Item Response Theory for Efficient Human Evaluation of Chatbots

João Sedoc
New York University
jsedoc@stern.nyu.edu

Lyle Ungar
University of Pennsylvania
ungar@cs.upenn.edu

Abstract

Conversational agent quality is currently assessed using human evaluation, and often requires an exorbitant number of comparisons to achieve statistical significance. In this paper, we introduce Item Response Theory (IRT) for chatbot evaluation, using a paired comparison in which annotators judge which system responds better to the next turn of a conversation. IRT is widely used in educational testing for simultaneously assessing the ability of test takers and the quality of test questions. It is similarly well suited for chatbot evaluation since it allows the assessment of both models and the prompts used to evaluate them. We use IRT to efficiently assess chatbots, and show that different examples from the evaluation set are better suited for comparing high-quality (nearer to human performance) than low-quality systems. Finally, we use IRT to reduce the number of evaluation examples assessed by human annotators while retaining discriminative power.

1 Introduction

One of the main problems in conversation dialog modeling is evaluation. Unlike in machine translation and task-driven dialog, automated metrics for non-task driven open-domain generative conversational models (chatbots) seem not to correlate well with human judgments (Liu et al., 2016; Tao et al., 2017; Lowe et al., 2017). While the creation of new automatic metrics is an extremely active area of research (Liu et al., 2016; Tao et al., 2017; Lowe et al., 2017; Novikova et al., 2017; Galley et al., 2015; Sugiyama et al., 2019), human annotations are currently the gold standard for assessing model improvements. Prior work mainly uses straightforward approaches, such as a two-sided t-test or binomial tests (e.g., (Serban et al., 2016; Li et al., 2015; Asghar et al., 2017; Ghazvininejad et al., 2018; Li et al., 2019b)), or pairwise bootstrap test (e.g. Baheti et al. (2018)). These methods do not assess or incorporate the effectiveness of prompts (conversational chunks used for evaluation). Given that human evaluation is necessary, it is desirable to discriminate the performance of two different systems with minimal cost.

		Annotator			Net
		1	2	3	Rating
prompt	*are you a leader or a follower?*				
sys A	leader	-	-	-	0
sys B	follower				(tie)
prompt	*do you run faster if someone chases you?*				
sys A	i don't run	A	-		0
sys B	no, i don't.			B	(tie)
prompt	*what do you think about creativity?*				
sys A	i think that's a very good quality.				3
sys B	i have a lot of ideas and a deal.	B	B	B	
prompt	*tell me something about your house?*				
sys A	my house is bigger than yours.	A	A	A	-3
sys B	i'm not sure				
prompt	*what is immoral?*				
sys A	the fact that you have a child.	-	-		1
sys B	it is a sin.			B	
prompt	*Are you a robot?*				
sys A	and you are?				3
sys B	i am a robot	B	B	B	
prompt	*spiderman*				
sys A	i'm not batman	A			1
sys B	spiderman spiderman		B	B	

Table 1: Comparison of two system responses and aggregate of three human annotator ratings. For each prompt Net Rating $= \sum_k^{annotators}(w_k^{sysB} - w_k^{sysA})$ where w_i^{sysA} is 1 if annotator i rated system A better and 0 otherwise, and similarly for system B. "-" indicates a tie vote.

In this paper, we present the use of Item Response Theory (IRT) (Lord et al., 1968) to compare chatbot models using a head-to-head paired experimental (A/B test) design (e.g. Table 1), which allows for statistical significance testing and item importance identification. IRT is traditionally used to assess student "ability" based on their answers ('responses') to test questions ('items') and, simultaneously, to determine how informative each question is. Throughout this paper we use the analogy of **student** $\sim$ **A/B chatbot comparison** and **question** $\sim$ **prompt**. We apply IRT

Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP), pages 21–33,
November 20, 2020.

to assess chatbot model performance based on human evaluations of chatbot responses to prompts, while simultaneously assessing how informative each prompt is.

IRT is a latent variable Bayesian model, with relative chatbot model quality (or student ability) being latent variables that probabilistically produce observable responses (one chatbot response to a prompt being judged as better than another, or a student answering a question correctly or wrong). IRT is widely used in psychometric studies (Embretson and Reise, 2013), and for paired comparison in psychological studies (Maydeu-Olivares and Brown, 2010). However, it is almost entirely ignored in natural language processing (NLP), with the exception of Hopkins and May (2013); Lalor et al. (2016); Otani et al. (2016); Lalor et al. (2019); Dras (2015).

Recent work has criticized the statistical methodology used in NLP and called for use of better statistical methods (Dror et al., 2018). Here, we present IRT as a powerful method for statistical assessment of model improvements. IRT not only **assesses the relative quality between two systems**, but also **assesses the usefulness of a prompt** in comparing systems. We show that IRT can filter and choose a subset of prompts from the evaluation set efficiently, i.e. with little loss in statistical power (Figure 2), and that IRT finds different prompts to be useful for assessing high quality vs. low quality chatbots.

