Q^2: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering

Or Honovich[1] **Leshem Choshen**[1] **Roee Aharoni**[2] **Ella Neeman**[1]
Idan Szpektor[2] **Omri Abend**[1]
[1]The Hebrew University of Jerusalem; [2]Google Research
or.honovich@gmail.com
{roeeaharoni,szpektor}@google.com

Abstract

Neural knowledge-grounded generative models for dialogue often produce content that is *factually inconsistent* with the knowledge they rely on, making them unreliable and limiting their applicability. Inspired by recent work on evaluating factual consistency in abstractive summarization, we propose an automatic evaluation metric for factual consistency in knowledge-grounded dialogue using automatic question generation and question answering. Our metric, denoted Q^2, compares answer spans using natural language inference (NLI), instead of token-based matching as done in previous work. To foster proper evaluation, we curate a novel dataset of dialogue system outputs for the Wizard-of-Wikipedia dataset, manually annotated for factual consistency. We perform a thorough meta-evaluation of Q^2 against other metrics using this dataset and two others, where it consistently shows higher correlation with human judgements.

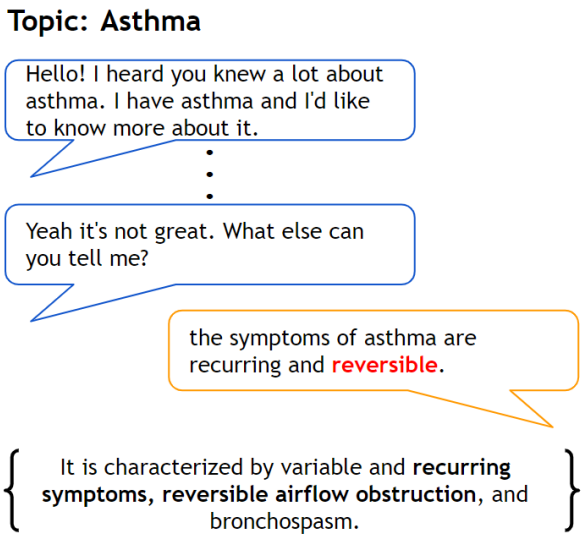


Figure 1: An example from our dataset. Human messages are in Blue, the generated response is in Orange and the grounding knowledge is in Black at the bottom. The factual inconsistency is marked in Red.

1 Introduction

Generative conversational agents show remarkable progress lately (Shuster et al., 2020; Adiwardana et al., 2020). Yet, generative dialogue models that are grounded by external knowledge sources still struggle to be consistent with that knowledge. Their output is often incompatible with the given knowledge or even completely "hallucinated" (Roller et al., 2020). Figure 1 depicts such inconsistency by the dialogue system of Shuster et al. (2020) when evaluated on the Wizard of Wikipedia dataset (Dinan et al., 2019). Since inconsistent generated text is usually fluent and well-formed, these outputs could mislead users with false information, limiting the applicability of such systems.

Factual inconsistency is often overlooked by evaluation methods for text generation (Celikyilmaz et al., 2020). Evaluation approaches that address this gap were recently proposed for tasks like machine translation and abstractive summarization (Sellam et al., 2020; Xu et al., 2020; Goodrich et al., 2019). Yet, evaluating grounded dialogues poses additional challenges, since dialogue outputs may refer to the dialogue history and include personal opinions, questions to the user, and general "chit-chat", whose consistency with external knowledge is mostly irrelevant. Additionally, many of those metrics require gold-label human-constructed references, while dialogue is an open-ended task – making it less suitable for reference-based evaluation.

In this work, we propose an automatic metric for evaluating the factual consistency of generative open-domain knowledge-grounded dialogue systems which does not require gold-label reference responses. Our metric, denoted Q^2, pairs automatic question generation (QG) and question answering (QA) for dialogue generation evaluation, inspired by recent work on factual consistency evaluation in abstractive summarization (Durmus et al., 2020;

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7856–7870
November 7–11, 2021.

Wang et al., 2020). Q^2 first takes a given generated response as input, and generates questions whose answers are informative spans in the response, using a QG system. It then employs a QA system to find corresponding answer spans in the knowledge that the response should be grounded in. The evaluation score reflects the similarity between each informative response span and its corresponding answer span from the knowledge, for each generated question.

Unlike previous QG/QA approaches, which used token-based matching to compare answer spans, we propose a novel comparison method using natural language inference models (NLI; Dagan et al., 2006) that is more robust to lexical variability. In addition, while QG/QA based methods showed promising results for summarization evaluation, our work is the first to apply them to knowledge-grounded dialogues, which hold distinct properties compared to other grounded generation tasks; Mixing different types of utterances such as knowledge, personal statements and chit-chat in a single response is unique to dialogue and is well addressed by our metric given its modular nature and robustness to lexical variability.

We assess Q^2 against other reference-response-free metrics on three dialogue benchmarks: Wizard of Wikipedia (WoW; Dinan et al., 2019), Topical-Chat (Gopalakrishnan et al., 2019) and Dialogue NLI (DNLI; Welleck et al., 2019). To foster proper evaluation, we curate a new dataset of dialogue system responses using the WoW dataset, manually annotated for factual consistency. Q^2 reaches significantly higher correlations with human judgments on all datasets compared to the other metrics, demonstrating its potential as an evaluation framework for grounded dialogue generation.

