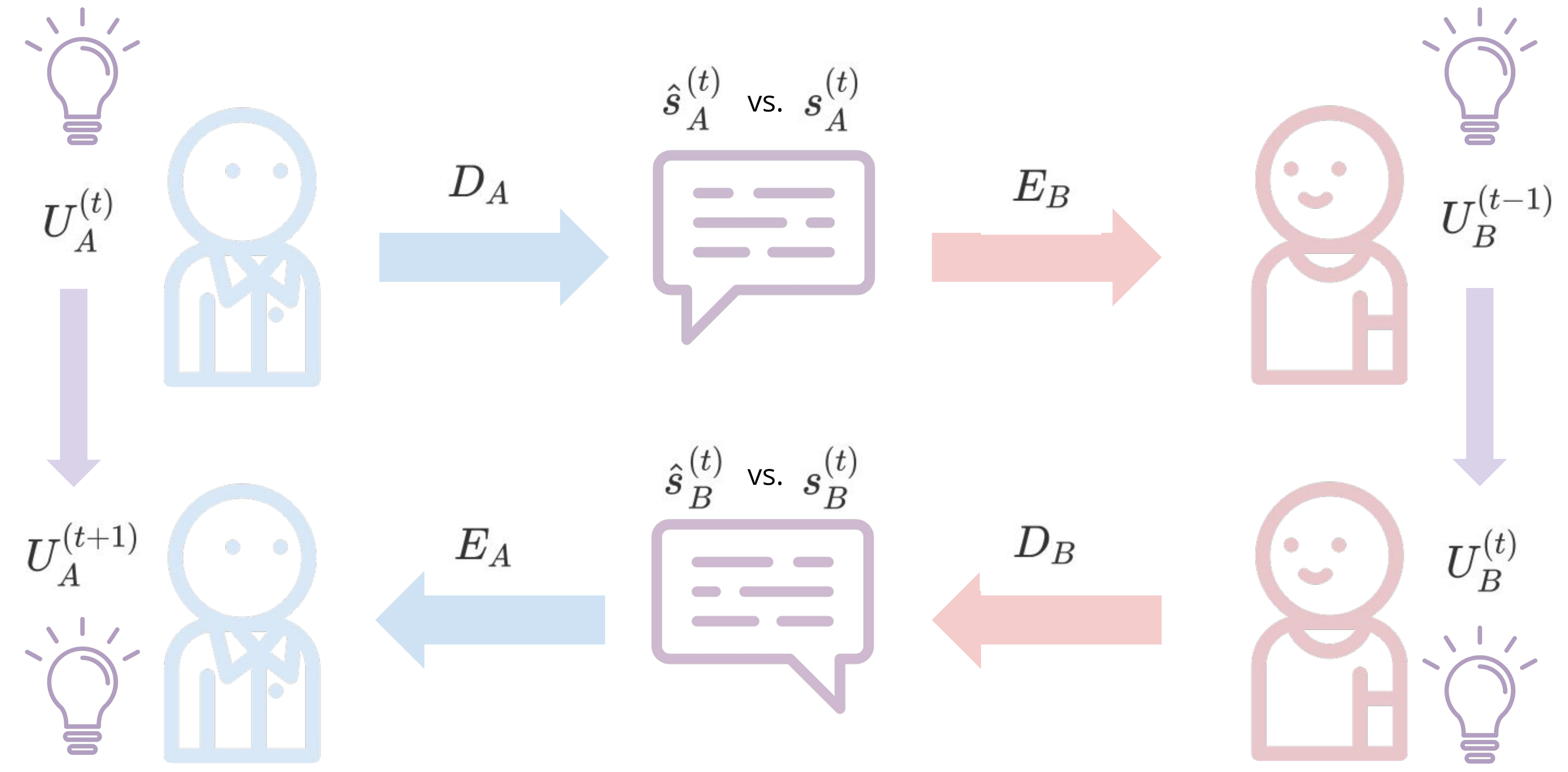
「Specifications」

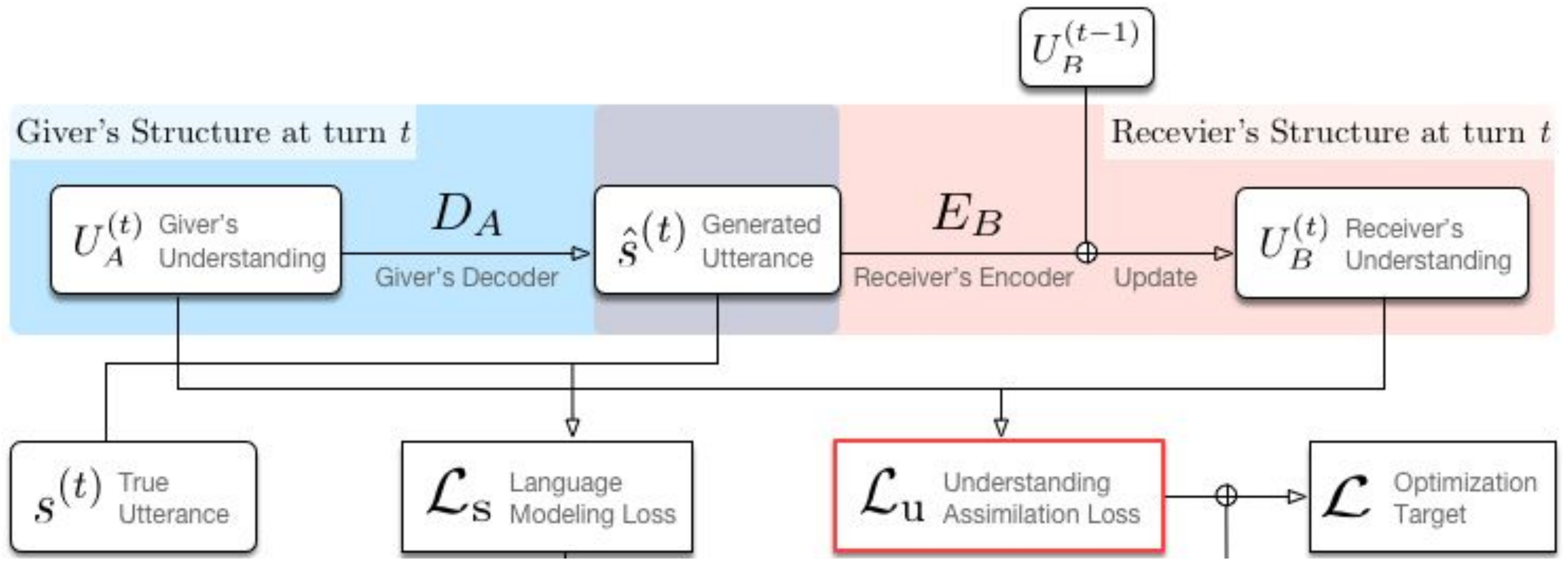
- ConvAI2 Dataset: consists of 164,356 utterances in over 10,981 dialogs
- The encode and decoder of the model is implemented using PyTorch and ParlAI Framework with 2-layer GRUs and a hidden size of 128

「Discussion」

- *ConvAI2 Dataset: Short Conversations*
  - Each episode of the conversation is quite short, therefore, it is hard for the model to learn the pattern of the understanding fast enough before the conversation ends
- *Representation of the Understanding*
  - The current model implementation uses hidden size for initializing the understanding, yet sizes can be variable and larger size may allow storing richer information
- *Sampling for the Understanding Update*
  - Current update is the average of previous understanding and the hidden state output. But what is a better way to refresh people's mind?
