End-to-end dialogue systems

Milica Gašić

Dialogue Systems and Machine Learning Group,
Heinrich Heine University Düsseldorf

End-to-end models

Chatbots

Disadvantages of modular approach

- Each module necessitates labeled data:
  - ASR transcriptions
  - Semantic decoding labels
  - Dialogue act specification and rewards
  - NLG labels
  - TTS labels
- The abundance of data from chatting platforms and/or human-human speech cannot be used in this set-up.
- Defining labeling scheme and performing labeling is a very costly and time-consuming process.
- Unsupervised and semi-supervised learning is very valuable in this respect, but typically not as accurate as supervised learning.

End-to-end modelling

- Deep learning has made a revolution across the AI spectrum: computer vision, speech, NLP, ...
- It learns from huge amounts of data
- Traditional models require careful feature engineering and intermediate labels
- Deep learning uses raw features directly.

Advantages from learning from raw input

- Removes the need for defining features.
- Removes the need for labeling.
- Has the potential to extract better features - the ones that really aid learning and not the ones for which a human thinks aid learning.

End-to-end dialogue modelling

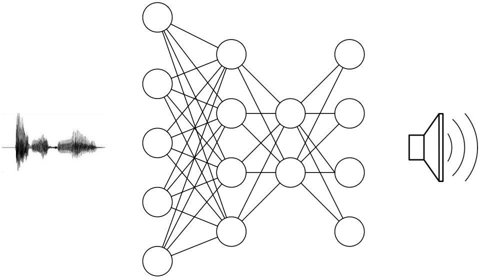
- Human brain takes speech as input and produces speech as output
- If we see human brain as a giant neural network, can we build a dialogue system as an end-to-end neural network without explicit intermediate modules?

Human brain vs artificial neural network

- Neurons have a much more complicated structure than neural networks building blocks.
- The way electric signals are passed through is different to gradient descent.
- We also know that different parts of the brain are responsible for different tasks, eg. language, emotions etc.
- Still, it is the best learning system we know and we would like to draw inspiration from it.

End-to-end neural network-based dialogue systems

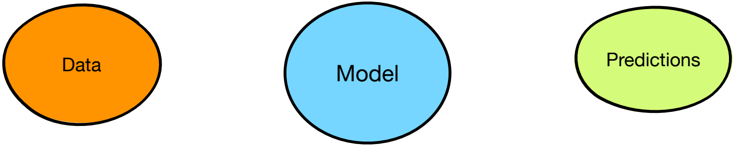
- It is possible to build each component of a dialogue system using a neural network
- Is it possible to build a dialogue system which is one giant neural network trained end-to-end?
- In theory we can simply propagate gradients.



End-to-end dialogue modelling

- To date there are still no attempts to build end-to-end speech dialogue system although there is remarkable success with end-to-end speech recognition and synthesis.
- Still end-to-end text dialogue modelling is a very active area of research

End-to-end neural network-based dialogue systems



- Dialogues: system and user utterances
- Dialogue rewards

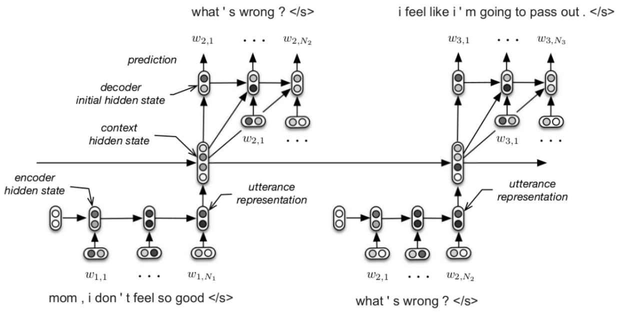
- Sequence-to-sequence learning model
- Deep reinforcement learning model

- System responses

Chatbots

- End-to-end modelling has first been applied to chatbots.
- These are systems that are not necessarily goal-driven but rather used for chit-chat and entertainment.
- The main reason is the sheer availability of data.
- In their development virtually no dialogue theory is applied, everything is learned from data.