To summarize, our contributions in this work are three-fold: (1) We develop a novel framework for evaluating the factual consistency of knowledge–grounded, open-domain dialogue systems, incorporating question generation, question answering and NLI models. (2) We construct a first-of-its-kind dataset of knowledge-grounded dialogue system outputs manually annotated for factual consistency, fostering future work on the subject. (3) We validate the effectiveness of our metric in comparison to previous approaches through various experiments with three dialogue benchmarks, where it obtains higher correlation with human judgements.[1]

[1]Our code and dataset are available in: http://github.com/orhonovich/q-squared

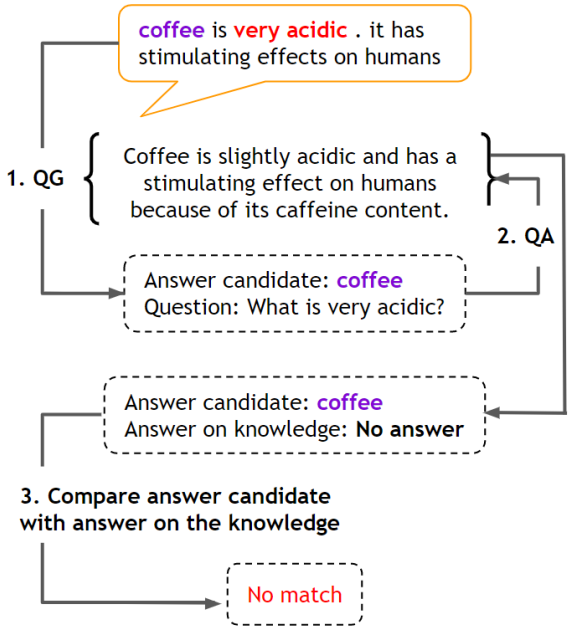


Figure 2: The Q^2 pipeline: (1) For a response, select answer candidates; then generate a question for each candidate using QG. (2) Use QA to answer each question based on the grounding knowledge. (3) Compare the answer candidate with the knowledge answer span.

2 Evaluating Factual Consistency

Formally, an evaluation metric for factual consistency in generative dialogue receives as input a dialogue history h, a textual knowledge source k, and a response r from a dialogue model (assumed to be generated conditioning on h and k). The goal is to score the model's output r so as to reflect its consistency with its grounding source k. We next introduce our metric, denoted Q^2, which suggests that factual questions that have answers in the generated response should have similar answers in the grounding knowledge source, while differences between answers from the response and the knowledge point at factual inconsistencies. This follows the intuition in Wang et al. (2020); Durmus et al. (2020) for evaluating abstractive summarization.

Q^2 iterates over all informative spans a_i^r in r. For each a_i^r, Q^2 uses a QG system to generate questions q_{i_j} whose answer is a_i^r. For each question q_{i_j}, Q^2 uses an extractive QA system to mark an answer span $a_{i_j}^k$ from k. Q^2 then measures the similarity of a_i^r and $a_{i_j}^k$ and aggregates the similarity scores for all questions as the factual consistency score of r. Figure 2 depicts this procedure. We next detail each component in our metric.

Question Generation. First, we mark informative spans in the response r to serve as target answer spans for the QG system. To this end, we mark all named entities and noun phrases in r using spaCy.[2] For example, in "*coffee is very acidic*" we mark '*coffee*' as an informative span. Then, a QG system takes each informative span a_i^r and the response r as input and generates the corresponding questions q_{i_j} for which a_i^r should be the answer. In our example, a generated question for the informative span '*coffee*' and the response in Figure 2 is "*What is very acidic?*". We use T5-base (Raffel et al., 2020) fine-tuned on SQuAD1.1 (Rajpurkar et al., 2016) as the QG model.[3]

As suggested by Wang et al. (2020), we use beam search decoding, taking the top-n generated questions for a_i^r. We set $n = 5$ and test two variants of generating multiple questions. In the first, we use all n questions for a_i^r. In the second variant, we only take the top-ranked question that passed the filtering stage for a_i^r (see "Question Filtering" below). We observed similar trends for both variants, and therefore only report the results of the second variant. To increase the diversity of the generated questions, we tried sampling-based methods (Fan et al., 2018; Holtzman et al., 2020), but obtained inferior results that are not reported in this paper.

Question Answering. To mark the answer span $a_{i_j}^k$ in the knowledge k for question q_{i_j}, we use the Albert-Xlarge model (Lan et al., 2020) fine-tuned on SQuAD2.0 (Rajpurkar et al., 2018).[4] This model can also determine that no answer can be found in the paragraph. This is important in Q^2, since question q_{i_j} generated for a completely hallucinated content a_i^r should have no answer in k.

Answer Similarity and Final Scores. The last step in Q^2 assesses the similarity between answers a_i^r and $a_{i_j}^k$. To be robust to lexical variability between the response and the knowledge, e.g. "*US*" vs. "*United States*" or "*a book series*" vs. "*a set of novels*", we measure the answer span similarity using an NLI model. We use RoBERTa (Liu et al., 2019) fine-tuned on SNLI (Bowman et al., 2015) as implemented in AllenNLP (Gardner et al., 2017).

For span pairs a_i^r and $a_{i_j}^k$ that match perfectly at the token-level, we assign a score of 1. For each span pair a_i^r and $a_{i_j}^k$ that do not match perfectly at the token-level, we run the NLI model with $a_{i_j}^k$ as the premise and a_i^r as the hypothesis. To add context for the NLI model, each answer is concatenated after the question q_{i_j}. For example, for the question "*Where were the Red Hot Chili Peppers formed?*", the response answer "*LA*", and the knowledge answer "*Los Angeles*", we run the NLI model with: "*Where were the Red Hot Chili Peppers formed? Los Angeles*" as the premise, and with "*Where were the Red Hot Chili Peppers formed? LA*" as the hypothesis. Our use of NLI differs from prior use of NLI in dialogue evaluation, where it was applied in an end-to-end manner (Welleck et al., 2019; Pang et al., 2020). We set q_{i_j}'s score to be 1 for the case of entailment and 0 for contradiction or for cases where the QA model produced no answer. In the neutral case, we take the answers token-level F1 score, as in Wang et al. (2020).

Finally, the match scores for all answer pairs are averaged to yield a response-level score, and the response-level scores are averaged to yield a system-level Q^2 score.

Question Filtering. To alleviate errors made by the automatic QG and QA models, we follow the validation step in Wang et al. (2020); We run the QA model to answer q_{i_j} with the response r as the input paragraph, and require the answer to be identical to the answer span a_i^r which was used to generate q_{i_j}. If this is not the case, q_{i_j} is discarded.

As we evaluate factual consistency, we wish to ignore opinionated parts of the response which are not factual. Hence, we filter out questions that include the personal pronouns "*I*" or "*you*" as the subject, as well as questions that mention the possessive pronouns "*my*" or "*your*".