- *Pitfall of minimizing understandings' discrepancy*
  - The model may be fooled by the optimization target to generate similar decoders' output, leading to great instability for gradient descent

「Model」





△ Architecture of UndNet

The UndNet framework views conversation as a turn-based activity. At turn t :

- We compute the Earth Mover's Distance [2] as **Understanding Similarity Loss** to measure the discrepancy between participants' understanding

$$\mathcal{L}_u = \mathcal{D}_{\mathrm{EMD}}(U_A^{(t)}, U_B^{(t)})$$

- **Language Modeling Loss** adopts Cross Entropy Loss to assess response accuracy

$$\mathcal{L}_S = D_{\mathrm{CE}}(\hat{s}_A^{(t)}, s_A^{(t)}) + D_{\mathrm{CE}}(\hat{s}_B^{(t)}, s_B^{(t)})$$

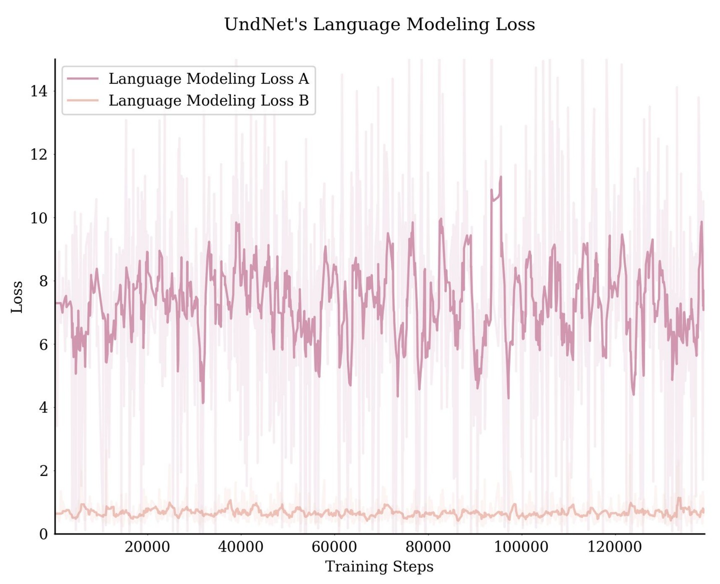
- **Optimization target** is thus the combination of all losses