Our core contribution is showing how Item Response Theory (IRT) can be used for open-domain social conversational agent (chatbot) comparison. In particular, we showcase the use of IRT in comparing multiple models for neural conversational agents. Finally, we show the utility of IRT for reducing the data collection required to evaluate chatbots by filtering evaluation set prompts. To our knowledge, this is the first work to apply IRT to chatbot evaluation and to use IRT for prompt selection in the evaluation of NLP systems.

2 Related Work

The structure of our chatbot evaluation is a comparison of two chatbots responses to each prompt. This form of head-to-head pairwise block (multiple evaluations shown to one annotator) comparison dates back at least to Thurstone (1927). Subsequently, the Bradley-Terry (BT) model has become the most common model for pairwise block comparison experiments (Bradley and Terry, 1952). Dras (2015) describes further extensions and application of the BT model to machine translation. Extended BT models can correct for dependent categorical object covariates (correlated examples) as well as subject covariates (annotator ratings) (Cattelan, 2012). As Dras (2015) points out, the BT model and IRT are similar in formulation, but IRT additionally estimates the difficulty of each item using a latent variable Bayesian model. Fixed effect BT models (Borenstein et al., 2010) or bootstrapping (Koehn, 2012) could be used to compare chatbots, but IRT's ability to assess prompts is more attractive for this task where every annotation has a non-trivial cost.

An alternative straightforward approach to assess usefulness (validity) of a prompt is item-total correlation (ITC; Guilford (1953)). However, ITC does not take the student's ability into account. In general, IRT is preferred over ITC due to the more expressive formulation. ITC is mostly used for survey analysis instead of testing. However, as a sanity check, we find that indeed prompts extremely low in discriminative power (according to IRT) also have a low item-total correlation.

There is surprisingly little work on improving statistical significance testing or prompt selection in chatbot evaluation. While this is less true for machine translation, only two prior works have used IRT for model assessment (Hopkins and May, 2013; Otani et al., 2016). Our work applies IRT in a similar fashion as Otani et al. (2016), but to chatbot evaluation instead of machine translation system evaluation. We differ from Hopkins and May (2013) and Otani et al. (2016) as follows: 1. We do pairwise comparison instead of requiring baselines - this allows for improved prompt selection as models improve. Their method is focused on WMT (batch/competition) settings whereas our work focuses on perpetual evaluation. 2. We aggregate annotators - which creates much more stable predictions (their graded mean is 1-baseline, 2-tie, 3-win) whereas ours ranges from [-3,3]. 3. We explicitly assume independence of prompts and account for their correlation and thus do not overstate significance. 4. We use IRT to reduce the total number of comparisons; Otani et al. (2016) suggest this for future work.

IRT has also been applied in NLP for dataset filtering (Lalor et al., 2016). Lalor et al. (2019) uses IRT to efficiently subsample training data based on

the difficulty. We differ from Lalor et al. (2019) on prompt selection: 1. We select individual prompts based on evaluations using the discriminative ability of the prompt–not just the item difficulty. 2. We use model win rank instead of item difficulty for selecting prompts for "better" models. Both of these yield more informative prompts. Kulikov et al. (2018) use a Bayesian approach for testing for significance in interactive evaluation; however, the correlation between items is not taken into account. As in Otani et al. (2016), IRT allows us to directly compare distributions; however, the correlation between the prompts still needs to be accounted for in order not to overstate significance.

Machine Translation Much effort has been placed in machine translation for correlating human annotator judgements with automatic metrics; however, Lowe et al. (2017) showed that automatic machine translation evaluation methods do not correlate with human judgments of open-domain conversational agents. This may be due to the fact that in machine translation there is a one-to-one semantic equivalence between reference and system output, whereas this is not true in the chatbot setting. Nonetheless, relevant prior work on assessing human evaluation in machine translation is relevant to chatbot evaluation. In machine translation, shared tasks offer standard evaluation sets and workshops, which have yielded standardized results (Callison-Burch et al., 2007, 2011).

Since 2015, the Workshop on Machine Translation (WMT) uses TrueSkill (Herbrich et al., 2007) for model ranking. TrueSkill can also be applied to chatbot evaluation. Sakaguchi et al. (2014) used it to efficiently pair machine translation systems and compared them using random subsets of data. They show that their non-parametric method is empirically superior in accuracy to Hopkins and May (2013). However, this comparison is limited since the non-parametric might focus only on one axis of difference similar to stochastic gradient descent. Returning to our student analogy, in an example of students taking the SAT (an English *and* a Math test), the TrueSkill method might focus on only the Math portion to discriminate between students, whereas, IRT would use both portions. Trueskill does not select examples using item utility.

Otani et al. (2016) and Hopkins and May (2013) applied IRT to machine translation. IRT is more important in chatbot evaluation than in machine translation as human evaluation is rarely reported in machine translation papers (e.g. (Sutskever et al., 2014; Vaswani et al., 2017)), but is rarely omitted in chatbot comparison (e.g. Liu et al. (2016); Serban et al. (2016); Li et al. (2017); Baheti et al. (2018); Li et al. (2019b); Zhang et al. (2019); Adiwardana et al. (2020)). Comparison of conversational generative agents using next utterance generation is in many ways similar to the evaluation of machine translation (MT); however, differentiating between chatbot models is uniquely challenging; many more responses than translations are plausible. Automated evaluation of MT is vastly better than of chatbots (Liu et al., 2016). The higher costs of human evaluation strongly encourage the use of more powerful statistical models such as IRT.

