Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark

Nouha Dziri[*†♢] **Hannah Rashkin**[*§] **Tal Linzen**[§♣] **David Reitter**[§]

[♢]University of Alberta, Canada [§]Google Research, USA [♣]New York University, USA
dziri@cs.ualberta.ca {hrashkin,linzen,reitter}@google.com

Abstract

Knowledge-grounded dialogue systems powered by large language models often generate responses that, while fluent, are not *attributable* to a relevant source of information. Progress towards models that do not exhibit this issue requires evaluation metrics that can quantify its prevalence. To this end, we introduce the Benchmark for Evaluation of Grounded INteraction (BEGIN), comprising 12k dialogue turns generated by neural dialogue systems trained on three knowledge-grounded dialogue corpora. We collect human annotations assessing the extent to which the models' responses can be attributed to the given background information. We then use BEGIN to analyze eight evaluation metrics. We find that these metrics rely on spurious correlations, do not reliably distinguish attributable abstractive responses from unattributable ones, and perform substantially worse when the knowledge source is longer. Our findings underscore the need for more sophisticated and robust evaluation metrics for knowledge-grounded dialogue. We make BEGIN publicly available at https://github.com/google/BEGIN-dataset.

1 Introduction

Neural language models (Bengio et al., 2000; Vaswani et al., 2017; Radford et al., 2019, inter alia) often form the backbone of open-ended dialogue systems (Wolf et al., 2019; Zhang et al., 2020b; Roller et al., 2021; Adiwardana et al., 2020). Utterances sampled from such language models sound natural, as reflected in these systems' high scores in human evaluations focused on measures such as ''engagingness'' or ''human-likeness'' (See et al., 2019). While fluent, however, the responses generated by these systems often contain statements that are not supported by the evidence available to the system; such statements are sometimes referred to informally as ''hallucinations'' (Tian et al., 2019; Maynez et al., 2020a; Dziri et al., 2021; Shuster et al., 2021; see Figure 1 for an example). This issue is particularly salient for knowledge-grounded dialogue systems, which are expected to interact with a user in an open-ended fashion while conveying information that is *attributable* to external identifiable sources. In this work, we develop a benchmark that can be used to assess attribution in knowledge-based dialog systems; following Rashkin et al. (2021a), we define an attributable response[1] as one connected to textual evidence that supports the entirety of the response.

A number of modeling approaches have recently been proposed to increase attribution in knowledge-grounded dialog systems (Rashkin et al., 2021b; Shuster et al., 2021; Dziri et al., 2021, 2022a). Progress in this area crucially relies on metrics that can measure the attribution of the text generated by the system; and indeed, recent work has developed automated metrics with relatively high correlations with human annotations, potentially paving the way for alternatives to expensive human evaluations (Honovich et al., 2021; Dziri et al., 2021, 2022a). Yet our understanding of these recently proposed metrics, as well as more established ones, remains limited, for two reasons. First, comparisons between automated metrics and human judgments rely on small-scale datasets with a few hundred examples. This results in high variance in our estimate of the correlation coefficient and a limited ability to measure performance on infrequent example types (Gehrmann et al., 2021).

Second, the correlation with human scores does not sufficiently determine the efficacy and

[*]Equal contribution.
[†]Work done while at Google Research.

[1]Attribution is sometimes referred to as faithfulness (Cao et al., 2018; Durmus et al., 2020, inter alia).

Transactions of the Association for Computational Linguistics, vol. 10, pp. 1066–1083, 2022. https://doi.org/10.1162/tacl_a_00506
Action Editor: Hang Li. Submission batch: 1/2022; Revision batch: 6/2022; Published 9/2022.

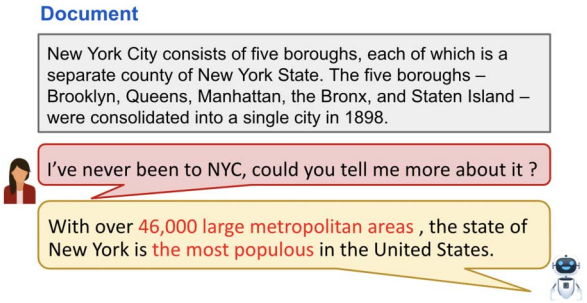


Figure 1: An example of a response generated by the GPT2 language model fine-tuned on the Wizard of Wikipedia dataset (Dinan et al., 2019). The phrases in red are ''hallucinations'' unsupported by the background document.

robustness of automatic metrics produced by neural networks: such learned metrics—like other properties learned by neural networks—can be susceptible to spurious correlations that fail to generalize to more challenging cases. To address these limitations, we introduce a large-scale resource, the Benchmark for Evaluation of Grounded INteraction (BEGIN), for meta-evaluation of metrics designed to evaluate grounded dialogue. In other words, the goal of this benchmark is to determine to what extent current evaluation metrics fulfill their purpose.

We define a taxonomy dividing knowledge-grounded dialogue responses into three broad categories—*fully attributable*, *not fully attributable*, and *generic*—and ask humans to classify a large set of utterances produced by dialogue systems with this taxonomy. The motivation for the *generic* category we introduce—which is assigned to utterances such as ''*Sorry, I'm not sure about this topic*''—is the intuition that evaluation metrics should not treat the basic elements of a natural-sounding conversation, such as backchanneling or acknowledgment (Grice, 1989; Stiles, 1992; Bunt et al., 2020), as equally undesirable as a misleading unattributable statement. In real-world scenarios, it is preferable for a model to acknowledge its ignorance instead of producing hallucinated content which may lead to the spread of disinformation.

Using this taxonomy, we then collect high-quality human annotations for 12k examples generated by four language-model-based dialogue systems, each trained on three different knowledge-grounded dialogue corpora. Examples of machine-generated responses along with labels are presented in Table 1. We use this benchmark to evaluate multiple existing automatic metrics including word-overlap measures, embedding-based measures, metrics based on Question Answering (QA) systems, and ones based on Natural Language Inference (NLI). We also propose a classifier trained on an adversarially generated dataset we create. We find that all metrics inadequately measure attribution and all rely on spurious correlations to a large extent. In particular, the metrics tend to misidentify cases that are attributable but highly abstractive, as well as cases that are not fully attributable but use multiple words from the evidence document (i.e., unattributable but extractive). We also find that the metrics fail to measure attribution under distribution shift, scoring responses that pertain to relatively long knowledge sources the lowest. These results are in line with the robustness issues reported for other natural language generation metrics, despite the high correlation of those metrics with human judgments (Durmus et al., 2022; Gehrmann et al., 2021; Gabriel et al., 2021; Yeh et al., 2021). We hope that BEGIN will facilitate progress toward more robust metrics for grounded dialogue response generation.