Hierarchical Recurrent Encoder-Decoder for dialogue [Serban et al., 2015]



Hierarchical Recurrent Encoder-Decoder for dialogue

encoder RNN maps each utterance to an utterance vector

context RNN keeps track of past utterances by processing iteratively each utterance vector; essentially maps dialogue turns into a dialogue vector

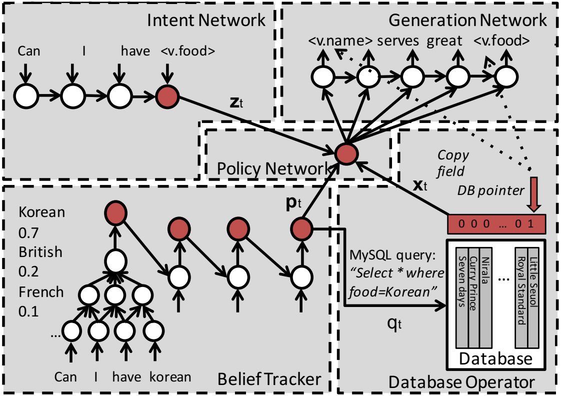
decoder RNN takes the hidden state of the context RNN and produces a probability distribution over the tokens in the next utterance

This model can be pre-initialised using a data set of a similar structure but not necessarily dialogue (eg QA). Also, the words can be represented as pretrained word embeddings.

Memory networks for end-to-end goal oriented dialogue [Bordes et al., 2017]

- By first writing and then iteratively reading from a memory component (using hops) that can store historical dialogues and short-term context to reason about the required response, they have been shown to perform well on those tasks

Seq2Seq model with additional supervision [Wen et al., 2017]



- Belief tracker trained separately
- Intent network and generation network trained end-to-end using the supervision signal from the belief tracker and the database

Seq2Seq model with additional supervision [Wen et al., 2017]

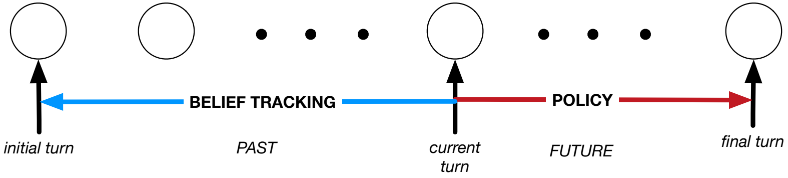
- Strictly speaking this model is not end-to-end!
- The reason is that it still necessitates intermediate labels for training the belief tracker.
- It is end-to-end trainable: everything is differentiable and the gradient can be propagated.
- This is an important property as it means that information of one part of the network can inform another part of the network.
- This is not normally the case in modular approaches.

Mem2Seq end-to-end model [Madotto et al., 2018]

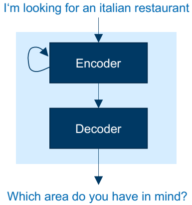
- The model augments the existing MemNN framework with a sequential generative architecture, using global multihop attention mechanisms to copy words directly from dialogue history or KBs.
- Combines multi-hop attention mechanisms with the idea of pointer networks, which allows us to effectively incorporate KB information.

What are all these models missing?

Core properties of goal-oriented dialogue

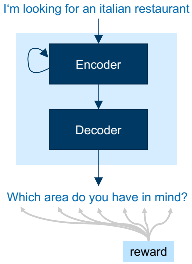


Most end-to-end dialogue models do not incorporate RL



- RL is essential for ensuring goal directed behaviour
- Without RL the models only imitate what they see in data, they do not perform any planning.

Word-level RL for end-to-end models



- Each word is treated as an action
- Huge action space
- Long trajectory
- Optimising language coherence and reward at the same time can lead to divergence

Theory: Variational autoencoder

- Autoencoders encode the input into lower-dimensional latent features
- These features should allow reconstruction of the input
- However, mapping between input and features is deterministic
- Can we modify the model such that we can generate more data from it?
- **Instead of deterministic mapping, VAE models the distribution of the latent variable**

Theory: Variational autoencoder - latent variabe

- We assume there is a variable that governs the generation of the output.
- This could be intent or an image type.
- We try to capture its distribution.
- We do not have labels for this variable therefore it is latent (hidden).