Lack of Valid Questions. For some responses, no valid questions are generated – i.e. all generated questions fail to pass the above filtering process. We use our NLI model as a fallback in such cases by taking its end-to-end prediction with k as the hypothesis and r as the premise. We set the score to be 1 in case it predicts entailment, 0 for contradiction, and 0.5 for the neutral case.

3 Evaluation Benchmarks

3.1 Wizard of Wikipedia

The Wizard of Wikipedia dataset (WoW; Dinan et al., 2019) contains dialogues in which a bot needs to respond to user inputs in a knowledgeable way.

[2] https://spacy.io/

[3] https://huggingface.co/mrm8488/t5-base-finetuned-question-generation-ap

[4] https://huggingface.co/ktrapeznikov/albert-xlarge-v2-squad-v2

Topic	Response	Knowledge
Coffee	coffee is very acidic. it has stimulating effects on humans.	Coffee is slightly acidic and has a stimulating effect on humans because of its caffeine content.
French cuisine	in that time italian cuisine was influenced by french cuisine	During that time, French cuisine was heavily influenced by Italian cuisine.
Madonna	she was born in 1968 and raised in new york city.	Born and raised in Michigan, Madonna moved to New York City in 1978 to pursue a career in modern dance.
Sephora	me too! it's an american fashion company founded in 1854.	Sephora is a French chain of cosmetics stores founded in 1969.

Table 1: Examples for factually inconsistent responses from our dataset. Factual inconsistencies are marked in red, with their corresponding parts in the knowledge marked in blue. The first two examples are outputs of the *dodeca*Dialogue system, and the last two are outputs of MemNet.

Each response should be grounded on a sentence from Wikipedia that is relevant to the conversation topic. Since this dataset does not contain explicit annotations for factual consistency of dialog responses, we construct a new dataset with such annotations for dialogues based on the WoW dataset as detailed in Section 4.

3.2 Topical-Chat

Topical-Chat (Gopalakrishnan et al., 2019) is a human-human knowledge-grounded conversation dataset. Each dialogue is accompanied by relevant Wikipedia pages, Washington Post articles and fun-facts from Reddit. Mehri and Eskenazi (2020) introduced USR, an evaluation metric that measures different aspects required from dialogue systems. To test USR, they collected human annotations on four different system responses and two human-generated responses for 60 dialog contexts from Topical-Chat. Each response was scored on a "Uses Knowledge" category, among others. Since a model that properly uses the knowledge is expected to use it in a factually consistent manner, we find it interesting to measure Q^2's correlation with the human judgements for this category.

3.3 Dialogue NLI

Dialogue NLI (DNLI; Welleck et al., 2019) is a dataset based on the Persona-Chat dialogue task (Zhang et al., 2018). It consists of pairs including either a personality description sentence or an utterance from the dialogue history (the *premise*) and a subsequent dialogue utterance (the *hypothesis*). Each pair is labeled as entailing, neutral, or contradicting. A contradiction may be a clear logical contradiction, e.g. "*I have a dog*" vs. "*I do not have a dog*", but can also be two utterances that are not likely to be said by the same persona although they are not strict logical inconsistencies, e.g. "*i'm a manager*" vs."*i'm a doctor*". Using this dataset, we test whether Q^2 can measure consistency when the grounding "knowledge" is a persona sentence or the previous dialogue history.

4 Dataset Creation and Annotation

To directly evaluate Q^2, we need an annotated dataset of knowledge-grounded dialogue responses and their factual consistency with respect to a given knowledge. To obtain this, three of the paper's authors annotated the factual consistency of a random sample of responses from the following dialogue systems on the WoW validation set: (1) **MemNet**, which is the model suggested by Dinan et al. (2019) for WoW. (2) ***dodeca*Dialogue**, which is the multi-task model fine-tuned on WoW in the *dodeca*Dialogue benchmark (Shuster et al., 2020), as available in ParlAI[5] (Miller et al., 2017). For both systems, we used beam search decoding with a beam size of 10, a beam block size of 3 and a context block size of 3 to generate responses.

The annotators went through the responses until 150 examples of factually inconsistent responses were annotated for each system (300 in total), and then repeated the process and annotated the same number of factually consistent responses. The annotators skipped factually consistent responses containing only general chit-chat with no reference to the grounding knowledge, such as "*Hi, how are you?*". For factually inconsistent responses, they selected challenging examples in which the text seemed clear and coherent. For each of the 600 extracted sentences, the annotation was extended to cover the outputs of both systems, resulting in 544 dialogue contexts and 1,088 annotated responses (due to overlaps). Out of the 544 contexts, 186 (34.2%) were marked as inconsistent in the *dodeca*Dialogue system and 274 (50.36%) in the MemNet system. The number of dialogue contexts and responses collected is comparable with those of other recently published datasets for dialogue evaluation, such as in Mehri and Eskenazi (2020); Pang et al. (2020); Zhao et al. (2020).

[5] https://parl.ai/docs/zoo.html

system	data	# questions	Q^2	Q^2 w/o NLI	% no answer	E2E NLI	Overlap(r, k)	BLEU	BERTScore
dodeca	Inconsistent	328	0.238	0.159	54.88%	0.5	0.299	3.355	0.179
	Consistent	341	0.696	0.516	15.25%	0.723	0.426	5.136	0.291
	Random sample	258	0.496	0.349	29.84%	0.573	0.325	3.788	0.164
MemNet	Inconsistent	324	0.135	0.123	62.04%	0.37	0.270	7.490	0.145
	Consistent	352	0.756	0.661	9.94%	0.717	0.526	20.145	0.376
	Random sample	268	0.448	0.387	32.09%	0.537	0.337	11.654	0.183

Table 2: Q^2 and baseline scores on the annotated system responses from WoW.

To evaluate the quality of the constructed dataset, 100 responses were sampled, and each annotator labeled them as consistent or inconsistent. The agreement level between annotators, measured by Fleiss' kappa, resulted in 0.853, representing high inter-annotator agreement. Table 1 shows factually inconsistent responses from this dataset. Detecting some of these inconsistencies requires identifying subtle semantic divergences from the facts expressed by the knowledge.

