Analysing methods of Neural Text Generation to refine conversations

Video Link- https://www.youtube.com/watch?v=H_P9B3YIIdI

CSE-4022- Natural Language Processing

Fall Semester 2021-22
Slot – B2+TB2

Under the guidance of,

Ilakiyaselvan N

Assistant Professor (Senior)

SCOPE

TEAM MEMBERS:

Mihir Antwal (19BCE1641)

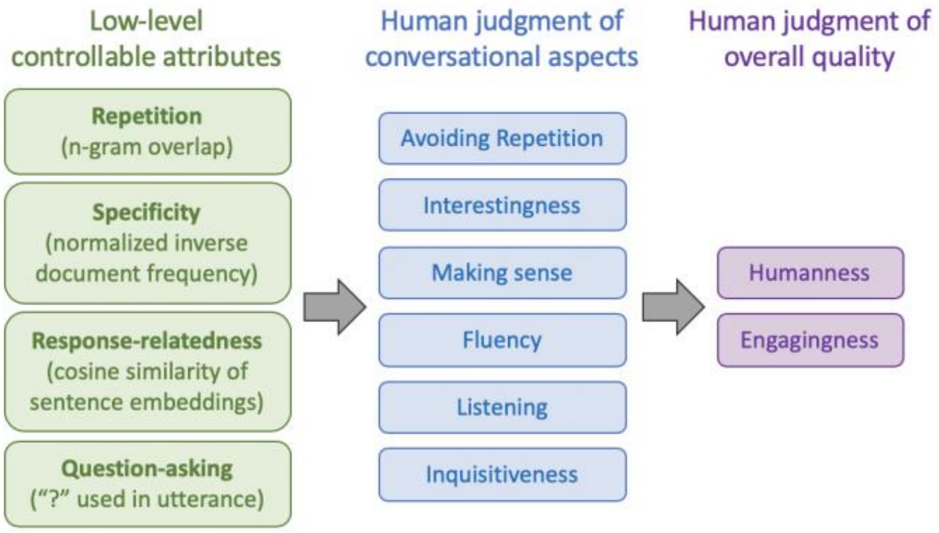
Sam Methuselah (19BCE1698)

ABSTRACT

A good conversation requires balance – between simplicity and detail; staying on topic and changing it; asking questions and answering them. Although dialogue agents are commonly evaluated via human judgments of overall quality, the relationship between quality and these individual factors is less well-studied. In this work, we examine two controllable neural text generation methods, conditional training and weighted decoding, in order to control four important attributes for chitchat dialogue: repetition, specificity, response-relatedness and question-asking. We conduct a large-scale human evaluation to measure the effect of these control parameters on multi-turn interactive conversations on the Personal Chat task. We provide a detailed analysis of their relationship to high-level aspects of conversation, and show that by controlling combinations of these variables our models obtain clear improvements in human quality judgments.

INTRODUCTION

Neural generation models for dialogue, despite their ubiquity in current research, are still poorly understood. Well known problems, such as the genericness and repetitiveness of responses, remain without a solution. Strikingly, the factors that determine human judgments of overall conversation quality are almost entirely unexplored. Most works have been limited to the next utterance prediction problem, whereas a multi-turn evaluation is necessary to evaluate the quality of a full conversation. In this work we both (i) conduct a large-scale study to identify the fine-grained factors governing human judgments of full conversations, and (ii) develop models that apply our findings in practice, leading to state-of-the-art performance. Specifically, we identify and study eight aspects of conversation that can be measured by human judgments, while varying four types of low-level attributes that can be algorithmically controlled in neural models; see the figure. To control the low-level model attributes, we consider two simple but general algorithms: conditional training, in which the neural model is conditioned on additional control features, and weighted decoding, in which control features are added to the decoding scoring function at test time only.


Low-level controllable attributes
Human judgment of conversational aspects
Human judgment of overall quality
Repetition (n-gram overlap)
Specificity (normalized inverse document frequency)
Response-relatedness (cosine similarity of sentence embeddings)
Question-asking ("?" used in utterance)
Avoiding Repetition
Interestingness
Making sense
Fluency
Listening
Inquisitiveness
Humanness
Engagingness