3 Chatbot Evaluation

Recently researchers tend to evaluate their methodological improvements relative to a sequence-to-sequence (Seq2Seq) baseline (Sutskever et al., 2014), as proposed for utterance generation by Shang et al. (2015); Vinyals and Le (2015); Sordoni et al. (2015) as well to compare against each other. While crowd-sourcing experiments are relatively cheap, the lack of automatic metrics means that every change in model architecture requires new evaluations. Our goal is efficient and cost-effective model assessment. Ideally, chatbots would be interactively evaluated, but due to the high cost, next utterance simulation is used as a surrogate. Although next utterance generation is a more artificial task, Logacheva et al. (2018) observed a Pearson correlation of 0.6 between conversation-level and utterance-level ratings.

Human judgments are often inconsistent for non-task driven chatbots, since there is no clear objective, which leads to low inter-annotator agreement (IAA) (Sedoc et al., 2019; Yuwono et al., 2019). However, Amidei et al. (2019) point out that even with low IAA we can still find statistical significance. There are further tensions between local coherence assessments using standard evaluation sets and human interactive evaluation. These issues are exacerbated for non task-driven dialog systems, as there is rarely a single "correct" response, leading to more local minima. Thus, there is a need to obtain the maximum possible

statistical power at the minimal possible cost.

Novikova et al. (2018) found that relative rankings yield more discriminative results than absolute assessments when evaluating natural language generation. Recent work of Li et al. (2019a) introduce both human-bot as well as self-chat for interactive evaluation and show that this is more effective than conversation-level Likert scales.

4 IRT for Chatbot Evaluation

We pose chatbot human evaluation as an Item Response Theory (IRT) problem, similar to the approach of Otani et al. (2016). Again, throughout this section we consider the analogy of **student $\sim$ A/B chatbot comparison** and **question $\sim$ prompt**. In the context of educational testing, we are seeking to find the ability of a student and the effectiveness of exam questions (e.g. SAT exam) which in our setting is the comparative difference in pairs of chatbots.

As seen in Table 1, we sum the wins minus losses for each human evaluation of a pair of chatbot systems for each prompt; this net rating ranges between $[n, -n]$ where n is the number of annotators. In the student analogy, this is equivalent to an exam question worth $2n$ points. This is a well-studied problem, the so called the "graded mean" formulation of IRT (Samejima, 1969).

We first introduce the graded mean formulation of IRT required to estimate the relative assessment of chatbots and the discriminative power of the prompts. Subsequently, we describe the exact problem formulation in our setting.

4.1 Item Response Theory

The core idea behind IRT is that the probability that student i gets each question (item) j correct depends both on the ability of the student and the difficulty of the question. IRT aims to assess a latent ability trait θ_i for each student i from their answers u_j^i to items j, and, simultaneously, to determine how informative each item j is. This informativeness depends on the ability of the student; one wants to give harder questions to good students and easier questions to weaker students. IRT is a latent variable Bayesian model that can be estimated via expectation maximization (EM) or variational inference. For a comprehensive exposition of IRT see Embretson and Reise (2013).

More formally, we use the *graded mean* IRT model in which the probability that a student i obtains a score above c (the "rated scale assignment") for question j (Andrich, 1978). $P_{ijc}(\theta_i)$, the probability that student score (or aggregate chatbot rating), $u_j^i > c$, is given by

$$\begin{aligned} P_{ijc}(\theta_i) &= P_{ij}(u_j^i \geq c \mid \theta_i, b_j, \alpha_j) \\ &= \sigma(\alpha_j(\theta_i - b_{jc})) \\ &= \frac{1}{1 + \exp(-\alpha_j(\theta_i - b_{jc}))}, \end{aligned}$$

where σ is the logistic function. b_{jc} is the item (j-th question) difficulty for the score c (e.g. to score 4 or more points out of 6 on an exam question), α_j is the slope or item's *discrimination* (measuring how informative the question is for measuring the student's ability), and θ_i is the latent *ability* of student i.[1] Better questions (higher α_j) allow investigators to determine which student is better with fewer questions. We will use this same model to test which chatbot is better using fewer prompts.

In order to make this model generative, we can define

$$\begin{aligned} P_{ij}(u_j^i = c \mid \theta_i, b_j, \alpha_j) = P_{ij}(u_j^i \geq c - 1 \mid \theta_i, b_j, \alpha_j) \\ - P_{ij}(u_j^i \geq c \mid \theta_i, b_j, \alpha_j). \end{aligned}$$

If $c \in [-3, 3]$ then $P_{ij-3}(\theta_i) = 1$ and $P_{ij4}(\theta_i) = 0$. IRT is a latent variable Bayesian model, where θ_i, b_j, and $\log(\alpha_j)$ have priors from a normal distribution. The model is estimated by gradient descent.