5 Experiments and Results

To evaluate Q^2 as a metric we performed the following experiments for each dataset.

5.1 Wizard of Wikipedia

Absolute Scores. Table 2 presents the Q^2 score for the different sets of annotated system responses, as well as for 150 randomly selected system responses. We additionally report the total number of generated questions (*after* filtering) for each set and the percentage of generated questions that had no answer in the knowledge. We denote our metric score by "Q^2", while "Q^2 w/o NLI" is an ablated variant obtained by dropping the NLI component and using the fallback token-level F1 instead, similarly to Wang et al. (2020).

As we would expect from a metric measuring factual consistency of generative dialogue systems, the Q^2 score is indeed always highest for the consistent outputs, lowest for the inconsistent outputs, and in-between for random samples. Assessing answer similarity using NLI results in higher absolute scores for both inconsistent and consistent responses, and by a larger margin for the latter.

Baselines. As baseline metrics, we first take the F1 token-level overlap of r with k as done in WoW (Dinan et al., 2019). We also use BLEU and BERTScore (Zhang et al., 2020) with the response r as the output, and the knowledge k as the reference. As our last baseline we run the NLI model described in §2 in an end-to-end manner, taking k as the premise and r as the hypothesis. We set the score to be 1 for the case of entailment and 0 for contradiction. In the neutral case, we set the score to be 0.5. The exact same settings are used as a fallback for Q^2 when no valid questions are generated. As Table 2 shows, the scores for the consistent data are higher than the scores for the inconsistent data for all baselines. However, in most cases, the score differences between the inconsistent data and the random samples are small, indicating that Q^2 better separates general responses from inconsistent ones.

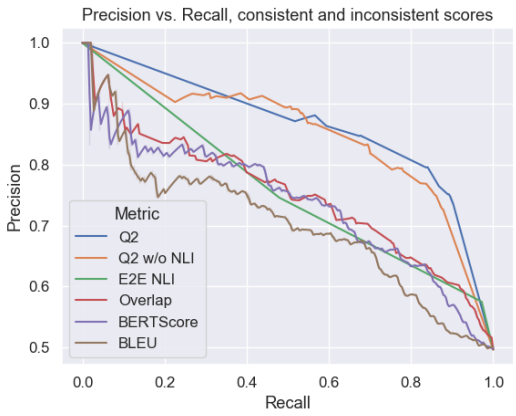


Figure 3: Precision-Recall curves for different response level score thresholds, calculated using the *dodeca* and MemNet consistent and inconsistent examples.

Response-Level Evaluation. To find if Q^2 can be used to automatically separate between consistent and inconsistent responses at the more granular, single response level, we report in Figure 3 the Precision/Recall curve of consistent responses for various response-level score thresholds for each evaluated metric on the WoW annotated data.

As Figure 3 shows, both Q^2 variants obtain higher precision and recall in comparison to the other metrics throughout the threshold values, suggesting that Q^2 is better at automatically separating between consistent and inconsistent examples at the response level. We additionally report in Table 3 the consistent and inconsistent Precision and Re-

Data split	Metric	Precision	Recall	F1
Inconsistent	Q^2	**73%**	86.7%	**0.793**
	Q^2 w/o NLI	67.1%	**91%**	0.772
	E2E NLI	61.2%	83.7%	0.707
Consistent	Q^2	83.5%	**67.9%**	**0.749**
	Q^2 w/o NLI	**85.9%**	55.2%	0.672
	E2E NLI	74.1%	46.8%	0.574

Table 3: Precision-Recall values for consistent and inconsistent response detection, using a threshold of 0.5 for the binary decision.

call values for a threshold of 0.5. Responses with a score of 0.5 or below are classified as inconsistent and vice versa. The accuracy of the binary decision using this threshold is 77.3% for Q^2, 73.1% for Q^2 without the NLI-based answer spans comparison, and 65.3% for the end-to-end NLI. We note that the threshold was arbitrarily selected for the purpose of demonstrating Q^2's ability in separating consistent from inconsistent content, and properly tuning it by splitting the data into development and test sets may improve the results further.

System-Level Evaluation. We measure the correlation of each metric with human judgments for systems with varying inconsistency levels. To simulate such systems, we follow the method of Graham and Liu (2016) for MT evaluation. We first take dialogue contexts for which we have both a consistent and an inconsistent response, leaving us with 244 dialogue contexts (and 488 responses). We then bootstrap (Efron, 1987) by sampling 350 contexts (with repetition) for each simulated system i, ensuring that each system output contains $c_i\%$ factually inconsistent responses. Finally, we compute the system-level score for each system and the correlation between those scores and the human annotations. We repeat this 1000 times and report average correlation and confidence intervals for each metric.

We take $c \in [0.05, 0.1, 0.15, 0.2, 0.25]$ as inconsistent response proportions for the simulated systems, and measure the Spearman correlation of Q^2 and the four baseline metrics with the human judgment scores of each system. The results are detailed in Table 4. Q^2 obtains an average correlation of 0.9798, while the end-to-end NLI baseline, overlap, BERTScore, and BLEU obtain lower correlations of 0.9216, 0.878, 0.8467 and 0.3051, respectively. This suggests that Q^2 is better in evaluating factual consistency at the system-level.

	Avg. Correlation	Lower CI	Upper CI
Q^2	**0.9798**	0.9	1
Q^2 w/o NLI	0.9711	0.9	1
E2E NLI	0.9216	0.6669	1
Overlap(r, k)	0.878	0.5	1
BERTScore	0.8467	0.4	1
BLEU	0.3051	-0.7	1

Table 4: Results for system level evaluation, taking systems with varying degrees of inconsistent outputs, and measuring the correlation between each system-level score and the human judgements.

Metric	Spearman	Pearson
Q^2	**0.4579**	**0.4698**
Q^2 w/o NLI	0.3933	0.4105
USR (best)	0.4468	0.3175
METEOR	0.3909	0.3328

Table 5: Correlation with human judgments for the "Uses Knowledge" category for different metrics. "USR (best)" stands for the best result achieved by Mehri and Eskenazi (2020) for each category.