2 Task, Datasets, and Models

In knowledge-grounded response generation, the system is given a dialogue history $\mathcal{H} = (u_1, \ldots, u_{n-1})$, and knowledge $\mathcal{K}_n = (k_1, \ldots, k_j)$ at turn n, and is expected to generate a response $\bar{u}_n$ that is coherent with $\mathcal{H}$ and attributable to a non-empty subset $M_n \subset \mathcal{K}_n$. Similar to the conversational QA task (Choi et al., 2018; Reddy et al., 2019), the system is expected to use knowledge to respond to the user query. However, since the previous utterance may be an open-ended statement rather than a direct question (see the second and third examples in Table 1), there is a wider range of possible types of informative replies compared to the conversational QA task.

BEGIN consists of responses generated by language-model-based systems trained to perform this task. This section describes the models we train on this task and the corpora we use to train them.

2.1 Dialogue Datasets

For all three datasets, we use the training portion to train the model, the development set to tune

	Category	Frequency	Example	
WoW	Fully Attributable	23.5%	Knowledge:	**Raised in Compton, California, Lamar embarked on his musical career as a teenager** under the stage name k-dot, releasing a mix-tape that garnered local attention and led to his signing with indie record label top dawg entertainment.
			History:	oh, yeah. I believe I've heard about that. The album has to be good then for sure. Do you know anything else about him or rap?
			T5 Response:	Oh yes, I know that the singer was raised in Compton, California and started on his musical career as a teenager.
CMU-DoG	Generic	6.4%	Knowledge:	Throughout the film, the narrative is intercut with scenes from depositions taken in the Winklevoss twins' and Saverin's respective lawsuits against Zuckerberg and Facebook. . . . After everyone leaves, Zuckerberg is shown sending a friend request to Albright on Facebook and then refreshing the webpage every few seconds as he waits for her response. . .
			History:	The ending seems cools. Like after all of that happens he just goes to hang out with his kids.
			GPT2 Response:	What else do you know about Zuckerberg?
TopicalChat	Not Attributable	70.1%	Knowledge:	Television (TV) is a telecommunication medium used for transmitting moving images in monochrome (black and white) .. . . Ringo Starr, George Carlin, and Alec Baldwin have all narrated Thomas the Tank Engine for at least 52 episodes each. There are over 750 episodes of ''Pokémon'': That's about 50 a year for 16 years straight.
			History:	no, I haven't watched Pokemon. My sons use to play with when they were little. Now, they use their phones looking for Pokemon in the neighborhood. It is funny.
			DoHA Response:	Pokemon appeared in TV in 1952 and it was created by Ringo Starr and directed by George Carlin. In total, there were 100 episodes.

Table 1: Examples of each of the three categories of responses included in BEGIN. For each category, we provide an example drawn from one of the four models trained on one of the three corpora (of course, all 12 models generated all three types of responses). The dialogue corpus used to train the model and generate the response is listed vertically. Text highlighted in green indicates information that is attributable to the knowledge; text in blue does not convey any information; and text in red is hallucinated and cannot be attributed to the knowledge.

hyperparameters, and the test set to generate the responses that are then annotated and included in the final BEGIN benchmark.

Wizard of Wikipedia (WoW) WoW dialogue (Dinan et al., 2019) takes place between a Wizard and an Apprentice. The Wizard is tasked with providing information about a particular topic and the Apprentice, in turn, is expected to seek more information. At each turn of the conversation, the Wizard is presented with passages from Wikipedia and chooses a span from the document—typically one or two sentences—that serves as evidence supporting their response. We omitted examples where the Wizard did not explicitly select a passage as evidence for the response or where there was no dialogue history. We also use the ''unseen'' topic portion of the test data. Overall, we

used 82722 training examples, 8800 development examples, and 3902 test examples.

CMU-DoG The CMU-DoG dataset (Zhou et al., 2018) consists of conversations about films. Each response is expected to be grounded in a section from Wikipedia. Workers can have either asymmetric or symmetric roles. In the asymmetric setting, one worker is asked to persuade the interlocutor to watch the movie using arguments from the document where only the persuader has access to the document. In the symmetric role, workers discuss together the content of the document. In total, there are 78136, 13800, and 13796 grounded responses (training/dev/test).

TopicalChat TopicalChat (Gopalakrishnan et al., 2019) consists of dialogues about a variety of topics. Workers are provided relevant facts from Reddit, Wikipedia, and news articles. Analogous to CMU-DoG, the data collection protocol consists of two scenarios. In the symmetric scenario, workers have access to the same knowledge source; in the asymmetric scenario, they have access to different sources. They are asked to use the information from the documents to chat knowledgeably about the topic. In total, the dataset has 134572, 8790, and 8081 grounded responses (training/dev/test).

2.2 Dialogue Models

We consider the outputs of four different dialogue systems; by selecting a relatively wide range of systems, we hope to encounter a range of attribution errors. Two of the systems are based on plain language models, GPT2-base (Radford et al., 2019) and T5-base (Raffel et al., 2020). The remaining two systems, DoHA (Prabhumoye et al., 2021) and CTRL-DIALOG (Rashkin et al., 2021b), are specifically designed as knowledge-grounded dialogue systems. DoHA augments a BART-based conversational model (Lewis et al., 2020) with a two-view attention mechanism that handles the encoded document and the dilaogue history separately during generation. CTRL-DIALOG augments T5-base with control tokens (Keskar et al., 2019) that guide the generation towards less subjective and more grounded content. We trained these models to generate responses based on a concatenation of two inputs: an evidence span (the knowledge snippet) and the dialogue history (we only use the previous turn u_{n-1}).

3 Annotations

We next describe the human annotations we collected for the utterances generated by the models described in Section 2.

