Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation

Mingkai Deng[1*], **Bowen Tan**[1*], **Zhengzhong Liu**[1,2], **Eric P. Xing**[1,2,3], **Zhiting Hu**[4]
[1]Carnegie Mellon University, [2]Petuum Inc., [3]MBZUAI, [4]UC San Diego
([*]equal contribution)
{mingkaid,btan2,liu,epxing}@andrew.cmu.edu, zhh019@ucsd.edu

Abstract

Natural language generation (NLG) spans a broad range of tasks, each of which serves for specific objectives and desires different properties of generated text. The complexity makes automatic evaluation of NLG particularly challenging. Previous work has typically focused on a single task and developed individual evaluation metrics based on specific intuitions. In this paper, we propose a unifying perspective that facilitates the design of metrics for a wide range of language generation tasks and quality aspects. Based on the nature of information change from input to output, we classify NLG tasks into compression (e.g., summarization), transduction (e.g., text rewriting), and creation (e.g., dialog). The *information alignment*, or overlap, between input, context, and output text plays a common central role in characterizing the generation. Using the uniform concept of information alignment, we develop a family of interpretable metrics for various NLG tasks and aspects, often without need of gold reference data. To operationalize the metrics, we train self-supervised models to approximate information alignment as a prediction task. Experiments show the uniformly designed metrics achieve stronger or comparable correlations with human judgement compared to state-of-the-art metrics in each of diverse tasks, including text summarization, style transfer, and knowledge-grounded dialog. With information alignment as the *intermediate representation*, we deliver a composable library for easy NLG evaluation and future metric design.[1]

1 Introduction

Natural language generation (NLG) refers to the broad set of tasks that produce fluent text from input data and other contextual information. The diverse tasks serve for vastly different uses in practice. For example, *summarization* compresses a source article into a short paragraph containing the most important information; *translation* transduces content expressed in one language into another; and a *chatbot* creates novel responses to drive the conversation. Recent years have seen remarkably fast progress in improving and making new models for NLG tasks. However, evaluation of NLG has long been considered difficult (Kryscinski et al., 2019; Mathur et al., 2020): human evaluation is often prohibitively expensive and slow, while accurate automatic evaluation is challenging given the complexity of text modeling and the diverse aspects to be measured for different NLG tasks.

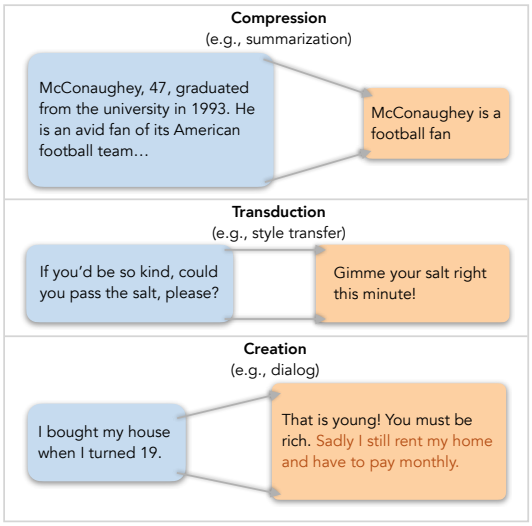


Figure 1: Illustration of three categories of NLG tasks in terms of information change. Task input is in blue box and output in orange box. Text in red in the dialog output box represents newly created information.

Previous work has developed a large variety of automatic metrics. A popular general strategy is to measure the similarity of generated text against human-written references, such as the classical BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and more recent variants based on neural models (e.g., Zhang et al., 2020a; Sellam et al., 2020).

[1]Code available at https://github.com/tanyuqian/ctc-gen-eval

However, an NLG task typically involves multiple desirable properties (e.g., consistency, conciseness, richness) that may have different priorities and need trade-off depending on the application scenarios (Hashimoto et al., 2019; Mir et al., 2019; Mehri and Eskenazi, 2020b; Gehrmann et al., 2021). Thus a single score without multi-aspect interpretability is often inadequate to characterize generation quality. A growing number of recent works have proposed aspect-based metrics for popular tasks such as summarization (Kryściński et al., 2019; Wang et al., 2020) and dialog (Mehri and Eskenazi, 2020b; Nie et al., 2020). Those metrics are typically each designed for individual tasks and aspects, based on specific intuitions. The lack of a common theoretical ground makes it difficult to share the evaluation strengths across the diverse NLG problems, and fails to offer guidance to metric design for emerging tasks and aspects.

In this paper, we propose a more unifying perspective of NLG evaluation through the lens of information change, which offers a general framework to measure many key aspects of NLG tasks. In particular, based on the practical use of NLG, each task can be seen as one of (1) *compression* to express salient information in concise text, such as summarization and image captioning; (2) *transduction* to transform text while preserving content precisely, such as translation and style transfer; and (3) *creation* to produce new content from input context, such as dialog and story generation. A common concept underlying the three broad categories is *information alignment*, which we define as the extent to which the information in one generation component is grounded in another. Here the generation components include input, output, additional context, and references when available.

Inspired by recent work on model-based evaluation, we adopt contextualized language models to measure information alignment. We then demonstrate the framework by devising a family of highly intuitive metrics for three representative tasks (*aspects*) in each category uniformly in terms of information alignment, including summarization (*relevance* and *consistency*), style transfer (*content preservation*) and knowledge-based dialog (*engagingness* and *groundedness*), respectively. Experiments show that the uniformly designed metrics robustly outperform or compete with state-of-the-art metrics specifically designed for each task, in terms of correlations with human judgement. We also study different implementations of the central information alignment estimation model, showing that improved alignment measure leads to better evaluation quality across all the tasks/aspects.

2 Related Work

Task- and Aspect-Specific NLG Evaluation. Canonical automatic evaluation (Papineni et al., 2002; Lin, 2004) often compute a single score measuring some forms of similarity between outputs and human-written references. The later-emerged learning-based approaches aggregate multiple features to regress on human-rated quality scores for different tasks (Lowe et al., 2017; Peyrard et al., 2017; Sellam et al., 2020). Researchers also identified that a single evaluation score cannot account for the variety of quality factors that exist in multi-faceted NLG applications. A number of metrics were then proposed for specific tasks, either to evaluate multiple aspects (Mehri and Eskenazi, 2020b; Egan et al., 2021) or to focus on one particular aspect (Kryściński et al., 2019; Mehri and Eskenazi, 2020a; Nie et al., 2020; Durmus et al., 2020; Wang et al., 2020). Our framework continues this line of research to produce interpretable metrics for multiple aspects. While recent evaluation frameworks each discussed the key evaluation aspects of one NLG task (Venkatesh et al., 2018; Mir et al., 2019; Yamshchikov et al., 2020; Fabbri et al., 2021), our framework provides a unified methodology that facilitates metric design for all the three main categories of tasks. We also highlight that all of metrics (except for the relevance metric for summarization) are reference-free once trained.

Several emerging NLG benchmarks (Gehrmann et al., 2021; Liu et al., 2021) collected existing metrics for various tasks, whereas we aim at developing new unified metrics with stronger performance. Belz et al. (2020) proposed a categorization for different NLG quality aspects. Our general framework covers all the described types of quality.