4.2 Problem Setting

IRT can be easily repurposed for chatbot evaluation. Rather than assessing individuals i based on their answers to exam questions j, we assess the relative rating (preference) between two chatbot models i based on their responses to conversational prompts j. Instead of teachers (or ETS) grading the students' answers, human raters now rate the chatbot responses. The overall score for a chatbot for each item is the accumulated annotator preferences for that chatbot over its competitor. The score for chatbot B compared against chatbot A for item j is

$$u_j^{B/A} = \sum_{k=1}^{\text{num annotators}} (w_{kj}^B - w_{kj}^A),$$

[1] Our formulation is slightly simpler than the canonical graded mean formulation since c is a fixed finite number. Thus, the asymptotes for the item response function (IRF) need not be estimated.

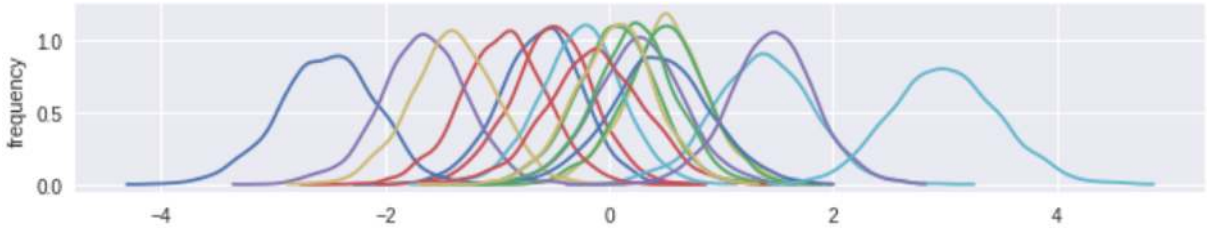

Figure 1: Each curve shows the estimated distribution of difference (inverse logit) in assessed quality between a pair of two different chatbot models produced by our Bayesian IRT model. The mode of each curve is the expected value of the quality difference, and zero means that the models are believed to be equally good.

where $w_{kj}^A = 1$ and $w_{kj}^B = 0$ if for prompt j the k-th annotator chose model A as having a better response; values are reversed if model B was preferred (see examples in Table 1).[2] The resulting *ability score* $\theta_i \in \mathbb{R}$ is then the relative "ability" (i.e. assessed quality) of models i =A vs B. Figure 1 shows a distribution of ability across multiple pairwise comparisons of models.

A critical difference between our formulation and that of Otani et al. (2016) is that we explicitly account for the independence of prompts, and do not model individual annotators k. Estimating a model of individual annotators would require many annotations for each annotator, which is not practical for estimator convergence.

IRT gives an optimal way to combine item results (given the modeling assumptions). It is flexible in that one need not make comparisons for all items for all chatbot pairs. In order to avoid overstating statistical significance, we group covariate prompts using a simple correlation filter (> 0.6) over all experiments.[3] In order to keep the net rating in $[-3, 3]$, we average the scores in the group. Note that this is the most conservative possible choice. We further control for multiple testing error by analyzing all comparisons simultaneously (Miller, 1981). As more comparisons are made, more information is revealed about the prompts in the evaluation dataset.

5 Experimental Details

While human evaluation remains the gold standard for dialog research, the design of human evaluation experiments is far from standard. We restrict our analysis to designs where the annotator is shown a prompt and two possible responses and then asked to select the better one or specify a tie. We follow the setup of Sedoc et al. (2019) (see the Appendix for instruction to Amazon Mechanical Turk crowd workers).

5.1 System Descriptions

We conducted a series of experiments to establish high-quality baselines for several popular training sets to show the efficacy of our proposed method. We compared our baselines against the OpenNMT benchmark for dialog systems[4]; Cakechat[5], which is a reimplementation of the hierarchical encoder-decoder model (HRED) (Serban et al., 2016); and the Neural Conversation Model's (NCM) released responses from Vinyals and Le (2015). Cakechat was trained on Twitter data, and NCM and OpenNMT benchmark were trained on movie subtitle data from OpenSubtitles (Tiedemann, 2012). We also evaluated two state-of-the-art Transformer base models: DialoGPT[6] medium (Zhang et al., 2019) and Blender (2.7B)[7] (Roller et al., 2020). Two human baselines created by Sedoc et al. (2019) were used.

All other models were trained with OpenNMT-py (Klein et al., 2017) Seq2Seq implementation with its default parameters: two layers of LSTMs with 512 hidden neurons for the bidirectional encoder and the unidirectional decoder. We trained several models and chose the best using non-exhaustive human evaluation.[8] *OpenNMT_Seq2SeqAttn* is trained using OpenSubtitles (Tiedemann, 2012) and *Seq2SeqAttn_OpenSubtitles_Questions* is trained using pairs where the first utterance ends in a

[2] If the number of annotators is variable, then we scale u_j^i to a fixed range which here we set to $[-3, 3]$.

[3] We calculate the correlation of judgments u_j^i between all prompts over all annotators and evaluations.

[4] http://opennmt.net/Models-py/

[5] https://github.com/lukalabs/cakechat from Replika.ai.