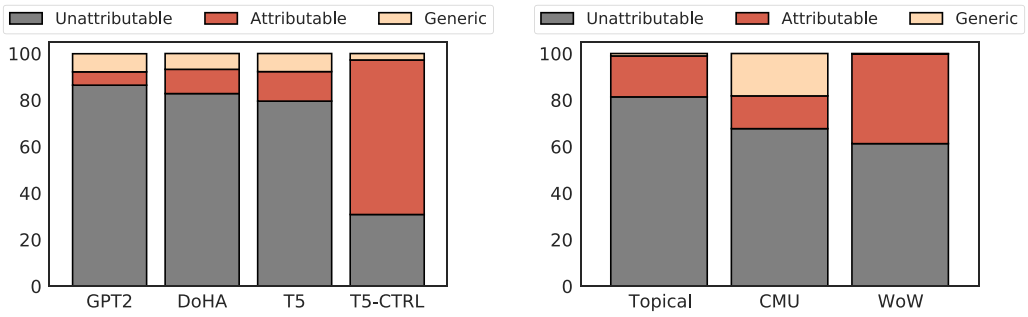
3.1 Taxonomy of Response Types

We classify responses into three broad categories:

Fully Attributable These are responses that convey information that can be completely supported by the provided document; this property has been referred in the literature to as faithfulness (Rashkin et al., 2021b; Maynez et al., 2020b; Dziri et al., 2021; Durmus et al., 2020) and attribution (Rashkin et al., 2021a). In our annotation set-up, we use similar definitions to the Attributable to Identifiable Source (AIS) framework of Rashkin et al. (2021a). The full framework in that paper consists of a two-stage annotation process in which annotators first filter out responses that are deemed to be too vague or ill-formed to be evaluated for attribution. Since Rashkin et al. (2021a) found that more than 90% of the conversational responses in their study were interpretable, we have our annotators focus solely on attribution.

Not Attributable These are responses that contain at least some information that cannot be verified given the evidence, regardless of whether that information is factually true in the real world. This includes statements that are relevant but not fully supported by the background information (hallucinations), statements that explicitly contradict the background information, and off-topic responses about information completely external to the evidence sources. In a pilot study we attempted to separate these three subcategories, but the boundaries between them turned out to be difficult to define and annotate.

Generic Responses that fall into this category are general enough to fit into a large number of possible contexts (Li et al., 2016). Examples include ‘‘I don’t know about that’’ and ‘‘Hello there!’’. Even when the responses are ostensibly about the same topic as the document, they are vague and do not provide new information. Nevertheless, such responses may be useful for various conversational purposes: back-channeling, expressing uncertainty, or diverting the conversation from ambiguous or controversial topics.



(a) Breakdown by model class.

(b) Breakdown by dialogue corpus used for training.

Figure 2: Breakdown of BEGIN response categories across models (left) and training corpora (right).

3.2 Collecting Prompt-Query-Reply Triples

As described in Section 2, we collect data using outputs from four models—T5, GPT2, DoHA, and CTRL-DIALOG. We train a version of each model on each of the three datasets (WOW, TOPICALCHAT, and CMU-DOG) and generate responses using the test portion of the dataset. For more details on training and hyperparameters, refer to Appendix B. We select at least 1000 examples from each dataset-model pair. We filter and remove toxic responses using the Google Perspective API. This yields 12288 examples in total.

3.3 Annotating Prompt-Query-Reply Triples

We present annotators with a knowledge snippet $\mathcal{K}$, the previous turn u_{n-1} and a generated response $\bar{u}_n$, and ask them to select which of the three categories fits $\bar{u}_n$ best. For the exact annotation instructions, see Appendix A. To obtain high quality data, we assign three annotators to each example and report results based on majority vote. We exclude examples where each of the three annotators assigned a different category, making it impossible to compute a majority vote.

Annotation Quality To ensure that the annotators understood the task, we use the following manual quality control procedure. In the first stage, we train the annotators by running two pilot annotation batches ($\sim$100 examples each). After each batch, we manually grade the answers for compliance with instructions, and provide feedback explaining any misconceptions. After the training stage, we launch the main annotation round for the full set of 12k examples. During this round, we intermittently check responses after every 3k completed annotations to examine the annotation quality. This procedure resulted in high inter-annotator agreement (a Krippendorff's alpha of 0.7).

3.4 Dataset Analysis

BEGIN is intended as a test benchmark; as such, it does not have a training portion: We only create development (10%) and test (90%) partitions. We include examples from BEGIN in Table 1 along with the label breakdown. Overall, the models generated a substantial number of unattributable responses (70%). As Figure 2(a) shows, this proportion was higher for GPT2, DOHA, and T5, whereas CTRL-DIALOG generated the lowest proportion of unattributable responses (30.8%). This indicates that CTRL-DIALOG, which is explicitly designed to discourage unattributable responses, is moderately successful at its goal. Figure 2(b), which breaks the results down by training corpus, shows that models trained on TOPICALCHAT produce the highest amount of unattributable responses followed by CMU-DOG and WOW. This is consistent with recent analyses on WOW, CMU-DOG, and TOPICALCHAT which revealed that more than 60% of the ground-truth responses are unattributable to the knowledge (Dziri et al., 2022b; Rashkin et al., 2021a).

3.5 The Need to Measure Attribution

Our analysis of the responses produced by the systems we trained highlights the potential pitfalls of language-model-based dialogue systems, especially when deployed in real-world scenarios across a broad range of domains where hallucinations pertaining to vital information may produce undesirable user experiences—e.g., healthcare (Laranjo et al., 2018; Jovanović et al., 2021) and education (Yang and Evans, 2019; Kochmar et al., 2021)—and underscores the need for progress on

both the modeling and the evaluation side. Neural dialogue systems are optimized to mimic the distributional properties of the human-generated dialogue corpus used to train them. Because humans often include unattributable information in their utterances, language models trained on those corpora can replicate and perhaps even amplify the prevalence of unattributable responses at test time (Kang and Hashimoto, 2020; Dziri et al., 2022b). These findings call for robust evaluation metrics to uncover actionable insights about best practices of using such models and benchmarks. We hope that BEGIN will, as an evaluation benchmark, promote a strict standard for evaluation metrics, laying the ground for trustworthy dialogue systems.

4 Evaluating Evaluation Metrics

We next use BEGIN to evaluate a range of evaluation metrics. In §4.1 we list the untrained metrics we use as well as metrics trained on existing resources, and in §4.2 we describe a training set that we designed to train a classifier for the three response categories. We then describe the extent to which these metrics align with the BEGIN categories and analyze the metrics' robustness.

4.1 Metrics

Lexical Overlap Metrics This category includes n-gram-based metrics that compare the lexical similarity between the response $\bar{u}_n$ and the knowledge $\mathcal{K}$.[2] We consider BLEU-4[3] (Papineni et al., 2002), ROUGE-L[4] (Lin, 2004), and F1, which measures the word-level lexical overlap between $\bar{u}_n$ and $\mathcal{K}$.

Semantic Similarity Metrics These metrics compare the *semantic* similarity between $\bar{u}_n$ and $\mathcal{K}$. We consider BERTScore (Zhang et al., 2020a), which computes the similarity between $\bar{u}_n$ and $\mathcal{K}$ based on the cosine similarity of the sentence embeddings, as well as BARTScore (Yuan et al., 2021) and BLEURT (Sellam et al., 2020); for implementation details, see Appendix C.