Text-to-Text Information Alignment. Measuring information overlap between texts is a recurring theme in designing NLG evaluation metrics. It has typically been approximated by n-gram overlap (Papineni et al., 2002; Popović, 2015), synonym matching (Banerjee and Lavie, 2005) and embedding similarities (Kusner et al., 2015). Recently, pre-trained models (Devlin et al., 2019) were introduced to improve token-level embedding matching (Zhang et al., 2020a) and leverage extrinsic

capabilities such as question answering (Eyal et al., 2019; Wang et al., 2020) and entailment classification (Falke et al., 2019; Kryściński et al., 2019; Zhou et al., 2020) to align variable spans and entire sentences. Egan et al. (2021) proposed automatic Shannon Game (Hovy and Lin, 1998) to measure the decrease of the information one can gain from a document after observing its summary; Peyrard (2019) conducted a theoretical analysis to characterize the information change among source document, background knowledge and summaries. These methods are often restricted to a single task, while we offer a general framework adaptable to a wide range of tasks and aspects.

3 A Unified Evaluation Framework

We present the new framework that offers a common foundation for characterizing diverse NLG tasks and leads to a set of interpretable metrics for evaluating their key aspects.

As discussed in §1, NLG tasks can be categorized as performing compression, transduction, or creation based on changes in conveyed information from input to output. For a **compression** task (e.g., summarization), the goal is to concisely describe the most important information in the input (e.g., a document). That is, the output should only contain content from the input, namely *"consistency"* (Cao et al., 2018; Kryscinski et al., 2019; Zopf et al., 2016; Peyrard, 2019), and the included content must be salient, namely *"relevance"* (Nenkova and Passonneau, 2004; Zopf et al., 2016). Intuitively, with an "information alignment" measure that assesses how the information in a generated output overlaps with that in the input (and in references that offer clues for salience), we can readily evaluate the two key aspects. The same intuition applies to **transduction** tasks (e.g., style transfer), where the output must preserve the input content precisely. The evaluation of *"preservation"* (Mir et al., 2019) thus also boils down to measuring the information alignment between input and output. A **creation** task (e.g., dialog) generates output that adds on top of input (e.g., dialog history) new information (e.g., from external knowledge). Information alignment between the output, input, and external sources is thus essential for evaluating how well the created content *engages* with the context (Venkatesh et al., 2018; See et al., 2019) and how meaningful the content is by *grounding* to the external sources (Dinan et al., 2019a; Smith et al., 2020).

From the above perspective, information alignment arises as a common central component that connects evaluations across the tasks. A single accurate alignment prediction model would enable us to reliably evaluate many relevant aspects in various applications.

Next, we first present our definition of information alignment (§3.1); then describe the details of how the aspect metrics for compression, transduction, and creation are built on the alignment (§3.2-3.4); we finally discuss different effective implementations of the underlying alignment estimation model based on neural networks (§3.5).

3.1 Preliminaries

For an NLG task, let $\mathbf{x}$ be the input, $\mathbf{c}$ be any other additional context, and $\mathbf{y}$ be the output text generated conditioning on $\mathbf{x}$ and $\mathbf{c}$. For example, in knowledge-based dialog, $\mathbf{x}$ is the dialog history, $\mathbf{c}$ is external knowledge such as a Wikipedia article, and $\mathbf{y}$ is the response. In the current work, we assume both $\mathbf{x}$ and $\mathbf{c}$ to be text, but the general framework is also applicable when $\mathbf{x}$ and $\mathbf{c}$ are in other modalities (e.g., images, tables), as long as we can measure their information alignment with $\mathbf{y}$ as defined below (e.g. using cross-modal models). In some tasks, gold standard output written by human is available, which we denote as $\mathbf{r}$.

As above, information alignment is the central module for NLG evaluation. We consider the alignment from arbitrary text $\mathbf{a}$ to $\mathbf{b}$ as token-level soft alignment. More formally:

Definition 3.1 (Information Alignment)**.** Let $\mathbf{a}$ be a piece of text of length N; $\mathbf{b}$ be arbitrary data. The information alignment from text $\mathbf{a}$ to $\mathbf{b}$ is a vector of alignment scores:

$$\textit{align}(\mathbf{a} \rightarrow \mathbf{b}) = \langle \alpha_1, \alpha_2, \ldots, \alpha_N \rangle, \quad (1)$$

where $\alpha_n \in [0, 1]$ is the confidence that the information of the n-th token in $\mathbf{a}$ is grounded by $\mathbf{b}$, i.e., the n-th token aligns with $\mathbf{b}$.

Note that the alignment is "one-directional" from $\mathbf{a}$ to $\mathbf{b}$: it does not measure how $\mathbf{b}$ aligns to $\mathbf{a}$. We next show how the alignment scores can be used to define intuitive metrics for various tasks. Besides, the fine-grained alignment scores also offer a certain level of interpretability for the resulting metrics, as illustrated by the example in Table C.1.

3.2 Evaluation of "Compression" Tasks

We discuss compression evaluation in the context of text summarization, an extensively studied task

for evaluation in previous work. The task aims to extract the most important information from document $\mathbf{x}$ and express it in summary $\mathbf{y}$. As above, *consistency* and *relevance* have been widely identified as key aspects to characterize the content quality of generated summaries (Cao et al., 2018; Kryscinski et al., 2019; Zopf et al., 2016; Peyrard, 2019). We propose our metrics below.

Consistency We adopt the prevailing definition of consistency (Cao et al., 2018; Kryscinski et al., 2019), which dictates that the summary $\mathbf{y}$ should only contain information from $\mathbf{x}$ (instead of other sources or hallucinations). The aspect is also referred to as "factual correctness" or "faithfulness" in previous work[2]. For $\mathbf{y}$ to be fully consistent, all tokens in $\mathbf{y}$ should align with $\mathbf{x}$. Therefore, we can straightforwardly devise the consistency metric based on the information alignment defined above:

$$\textsc{Consistency}(\mathbf{y}, \mathbf{x}) = \text{mean}\left(\textit{align}(\mathbf{y} \rightarrow \mathbf{x})\right), \quad (2)$$

which is the average alignment scores of tokens in $\mathbf{y}$ w.r.t. $\mathbf{x}$. Our metric offers a simpler solution than the recent QA-based metrics (Scialom et al., 2019; Durmus et al., 2020; Wang et al., 2020) that compare the answers extracted from $\mathbf{y}$ and $\mathbf{x}$ by a Question-Answering system, and is more interpretable than the black-box consistency classification models (Falke et al., 2019; Kryściński et al., 2019; Maynez et al., 2020). We also achieve stronger empirical performance (§4.1).