RELATED WORK

Dialogue

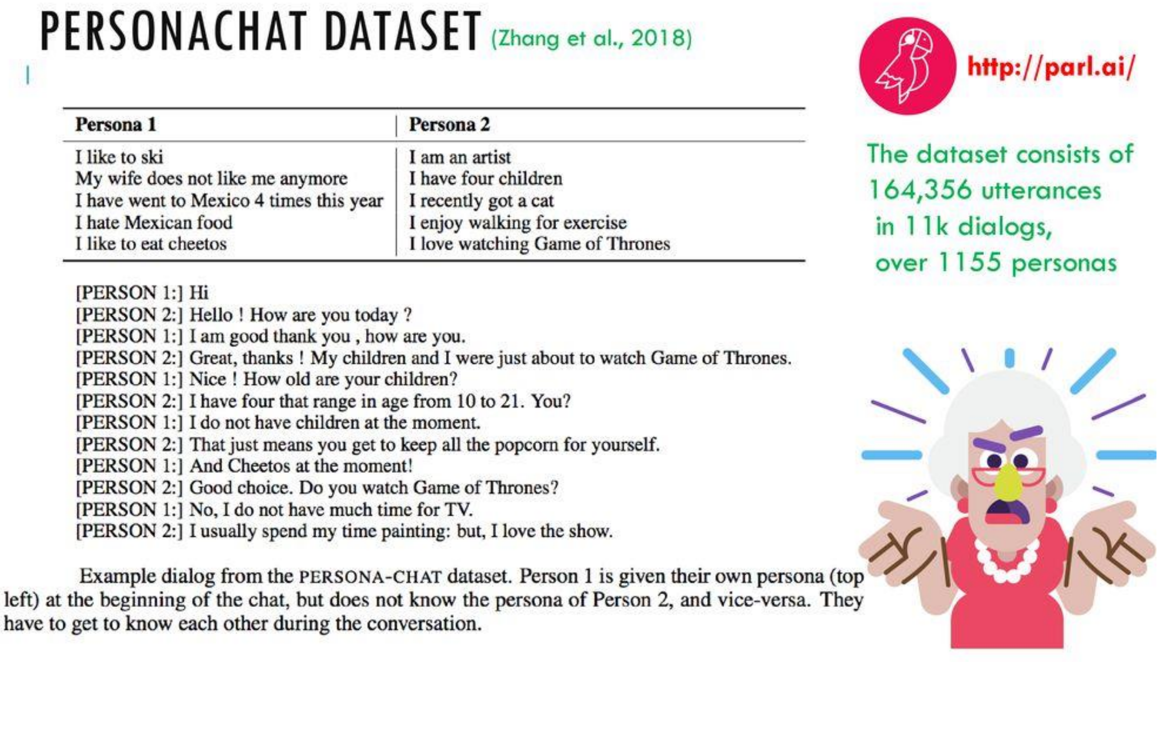
Dialogue evaluation is relatively well understood in goal-oriented tasks, where automated approaches can be coded by measuring task completion. Task success combined with dialogue cost can be linked to human judgments like user satisfaction via the PARADISE framework. However, in chitchat tasks, which we study in this work, automatic metrics and their relation to human ratings are less well-understood. While word-overlap metrics are effective for question answering and machine translation, for dialogue they have little to no correlation with human judgments - this is due to the open-ended nature of dialogue. There are more recent attempts to find better automatic approaches, such as adversarial evaluation and learning a scoring model, but their value is still unclear. Nevertheless, a number of studies only use automatic metrics, with no human study at all. Other works do use human evaluations, typically reporting just one type of judgment (either quality or appropriateness) via a Likert scale or pairwise comparison. Most of those works only consider single turn evaluations, often with a shortened dialogue history, rather than full multi-turn dialogue. A more comprehensive evaluation strategy has been studied within the scope of the Alexa prize by combining multiple automatic metrics designed to capture various conversational aspects (engagement, coherence, domain coverage, conversational depth and topical diversity). Though these aspects have some similarity to the aspects studied here, we also focus on lower-level aspects (e.g. avoiding repetition, fluency), to understand how they correspond to both our controllable attributes, and to overall quality judgments.

Controllable neural text generation-

Researchers have proposed several approaches to control aspects of RNN-based natural language generation such as sentiment, length, speaker style and tense. In particular, several works use control to tackle the same common sequence-to-sequence problems we address here (particularly genericness and unrelated output), in the context of single-turn response generation. By contrast, we focus on developing controls for, and human evaluation of, multi-turn interactive dialogue – this includes a new method to control attributes at the dialogue level rather than the utterance level. In this work, we require a control method that is both general-purpose (one technique to simultaneously control many attributes) and easily tunable (the control setting is adjustable after training). Given these constraints, we study two control methods: conditional training and weighted decoding. To our knowledge, this work is the first to systematically compare the effectiveness of two general-purpose control methods across several attributes.