Theory: Variational autoencoder

Input x and latent variable z

recognition network Encoder maps input x to a distribution $q_\phi(z|x)$

generation network Decoder generates new data conditioned on z $p_\theta(x|z)$

Distribution of latent variable z

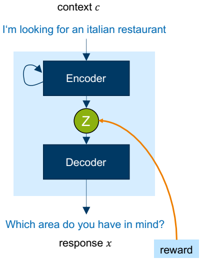
- True posterior $p_{\theta'}(z|x)$ is not known
- Prior $p_{\theta''}(z)$ initial assumption of how z is distributed

VAE loss function: evidence lower bound (ELBO)

$$
\begin{aligned}
\log p(x) &= \mathbb{E}_{z \sim q_\phi(z|x)} \log p_{\theta''}(x) \\
&= \mathbb{E}_z \log \frac{p_\theta(x|z) p_{\theta''}(z)}{p_{\theta'}(z|x)} \\
&= \mathbb{E}_z \log \frac{p_\theta(x|z) p_{\theta''}(z)}{p_{\theta'}(z|x)} \frac{q_\phi(z|x)}{q_\phi(z|x)} \\
&= \mathbb{E}_z \log p_\theta(x|z) + \mathbb{E}_z \log \frac{p_{\theta''}(z)}{q_\phi(z|x)} + \mathbb{E}_z \log \frac{q_\phi(z|x)}{p_{\theta'}(z|x)} \\
&= \mathbb{E}_z \log p_\theta(x|z) - \mathbb{E}_z \log \frac{q_\phi(z|x)}{p_{\theta''}(z)} + \mathbb{E}_z \log \frac{q_\phi(z|x)}{p_{\theta'}(z|x)} \\
&= \mathbb{E}_z \log p_\theta(x|z) - \mathbb{KL}(q_\phi(z|x)||p_{\theta''}(z)) + \mathbb{KL}(q_\phi(z|x)||p_{\theta'}(z|x)) \\
&\geq \mathbb{E}_z \log p_\theta(x|z) - \mathbb{KL}(q_\phi(z|x)||p_{\theta''}(z))
\end{aligned}
$$

If we maximize the right hand side we maximize the left hand side too.

Latent action RL in end-to-end dialogue systems [Zhao et al., 2019]



- Train a variational model to infer a latent space between encoder and decoder to serve as the action space
- x is the response for a given context c
- Modified evidence lowerbound (ELBO), i.e. lite ELBO avoids distribution mismatch between training and testing, since x is not present during testing

$$L_{\text{full}}(\theta) = \mathbb{E}_{q_\theta(z|x,c)}[\log p_\theta(x|z)] - \mathbb{KL}(q_\theta(z|x,c)||p_\theta(z|c))$$

$$L_{\text{lite}}(\theta) = \mathbb{E}_{p_\theta(z|c)}[\log p_\theta(x|z)] - \beta\mathbb{KL}(p_\theta(z|c)||p(z))$$

Latent action RL in end-to-end dialogue systems

Benefits:

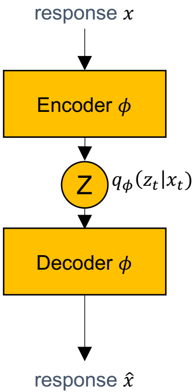
- Shortening the dialogue trajectory
- Decouples decision making and language generation

Latent action RL in end-to-end dialogue systems

Shortcomings:

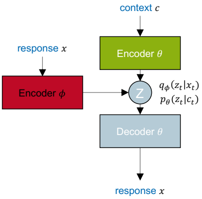
- Optimises latent space with an uninformed prior
- Does not consider the distributions w.r.t. dialogue responses
- Latent space is modelled conditioned on the context only
- Unclear whether the variables effectively encode action information

LAVA: Latent Action Space via VAE [Lubis et al., 2020]