Relevance As one of the most heavily studied aspects of summarization, relevance concerns how well the summary $\mathbf{y}$ retains important information in $\mathbf{x}$ (Nenkova and Passonneau, 2004; Zopf et al., 2016). As in previous work, the "importance" of information can be determined by human-written reference summaries $\mathbf{r}$. That is, a piece of information is considered important if it is mentioned in a reference. The intuition can readily be captured by the information alignment $\textit{align}(\mathbf{r} \rightarrow \mathbf{y})$ that measures the extent to which information in reference $\mathbf{r}$ is covered by the summary $\mathbf{y}$. Additionally, we account for the criterion that any information in $\mathbf{y}$ should be precise, i.e., consistent with $\mathbf{x}$. Combining the two considerations, the full definition of our relevance metric conveys the intuition that a fully relevant summary $\mathbf{y}$ should *achieve both and balance* reference-alignment and consistency:

$$\textsc{Relevance}(\mathbf{y}, \mathbf{x}, \mathbf{r}) = \text{mean}\left(\textit{align}(\mathbf{r} \rightarrow \mathbf{y})\right) \times \text{mean}\left(\textit{align}(\mathbf{y} \rightarrow \mathbf{x})\right), \quad (3)$$

which is the product of both components. Traditional reference-based metrics consider only the reference text (rather than the input). For example, ROUGE (Lin, 2004) can be seen as measuring the alignment between $\mathbf{y}$ and $\mathbf{r}$ where the alignment is defined by text matching. Our metric, with the combination of both reference and input, plus better alignment modeling (§3.5), greatly outperforms those previous metrics (§4.1).

3.3 Evaluation of "Transduction" Tasks

We take style transfer as the example task to discuss semantic preservation of transduction tasks. The aim of style transfer is to generate text $\mathbf{y}$ that changes one or more stylistic attributes (e.g., formality) of source text $\mathbf{x}$ and completely preserve its style-independent information (Hu et al., 2017; Shen et al., 2017). Measuring content preservation is the core yet challenging problem for the evaluation.

Preservation A transduction result $\mathbf{y}$ is required to contain *all and only* information from $\mathbf{x}$. In other words, all tokens in $\mathbf{y}$ should align with $\mathbf{x}$, and *vice versa*. Considering the former to be the "precision" of the $\mathbf{y}$ information w.r.t $\mathbf{x}$, and the latter the "recall", we naturally arrive at the following "F1"-style definition of the preservation metric:

$$\textsc{Preservation}(\mathbf{y}, \mathbf{x}) = \frac{\text{mean}\left(\textit{align}(\mathbf{y} \rightarrow \mathbf{x})\right) \times \text{mean}\left(\textit{align}(\mathbf{x} \rightarrow \mathbf{y})\right)}{\text{mean}\left(\textit{align}(\mathbf{y} \rightarrow \mathbf{x})\right) + \text{mean}\left(\textit{align}(\mathbf{x} \rightarrow \mathbf{y})\right)}, \quad (4)$$

which is the harmonic mean of the two directions of information alignment. Note that the two-way alignments differ from the "consistency" and "relevance" metrics in compression where we have only required output $\mathbf{y}$ to align with input $\mathbf{x}$. Our experiments show that it is crucial to account for alignments in both directions for transduction (§4.2).

3.4 Evaluation of "Creation" Tasks

We formulate aspects of creation tasks using the example of knowledge-grounded dialog generation. In this task, an agent generates text $\mathbf{y}$ as a response to conversation history $\mathbf{x}$ while exhibiting information from knowledge context $\mathbf{c}$, e.g., an external document (Qin et al., 2019; Guo et al., 2018) or a set of facts (Dinan et al., 2019b; Zhang et al.,

[2]For the aspects studied in this paper, we summarize in Table B.1 the alternative names that used in previous work.

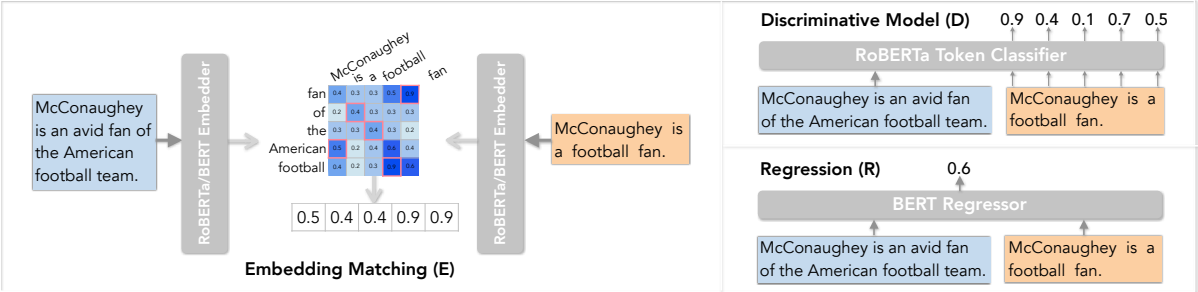


Figure 2: We study three effective ways of information alignment prediction, i.e., embedding matching (left), discriminative model (upper right) and regression (lower right). The figure illustrates the estimation of alignment from output to input.

2018). For the agent, sustaining an engaging conversation is considered an essential skill (Venkatesh et al., 2018; Guo et al., 2018; Mehri and Eskenazi, 2020b). Besides, the generated response must be grounded in the knowledge context by referring to its information as often as possible (Dinan et al., 2019a; Smith et al., 2020). We devise metrics for the two central aspects, respectively.

A crucial property of creation tasks is that the agent is allowed to create new information beyond the input and context. Thus, to aggregate the information alignment vector, it is more suitable to consider the *total volume* rather than the density. That is, we would use $\mathrm{sum}(\cdot)$ instead of the previous $\mathrm{mean}(\cdot)$ to aggregate token-level alignment scores.

Engagingness We adopt the common definition of engagingness (e.g., Mehri and Eskenazi, 2020b), namely, the response should not be generic or dull (e.g., *"I don't know"*), but engages the partner in conversation, such as presenting an interesting fact. Therefore, an engaging response $\mathbf{y}$ should provide high volume of information that acknowledges both the history $\mathbf{x}$ to engage the partner and the context $\mathbf{c}$ which we assume contains relevant facts. This naturally leads to the following metric definition:

$$\textsc{Engagingness}(\mathbf{y}, \mathbf{x}, \mathbf{c}) = \mathrm{sum}\left(\mathit{align}(\mathbf{y} \rightarrow [\mathbf{x}, \mathbf{c}])\right), \quad (5)$$

where we concatenate the history $\mathbf{x}$ and knowledge context $\mathbf{c}$, and measure the extent of response $\mathbf{y}$'s acknowledgement of the information. Previous works have devised various metrics for the aspect, ranging from measuring response-topic consistency (Guo et al., 2018), conversation length (Venkatesh et al., 2018), retrieval of reference responses (Mehri and Eskenazi, 2020b), etc. Our metric is cleanly defined in line with all other metrics we developed, and shows stronger human correlation than previous designs.

Groundedness As a widely studied aspect of knowledge-based dialog, groundedness measures how well the response refers to the knowledge context (Dinan et al., 2019b; Qin et al., 2019; Mehri and Eskenazi, 2020b). Straightforwardly, the aspect can be evaluated with the following metric:

$$\textsc{Groundedness}(\mathbf{y}, \mathbf{c}) = \mathrm{sum}\left(\mathit{align}(\mathbf{y} \rightarrow \mathbf{c})\right), \quad (6)$$

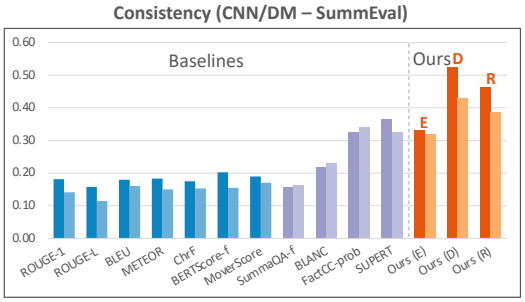
which measures the alignment between the response $\mathbf{y}$ and knowledge context $\mathbf{c}$.