[6] https://github.com/microsoft/DialoGPT

[7] https://parl.ai/

[8] We experimented with whether or not to use pre-trained word embeddings, the impact of optimizer stochasticity, and various types of data preprocessing.

question mark and the second does not. Finally, *Seq2SeqAttn_Twitter* was trained on Twitter micro-blogging data as originally done by Ritter et al. (2010).[9] All of the data was extracted and tokenized using ParlAI (Miller et al., 2017).[10]

5.2 Selection of Evaluation Set

Our evaluation set is the list of 200 questions released by Vinyals and Le (2015) in their seminal work on neural conversational models using a standard Seq2Seq framework borrowed from machine translation. The evaluation set is hand-crafted and there are several correlated examples, such as the prompts *are you a follower or a leader ?* and *are you a leader or a follower ?* This quality is not unique to this evaluation dataset.

5.3 Human Evaluation Details

The evaluation prompts are split into blocks (currently defaulted to 10)[11]. We used the same experimental setup as Sedoc et al. (2019). The overall inter-annotator agreement (IAA) varies depending on the vagueness of the prompt as well as the similarity of the models. The overall IAA as measured by Fleiss' kappa (Fleiss, 1971) varies between .2 to .54 if we include tie choices. As Dras (2015) note, there is little agreement in the community on how to handle tie choices. Our IAA is similar to the findings of Yuwono et al. (2019) who also found low inter-annotator agreement when assessing conversational turns.

Unfortunately, "bad" workers accounted for roughly seven percent of all annotations, which we remove from our results. To identify such workers, we examine the worker annotation against the other two annotations. We remove annotators whose correlation is not statistically significantly greater than 0. It is important to note two things 1) the two annotations are likely more than two other workers since we have a minimum of 3 annotators and a maximum of 60, and 2) unless the "bad" worker is adversarial (i.e. labeling the opposite of the correct judgment) and instead just randomly labels, then the annotator will lower inter-annotator agreement, but IRT will not be significantly affected (Hopkins and May, 2013). However, "bad" workers will create bias in the estimate of mean difference (a.k.a. ability) of models to be closer to 0 (see the Appendix for further details).

6 Results

We used IRT to compare multiple neural models for their relative strength. Furthermore, we also included human baselines in our model comparison. Finally, we assessed the discriminative quality of the hand-crafted prompts from Vinyals and Le (2015).

6.1 Model Comparison Results

A comparison of the models described in section 5.1 is in Table 3 (all model comparisons are in the Appendix).[12] By analyzing the significance of all of the models at once using IRT, we can correct for multiple testing (Miller, 1981). I.e., given multiple comparisons, by chance a comparison might look statistically significant if naively using a p-value of 0.05.

Overall, there is a roughly uniform distribution of ratings (see the appendix for more detail). The grade is from -3 to 3 since there are 3 annotators per prompt for all but one experiment.

As seen in Table 3 the NCM (Vinyals and Le, 2015) model performance cannot be matched by any other model, even though all models are based on Seq2Seq. This indicates that either baseline models are difficult to properly train and parameterize, or that the NCM model may be overfit for the evaluation set. Interestingly, there are not enough ratings to evaluate whether NCM is worse than our human baselines. NCM also seems to outperform both Blender as well as DialoGPT; however, these results are not statistically significant. Blender is designed for multi-turn interactions, so single-turn prompts may not be a fair comparison.

Note, that IRT does not yield a total ordering of systems. In pairwise comparisons between Cakechat and Seq2SeqAttn_Twitter and Seq2SeqAttn_OpenSubtitles, Cakechat is superior to Seq2SeqAttn_Twitter. However, Seq2SeqAttn_OpenSubtitles is almost statistically significantly better than Cakechat, while Seq2SeqAttn_Twitter and Seq2SeqAttn_OpenSubtitles are rated to have equivalent performance. One possible rea-

[9] From https://github.com/Marsan-Ma/chat_corpus/raw/master/.

[10] https://github.com/facebookresearch/ParlAI

[11] We used the code from ChatEval https://github.com/chateval/chateval/

[12] We used pyStan for our IRT. Our code is available on Google Colab.

System A	System B	Mean Δ Ability	Std Δ Ability
Human2	Human1	-0.356	0.256
Human2	Seq2SeqAttn_Twitter	-2.760*	0.291
Human2	Seq2SeqAttn_OpenSubtitles_Ques	-2.015*	0.265
Human2	NCM	-0.377	0.324
Human1	Seq2SeqAttn_Twitter	-1.980*	0.269
Human1	NCM	0.224	0.262
NCM	DialoGPT	-0.223	0.245
NCM	Blender (2.7B)	-0.347	0.256
NCM	Cakechat	-0.715*	0.261
NCM	Seq2SeqAttn_Twitter	-1.426*	0.274
NCM	OpenNMT_Seq2SeqAttn	-1.034*	0.287
Cakechat	Seq2SeqAttn_Twitter	-0.529*	0.268
Cakechat	OpenNMT_Seq2SeqAttn	0.125	0.262
Cakechat	Seq2SeqAttn_OpenSubtitles	0.460	0.281
Seq2SeqAttn_OpenSubtitles	Seq2SeqAttn_OpenSubtitles_Ques	0.295	0.274
Seq2SeqAttn_OpenSubtitles	Seq2SeqAttn_Twitter	0.052	0.274
Seq2SeqAttn_OpenSubtitles	OpenNMT_Seq2SeqAttn	0.177	0.318