5.2 Topical-Chat

Mehri and Eskenazi (2020) evaluated the correlation of their suggested metric, USR, as well as other existing automatic metrics, against human judgments on the Topical-Chat dataset (Gopalakrishnan et al., 2019). We note that in 8 out of the 60 examined dialogue contexts, no knowledge was used (the original dataset contains a "no fact" option). We thus experimented only with the 52 knowledge-grounded dialogue contexts. We follow the settings of Mehri and Eskenazi (2020), which used only 5 responses (out of the 6 annotated per response), leaving out the original human response that was collected by Gopalakrishnan et al. (2019). Accordingly, we are left with 260 responses. Table 5 presents their reported correlation results for the "Uses Knowledge" category, as well as the correlation of Q^2 with the same human judgments. Q^2 demonstrates an improvement in this category that is statistically significant with $p < 0.001$ compared to the baselines. The contribution of the NLI component on this dataset resulted in even higher gains in terms of correlation in comparison to the WoW experiments, again showing the benefit of using our more intricate span comparison method.

5.3 Dialogue NLI

We test Q^2's applicability for measuring persona consistency and self-consistency between dialogue utterances, as described in §3.3. We calculate the Q^2 score for each persona-utterance or utterance-

Model	Accuracy
Q^2	**74.49%**
Baseline – NLI only	67.42%
InferSent SNLI	47.03%
InferSent Hyp. Only	51.52%

Table 6: Accuracy on the DNLI dataset, Test Gold.

utterance pair and choose a threshold of 0.1 for predicting entailment or contradiction by tuning on the development set. Since a dialogue utterance should be grounded in the personality description or in the conversation's history, we treat neutral claims as inconsistent, and expect Q^2 to address them as contradictions. As DNLI aims at testing persona consistency, we avoid filtering out questions that include personal or possessive pronouns.

Table 6 presents Q^2's accuracy on the Test Gold split of DNLI, compared to other zero-shot methods. Our first baseline uses the NLI model in Q^2 in the end-to-end manner described above ("Baseline – NLI only"), which is similar to the approach of Welleck et al. (2019); Pang et al. (2020). To be comparable with Q^2's binary decision, we allow neutral claims to be predicted as either neutral or contradicting. We also show results from zero-shot methods reported in Welleck et al. (2019): a model that uses the hypothesis sentence only ("InferSent Hyp. Only") and a model trained on the SNLI dataset but evaluated on DNLI ("InferSent SNLI"). Q^2 performs better than the end-to-end NLI baselines, indicating that our QG/QA approach with NLI is more robust than simply applying end-to-end NLI with full sentences or passages.

5.4 Analysis

The results on the three datasets demonstrate Q^2's zero-shot, reference-response-free capability to generalize to various dialogue tasks that require evaluation of factual consistency. To shed more light on our approach we performed the following qualitative and quantitative analyses.

Robustness to Underlying Model Quality. The performance of Q^2 depends on the different components used throughout the pipeline, i.e., the QG, QA, and NLI models. To demonstrate that Q^2 is robust to the quality of these models, we experiment with using smaller models in the pipeline. First, we replace the T5-base model for question generation with a T5-small model, again fine-tuned on SQuAD1.1. Next, we replace the Albert-Xlarge QA model with Albert-base, similarly fine-tuned on SQuAD2.0 for question answering.

	Avg. Correlation	Lower CI	Upper CI
Original Q^2	**0.9798**	0.9	1
T5-small	0.9722	0.9	1
Albert-base	0.9797	0.9	1

Table 7: Correlations with human judgements when using a smaller QG or a smaller QA model.

Data	Model	Coverage	Q^2	Q^2 w/o NLI
dodeca inconsistent	Original	92.67%	0.238	0.159
	T5-small	90.67%	0.198	0.143
	Albert-base	92%	0.293	0.213
dodeca consistent	Original	94%	0.696	0.516
	T5-small	90.67%	0.601	0.44
	Albert-base	92.67%	0.709	0.534
MemNet inconsistent	Original	94.67%	0.135	0.123
	T5-small	90%	0.104	0.099
	Albert-base	94%	0.189	0.134
MemNet consistent	Original	92.67%	0.756	0.661
	T5-small	88.67%	0.705	0.613
	Albert-base	89.33%	0.791	0.7

Table 8: Q^2's results on WOW when using a smaller QG or a smaller QA model. Coverage refers to the questions coverage, i.e., the percentage of responses for which Q^2 generated at least one valid question.

As Table 7 shows, the correlations with human judgments are barely influenced by using smaller QG/QA models, showing the robustness of our method to changes in the underlying models. Table 8 presents the absolute scores of the smaller models on the WOW dataset, as well as each variant's question coverage, defined as the percentage of responses for which Q^2 generated at least one valid question, not resorting to the end-to-end NLI fallback. While the question coverage slightly decreases when using smaller models, the gap between consistent and inconsistent scores remains unaffected. As we expected, a smaller QG model results in lower Q^2 scores, for all data splits. Surprisingly, using a smaller QA model had the opposite outcome - higher Q^2 scores in all cases.

Regarding domain robustness of the undelying models, while the QG and QA models were trained on a dataset collected from Wikipedia and are therefore suited for WOW's domain, these models work well even when the grounding source is not Wikipedia. This is the case in Topical-Chat, in which each dialogue is accompanied by Washington Post articles and fun-facts from Reddit in addition to pages from Wikipedia; and in the DNLI dataset, which deals with persona and self-consistency of dialogue systems and does not contain any references to Wikipedia.

Lack of Valid Questions. For some responses, Q^2 does not generate any valid questions. When testing the extent of this phenomenon in the inconsistent vs. consistent samples collected based on the MemNet and *dodeca*Dialogue outputs, a similar proportion of around 6-8% responses had no valid questions. The proportion of such responses in the randomly sampled examples is much higher – around 20%. As mentioned in §2, we handle such cases using an end-to-end NLI fallback.