3.5 Implementation of Alignment Estimation

We have presented the metrics for a range of key aspects in different tasks, building on the core information alignment measure (Definition 3.1). We next discuss different effective implementations for measuring the alignment scores between text, including embedding matching, discriminative model, and regression, all based on powerful pretrained language models (Figure 2).

Embedding Matching (E) One simple way to estimate the alignment vector $\mathit{align}(\mathbf{a} \rightarrow \mathbf{b})$ is by matching the embeddings of tokens in the two sequences. Specifically, we use either pretrained BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) to extract contextual embedding for each token in $\mathbf{a}$ and $\mathbf{b}$, normalize each embedding vector to unit norm, and then use greedy matching following (Corley and Mihalcea, 2005; Zhang et al., 2020a). That is, the alignment score of each token in $\mathbf{a}$ is defined as its maximum cosine similarity with the tokens in $\mathbf{b}$. We found in our empirical studies (§4) that the E method seems to work better when $\mathbf{a}$ and $\mathbf{b}$ have similar volume of information (so that one-to-one token matching is suitable).

Discriminative Model (D) To estimate the information alignment from arbitrary text $\mathbf{a}$ to $\mathbf{b}$, we formulate the problem as sequence tagging, for which we train a model that labels each token in $\mathbf{a}$



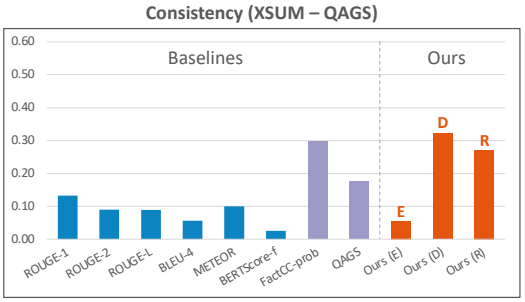


Figure 3: Correlations with human judgement on *consistency* in summarization. `E` denotes our metrics using embedding-matching alignment estimation, `D` using discriminative-model and `R` using regression. Reference-based metrics are in blue, reference-free metrics in purple, and our metrics in red/orange. For SummEval annotation data (Fabbri et al., 2021) **(left)**, we report Pearson (left, dark color) and Spearman (right, light color) correlations for each metric. For QAGS annotation data (Wang et al., 2020) **(right)**, only Pearson correlations were available for baselines.

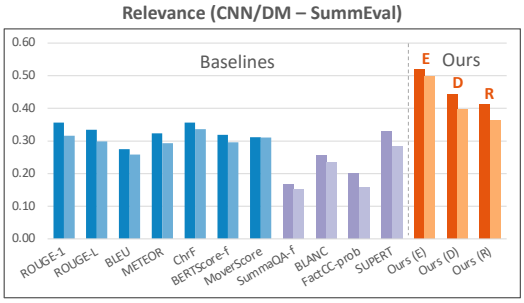


Figure 4: Correlations with human judgments on relevance in summarization. Reference-based metrics are computed using all 11 references for each example provided in the data. The plot format follows Figure 3.

Relevance	$\mathbf{r} \to \mathbf{y}$	$\mathbf{y} \to \mathbf{x}$	$+$	$\times$ (Ours)
Align (E)	0.4705	0.4381	0.5195	**0.5198**
Align (D)	0.4184	0.2834	0.4308	**0.4423**
Align (R)	0.4050	0.2688	0.3861	**0.4115**

Table 1: Ablation Studies: Pearson correlations for different variants of our *relevance* metric (Eq.3) using different components and combination strategies. $\mathbf{r} \to \mathbf{y}$ corresponds to $\text{mean}\,(align(\mathbf{r} \to \mathbf{y}))$ and similarly for $\mathbf{y} \to \mathbf{x}$; $+$ sums the two components and $\times$ is our design that takes the product.

with 1 if it aligns with $\mathbf{b}$, and 0 otherwise. The predicted probability of label 1 for each $\mathbf{a}$ token serves as the alignment score. We base our model on RoBERTa and train with automatically constructed weak supervision data. Appendix §A describes all details. For example, to learn to estimate the alignment of the output $\mathbf{y}$ to the input in an NLG task, we use the training corpus of the task: for each output $\mathbf{y}$, we perturb it by masking out portions of tokens and using a pretrained BART (Lewis et al., 2020) model to fill in the blanks (Zhou et al., 2020). The BART model is not conditioning on any input context (e.g., $\mathbf{x}$), so the infilled tokens can be considered to not align with the input. We do the masking by first applying constituency parsing to the text and then randomly masking out a subtree of the parsing. Besides the infilling data, we also augment the training with paraphrasing data. That is, we apply a paraphrasing model to $\mathbf{y}$, and treat all tokens in the paraphrases as alignment to the input. Note that $\mathbf{y}$ need not be the gold output, but can also be any automatically constructed output as long as it is guaranteed to align fully with the input. For example, an output $\mathbf{y}$ by an extractive summarization model aligns fully with the input article. We will see more examples in our experiments.

Aggregated Regression (`R`) Instead of estimating the per-token alignment vector as defined in Eq.(1), we may also directly estimate the single aggregated alignment score such as $\text{mean}\,(align(\mathbf{a} \to \mathbf{b}))$ (or sum). This is because all the metrics proposed above have only used the aggregated score. To this end, we train a regression model using the same weak supervision data for `D`, with the aggregated alignment score as the regression target. Similar to Sellam et al. (2020), in our experiments, we implement the regression model with BERT (Devlin et al., 2019). In particular, we initialize the regression model with the intermediate `BERT-base-midtrained` model weights provided by Sellam et al. (2020). We note that the aggregated estimation method may not be applicable to future metrics in our evaluation framework when fine-grained per-token alignment is required.

4 Experiments

We evaluate the proposed metrics on commonly used human annotation datasets for summarization (§4.1), style transfer (§4.2) and dialog (§4.3), and study the effect of information alignment accuracy on the performance of metrics (§4.4).

Evaluation Criteria To measure a metric's performance on an aspect, we compute the sample-level correlation between the metric scores and human judgments on generation samples. We also evaluate system-level correlation (based on the

ranking of comparison systems) as the secondary criterion (Mathur et al., 2020) and report results in the appendix, which typically exhibits the same patterns as sample-level correlation. We measure Pearson and Spearman correlations whenever applicable. We also report Kendall-Tau correlation in the appendix when available.

4.1 Experiments for "Compression" Metrics

Datasets For the *consistency* aspect, we follow previous studies and evaluate metrics using human annotations from two commonly-used sources: **(1)** SummEval (Fabbri et al., 2021) on the CNN/DM summarization dataset (Hermann et al., 2015; Nallapati et al., 2016). The annotation dataset contains 1,600 examples from 16 summarization systems; **(2)** QAGS (Wang et al., 2020) (which names the aspect "correctness") on the XSUM dataset (Narayan et al., 2018), another summarization task with strong abstractive property. The dataset contains 235 outputs from a fine-tuned BART model (Lewis et al., 2020). The QAGS dataset also contains another 239 outputs for CNN/DM, for which we report results in Table D.4 in the appendix.