Table 2: The mean and standard deviation of "ability" (inverse logit) of paired comparisons of various models, where overlap with zero indicates no difference. Larger positive indicates that System B is superior in terms of rating by human annotators and similarly smaller negative numbers mean that System A is superior. (* shows significant differences $p < 0.05$ and better system is in bold.)

son for this might be that both Cakechat and Seq2SeqAttn_Twitter are trained on Twitter, so their model responses are more directly comparable.

6.2 Evaluation Set Selection

In order to minimize the numbers of evaluations required to assess the relative performance of models, we first removed redundant prompts, and then used IRT to select the prompts that were most discriminative.

IRT evaluates the discriminative ability of each prompt independently, so first we analyzed the correlation structure of responses over all evaluations and removed redundant prompts. By construction, the NCM evaluation set has correlated examples such as *my name is david . what is my name ?* and *my name is john . what is my name ?* Most models generate similar responses to both examples, and as a result, human judgments will correlate. Thus, we can use a smaller evaluation set while achieving similar significance. Defining redundancy as a correlation > 0.6 removes 6 out of 200 prompts.

To test the effect of using IRT to select prompts, we use a leave-one-out design, i.e. we keep 19 model comparisons and then select a subset of prompts with the most discriminative power for the 20th out-of-sample comparison. It is important to note that the most discriminative prompts (α_j) are usually not the most difficult ones ($b_j c$). This is different from Lalor et al. (2019) who use training example difficultly.

Figure 2 shows the change in the standard error of the ability estimates as we reduced the number of prompts. Our main result is that selecting just 100 of the 200 prompts using IRT maintains the same standard error, while selecting 100 random prompts gives a significantly higher error. Thus, using IRT allows us to reliably compare methods using fewer prompts.

Different Prompts for Better Students Finally, we assessed the effect of model quality on chatbot evaluation. Intuitively, one wants harder questions for better students. Similarly, an example such as *my name is david . what is my name ?* is an easier prompt than *what is the purpose of being intelligent ?* However, two models that are closer to human parity will only be distinguishable by the latter example. Similarly, for models further from human performance, both would perform poorly for example *OpenNMT Seq2Seq: I don 't know .* and *CakeChat: i ' m not sure what to say .* Using

IRT, we were able to validate this intuition across multiple models.

We split systems into two categories "better" - (NCM, DialoGPT, Blender, and Cakechat) and the other systems (e.g. OpenNMT) by sorting using mean Δ ability (Table 3). For each set of chatbots, we re-estimate the ability and item difficulty using only the subset of comparisons within each category (i.e., better chatbots are only compared against other better chatbots). We report the average standard error of difference of ability estimates of the left-out comparisons when using IRT with the most discriminative prompts. Thus, different prompts are selected for the better chatbots than for the others. The number of prompts was reduced while maintaining discriminative power as measured by standard error of discriminative ability (Figure 2); using prompts customized to each group yields lower standard error than using the globally "best" prompts. As the number of models increases, such filtering based on model quality further improves samplewise efficiency. IRT prompt selection using model quality allows us to dynamically update the evaluation set to adapt to better models.

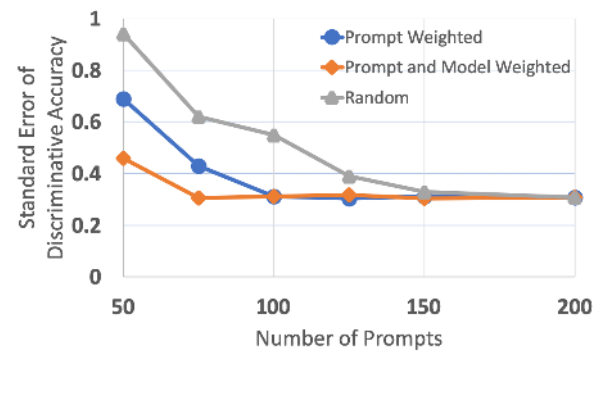

Figure 2: Standard error of discriminative accuracy as a function of the number of prompts. We compare selecting random subset (Random) to selecting prompts (Prompt Weighted), and both prompt difficulty and model performance (Prompt and Model Weighted).

Our work generalizes beyond the evaluation set from Vinyals and Le (2015). While other evaluation sets, such as random subsets of Twitter or OpenSubtitles may have fewer covariate prompts, there are many examples where further conversational context is required causing the prompts to have low discriminative power. For example, the prompt from the Twitter evaluation set (Sedoc et al., 2019), *Not really* is difficult to respond to without conversational context causing the prompt to have low discriminative power. Also, our method is not limited to single-turn prompts; however, for this case study, we focus on the available evaluation set. Multi-turn prompts such as **A:** *Was this useful to you?* **B:** *Yes* **A:** *Ok* are not very useful since almost any future response is valid. Initial results show that we can use IRT to automatically filter such uninformative prompts instead of hand-curating an evaluation set.