THE PERSONACHAT DATASET

PersonaChat is a chitchat dialogue task involving two participants (two humans or a human and a bot). Each participant is given a persona – a short collection of personal traits such as I'm left handed or My favourite season is spring – and are instructed to get to know each other by chatting naturally using their designated personas, for 6–8 turns. The training set contains 8939 conversations and 955 personas, collected via crowd workers, plus 1000 conversations and 100 personas for validation, and a similar number in the hidden test set. The PersonaChat task was the subject of the NeurIPS 2018 ConvAI2 Challenge, in which competitors were first evaluated with respect to automatic metrics, and then with respect to human judgment via the question "How much did you enjoy talking to this user?" on a scale of 1–4.



PERSONACHAT DATASET (Zhang et al., 2018)

http://parl.ai/

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

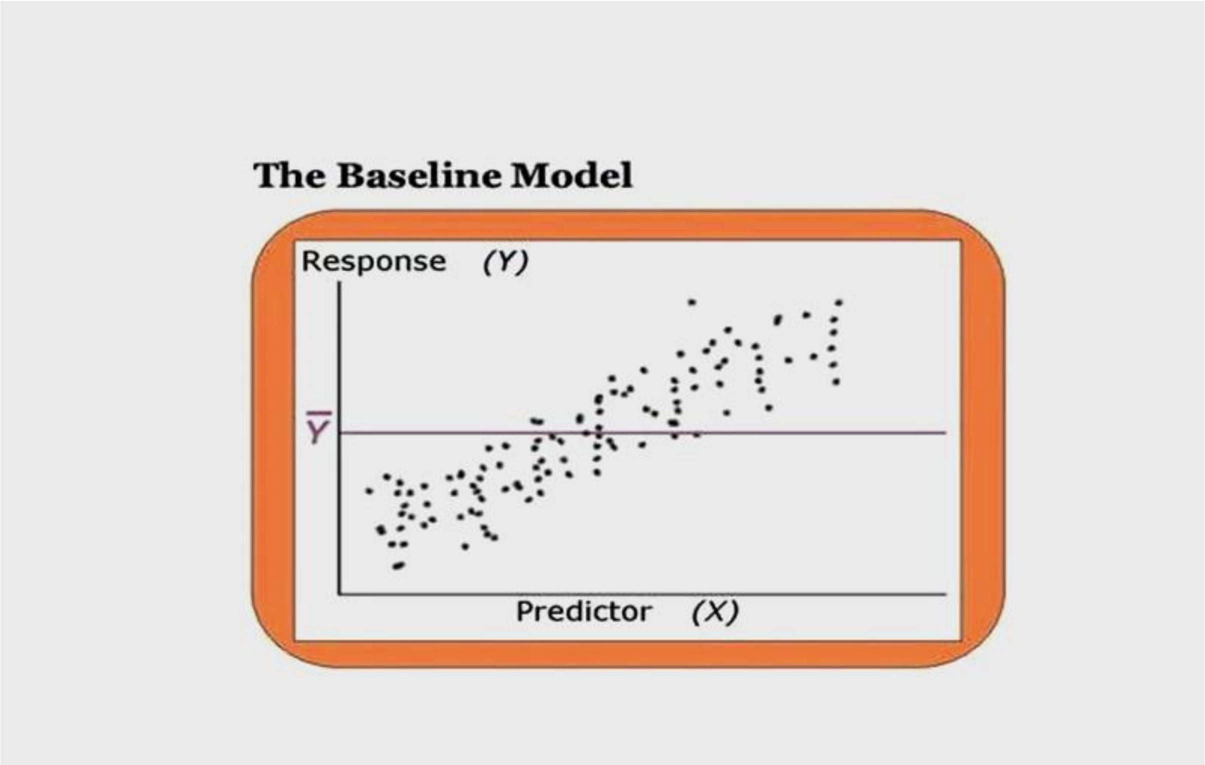
[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Example dialog from the PERSONA-CHAT dataset. Person 1 is given their own persona (top left) at the beginning of the chat, but does not know the persona of Person 2, and vice-versa. They have to get to know each other during the conversation.

The dataset consists of 164,356 utterances in 11k dialogs, over 1155 personas

BASELINE MODEL

Our baseline model is a 2-layer LSTM sequence-to-sequence model with attention. On any dialogue turn, the input x to the encoder is the entire dialogue history (separated using unique speaker-identifying tokens), with the model's own persona prepended. Conditioned on this input sequence x, the decoder generates a response y. Except when stated otherwise, all our models decode using beam search with beam size 20. We initialized the word embedding matrix with 300-dimensional GloVe embeddings. Using the ParlAI framework, we pretrained the model on a dataset of 2.5 million Twitter message-response pairs,1 then fine-tuned it on PersonaChat. On the PersonaChat validation set, the baseline model has a perplexity of 26.83 and F1 of 17.02, which would have placed us 4th out of 26 models in the ConvAI2 competition. We attempt to improve over this baseline using control.