For *relevance*, we test our metric on the respective annotations from SummEval on CNN/DM.

Baselines and Setup For baselines, we include commonly-used metrics reported in previous papers, ranging from reference-based metric, such as ROUGE, BLEU and BERTScore (Zhang et al., 2020a), to reference-free ones, such as SummaQA (Scialom et al., 2019) based on QA and FactCC (Kryściński et al., 2019) based on sentence classification. For our metrics, we use `RoBERTa-large` for the embedding-matching (`E`) alignment since it was pre-trained on the CommonCrawl News dataset (Nagel, 2016) that is close to the summarization domains. For the discriminative-model (`D`) alignment, we train two `RoBERTa-large` token classifiers to compute $\textit{align}(\mathbf{y} \rightarrow \mathbf{x})$ and $\textit{align}(\mathbf{r} \rightarrow \mathbf{y})$, respectively, with training data automatically constructed for CNN/DM and XSUM according to Appendix §A.1. For the regressive (`R`) alignment, we train the `BERT` models (§3.5) to estimate the respective mean alignment scores.

Results We present the consistency results in Figure 3. On CNN/DM, our metrics based on the trained alignment models (`D` and `R`) both clearly outperform previous metrics. On XSUM, our `D`-based metric also achieves the best performance.

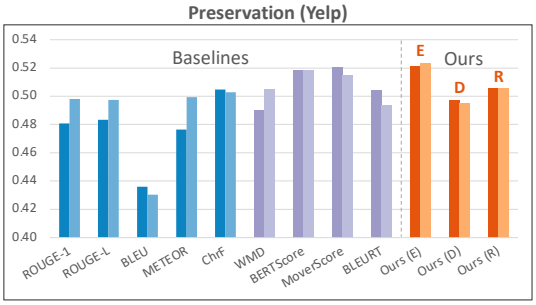


Figure 5: Human correlations on style transfer *preservation* aspect. Lexical-matching metrics are in blue. Embedding- or model-based similarity metrics are in purple, and ours are in red/orange.

Preservation	$\mathbf{y} \rightarrow \mathbf{x}$	$\mathbf{x} \rightarrow \mathbf{y}$	$\mathbf{y} \Leftrightarrow \mathbf{x}$ (Ours)
Align (E)	0.4989	0.5078	**0.5216**
Align (D)	0.4481	0.4608	**0.4974**
Align (R)	0.4744	0.4823	**0.5060**

Table 2: Ablation Studies: Pearson correlations for variants of *preservation* metric (Eq.4) accounting for different directions of information alignment. $\mathbf{y} \rightarrow \mathbf{x}$ corresponds to mean $(\textit{align}(\mathbf{y} \rightarrow \mathbf{x}))$ and similarly for $\mathbf{x} \rightarrow \mathbf{y}$; Our $\mathbf{y} \Leftrightarrow \mathbf{x}$ is harmonic mean of alignments in both directions.

The `E`-based metric sees a catastrophic drop in correlations, which is likely due to the higher abstractiveness of XSUM summaries that renders embedding matching inadequate. The sentence-classifier based `FactCC` metric (Kryściński et al., 2019), which is trained to distinguish paraphrases from artificially perturbed sentences, also achieves a decent correlation on XSUM. However, it seems unable to effectively model the summaries on CNN/DM that tend to be longer and richer in information, and thus produces a lower correlation.

Figure 4 shows the results for relevance on CNN/DM. Our metrics strongly outperform all other baselines, showing that accounting for alignments with both references and the input article (Eq.3) is superior to only considering the references (metrics in blue in the figure) or the input article (metrics in purple). This is further validated by the ablation studies in Table 1, which demonstrate that multiplying the two alignments, which emphasizes joint and balanced achievement of both, improves the correlations compared to individual alignments or simply summing them together. Figure 4 also shows our `E`-based implementation performs better than the `D`- and `R`-based variants, likely because the metric involves alignment between generation and references which tend to have similar information volume and thus favor one-to-one token mapping. We observe similar patterns in transduction below.

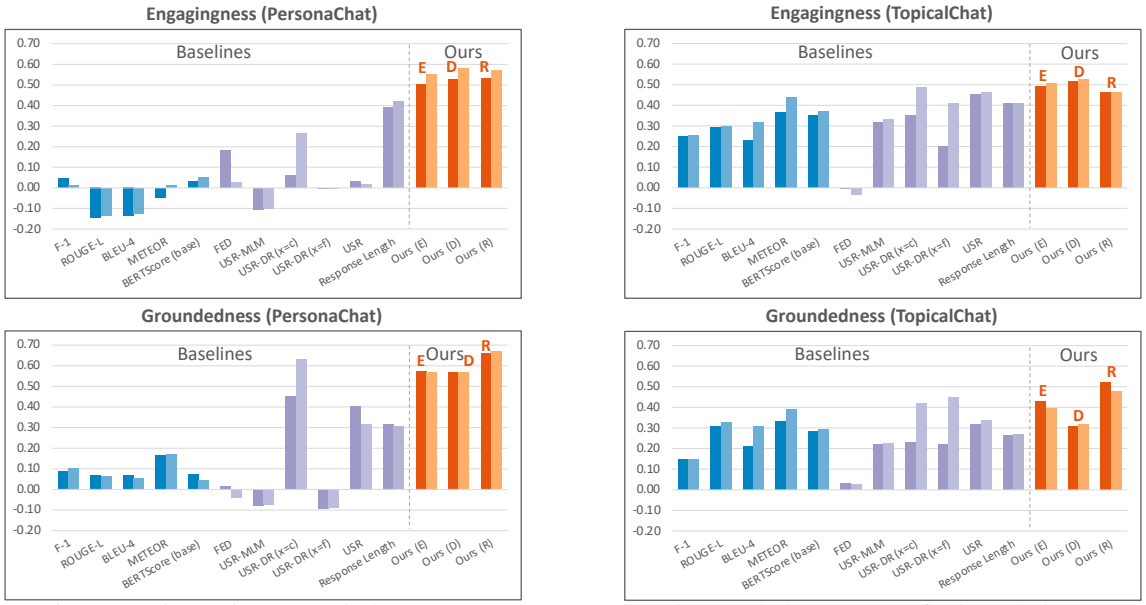


Figure 6: Correlations with human judgement on *engagingness* and *groundedness* aspects for knowledge-grounded dialog. The plot format is the same as Figure 3.

4.2 Experiments for "Transduction" Metrics

Datasets We apply our *preservation* metric to evaluating text style transfer, where we use the human annotations from (Mir et al., 2019) on the Yelp sentiment transfer data (Shen et al., 2017)[3]. The dataset contains 8,784 outputs from 12 systems.

