Improving open domain Conversational Dialogue systems(ChatBots)

* **Nitin Kishore Sai Samala (30168428)**

# Abstract

Conversational Agents or Dialog Systems, are a hot topic.Many companies are hoping to develop bots to have natural conversations indistinguishable from human ones, and are using NLP and Deep Learning techniques to make this possible. They start off by outsourcing their conversations to human workers and promise that they can “automate” it once they’ve collected enough data. Retrieval-based models use a repository of predefined responses and a heuristic to pick an appropriate response based on the input and context while Generative models create new responses from scratch. A retrieval-based open domain system is obviously impossible because you can never handcraft enough responses to cover all cases. A generative open-domain system is almost Artificial General Intelligence (AGI) because it needs to handle all possible scenarios. We’re very far away from that as well. There are few challenges, most of which being active research areas, when building conversational agents like

* **Not incorporating Context**, or basically keeping track of what has been said and what information has been exchanged to produce more sensible responses
* **Coherent Personality/Speaker Inconsistency**, since systems learn to generate linguistic plausible responses, but aren’t trained to generate semantically consistent answers to identical inputs
* **Lack of Intention and Diversity**, resulting in generic responses like “That’s great!” or “I don’t know” that work for a lot of input cases

This leaves us with problems in restricted domains where both generative and retrieval based methods are appropriate. The longer the conversations and the more important the context, the more difficult the problem becomes. In this project I aim to survey and draw a contrast between retrieval and generative models for short conversations in both open and closed domains.

# Existing model and improvements

Deep Learning architectures like Sequence to Sequence are uniquely suited for generating text and researchers are hoping to make rapid progress in this area. However, we’re still in a nascent phase of building generative models that work reasonably well. Production systems are more likely to be retrieval-based for now. To tackle the challenges mentioned above, building end-to-end dialogue systems using Generative Hierarchical Neural Network Models and using attention with intention seems to improve the context related issues. These are usually trained on a lot of data from multiple different users, so explicitly modeling a personality would reduce the inconsistency in the outputs. To prevent generic responses, artificially promoting diversity through various objective functions can be pursued.

Naturally each of these tasks would have a compatible dataset and models. For this project, keeping the timeframe in mind I wish to explore and survey different solutions and attempt implementing some of these ideas in PyTorch and TensorFlow and combine them to improve and add to existing neural conversational models for open domain dialogue.

# Datasets

I have access to the Weibo Dataset and I intend to use ParlAI for bAbI Dialog tasks, Ubuntu Dialog, Open Subtitles, Cornell Movie datasets. I will choose a final dataset for the combined model after exploring the individual models with these.

# Evaluation

The ideal way to evaluate a conversational agent is to measure whether it is fulfilling its task, e.g. solve a customer support problem, in each conversation but such labels are expensive to obtain because they require human judgment and evaluation. Sometimes there is no well-defined goal, as is the case with open-domain models. Common metrics such as BLEU that are used for Machine Translation and are based on text matching aren’t well suited because sensible responses can contain completely different words or phrases. Researchers find that none of the commonly used metrics really correlate with human judgment. If necessary, I would prefer to conduct a human evaluation using crowdsourced annotators or have few volunteers grade responses to specific set of inputs.

# References

[1] Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, Joelle Pineau, **How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation,**

[2] Ryan Lowe, Nissan Pow, Iulian Serban, Joelle Pineau, **The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems**

[3] Rudolf Kadlec, Martin Schmid, Jan Kleindienst, **Improved Deep Learning Baselines for Ubuntu Corpus Dialogs**

[4] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, Bill Dolan, **A Persona-Based Neural Conversation Model**

[5] Oriol Vinyals, Quoc Le, **A Neural Conversational Model**

[6] Kaisheng Yao, Geoffrey Zweig, Baolin Peng, **Attention with Intention for a Neural Network Conversation Model**

[7] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, Bill Dolan, **A Diversity-Promoting Objective Function for Neural Conversation Models**

[8] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, Joelle Pineau, **Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models**