The higher proportion of such responses in the random samples indicates that lack of valid questions is more common in general chit-chat than in knowledge-grounded content. This raises the need to improve the identification and separation of general chit-chat responses from more "knowledgable" ones, which we plan to address in future work.

Another cause for low-quality questions that do not pass the filtering process is responses that contain pronouns referring to entities in the dialogue history – e.g. "*he won an album of his own in 2015*" requires resolving "*he*". Preliminary experiments with adding a coreference resolution step to our pipeline showed increased coverage, and we plan to further address this gap in future work.

Qualitative Analysis. To get a better impression of Q^2's operation, we give examples of how it operates in its various stages. Figure 2 presents an example for an inconsistent response, together with a generated question and the answer Q^2 obtained based on the knowledge. In this example, the question was unanswerable using the knowledge, thus the score for this question is 0. Indeed, this is the desired score, as the knowledge didn't mention that coffee is *very* acidic.

Another example for successful output is for the following response: "*i'm not sure about that but i do know that they are reliant on vulnerable species!*", generated by the *dodeca*Dialogue system when conversing about giant Pandas, while conditioning on the following knowledge paragraph: "*The giant panda is a conservation reliant vulnerable species.*". The response is clearly inconsistent with the knowledge as Pandas are reliant on conservation and not on vulnerable species. Here, Q^2 extracted "*vulnerable species*" as an informative span, and generated the question: "*What are they reliant on?*". The answer to this question using the knowledge was *"conservation"*, which resulted in assigning this question a score of 0.

These examples also demonstrate a major advantage of Q^2, being self-explanatory and interpretable. Other than the final score, Q^2 outputs the generated questions, the response-based answer spans and the answers the QA model predicted based on the knowledge, which can be used as an explanation to the assigned score or to highlight the potentially inconsistent text spans in the response.

Some errors of Q^2 are caused by generating questions for the chit-chat parts of responses. In a conversation regarding the color purple, the *dodeca*Dialogue system generated the response: "*purple is my favorite color. it's between red and blue.*", when the knowledge was: "*Purple is a color intermediate between blue and red.*" Even though the response used the knowledge faithfully, one out of two valid generated questions for it was "*What is purple?*", for which the response-based answer is "*my favorite color*", while the knowledge-based answer is, of course, different.

6 Related Work

Automatic Evaluation of Dialogue Systems. Automatically evaluating natural language generation is a notoriously difficult problem, especially when considering open-ended tasks such as dialogue. Standard token-matching metrics, such as BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005) in machine translation, or ROUGE (Lin, 2004) in summarization, were shown to have weak or no correlation with human judgements for dialogue (Liu et al., 2016; Lowe et al., 2017). Supervised assessment methods learn to predict human-like evaluation scores (Lowe et al., 2017), but they require a significant annotation effort for achieving training data. Recently, Mehri and Eskenazi (2020) and Pang et al. (2020) suggested to use large pretrained language models (Liu et al., 2019; Radford et al., 2019) to develop reference-response-free metrics for dialogue evaluation. Such LMs are also the backbone of the QG, QA and NLI models employed in Q^2.

Factual Consistency and Hallucinations. Factual consistency in summarization has attracted increasing attention in recent years (Maynez et al., 2020) both in improving factual consistency of abstractive summarization systems (Cao et al., 2018) and in evaluating the factual consistency of generated summaries (Goodrich et al., 2019; Kryściński et al., 2019; Xu et al., 2020). Factual inconsistency has been observed in neural machine translation (Lee et al., 2019) mainly when considering out-

of-domain scenarios (Koehn and Knowles, 2017; Wang and Sennrich, 2020; Müller et al., 2020).

Concurrently with our work, Dziri et al. (2021) introduced the Benchmark for Evaluation of Grounded INteraction (BEGIN). BEGIN consists of WoW-based dialogue turns annotated for factual consistency with respect to the grounding knowledge. BEGIN models the task of evaluating groundedness as an NLI task and examples are annotated with five labels: entailment, contradiction, hallucination, off-topic and generic, where the last three are all considered to be neutral from an NLI perspective. Also relevant to our work, Rashkin et al. (2021) showed that faithfulness in knowledge-grounded dialogues can be improved by using controllable features based on NLI model predictions.

Evaluation via Question Answering and Question Generation. QA-based evaluation metrics have been proposed as a means for measuring content coverage in text generation tasks. For example, Eyal et al. (2019) used QA models for abstractive summarization both as an evaluation metric and as an optimization criterion that improved the downstream ROUGE scores by manually constructing questions around entities in the source document. These metrics aim at assessing whether key information from the input documents is expressed in the summaries (Recall-oriented). Durmus et al. (2020) and Wang et al. (2020) suggested using QG and QA to identify factual inconsistencies in abstractive summaries, which is more Precision-oriented. Their approach is based on the intuition that if a summary is consistent with its source, questions asked on the summary and the source should result in similar answers. Recently, Scialom et al. (2021) suggested QuestEval, which combines the Recall and Precision oriented QG and QA approaches, obtaining a more robust metric for evaluating abstractive summaries which was adopted in the GEM shared task (Bosselut et al., 2021). To overcome the low scores assigned by the token-level F1 measure to semantically-identical answers that are lexically different, they use a measure of the QA confidence of answerability (Scialom et al., 2019), which is the complement of the probability that the QA model gives to the "no answer" prediction. This measure reflects the answerability independently of the way the answer is expressed, but does not take into account possible model hallucinations, and it is therefore only applied for the Recall-based component. Our suggested NLI-based answer comparison allows lexical variability in the Precision-based component as well.

Comparing to other automatic evaluation methods of abstractive summaries, the QG-QA based methods showed higher correlations with human judgments of factual consistency. To the best of our knowledge, our work is the first to apply a QG-QA approach for evaluating dialogue generation.