Baselines and Setup We compare with the common metrics used previously (Mir et al., 2019), and further include BERTScore (Zhang et al., 2020a), MoverScore (Zhao et al., 2019) and BLEURT (Sellam et al., 2020), the latest neural text similarity metrics. We use BLEURT out-of-the-box as we do not assume access to human scores for fine-tuning the evaluation models. For our metrics, we use `RoBERTa-large-MNLI` for embedding-matching (`E`) due to its fine-tuning on entailment detection which is close to the domain under study. For discriminative model (`D`), we train `RoBERTa-large` on Yelp alignment data created by paraphrasing and perturbing the inputs $\mathbf{x}$. For regression (`R`), we train to estimate the mean alignment score computed from the same dataset as `D`.

Results We present preservation results in Figure 5. Our metric (`E`) achieves competitive or better performance than all previous metrics. MoverScore (Zhao et al., 2019) as a strong baseline computes word mover's distance (Kusner et al., 2015) between input $\mathbf{x}$ and output $\mathbf{y}$ token embeddings. In contrast, our metric explicitly accounts for the two-way input-output alignments with an "F1"-style harmonic mean aggregation (Eq.4). Table 2 shows the two-way approach is effective and exhibits higher correlation compared to single-directional alignment, in line with the nature of transduction tasks. Similar to their relevance results in summarization, our `D`- and `R`-based implementations fall behind `E`, likely because token matching is more suitable for measuring alignments between two text pieces with similar information volume.

[3]It is arguable whether "sentiment" is part of style (Krishna et al., 2020). Here we just use the most common dataset.

Engagingness	Mean		Sum (Ours)	
	P	T	P	T
Align (E)	0.1502	0.3184	**0.5003**	**0.4937**
Align (D)	0.1821	0.3223	**0.5265**	**0.5163**
Align (R)	-0.0490	-0.0191	**0.5320**	**0.4653**

Table 3: Ablation Studies: Pearson correlations for our engagingness metric (Eq.5) with different alignment aggregation strategies. "Mean" takes the average of the alignment vector, and "Sum" is our designed metric that takes the sum. "P" and "T" refer to PersonaChat and TopicalChat, respectively.

4.3 Experiments for "Creation" Metrics

Datasets For the *engagingness* aspect, we use the latest human annotation data collected by (Mehri and Eskenazi, 2020b) (which names the aspect "interesting") on PersonaChat (Zhang et al., 2018) and TopicalChat (Gopalakrishnan et al., 2019), two knowledge-grounded dialog tasks with different forms of knowledge. The dataset contains 300 examples from 5 systems for PersonaChat, and 360 examples from 6 systems for TopicalChat. All turns preceding the current response $\mathbf{y}$ are treated as the history $\mathbf{x}$ (4.2 turns on average for PersonaChat and 5.1 turns for TopicalChat). The knowledge context

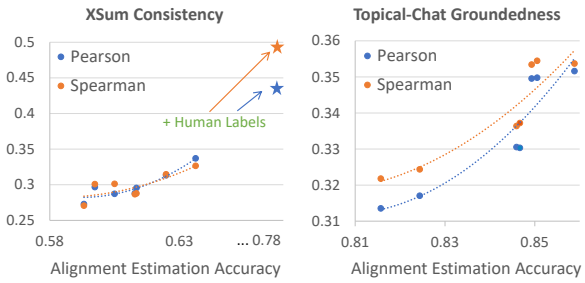


Figure 7: Effect of alignment estimation accuracy on metric performance.

$\mathbf{c}$ refers to the persona statements in PersonaChat and the knowledge snippets in TopicalChat.

For the *groundedness* aspect, we again use the human annotations from Mehri and Eskenazi (2020b) (which names the aspect "uses knowledge") on both PersonaChat and TopicalChat.

Baselines and Setup We compare with all the diverse metrics studied in (Mehri and Eskenazi, 2020b) and FED (Mehri and Eskenazi, 2020a), a set of latest unsupervised dialogue metrics based on the DialoGPT model (Zhang et al., 2020b). We use FED-Interesting from the original paper designed for engagingness and FED-Informative designed for groundedness, respectively. We also add a particularly simple baseline—response length, which as we show performs surprisingly well. For our metrics, we use `BERT-base` for embedding matching (`E`), `RoBERTa-large` token classifiers trained on $align(\mathbf{y} \rightarrow [\mathbf{x}, \mathbf{c}])$ and $align(\mathbf{y} \rightarrow \mathbf{c})$ for discriminative model (`D`), and `BERT-base` regressors on the sums of the respective alignment scores for regression (`R`). We create separate alignment datasets for PersonaChat and TopicalChat, as described in Appendix A.3.

Results We present the results for engagingness in the top two plots of Figure 6. Our metrics with different implementations all improve over previous methods by large margins on the two datasets. Many of the baseline metrics show decent correlations on TopicalChat, but fail on the PersonaChat corpus. This is likely because PersonaChat requires strong dependency of responses on the dialog history and knowledge context, thus metrics that do not directly model the dependency (e.g., USR-DR (Mehri and Eskenazi, 2020b) based on response retrieval) as ours struggle for accurate evaluation.

Noticeably, the simple `response length` performs consistently well on both datasets, far better than previous metrics on PersonaChat. The baseline can be considered as a special case of ours by setting alignment scores of all tokens to 1. The stronger correlations of our model-based metrics demonstrate the effect of accurate alignment.

Ablation studies in Table 3 shows that measuring the volume (sum) instead of the density (mean) of aligned information is crucial for the superior performance of our metrics, highlighting the unique characteristics of the "creation" task (§3.4).

The results for groundedness are shown in the bottom two plots of Figure 6. Our metrics again generally achieve strong correlations, with the `R`-based metric consistently outperforming other implementations, likely because the estimation of grounded information *volume* (sum) benefits from the expressivity of end-to-end models. This is indicated by the underperformance of the `D`-based metric, which is trained on the same data but aggregates token-level predictions with more structure.

We provide more empirical studies in Appendix §F. In particular, we found that besides the two core aspects, our alignment based method also achieves stronger human correlations than existing metrics on other dialog aspects, such as the *understandability* and *naturalness* of responses (Table F.6).

4.4 Ablation: higher alignment estimation accuracy, better correlation

We study how the accuracy of information alignment estimation influences the performance of metrics. We demonstrate a highly desirable pattern that higher alignment estimation accuracy can usually lead to better correlation. This indicates that improvement on the single alignment estimation model could immediately benefit a broad range of aspect metrics defined in our unified framework.

Specifically, we use the discriminative model (§3.5) for our study. First, we vary the number of training iterations to get different model checkpoints, and evaluate both the alignment estimation accuracy and the metric human correlation based on the checkpoints. We evaluate accuracy with the human-annotated token alignment labels on the XSUM summarization data Maynez et al. (2020). Figure 7 (left) shows the *consistency* metric achieves better correlation as the alignment accuracy increases. We do the same on TopicalChat dialog data and evaluate accuracy with our weak supervision data (since no human labels are available). Figure 7 (right) shows similar trends for the *groundedness* metric. Second, we further use part of XSUM human alignment annotations to finetune

the alignment model, and obtain even higher accuracy, which in turns gives better correlation for *consistency* evaluation (star marks in the figure).

5 Conclusions

We have proposed a general evaluation framework for NLG tasks categorized as compression, transduction, and creation. Based on the common central concept of information alignment between input, context, and output, we devised a family of interpretable metrics for the key aspects of diverse tasks (summarization, style transfer, and dialog) uniformly in terms of the alignment, most of which don't require human references. The uniformly designed metrics achieve superior or comparable human correlations compared to existing metrics.