7 Conclusion

We present a new method for incorporating IRT into chatbot evaluation and show that we can use IRT to adaptively and optimally weight prompts from the evaluation sets, eliminating less informative prompts. One of the strengths of our method is that prompt discriminative ability and difficulty are re-estimated as new evaluations are added. One can thus start with a larger evaluation set, such as a subset from the Cornell Movie Database (Danescu-Niculescu-Mizil and Lee, 2011) and continue refining the subset of the evaluation set. We showed that our method is effective with the NCM evaluation set. Applying it to the Cornell Movie Database evaluation set of Baheti et al. (2018), we found that we could reduce from 1000 to 150 prompts with negligible loss of accuracy. When evaluating a new model, one would start with a comparison, say against a human baseline on a large set of prompts, then against a similarly ranked model using an appropriate subset of prompts. After each evaluation, the accuracy of all comparisons will increase. IRT can also be used to adapt evaluation sets as chatbot models improve in performance, reducing annotation costs.

While our main exposition addresses single turn prompts for chatbot evaluation, our IRT model comparison method generalizes to many natural language generation tasks, including machine translation and text simplification. It also generalizes to multi-turn prompts, point-wise evaluation, pairwise conversational evaluation (e.g. ACUTE-Eval (Li et al., 2019a)), and interactive evaluations such as those of Kulikov et al. (2019).

Acknowledgments

We thank the reviewers for their insightful comments.

This work was partially supported by the Amazon AWS Cloud Credits for Research program. This work was supported in part by DARPA KAIROS (FA8750-19-2-0034). The views and conclusions contained in this work are those of the authors and should not be interpreted as representing official policies or endorsements by DARPA or the U.S. Government.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354, Tokyo, Japan. Association for Computational Linguistics.

David Andrich. 1978. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573.

Nabiha Asghar, Pascal Poupart, Xin Jiang, and Hang Li. 2017. Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 78–83. Association for Computational Linguistics.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating more interesting responses in neural conversation models with distributional constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3970–3980, Brussels, Belgium. Association for Computational Linguistics.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, 1(2):97–111.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Manuela Cattelan. 2012. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, pages 412–433.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Mark Dras. 2015. Evaluating human pairwise preference judgments. *Computational Linguistics*, 41(2):337–345.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392. Association for Computational Linguistics.

Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 445–450, Beijing, China. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Scott Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *AAAI*.

Joy P Guilford. 1953. The correlation of an item with a composite of the remaining items in a test. *Educational and Psychological Measurement*, 13(1):87–93.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill: a bayesian skill rating system. In *Advances in neural information processing systems*, pages 569–576.

Mark Hopkins and Jonathan May. 2013. Models of translation competitions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1424, Sofia, Bulgaria. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *ACL, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Philipp Koehn. 2012. Simulating human judgment in machine translation evaluation campaigns. In *International Workshop on Spoken Language Translation (IWSLT) 2012*.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.

Ilya Kulikov, Alexander H Miller, Kyunghyun Cho, and Jason Weston. 2018. Importance of a search strategy in neural dialogue modelling. *arXiv preprint arXiv:1811.00907*.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259, Hong Kong, China. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A Diversity-Promoting Objective Function for Neural Conversation Models.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Data Distillation for Controlling Specificity in Dialogue Generation.

Margaret Li, Jason Weston, and Stephen Roller. 2019a. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2019b. Dialogue generation: From imitation learning to inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6722–6729.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Varvara Logacheva, Mikhail Burtsev, Valentin Malykh, Vadim Poluliakh, Alexander Rudnicky, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, and Yoshua Bengio. 2018. A dataset of topic-oriented human-to-chatbot dialogues.

FM Lord, MR Novick, and Allan Birnbaum. 1968. Statistical theories of mental test scores.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *ACL*, pages 1116–1126. Association for Computational Linguistics.

Alberto Maydeu-Olivares and Anna Brown. 2010. Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, 45(6):935–974.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Rupert G Miller. 1981. Simultaneous statistical inference.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Naoki Otani, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2016. IRT-based aggregation model of crowdsourced pairwise comparison for evaluating machine translations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 511–520, Austin, Texas. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

Fumiko Samejima. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd ACL*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the NAACL-HLT*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Hiroaki Sugiyama, Toyomi Meguro, and Ryuichiro Higashinaka. 2019. Automatic evaluation of chat-oriented dialogue systems using large-scale multi-references. In *Advanced Social Interaction with Agents*, pages 15–25. Springer.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems.

Louis L Thurstone. 1927. A law of comparative judgment. *Psychological review*, 34(4):273.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. *Natural Language Dialog Systems and Intelligent Assistants*, 37:233–239.

Steven Kester Yuwono, Biao Wu, and Luis Fernando DHaro. 2019. Automated scoring of chatbot responses in conversational dialogue. In *9th International Workshop on Spoken Dialogue System Technology*, pages 357–369. Springer.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A Further Human Evaluation Details

Crowd workers are paid $0.01 per prompt, and on average it takes 1 minute to evaluate 10 choices with a maximum allowed time of 2 minutes. We used three evaluators per prompt, so, if there are 200 prompts, we have 600 ratings and the net cost of the experiment is $7.2. We chose 3 annotators since we can generalize enough for IAA and it is cost-effective.