7 Conclusion and Future Work

We presented Q^2, an automatic evaluation method for factual consistency in knowledge grounded dialogue. Q^2 employs question generation, question answering and NLI models, and does not require reference responses. To test our approach, we compiled a dataset of dialogue responses from two systems on the Wizard of Wikipedia dataset, which we annotated for factual consistency. Extensive experiments on this dataset, as well as on the Topical-Chat and DialogueNLI datasets, present strong results for Q^2 against various baselines. In future work we would like to map parts of a response to different types like chit-chat, persona and factual, in order to evaluate each against its appropriate source of truth. Other directions for future research are to apply Q^2 in additional tasks where factual consistency is essential, such as automated fact-checking (Thorne and Vlachos, 2018), and to use its evaluation signal to improve the factual consistency of generation models as proposed by Rashkin et al. (2021) or Nan et al. (2021).

Acknowledgements

This work was carried out as part of a Master Sponsored Research Agreement between the Hebrew University and Google, and was also supported by a gift from Google. Or Honovich was partially supported by a fellowship from the Hebrew University Center for Interdisciplinary Data Science Research.

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*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Transla-*

tion and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Antoine Bosselut, Esin Durmus, Varun Prashant Gangal, Sebastian Gehrmann, Yacine Jernite, Laura Perez-Beltrachini, Samira Shaikh, and Wei Xu, editors. 2021. *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Association for Computational Linguistics, Online.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, W. Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *AAAI*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The begin benchmark.

B. Efron. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82:171–185.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. 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 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *KDD '19*, KDD '19, page 166–175, New York, NY, USA. Association for Computing Machinery.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Yvette Graham and Qun Liu. 2016. Achieving accurate conclusions in evaluation of automatic machine translation metrics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–10, San Diego, California. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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 *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical Report*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online. 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, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Thomas Scialom, Paul-Alexis Dray, Gallinari Patrick, Lamprier Sylvain, Piwowarski Benjamin, Staiano Jacopo, and Wang Alex. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers

unite! unsupervised metrics for reinforced summarization models. 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 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470, Online. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Xinnuo Xu, Ondřej Dušek, Jingyi Li, Verena Rieser, and Ioannis Konstas. 2020. Fact-based content weighting for evaluating abstractive summarisation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5071–5081, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Online. Association for Computational Linguistics.

Data	Model	Coverage	Score
dodeca inconsistent	Q^2	92.67%	0.238
	-top-n	87.33%	0.265
	-filter personal	92.67%	0.243
dodeca consistent	Q^2	94%	0.696
	-top-n	85.33%	0.7
	-filter personal	90%	0.675
MemNet inconsistent	Q^2	94.67%	0.135
	-top-n	84.67%	0.153
	-filter personal	86%	0.139
MemNet consistent	Q^2	92.67%	0.756
	-top-n	85.33%	0.729
	-filter personal	88%	0.719

Table 9: Results for the ablations studies.

	Q^2	% no answer
Same dialogue	0.02	91.02%
Random dialogue	0	99.61%

Table 10: Results using randomly selected knowledge.

	Average # Characters	Average # Tokens
Inconsistent	70.84	15.79
Consistent	69.49	15.13
Random	69.44	15.86

Table 11: Average sentence length and average number of tokens per sentence in our collected dataset.

A Ablation Study

Table 9 presents the results of two ablations studies on Q^2. We show the scores obtained in these studies, as well as the question coverage, defined as the percentage of responses for which Q^2 generated at least one valid question, not resorting to the end-to-end NLI fallback.

First, we experiment with a different decoding strategy for generating questions. Instead of using beam search and taking the n top-ranked generated questions (see §2), we use greedy decoding, generating only one question per answer candidate. Next, we additionally drop the filtration of questions relating to personal statements and opinionated parts of the response.

Top-n Questions. Contrary to our expectations, When applying greedy decoding and taking a single question per an informative span, we inspect an increase for all data splits, except for the MemNet consistent responses. While the top-n decoding seems to be ineffective in terms of separating consistent responses from inconsistent responses, it is effective for improving the question coverage of Q^2.

Filtering Questions Relating to Personal Statements. As mentioned in §2, we filter questions that ask about personal statements expressed by the model. Examples of such questions are "What do I love?", which was generated given the text "I love cats" and the informative span 'cats'. Such text should not be evaluated for factual consistency and is allowed regardless of the knowledge. We report here the results for dropping this filtering step, on top of the previous experiment (applying greedy decoding). As Table 9 shows, when not removing such questions, scores are lower for all data splits. Naturally, the question coverage increases.

B Computing Infrastructure

We ran each experiment on 4 CPUs. For each data split (i.e., 150 responses), the runtime was $\sim 1.5-2$ hours. In future work, we plan to design a more efficient version of Q^2.

C Additional Experiments

Random Knowledge. We replace the knowledge k with randomly selected knowledge to test the sensitivity of our method to such adversarial cases. Two variants of knowledge selection are applied: In the first variant, we randomly select knowledge from the same dialogue, but from a different turn. In the second, we randomly select knowledge from a different dialogue. In both cases, we expect Q^2's score to be extremely low, as the knowledge should have little (in the first variant) to no (in the second variant) relation with r. Table 10 shows the results for using randomly selected knowledge; As expected, in both cases more than 91% of the generated questions had no answer in the knowledge, and this is more severe (99.6%) when using knowledge from a different dialogue.

Response Length. To test whether simple "surface markers" can differentiate consistent responses from inconsistent responses, we compare the average number of characters and the average number of tokens for responses in our dataset. As Table 11 shows, no strong differences were found for the *dodeca* system outputs. Similar results were obtained for the MemNet system.

D Additional Graphs

Figures 4 – 6 show the distribution of the response-level scores assigned by Q^2 and by the Overlap(r, k) baseline for the consistent and inconsistent data.

E Annotation Guidelines

[6] In this task, you will be presented with dialogues spanning various topics, conducted with a bot.

In each turn of the conversation, the bot was provided with a Wikipedia sentence relevant to the conversation topic and the current context of the conversation. The knowledge, or pieces of it, are integrated into the conversation.

Inconsistent responses collection You will be asked to detect bot responses that are *inconsistent* with the given knowledge. Such inconsistencies may include:

1. Information that was not at all mentioned by the knowledge.

2. Changes to the knowledge, resulting in information that was not expressed by it. Note that these changes may be subtle.