The unified framework not only offers structured guidance for the metric design of new aspects and tasks, but also opens up exciting possibilities for composable NLG evaluation. Following the more rigorous practices of software engineering, we may divide the process into modular components that can be improved, scaled, and diagnosed independently. We are excited to explore more in the future.

Acknowledgements

Bowen Tan is sponsored by US NGA NURI No. HM0476-20-1-0002 and the National Science Foundation under Grant No. IIS-16-17583, IIS-19-55532 and CNS-20-08248. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NGA or the U.S. Government.

References

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

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019a. The second conversational intelligence challenge (convai2).

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

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

Nicholas Egan, Oleg Vasilyev, and John Bohannon. 2021. Play the shannon game with language models: A human-free approach to summary evaluation.

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.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation.

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.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, and et al. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics.

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.

Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. 2018. Topic-based evaluation for conversational bots.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

Eduard Hovy and Chin-Yew Lin. 1998. Automated text summarization and the summarist system. In *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*, TIPSTER '98, page 197–214, USA. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

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

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

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.

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.

Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou, Ming Gong, Pengcheng Wang, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, Ruofei Zhang, Winnie Wu, Ming Zhou, and Nan Duan. 2021. GLGE: A new general language generation evaluation benchmark. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 408–420, Online. 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.

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.

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.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with dialogpt.

Shikib Mehri and Maxine Eskenazi. 2020b. Usr: An unsupervised and reference free evaluation metric for dialog generation.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. 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 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Nagel. 2016. Cc-news. http://web.archive.org/save/http://commoncrawl.org/2016/10/newsdataset-available.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2020. I like fish, especially dolphins: Addressing contradictions in dialogue modeling.

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.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, William B Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436.

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.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments.

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.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju.

2018. On evaluating and comparing open domain dialog systems.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries.

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2020. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric.

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. 2020a. Bertscore: Evaluating text generation with bert.

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 *ACL, system demonstration*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. 2020. Detecting hallucinated content in conditional neural sequence generation.

Markus Zopf, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. Beyond centrality and structural features: Learning information importance for text summarization. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 84–94, Berlin, Germany. Association for Computational Linguistics.

A Implementation of Alignment Estimation Models

We train our alignment models by constructing weakly supervised data using texts in the domain of evaluation. The data construction process can be divided into three steps:

1. Retrieve or generate a target sentence $\mathbf{y}_1$ given the desired input $\mathbf{x}$ (e.g., the document in summarization tasks). All tokens in $\mathbf{y}_1$ should be considered aligned with $\mathbf{x}$;

2. Sometimes, $\mathbf{y}_1$ consists of several original sentences from $\mathbf{x}$. In order to make our model non-trivial and more robust, we generate a pharaphrase $\mathbf{y}_2$ of $\mathbf{y}_1$ with a pretrained paraphrase generator[4];

3. After that, we mask some portion of $\mathbf{y}_2$, and use a `BART-large` model (Lewis et al., 2020) to infill those masks. Because the infilled content is generated without conditioning on $\mathbf{x}$, we label the infilled words as "not aligned" with $\mathbf{x}$ (BAD), and other words of $\mathbf{y}_2$ are labeled as "aligned" (OK);

Finally, $\mathbf{x}$, $\mathbf{y}_2$, and alignment labels on $\mathbf{y}_2$'s words are our desired training data.

Specially on our paraphrasing operation, in order to make the generated paraphrase different enough from the original text, we always generate 10 paraphrases and take the one with biggest edit distance with the sentence; and specially about our masking mechanism, we randomly mask some sub-trees in the constituency parsing tree of $\mathbf{y}_2$ with a pretrained parser [5]. The differences across tasks are the definitions of $\mathbf{x}$ and $\mathbf{y}_1$ in the step (1), as detailed below.

A.1 Compression: Summarization

Our training for $\textit{align}(\mathbf{y} \rightarrow \mathbf{x})$ in the summarization domain is reference-free. We use the document as $\mathbf{x}$, and generate its pseudo-summaries as $\mathbf{y}_1$ using a traditional unsupervised extractive summarizer based on TextRank (Mihalcea and Tarau, 2004). We don't use reference summaries because they can contain hallucinations that don't align with the article (Maynez et al., 2020). In an ablation study with XSUM Consistency data (Wang et al., 2020), training a `D` model using reference summaries leads to 0.2822 Pearson correlation compared to 0.3222 using auto-generated summaries, which is clearly lower. To train for $\textit{align}(\mathbf{r} \rightarrow \mathbf{y})$, we use the reference as both $\mathbf{x}$ and $\mathbf{y}_1$.

A.2 Transduction: Text Style Transfer

In this domain, we simply set $\mathbf{y}_1$ to be the original sentence $\mathbf{x}$.

A.3 Creation: Dialog

When training for $\textit{align}(\mathbf{y} \rightarrow [\mathbf{x}, \mathbf{c}])$, we use the reference response as $\mathbf{y}_1$ and the concatenation of $\mathbf{x}$ and $\mathbf{c}$ as the input. For models that predict $\textit{align}(\mathbf{y} \rightarrow \mathbf{c})$, we set the knowledge context $\mathbf{c}$ as the input, and randomly extract sentences from it as $\mathbf{y}_1$. For PersonaChat, we sample 1-3 sentences at random, whereas for TopicalChat, we only sample 1 sentence because its $\mathbf{c}$ tends to be long. When aggregating the alignment vectors, we remove stopwords according to NLTK (Bird et al., 2009) to focus on important words.

[4] https://huggingface.co/Vamsi/T5_Paraphrase_Paws

[5] https://github.com/nikitakit/self-attentive-parser

B Key Aspects

Task Category	Aspect	Alternative Names	Considered By
Compression	Consistency	Factual Correctness, Faithfulness, (No) Hallucination	(Wang et al., 2020), (Maynez et al., 2020), (Durmus et al., 2020), (Kryściński et al., 2019), (Fabbri et al., 2021), etc.
	Relevance	Content Selection, Importance	(Nenkova and Passonneau, 2004) (Peyrard, 2019) (Fabbri et al., 2021), etc.
Transduction	Preservation	Semantic Similarity	(Mir et al., 2019) (Yamshchikov et al., 2020)
Creation	Engagingness	Depth, (Not) Dull, Interestingness	(Venkatesh et al., 2018), (See et al., 2019) (Mehri and Eskenazi, 2020b) (Gopalakrishnan et al., 2019), etc.
	Groundedness	Persona Distinctiveness Knowledge Usage, Knowledge Injection	(Mehri and Eskenazi, 2020b) (Dinan et al., 2019a) (Smith et al., 2020), etc.

Table B.1: The key aspects discussed for each task category. Examples of prior work that considered each aspect as desirable properties are listed.