CONTROLLABLE TEXT GENERATION METHODS

Suppose we have a sequence-to-sequence model which gives $P(y|x) = \Pi tP(yt |x, y1, . . . , yt-1)$, the conditional probability of a response y (the model's next utterance) given input x (the context, which in our case includes the model's own persona and the dialogue history). Contrary to most previous work, which controls at the sentence level, we wish to control attributes of the output y at the dialogue level – meaning that a single control setting is used for a whole dialogue. For example, to control question-asking, we provide a control setting at the beginning of each dialogue (e.g. 20% questions or 70% questions) rather than providing a control setting for each utterance (e.g. is a question or isn't a question). With this approach, the sequence-to-sequence model is able to choose what value the controlled attribute should take for any particular utterance, but we are able to choose the overall distribution. We find that this approach works well – for example, the sequence-to-sequence model is generally good at detecting when to ask a question. In particular, this is easier than the alternative: developing a separate process to decide, for each utterance, whether to ask a question. We use two methods – which we call Conditional Training (CT) and Weighted Decoding (WD) – to control attributes of the output y at the dialogue level.

- ## Conditional Training (CT)

Conditional Training is a method to learn a sequence-to-sequence model P(y|x, z), where z is a discrete control variable. If the control attribute is naturally continuous (for example in our work, repetitiveness, specificity and response-relatedness), we use z to represent bucketed ranges. For a binary attribute like question-asking, z represents an overall probability. To train a CT model, we first automatically annotate every (x, y) pair in the training set with the attribute we wish to control (for example, whether y contains a question mark). During training, for each example we determine the corresponding z value (for continuous attributes, this simply means sorting into the correct bucket; for question-asking). Next, the control variable z is represented via an embedding (each of the possible values of z has its own embedding). For all our experiments, the embedding is of length 10; this was determined via hyperparameter tuning. There are several possible ways to condition the sequence-to-sequence model on z – for example, append z to the end of the input sequence, or use z as the START symbol for the decoder. We find it most effective to concatenate z to the decoder’s input on every step. Lastly, the CT model learns to produce y = y1, . . . , yT by optimizing the cross-entropy loss:

$$\text{LossCT} = -\frac{1}{T}\sum_{t=1}^{T} \log \mathrm{P}(y_t \,|\mathrm{x}, \mathrm{z}, y_1, \ldots, y_{t-1})$$

Our CT models are initialized with the parameters from the baseline sequence-to-sequence model P(y|x') (the new decoder parameters are initialized with small random values), then fine-tuned to optimize lossCT on the PersonaChat training set, until convergence of lossCT on the validation set.

- **Weighted Decoding (WD)**

Weighted Decoding is a decoding method that increases or decreases the probability of words with certain features. The technique is applied only at test time, requiring no change to the training method. A limitation of WD is that the controllable attribute must be defined at the word-level; any desired utterance-level attribute must be redefined via word-level features. In weighted decoding, on the tth step of decoding, a partial hypothesis y<t = y1 ,……, yt-1 is expanded by computing the score for each possible next word w in the vocabulary:

$$\text{score(w , y<t ; x) = score(y<t ; x) + log PRNN(w|y<t , x) +}\sum_{i} \text{ wi * fi(w ; y<t , x)}$$

Here, log PRNN(w|y<t,x) is the log-probability of the word w calculated by the RNN, score(y<t ; x) is the accumulated score of the already-generated words in

the hypothesis y<t, and fi(w; y<t, x) are decoding features with associated weights wi . There can be multiple features fi (to control multiple attributes), and the weights wi are hyperparameters to be chosen. A decoding feature fi(w; y<t , x) assigns a real value to the word w, in the context of the text generated so far y<t and the context x. The feature can be continuous (e.g. the unigram probability of w), discrete (e.g. the length of w in characters), or binary (e.g. whether w starts with the same letter as the last word in y<t). A positive weight wi increases the probability of words w that score highly with respect to fi ; a negative weight decreases their probability. Note that weighted decoding and conditional training can be applied simultaneously (i.e. train a CT model then apply WD at test time) – a strategy we use in our experiments.

REFERENCES

Websites-

www.google.com

www.youtube.com

https://parl.ai/

www.github.com

Books-

1) Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering output style and topic in neural response generation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2140–2150. Association for Computational Linguistics.

2) Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Video Link- https://www.youtube.com/watch?v=H_P9B3YIIdI