Question-Based Metrics We use Q^2 (Honovich et al., 2021), which computes a factuality score through asking and answering questions. Given a candidate response as input, Q^2 generates a corresponding question and identifies potential answer spans in the knowledge source $\mathcal{K}$ that can justify the question–answer pair (Durmus et al., 2020; Wang et al., 2020). It also computes an NLI-inspired similarity score between a candidate response and a predicted answer span in the knowledge source.

Inference-Based Metrics Finally, we study the performance of NLI-based models, trained either on gold NLI benchmarks or on adversarially augmented silver data that we generate. We first describe the metrics trained on gold NLI datasets; we discuss our adversarially augmented dataset (BEGIN-ADVERSARIAL) in §4.2. We use two transformer-based classifiers: T5-base (Raffel et al., 2020) and RoBERTa-large (Liu et al., 2019). We fine-tune them on MNLI (Williams et al., 2018) and the dialogue inference dataset DNLI (Welleck et al., 2019a). For both datasets, we map the labels (entailment, contradiction, neutral) to the labels (attributable, unattributable, generic) in BEGIN.

We also train classifiers on AugWow (Gupta et al., 2022), a synthetic dataset designed to evaluate factuality in dialogue systems. This dataset includes three categories: *Supported* responses that are fully verified by $\mathcal{K}$, *Refuted* responses that explicitly contradict $\mathcal{K}$, and responses with *Not Enough Information* (NEI), which do not contain enough information to be verified or refuted by $\mathcal{K}$. We map the labels (supported, refuted, NEI) to the labels (attributable, unattributable, generic) in BEGIN.

4.2 Adversarially Augmented Training Set

This section describes our curated silver training set (BEGIN-ADVERSARIAL) for NLI-based attribution classifiers. This dataset includes 8k ($\mathcal{K}$, $\mathcal{H}$, u_p) triples that fit into the three categories: attributable, generic, and unattributable.

Attributable Here we use the original human generated responses u_g from WOW. To avoid human responses that contain opinions or generic chit-chat, we only use response that do not use first-person pronouns and where at least 25% of the words in the response are contained in the evidence.

[2] Note that we do not compare the generated responses to the gold responses as they may be unattributable (Sec 3.4).

[3] https://github.com/mjpost/sacrebleu.

[4] https://github.com/google-research/google-research/tree/master/rouge.

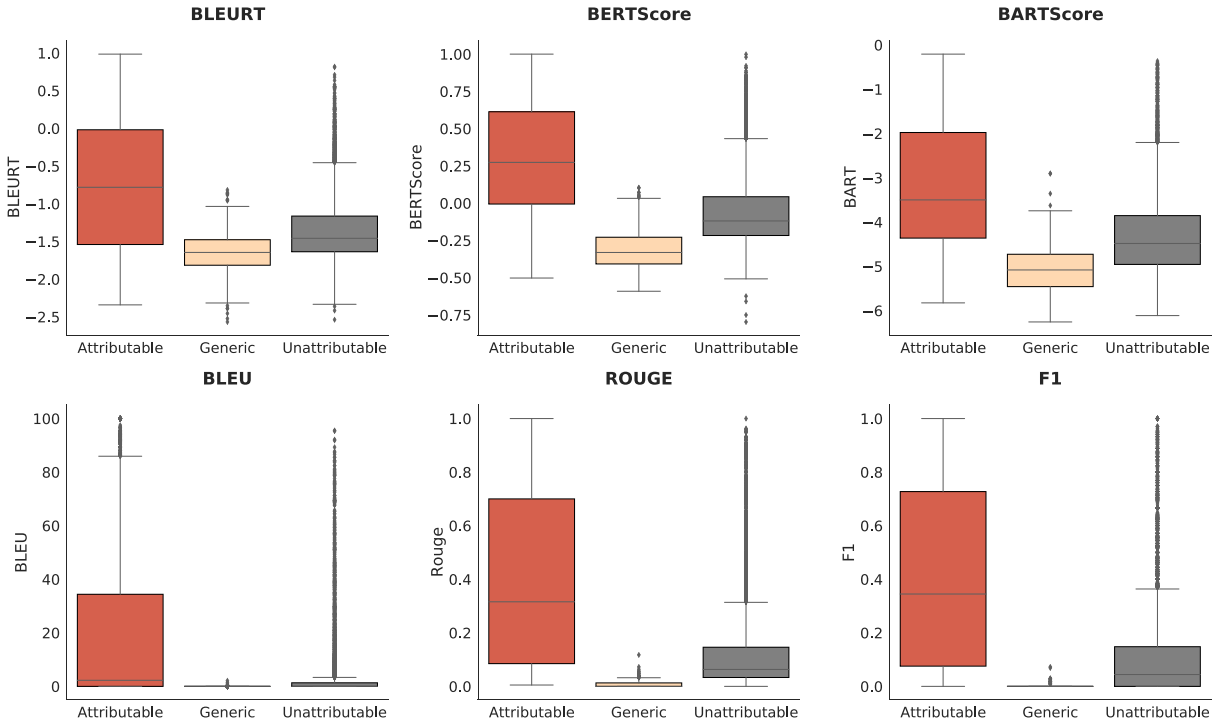


Figure 3: The distribution of scores assigned by semantic similarity metrics (upper row) and lexical overlap scores metrics (lower row) to the BEGIN test set.

Unattributable To generate examples that are likely to be unattributable, but are sufficiently challenging to distinguish from attributable ones as to be useful in training a classifier, we use multiple perturbation strategies. First, we directly perturb the knowledge spans $\mathcal{K}$ from the WOW test set and then feed them to GPT2 trained on WOW. We use three perturbation methods, each applied to a different $\mathcal{K}$. First, we swap the subject and the object of $\mathcal{K}$. Second, we replace up to two verbs with verbs of the same tense. Finally, we extract all mentioned entities from different dialogue examples using the SpaCy NER tagger (Honnibal et al., 2020), and replace up to two randomly chosen entities in the original $\mathcal{K}$ with entities of the same type. Manual inspection reveals that this usually results in responses that are hallucinations with respect to the original $\mathcal{K}$.

We also generate responses designed to specifically contradict $\mathcal{K}$, using two techniques. First, we directly negate the human response u_g from WOW using the English Resource Grammar parser (ERG; Flickinger et al., 2014). Second, we replace adjectives in u_g with their WordNet antonyms (Miller, 1994).