- VAE as pre-training
- Auto-encode dialogue responses
- VAE infers the distribution of the latent variables to be used to reconstruct the response
- Captures underlying generative factors of responses
- In a modular approach this is what a dialogue act would do
- Here we let the model find out what are possible dialogue acts

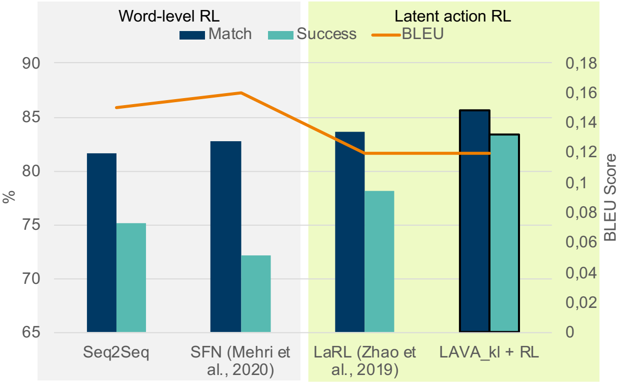
LAVA: Latent Action Space via VAE



- Use VAE and RG encoders in tandem during fine-tuning
- Newly initialized RG encoder
- Pre-trained VAE encoder to obtain an informed prior
- Optimise using informed prior

$$L_{\text{LAVA\_kl}}(\theta) = \mathbb{E}_{p_\theta(z|c)}[\log p_\theta(x|z)] - \beta \mathbb{KL}(p_\theta(z|c)||q_\phi(z|x))$$

Results



Word-level RL
Latent action RL
Match
Success
BLEU
90
85
80
75
70
65
%
0,18
0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
BLEU Score
Seq2Seq
SFN (Mehri et al., 2020)
LaRL (Zhao et al., 2019)
LAVA_kl + RL

Shortcomings of end-to-end approaches

- Only corpus based evaluation
- Utilises delexicalisation
- Best performing systems still utilise dialogue state information

More shortcomings of end-to-end approaches

- Lack of interpretability is the main problem of these approaches.
- In fact this is already a problem in statistical modular approaches.
- One cannot place guarantees on how the system will perform in each case.
- In end-to-end approaches this is further exacerbated: when the system fails there is almost no way of saying what caused it to fail.
- Interpretability and accountability are important considerations for machine learning.

Bias and ethics when learning from data

- All models that we presented learn from data.
- The less human intervention there is the more they will be governed from what is in the data.
- This means that there is no curating going on, if there is abusive or non-ethical behaviour exhibited in the data, the model will imitate it.
- This is exacerbated in end-to-end models as there is little opportunity to inspect what is happening inside the model.

Interaction

- A lot of advances have been made recently in terms of end-to-end learning.
- Still, due to all the shortcomings the use of end-to-end dialogue models is very limited.
- They are typically evaluated on measures such as BLEU.
- Almost no models have so far been tested in interaction with real users.

Summary

- Advances in deep learning enabled tackling dialogue as an end-to-end learning task.
- Early models treated dialogue as a purely supervised learning task.
- It is non-trivial to include RL in end-to-end models.
- Including RL achieves best success and match rates.

References I

Bordes, A., Boureau, Y.-L., and Weston, J. (2017).
Learning end-to-end goal-oriented dialog.
In *5th International Conference on Learning Representations*.

Lubis, N., Geishauser, C., Heck, M., Lin, H.-c., Moresi, M., van Niekerk, C., and Gasic, M. (2020).
LAVA: Latent action spaces via variational auto-encoding for dialogue policy optimization.
In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 465–479, Barcelona, Spain (Online). International Committee on Computational Linguistics.

References II

Madotto, A., Wu, C.-S., and Fung, P. (2018).
Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2015).
Hierarchical neural network generative models for movie dialogues.
arXiv preprint arXiv:1507.04808.

References III

Wen, T.-H., Vandyke, D., Mrkšić, N., Gašić, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2017).
A network-based end-to-end trainable task-oriented dialogue system.
In *EACL*.

Zhao, T., Xie, K., and Eskenazi, M. (2019).
Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1208–1218.

Credits

We thank Nurul Lubis for sharing her slides on Variational Autoencoders and LAVA.