Automatic design of conversational models from observation of human-to-human conversation - proposal

Contact: Petr Schwarz, Brno University of Technology, schwarzp@fit.vutbr.cz

Currently used conversation models (or dialog models) are mostly hand designed by data analysts as a conversation graph consisting of the system's prompts and the user's answers. The advanced conversation models [1, 2] are based on large language models fine-tuned on the dialog task, and still require significant amounts of training data. These models produce surprisingly fluent outputs but are not trustable because of hallucination (which can produce unexpected and wrong answers), and their adoption in commerce is limited. Our goal is to explore ways to design conversation models in the form of finite state graphs[1] semi-automatically or fully automatically from an unlabeled set of audio or textual training dialogs. Words, phrases, or user turns can be converted to embeddings using (large) language models trained specifically on conversational data [3, 4]. These embeddings represent points in a vector space and carry semantic information. The conversations are trajectories in the vector space. By merging, pruning, and modeling the trajectories, we can get dialog model skeleton models. These models could be used for fast data content exploration, content visualization, topic detection, and topic-based clustering, speech analysis, and mainly for much faster and cheaper design of fully trustable conversation models for commercial dialog agents. The models can also target some specific dialog strategies – the fastest way to reach a conversation goal (to provide useful information or sell a good or entertain users for the longest time). One promising approach to building a conversational model from data is presented in [4]. Variational Recurrent Neural Networks are trained to get discrete embeddings with a categorical distribution. The categories are conversation states. Then a transition probability matrix among states is calculated, and low probabilities are pruned out to get a graph.

Relation to the previous JSALT / ESPERANTO workshop 2022

Semantically aligned speech representations (extracted from SAMU-XLSR) were the core component in the speech-to-speech translation system effort exploiting multi-modal and multilingual data. The semantically aligned speech representations were also evaluated on spoken language understanding [7], showing the capabilities of speech representation models. However, these representations are obtained for each utterance independently and therefore not optimal. Bringing conversation context, for example, through some kind of Recurrent Neural Networks like in [3, 4], may be beneficial.

Research questions

1) What are the best embeddings to capture semantic information? How to ensure that embeddings will not overfit the data?
2) How to incorporate conversation context into the embeddings?
3) What is the best temporal resolution to capture the conversation? Words, phrases, sentences, the speaker turns … ?

[1] Models represented by finite state graphs are widely used in current commercial conversational systems. If these models are too restrictive, a mixture of continuous space and symbolic models that enable model verification by humans can be used.

4) How to merge trajectories (sequences of embeddings) to some data flows (conversation models)?
5) How to prune trajectories to remove outliers?
6) Can the outliers be important for some applications - to explore new info not known to users before?
7) Can the embeddings be language-independent?
8) How to build models usable for commercial conversation/dialog systems?
9) How do the embeddings differ for textual and audio inputs? Do we need to handle these inputs separately or is an automatic speech-to-text system enough? Or can we use multi-modal embeddings (audio + text)?
10) How to design some essential algorithms (the shortest path through the conversation)?
11) How to use the model for topic detection?
12) How to use the model for content exploration from unknown data?
13) How to visualize the content - projection of the trajectories to 2D?
14) How to detect known sub-dialogs? For example, to extend a library of hand-crafted conversation models.
15) How to present dialog states to humans? Though some characteristic sentences selected from training data? Using natural language generation?

Possible research directions

- Monolingual systems to multilingual systems
- Task-oriented conversations to open domain conversations
- From the development of individual components (ASR, dialog manager …) to joint training
- From purely data-driven approaches to grounded approaches (verified by humans)

Areas impacted by this research

- Contact centers (automatic dialog agents)
- Meeting analysis and minuting
- Increasing understanding and trust of AI technologies by business users
- Social chat-bots

Data

Although companies have access to goal-oriented speech dialog data from their clients often, publicly available and close to real research corpora is still a challenge. In the case of textual data, the MultiWOZ corpus is a common choice [8]. We plan to use it, and Paweł Budzianowski, one of its authors, expressed interest in joining this proposal. For audio, there is a recent effort led by Izhak Shafran and his team from Google, who extended MultiWOZ with an audio part and offered it to the research community. It is done through speech synthesis, reading, and paraphrasing. The data was used DSTC11 Challenge [9], Speech-Aware Dialog Systems Technology Challenge Track, which ended in November. Both data and the challenge are relevant to this proposal. Also, Izhak is open to being an external advisor for the team. In addition, Charles University and Brno University of Technology would like to collect a few “real” voice dialogs similar to MultiWOZ. We created a project for this supported by the European Humane AI project [10]. Then classical LDC data like Fisher, Swichboard, and also the People’s speech data set may be used.

People who expressed interest

- Petr Schwarz, Santosh Kesiraju, Lukas Burget, *Brno University of Technology, Czechia*
- Themos Stafylakis, *Omilia - Conversational Intelligence, Greece*
- Ondřej Dušek, Ondřej Bojar, Ondřej Plátek, *Charles University, Czechia*
- Paweł Budzianowski, Ivan Vulic, *PolyAI / University of Cambridge, United Kingdom*
- Petr Motlicek, *IDIAP, Switzerland*
- Miroslav Hlaváček, Tomáš Pavlíček, *Phonexia, Czechia*

References

[1] Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao: “Soloist: Building task bots at scale with transfer learning and machine teaching”, *Transactions of the Association for Computational Linguistics*, 9:807–824, https://doi.org/10.1162/tacl_a_00399

[2] Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, Ondřej Dušek: “AuGPT: Auxiliary Tasks and Data Augmentation for End-To-End Dialogue with Pre-Trained Language Models”, *3rd Workshop on Natural Language Processing for Conversational AI*, https://arxiv.org/abs/2102.05126

[3] Vojtěch Hudeček, Ondřej Dušek: “Learning Interpretable Latent Dialogue Actions With Less Supervision.”, *In Proceedings of AACL-IJCNLP, 2022 (to appear)*, https://doi.org/10.48550/arXiv.2209.11128

[4] Weiyan Shi, Tiancheng Zhao, and Zhou Yu: “Unsupervised Dialog Structure Learning”, *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 1797–1807, Minneapolis, Minnesota, Association for Computational Linguistics, 2019, https://aclanthology.org/N19-1178/

[5] Santosh Kesiraju, Oldřich Plchot, Lukáš Burget, and Suryakanth V. Gangashetty: “Learning Document Embeddings Along With Their Uncertainties.”, *IEEE/ACM Transactions on audio, speech and language processing*, vol. 2020, no. 28, pp. 2319-2332, ISSN 2329-9290 https://www.fit.vut.cz/research/publication/12343/

[6] Javier Miguel Sastre Martinez and Aisling Nugent: “Inferring Ranked Dialog Flows from Human-to-Human Conversations”, In *SigDial 2022*, https://www.youtube.com/watch?v=Fn8syqjZFCA

[7] G. Laperriere, V. Pelloin, M. Rouvier, T. Stafylakis, and Y. Estève: “On the Use of Semantically-Aligned Speech Representations for Spoken Language Understanding”, https://arxiv.org/abs/2210.05291 *(to appear in IEEE SLT 2022)*

[8] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, Milica Gašić: “MultiWOZ -- A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling”, https://arxiv.org/abs/1810.00278

[9] The Eleventh Dialog System Technology Challenge, https://dstc11.dstc.community/

[10] Humane AI – European Network of Human-centered Artificial Intelligence, https://www.humane-ai.eu/