Lastly, we gather responses that are off-topic with respect to the information in the $\mathcal{K}$. For a given context, we randomly select a WOW gold response that was based on different $\mathcal{K}$. To avoid easy-to-detect off-topic responses, we sample from conversations that were prompted by the same initial topic word as the target conversation.

Generic Generic responses are generated from the GPT2 model we trained on WOW, using a low softmax temperature of 0.4.

4.3 Results

In this section, we report the performance of automatic metrics on the BEGIN test set.

Lexical and Semantic Metrics The distribution of scores is shown in Figure 3. For all metrics, the median score of fully attributable responses is higher than that of generic and unattributable responses, as expected. In many individual cases, however, unattributable responses are scored quite highly, and there is some overlap in the distribution of scores across all three labels, particularly between generic and unattributable responses, indicating that it would be impossible to map these score ranges directly to the BEGIN label taxonomy. Higher scores do not always translate into more desirable response types: Even though a generic response would typically be preferable to an unattributable one in a knowledge-grounded

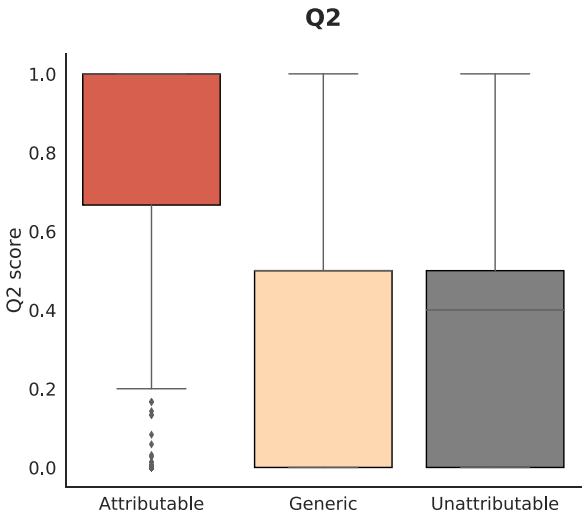


Figure 4: The distribution of Q^2 scores for each of the three example categories in the BEGIN test set.

	Test set			Dev set		
Finetuning data	P	R	F1	P	R	F1
T5						
MNLI	48.6	47.9	34.6	52.1	50.7	37.4
DNLI	40.8	56.5	25.6	41.6	59.2	28.6
AugWow	36.8	39.8	37.8	36.7	39.9	38.1
BEGIN-Adv.	46.7	47.4	**45.9**	47.2	47.1	**46.3**
+MNLI	46.9	49.3	45.3	47.6	49.4	46.1
RoBERTa						
MNLI	50.5	51.1	36.4	52.3	53.8	38.5
DNLI	40.2	46.6	27.2	34.9	46.1	29.2
AugWow	41.2	39.2	29.7	29.4	41.4	29.1
BEGIN-Adv.	42.6	46.1	41.1	49.2	45.8	41.1
+MNLI	44.8	45.9	**45.2**	44.9	45.6	**45.1**
Human	96.4	–	–	97.2	–	–

Table 2: Precision, recall, and F1 of the classifier-based metrics created by fine-tuning T5 and RoBERTa on NLI datasets, AugWow and our adversarial training set. Scores are macro-averaged across labels on the BEGIN test and dev sets.

dialogue system, the median scores are lower for generic responses than unattributable ones.

Q^2 Figure 4 shows a box plot for each BEGIN class using the Q^2 metric. As in the case of the lexical and semantic metrics, Q^2 scores are typically higher for attributable responses but indistinguishable between generic and unattributable responses.

Inference-Based Classifiers Table 2 reports the performance of the NLI-based classifiers on BEGIN. BEGIN-ADVERSARIAL substantially outperforms the classifiers trained on the gold datasets MNLI, DNLI, and AugWoW even though it is a significantly smaller resource than those datasets. We also use MNLI as an intermediate fine-tuning dataset before fine-tuning on BEGIN-ADVERSARIAL.[5] We find that intermediate task fine-tuning can be beneficial when RoBERTa is used as the pretrained model ($\uparrow 4.1$ on F1).

Overall, our adversarially generated dataset provides better supervision for detecting our taxonomy than NLI-style datasets. This can be attributed to the fact that NLI-style datasets are designed with a focus on detecting direct contradictions. By contrast, identifying unattributable responses requires detecting multiple types of unverifiable information including, but not limited to, contradictions. At the same time, none of the models exceed 46% F1 score, showing that there is still room for improvement compared to human performance (over 95% precision when comparing human annotations to the majority vote). Finally, T5 and RoBERTa have similar F1 scores despite differences in model size and pretraining corpora, suggesting that simply scaling up the pretrained model may not be sufficient to make progress on this problem.

4.4 Are Metrics Measuring Attribution or Extractivity?

Do the metrics perform similarly on both challenging and easier examples? We adopt a density metric from Grusky et al. (2018) to split the data into three groups—low, medium, and high density—based on the extent to which they reuse language from the knowledge sources. Density represents the average length of the text spans in the responses that are copied from the knowledge. Extractive (high density) responses reuse the same phrases as the knowledge source, whereas abstractive (low density) responses may express the same meaning using a paraphrase.

Results Figures 5 and 6 show the distributions across different levels of extractivity of the lexical and semantic metrics and the Q^2 score. We observe a common pattern across all metrics: high density responses for all categories (except *generic* on BLEURT) score the highest, followed by medium density and low density responses. The differences

[5] We did not observe a similar improvement when using DNLI as an intermediate task.