When marking a response as inconsistent, please:

1. Check if the response is clear and coherent. If not, ignore the response.

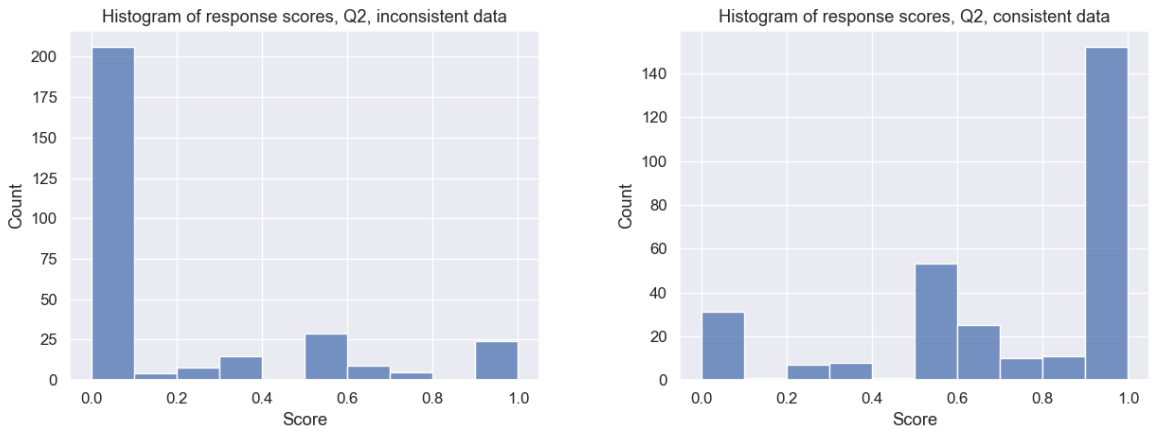
2. Ignore your background knowledge and focus on the information provided to the bot.

Consistent responses collection You will be asked to detect bot responses that are *consistent* with the given knowledge. When marking a response as consistent, please:

1. Check if the response is clear and coherent. If not, ignore the response.

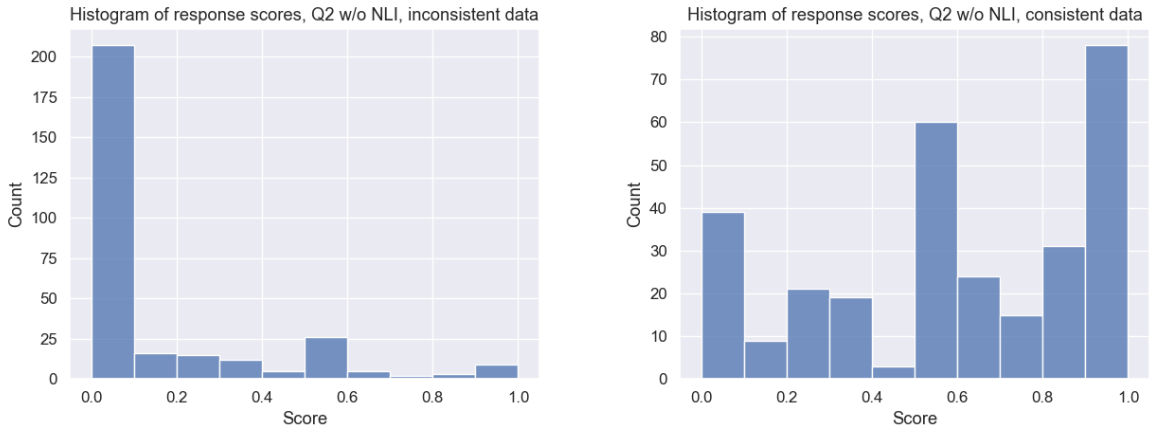
2. Select a response only if it uses the given knowledge. Ignore responses that are uninformative and only contain chit-chat.

[6] The guidelines are based on the insights provided by Durmus et al. (2020) regarding annotating faithfulness.



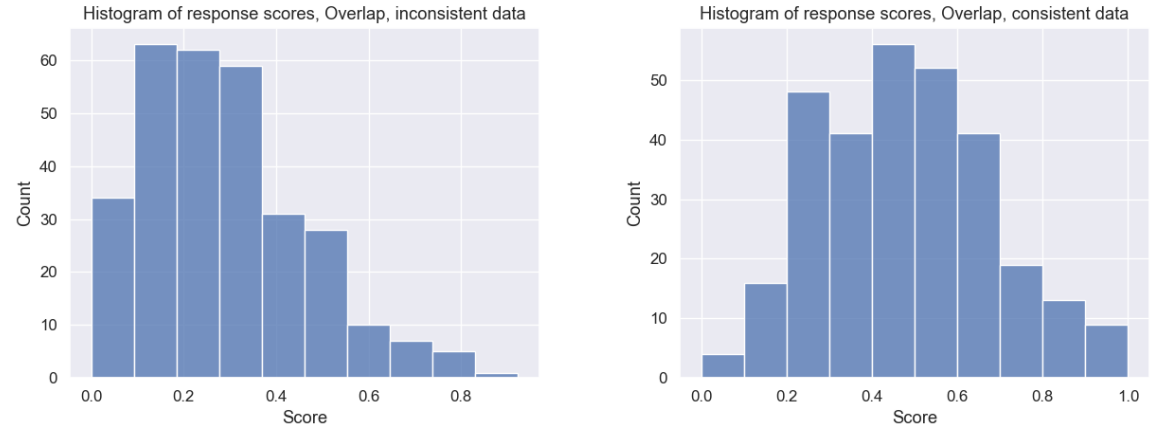
(a) (b)

Figure 4: Distribution of the response-level scores for Q^2. (a) Distribution for the inconsistent data. (b) Distribution for the consistent data.



(a) (b)

Figure 5: Distribution of the response-level scores for Q^2 w. token-matching. (a) Distribution for the inconsistent data. (b) Distribution for the consistent data.



(a) (b)

Figure 6: Distribution of the response-level scores for the overlap baseline. (a) Distribution for the inconsistent data. (b) Distribution for the consistent data.