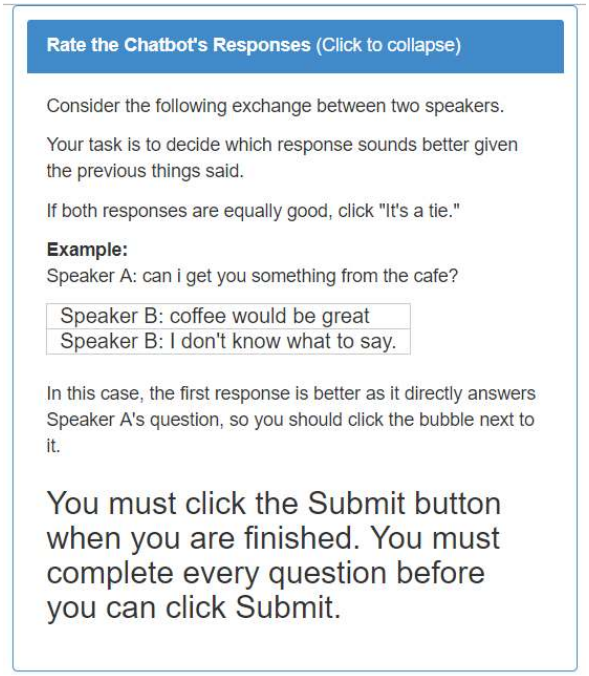

Figure 3: The instructions seen by AMT workers.

The instructions seen by AMT workers are shown in Figure 3.

We removed workers with a correlation below 0.05 with other annotators. For a worker identified as "bad", all annotations are removed. Including these workers only increases the standard error by 10%.

From the 200 NCM evaluation set prompts, each annotation task has 10 prompts; however, we do not pair the same 3 workers to the 10 prompts; instead we randomize the prompts shown, so worker 1 many compare prompts 1-10, while worker 2 compares prompts 2,3,5,7,9,11,13,17,19,23. As a result, the correlation between one worker and the others is more stable.

A full set of model comparisons on the Neural Conversation Model is available in Table 3.

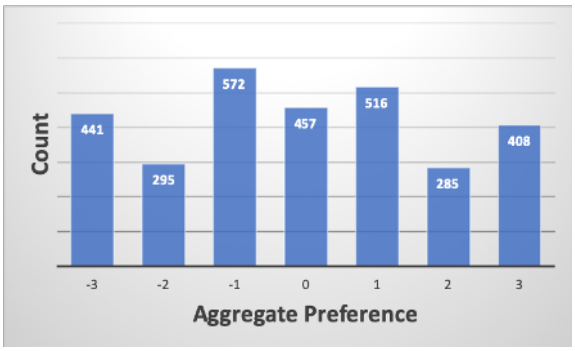

Figure 4: A histogram of aggregated preferences, $\sum_i \sum_j u_j^i$, across all prompts and model comparisons by all annotators.

A.1 Rating Distribution

Figure 4 shows a histogram of the grades over all experiments run.

System A	System B	Mean Δ Ability	Std Δ Ability
Cakechat	Seq2SeqAttn_Twitter	-0.529*	0.268
Cakechat	OpenNMT_Seq2SeqAttn	0.125	0.262
Seq2SeqAttn_OpenSubtitles	Cakechat	-0.460	0.281
Seq2SeqAttn_OpenSubtitles_wo_PTE	Seq2SeqAttn_OpenSubtitles	0.088	0.273
Seq2SeqAttn_Twitter_without_PTE	Seq2SeqAttn_Twitter	0.424	0.273
Cakechat	NCM	1.314*	0.310
Human1	Seq2SeqAttn_Twitter	-1.98*	0.269
Human1	Human2	0.356	0.256
NCM	Cakechat	-0.715*	0.261
NCM	Seq2SeqAttn_Twitter	-1.426*	0.274
NCM	OpenNMT_Seq2SeqAttn	-1.034*	0.287
NCM	Human1	-0.224	0.262
NCM	Human2	0.377	0.324
Seq2SeqAttn_OpenSubtitles	Seq2SeqAttn_OpenSubtitles	0.295	0.274
OpenNMT_Seq2SeqAttn	Seq2SeqAttn_OpenSubtitles	-0.177	0.318
Seq2SeqAttn_OpenSubtitles_Ques	Human2	2.015*	0.265
Seq2SeqAttn_OpenSubtitles	Seq2SeqAttn_Twitter	0.052	0.274
Seq2SeqAttn_Twitter	Human2	2.760*	0.291
NCM	DialoGPT	-0.223	0.245
NCM	Blender (2.7B)	-0.347	0.256

Table 3: Comparison of various models using IRT. Larger positive indicates that System B is superior in terms of rating by human annotators and similarly smaller negative numbers mean that System A is superior. (* shows significant differences.)