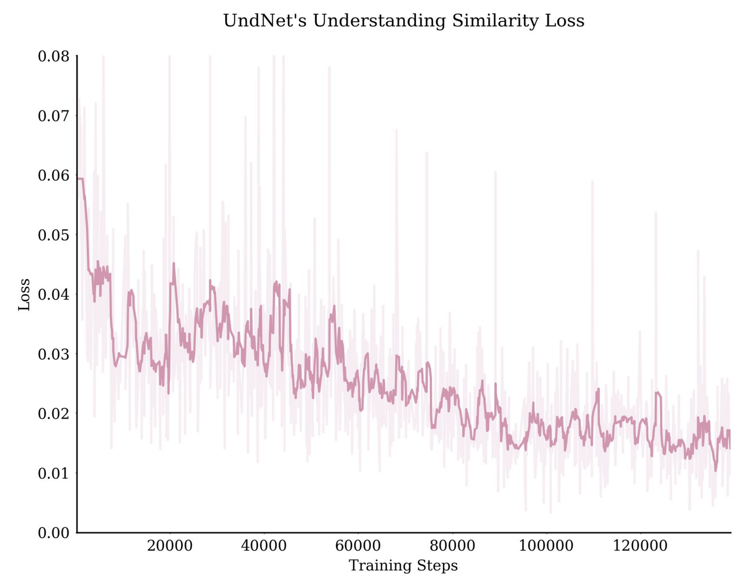
$$\mathcal{L} = \mathcal{L}_S + \lambda_u \mathcal{L}_u$$

「Experiments & Results」

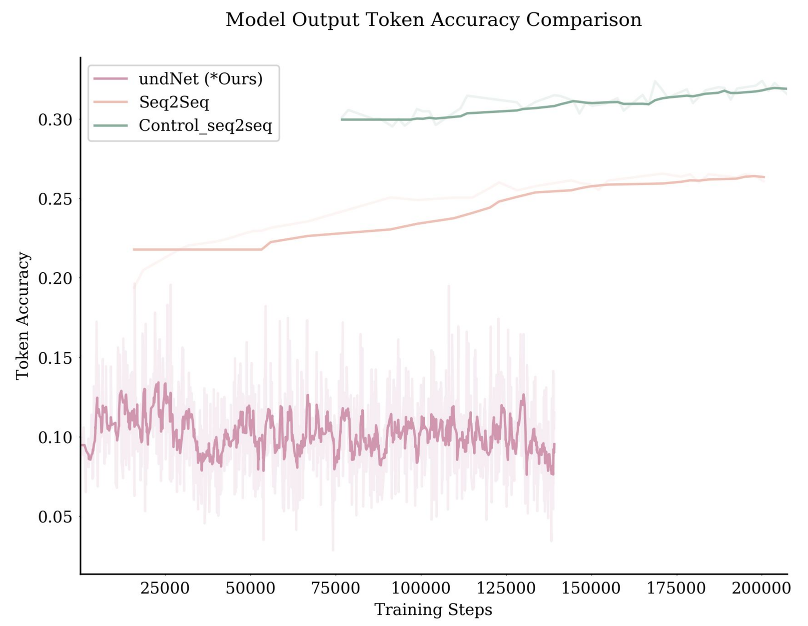
The UndNet shown in the experiments below is

- initialized with Seq2Seq ConvAI2 model weights
- the understanding units are initialized using persona description in each dialog
- Encoders and decoders of A and B are trained simultaneously in each train step





△ The continuous decreasing of understanding loss may show excessive training in forcing producing the same understanding between A and B, therefore leading to the fluctuation of the language modeling loss



△ The performance of UndNet is compared against baseline seq2seq model used in ConvAI2 competition and Controllable Dialogue Model [3].

- The frequent fluctuation of UndNet's token accuracy results from the understanding unites' re-initialization
- The model underperformance may due to the dominating understanding similarity loss and how to take advantage of its power is our next step

「References」

[1] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. IEEE Computational intelligence magazine, 13(3):55–75, 2018.

[2] Liqun Chen, Shuyang Dai, Chenyang Tao, Haichao Zhang, Zhe Gan, Dinghan Shen, Yizhe Zhang, Guoyin Wang,Ruiyi Zhang, and Lawrence Carin. Adversarial text generation via feature-mover's distance. In Advances in Neural Information Processing Systems, pages 4666–4677, 2018

[3] Abigail See, Stephen Roller, Douwe Kiela, Jason Weston. What makes a good conversation? How controllable attributes affect human judgments. NAACL 2019.