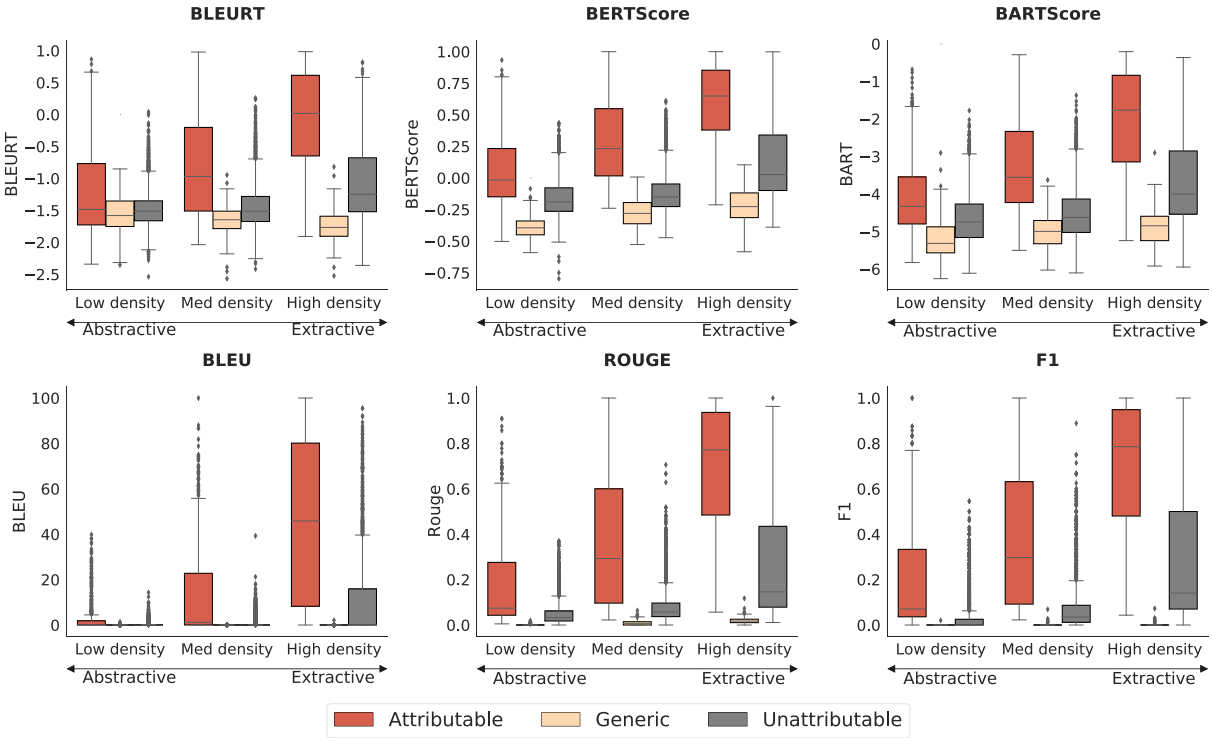


Figure 5: Scores assigned to each of the three BEGIN categories by semantic similarity metrics (upper row) and lexical overlap metrics (lower row), broken down by extractivity of the response (the extent to which it copies verbatim from the knowledge).

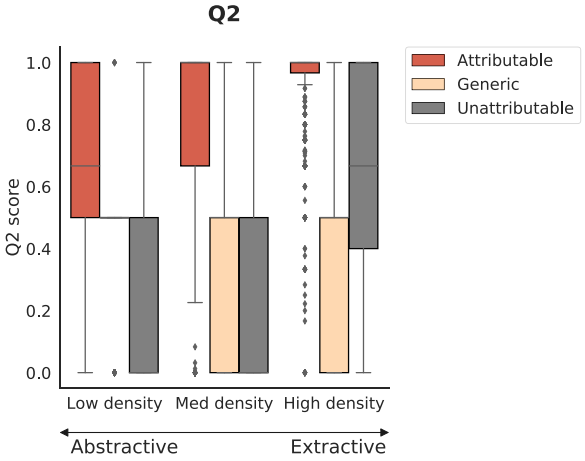


Figure 6: Q^2 scores across extractive and abstractive responses on BEGIN test.

between the scores of the attributable, generic and unattributable categories are more pronounced in the more extractive responses, and less in the abstractive cases. Only Q^2, though generally unable to separate generic examples, maintains a clear separation between attributable and unattributable examples in the abstractive cases. Moreover, extractivity strongly influences the score assigned to attributable examples; an attributable response is likely to be scored much lower by all of these metrics if it is abstractive. Even more strikingly, unattributable extractive responses score higher on average than attributable abstractive responses in all metrics.

We observe similar trends for the classifiers (Figure 7). The performance on classifying attributable responses is much higher in extractive cases than in abstractive ones. In contrast, the performance on unattributable responses is typically worse in the extractive cases. This pattern of results suggests that a response that is unattributable but has a high word overlap with the knowledge is very likely to be misclassified as attributable. In summary, we find that current metrics are relying on the spurious correlation between attribution and word overlap, and do not capture a deep understanding of the notion of attribution (cf. McCoy et al., 2019).

4.5 Robustness to Distribution Shift

We further investigate the robustness of the metrics under distribution shift. Figure 8 shows the distributions of both semantic and Q^2 scores across the data broken down by source. All

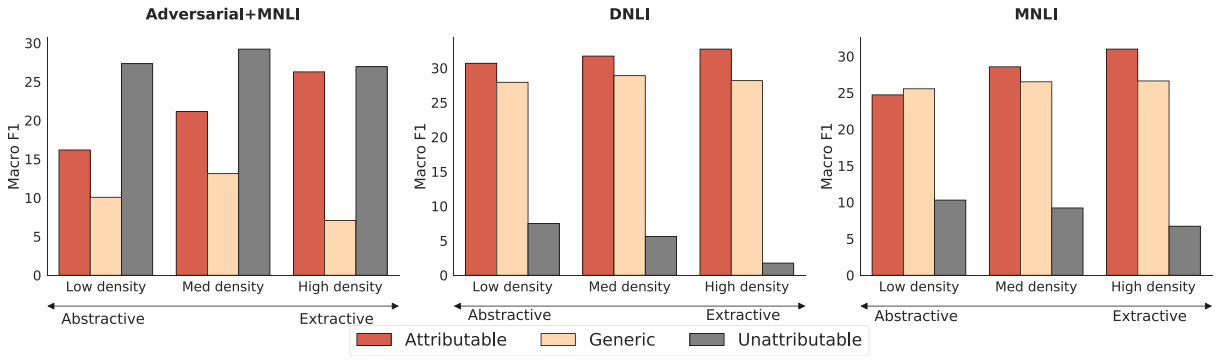


Figure 7: Comparison of F1 scores of RoBERTa-based classifiers on BEGIN categories with examples split by density (the extent to which the response copies verbatim from the knowledge).

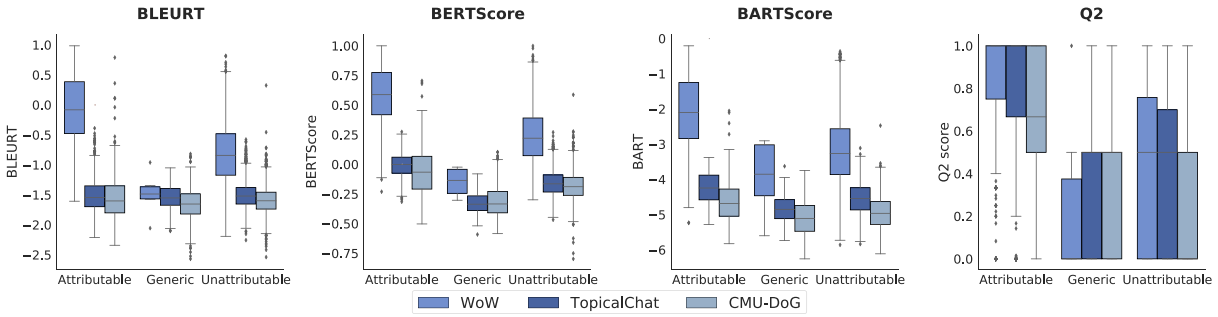


Figure 8: Scores of the semantic and Q^2 metrics across the three dialogue corpora we used to train our models.

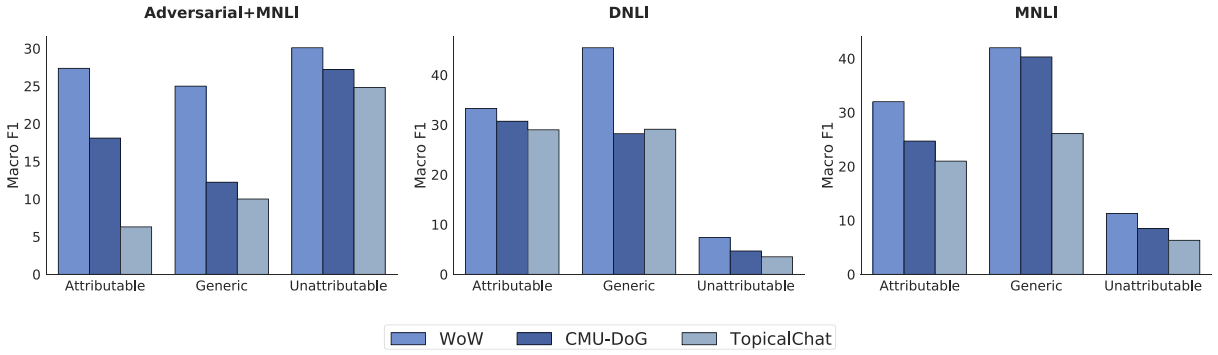


Figure 9: Comparison of F1 scores of RoBERTa classifiers on BEGIN categories with examples split by benchmark.

metrics[6] rate responses from WoW in all categories significantly higher than responses derived from CMU-DoG and TopicalChat. Concerningly, attributable responses generated based on CMU-DoG and TopicalChat receive nearly identical scores to unattributable responses. Likewise, the F1 scores of all the classifiers (Figure 9) are higher on the responses from WoW compared to the ones from CMU-DoG and TopicalChat. Specifically, classifiers tested on TopicalChat examples yield the worst F1 scores. For example, RoBERTA-MNLI's F1 score decreases by 10 points when tested on attributable responses from TopicalChat compared to WoW. In general, the metrics appear to perform poorly on datasets that have longer knowledge sources. TopicalChat has on average 271 words in $\mathcal{K}$, followed by CMU-DoG and WoW which have 215 words and 27 words, respectively. This shows that shorter knowledge spans correlates with higher metrics performance, pointing to the limited robustness of the metrics.

5 Related Work

Analysis of Evaluation Metrics in Natural Language Generation There is extensive interest in analyzing and meta-evaluating neural

[6]We observe similar results for lexical metrics.

language generation (NLG) evaluation metrics (Gehrmann et al., 2022, 2021), for various tasks including machine translation (Freitag et al., 2021; Mathur et al., 2020), data-to-text generation (Dhingra et al., 2019), summarization (Bhandari et al., 2020; Pagnoni et al., 2021; Durmus et al., 2020; Gabriel et al., 2021; Fabbri et al., 2021; Durmus et al., 2022), and dialogue generation (Yeh et al., 2021; Durmus et al., 2022). Most of these studies have compared reference-free and reference-based evaluation metrics to human evaluation. For example, Gabriel et al. (2021) measured the performance of automated metrics on summaries and compared certain dimensions such as sensitivity and high correlation with human scores. Fabbri et al. (2021) analyzed metrics in summarization and released human-annotated data for faithfulness across 16 summarization models. We perform a similar meta-evaluation of existing automatic metrics in the context attribution in knowledge-grounded responses. Closest to our work is Durmus et al. (2022), who found that reference-free evaluation metrics of summarization and dialogue generation rely heavily on spurious correlations such as perplexity and length.

Metrics in Knowledge-Grounded Response Generation In contrast to the significant progress achieved in evaluating many NLG tasks, the evaluation of grounded response generation is a nascent research area (Shuster et al., 2021; Rashkin et al., 2021a; Dziri et al., 2021). Yeh et al. (2021) conducted a comprehensive study of existing dialog evaluation metrics. They measured properties such as engagingness and relevance but did not investigate the faithfulness of responses. While hallucination is well studied in the context of summarization (Durmus et al., 2020; Maynez et al., 2020b; Nan et al., 2021; Falke et al., 2019), fewer researchers have looked into the problem of assessing hallucination in dialogue systems. Dziri et al. (2021) introduced a token-level critic that leverages a knowledge graph to identify hallucinated dialogue responses. Rashkin et al. (2021a) proposed a human evaluation framework to assess output of dialogue models that pertains to the external world and utilized their evaluation framework for conversational QA tasks. Dziri et al. (2022a) introduced a faithful benchmark for information-seeking dialogues and demonstrated that it can serve as training signal for a hallucination critic, which discriminates whether an utterance is faithful or not. An alternative approach for assessing faithfulness uses an auxiliary language understanding task, which measures whether a question answering system produces the same responses for the source document (Honovich et al., 2021). BEGIN as a testing benchmark should be useful in developing similar metrics further.

NLI and Adversarial Data for Grounded Dialogue Evaluation In this work, we also investigate the performance of classifiers trained on NLI data, extending prior work that has proposed using NLI as a framework for evaluating conversational consistency (Welleck et al., 2019b). Dziri et al. (2019) also used NLI to evaluate dialogue consistency. They generated a large-scale, noisy synthetic dataset of (premise, hypothesis) pairs tailored for dialogue, based on Zhang et al. (2018). We also explore training classifiers on adversarially augmented training data similar to concurrent work from Gupta et al. (2022) and Kryscinski et al. (2020), which proposed a synthetic dataset for determining whether a summary or response is consistent with the source document; this dataset was constructed by applying a number of syntactic transformations to reference documents (for a similar approach applied to NLI, see Min et al., 2020).

6 Conclusion

Contemporary knowledge-based dialogue systems that rely on language models often generate responses that are not attributable to the background knowledge they are expected to convey. We present BEGIN, a new benchmark to advance research toward robust metrics that can assess this issue. We use BEGIN to comprehensively evaluate a broad set of existing automatic metrics. We show that these metrics rely substantially on word overlap and fail to properly rank abstractive attributable responses as well as generic responses. They also struggle under distribution shift, assigning low scores to attributable responses grounded on long knowledge sources. We hope that this work will spur future research on building robust evaluation metrics for grounded dialogue systems.

Acknowledgments

We are grateful to the anonymous reviewers for helpful comments. We thank Dipanjan Das, Vitaly

Nikolaev, Sebastian Gehrmann, Roee Aharoni, Jennimaria Palomaki, Tom Kwiatkowski, Michael Collins, and Slav Petrov for helpful discussions and feedback. We also thank Ashwin Kakarla and his team for helping with the annotations.

A BEGIN Annotation Protocol

Each worker was given a document, previous turn in a conversation, and a generated response (either by T5, GPT2, DoHA, or CTRL-DIALOG). They were asked to evaluate the response as either fully attributable, not attributable, or too generic to be informative. They also were provided with multiple examples with explanations for each category. The exact instructions were as follows:

> Which of these best describes the highlighted utterance?
>
> - Generic: This utterance is uninformative (too bland or not specific enough to be sharing any new information)
> - Contains *any* unsupported Information: This utterance is sharing information that cannot be fully verified by the document. It may include false information, unverifiable information, and personal stories/opinions.
> - *All* information is *fully* supported by the document: This utterance contains only information that is fully supported by the document.

B Implementations

GPT2, T5 We implement these models using the TensorFlow Huggingface Transformers library (Wolf et al., 2020). During training, we use the Adam optimizer (Kingma and Ba, 2015) with Dropout (Srivastava et al., 2014) on a batch size of 32 with a learning rate of 6.25×10^{-5} that is linearly decayed. The maximum dialogue history length is set to 3 utterances. The model early-stops at epoch $\{6, 10, 10\}$ respectively for WoW, CMU-DoG, and TOPICALCHAT.

CTRL-DIALOG We reproduce the results from (Rashkin et al., 2021b), following the training details in that paper.

DoHA We use the code and the pre-trained model on CMU-DoG that are publicly available by the authors at their Github account.[7] For WoW and TOPICALCHAT, we follow closely the authors' training procedure described in Prabhumoye et al. (2021) and we train two models on both datasets.

For each dataset, we save the best model based on the validation set. We use nucleus sampling with $p = 0.9$.

[7] https://bit.ly/3bBup2M.

C Model-Based Metrics

Semantic Similarity Models We use BERTScore version 0.3.11. with the DeBERTa-xl-MNLI model (He et al., 2021), which is the recommended model as of the time of investigation. For BLEURT, we use the recommended BLEURT-20 checkpoint (Pu et al., 2021). For BARTScore, we use the latest publicly available checkpoint (accessed March 2022) from https://github.com/neulab/BARTScore.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR (arXiv preprint)*, abs/2001.09977.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.751

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 549–558, Marseille, France. European Language Resources Association.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*,

32(1). https://doi.org/10.1609/aaai.v32i1.11912

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1241

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1483

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net.

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. https://doi.org/10.18653/v1/2020.acl-main.454

Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. 2022. Spurious correlations in reference-free evaluation of text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1443–1454, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.102

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. 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 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1381

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022a. Faithdial: A faithful benchmark for information-seeking dialogue. *CoRR (arXiv preprint)*, abs/2204.10757.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.168

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaïane, and Siva Reddy. 2022b. On the origin of hallucinations in conversational models: Is it the datasets or the models? *CoRR (arXiv preprint)*, abs/2204.07931.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409. https://doi.org/10.1162/tacl_a_00373

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1213

Dan Flickinger, Emily M. Bender, and Stephan Oepen. 2014. ERG semantic documentation. Accessed on 2020-08-25.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context:

A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474. https://doi.org/10.1162/tacl_a_00437

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.42

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.gem-1.10

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *CoRR (arXiv preprint)*, abs/2202.06935.

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 *Proceedings of Interspeech 2019*, pages 1891–1895. https://doi.org/10.21437/Interspeech.2019-3079

Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1065

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.263

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. SpaCy: Industrial-strength natural language processing in Python. 10.5281/zenodo.1212303.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.619

Mlađan Jovanović, Marcos Baez, and Fabio Casati. 2021. Chatbots as conversational healthcare services. *IEEE Internet Computing*, 25(3):44–51. https://doi.org/10.1109/MIC.2020.3037151

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.66

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR (arXiv preprint)*, abs/1909.05858.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Ekaterina Kochmar, Dung Do Vu, Robert Belfer, Varun Gupta, Iulian Vlad Serban, and Joelle Pineau. 2021. Automated data-driven generation of personalized pedagogical interventions in intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, pages 1–27. https://doi.org/10.1007/s40593-021-00267-x

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.750

Liliana Laranjo, Adam G. Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y. S. Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258. https://doi.org/10.1093/jamia/ocy072

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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.

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. *CoRR (arXiv preprint)*, abs/1907.11692.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.448

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020a. 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. https://doi.org/10.18653/v1/2020.acl-main.173

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020b. 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. https://doi.org/10.18653/v1/2020.acl-main.173

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1334

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8–11, 1994*. https://doi.org/10.3115/1075812.1075938

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

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)*, pages 6881–6894, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.536

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.383

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W. Black, and Ruslan Salakhutdinov. 2021. Focused attention improves document-grounded generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4274–4287, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.338

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.58

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

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.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021a. Measuring attribution in natural language generation models. *CoRR (arXiv preprint)*, abs/2112.12870.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021b. 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. https://doi.org/10.18653/v1/2021.acl-long.58

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266. https://doi.org/10.1162/tacl_a_00266

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.24

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. 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 1702–1723, Minneapolis, Minnesota. 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. https://doi.org/10.18653/v1/2020.acl-main.704

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.320

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

William B. Stiles. 1992. *Describing Talk: A Taxonomy of Verbal Response Modes*. Sage Publications.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *CoRR (arXiv preprint)*, abs/1910.08684.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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. https://doi.org/10.18653/v1/2020.acl-main.450

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019a. 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. https://doi.org/10.18653/v1/P19-1363

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019b. 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. https://doi.org/10.18653/v1/P19-1363

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge

corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR (arXiv preprint)*, abs/1901.08149.

Shanshan Yang and Chris Evans. 2019. Opportunities and challenges in using AI chatbots in higher education. In *Proceedings of the 2019 3rd International Conference on Education and E-Learning*, ICEEL 2019, pages 79–83, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3371647.3371659

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

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. https://doi.org/10.18653/v1/P18-1205

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.30

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1076