C Alignment Prediction Example

DOCUMENT: **Darth vader** and imperial stormtroopers have invaded a denbighshire seaside town to welcome **the actor who plays the infamous villain.**
Spencer wilding, who **hails from rhyl**, was **the guest of honour** at **a special screening** of rogue one.
He had to muster all powers of the force to keep **his vader role** secret until the film's release. "it's a hell of a secret to keep," said wilding, who was cast as the body **actor** for the role.
"but when you're **a professional actor** - when you sign that black and white sheet of paper saying you cannot say a word... I'm true to my word and i didn't say anything."
Speaking to bbc radio wales' good morning wales programme, the 44-year-old said it proved a tricky task after rumours of the role leaked a year ago. "i've been having hundreds of people every day for a year asking me if i'm **vader**," he said. "if i had a pound for everyone who asked i'd be buying myself a new death star - and it'd be gold plated."
The 6ft 7in (2m) tall actor already has a string of hollywood appearances to his name, including guardians of the galaxy, green lantern, three harry potter films and the tv blockbuster game of thrones.
He said **the vader role** came from a regular casting call, first with a self-filmed tape, then a recall to pinewood studios. "it's very, very secretive. We didn't even know exactly what the character was and what film it was until we got there," he said.
"i opened up the curtain when i went in the dressing room and there he was - **vader**.
"anybody out there who got into that costume and got **an audition to be darth vader** alone is very exciting, so to pull the character off as well, it's like 'what!'
"i'm always pinching myself - i am definitely awake - it is not a dream, it is just another dream come true."
While the **actor** has the body role, just like his predecessor in the original **star wars films** david prowse, the voice of **lord vader** is **actor** james earl jones.
That did not stop wilding trying out the voice during filming.
"i'm not james earl jones - nowhere near him - but you know i got close to him i think, which helped the other **actors** - you know, you've got **vader** in front of you."

SUMMARY 1: (BART)	A	welsh	actor	who	plays	darth	vader	in	the
	0.94	0.79	0.98	1.00	0.99	0.98	0.99	0.99	0.84
	latest	star	wars	film	has	been	honoured	at	the
	0.69	0.80	0.84	0.92	0.97	0.91	0.89	0.83	0.56
	london	film	festival.						
	0.47	0.56	0.63						
SUMMARY 2: (REPETITION)	the	the	the	the	the	the	the	the	the
	0.83	0.61	0.56	0.53	0.49	0.48	0.50	0.53	0.57
	the	the	the	the	the	the			
	0.58	0.57	0.56	0.57	0.55	0.55			

Table C.1: An example of word-level alignment prediction using discriminative model (D) for an XSUM (Narayan et al., 2018) article. SUMMARY 1 is generated by BART (Lewis et al., 2020) and received a human consistency score of 0 according to Wang et al. (2020), meaning it contains hallucination; SUMMARY 2 is a repetition of "the". As the predictions show, our model assigns low scores to words in red, which either don't follow directly from the article ("latest", "the london film festival", "welsh"), or are meaningless repetitions ("the"s).

D All Summarization Results

Metric Name	Sample-Level Correlations			System-Level Correlations		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Reference-Based Metrics						
ROUGE-1	0.1811	0.1416	0.1114	0.6648	0.7441	0.5500
ROUGE-2	0.1583	0.1360	0.1069	0.6610	0.7794	0.6000
ROUGE-L	0.1578	0.1147	0.0899	0.5180	0.0882	0.1000
BLEU	0.1794	0.1607	0.1265	0.5872	0.0765	0.0500
METEOR	0.1832	0.1508	0.1182	0.7157	0.8441	0.6500
ChrF	0.1750	0.1535	0.1205	0.6446	0.8235	0.6167
CIDEr	0.0336	0.0075	0.0058	0.0676	-0.3618	-0.1833
S3-pyr	0.1624	0.1442	0.1135	0.4616	0.6676	0.5167
S3-rsp	0.1609	0.1490	0.1173	0.4758	0.6647	0.5000
SMS	0.2110	0.2384	0.1876	0.7136	0.8000	0.6000
BERTScore-f	0.2030	0.1547	0.1215	0.6318	0.0794	0.0333
MoverScore	0.1899	0.1707	0.1339	0.5616	0.0000	-0.0500
Reference-Free Metrics						
SummaQA-prob	0.1202	0.1328	0.1045	0.7545	0.8294	0.6667
SummaQA-f	0.1572	0.1635	0.1285	0.7198	0.8324	0.6333
BLANC	0.2183	0.2303	0.1807	0.6294	0.7706	0.6167
FactCC-prob	0.3256	0.3410	0.2703	0.7401	0.7990	0.6176
SUPERT	0.3665	0.3264	0.2587	0.7274	0.7912	0.6000
Our Metrics						
Ours (E) (BERT-base)	0.4359	0.3744	0.2974	0.7886	0.7794	0.5667
Ours (E) (RoBERTa-large)	0.3315	0.3202	0.2518	0.6166	0.7912	0.5833
Ours (D) (CNN/DM)	0.5240	**0.4293**	**0.3422**	0.9146	0.8324	0.6333
Ours (D) (XSUM)	**0.5314**	0.4273	0.3414	0.9089	0.6059	0.3833
Ours (R) (CNN/DM)	0.4626	0.3871	0.3071	0.8401	0.6000	0.4167
Ours (R) (XSUM)	0.4868	0.3896	0.3094	0.8473	0.5265	0.3500

Table D.1: Correlations of all metrics with Consistency aspect of CNN/DM, using annotations from (Fabbri et al., 2021). Reference-based metrics were calculated using 11 references. Our metrics based on the trained alignment models (`D` and `R`) both clearly outperform previous metrics.

Metric Name	Sample-Level Correlations			System-Level Correlations		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Reference-Based Metrics (1 Reference)						
ROUGE-1	0.3392	0.3337	0.2402	0.6089	0.6000	0.4667
ROUGE-2	0.2479	0.2581	0.1853	0.6378	0.6176	0.4333
ROUGE-L	0.3165	0.3153	0.2262	0.5544	0.4059	0.2500
BLEU	0.2135	0.2565	0.1842	0.5760	0.3912	0.2333
METEOR	0.3336	0.3268	0.2351	0.6267	0.7176	0.5000
ChrF	0.3310	0.3305	0.2378	0.6735	0.7559	0.5333
CIDEr	0.0457	-0.0205	-0.0149	0.3348	0.2765	0.1500
S3-pyr	0.3206	0.3087	0.2209	0.6126	0.6824	0.4833
S3-rsp	0.2913	0.2940	0.2107	0.6452	0.7853	0.5667
SMS	0.2461	0.2535	0.1798	0.7681	0.7618	0.5833
BERTScore-f	0.3041	0.2937	0.2102	0.5509	0.4206	0.2667
MoverScore	0.2850	0.2898	0.2077	0.5701	0.4735	0.3167
Reference-Free Metrics						
SummaQA-prob	0.1370	0.1474	0.1039	0.6894	0.8235	0.6333
SummaQA-f	0.1665	0.1528	0.1071	0.5217	0.4412	0.3333
BLANC	0.2552	0.2355	0.1679	0.4690	0.3529	0.3167
FactCC-prob	0.2009	0.1576	0.1109	0.3487	0.3162	0.2353
SUPERT	0.3282	0.2848	0.2036	0.4569	0.3618	0.3667
Our Metrics (1 Reference)						
Ours (E) (BERT-base)	0.3635	0.3359	0.2401	0.6052	0.7500	0.5833
Ours (E) (RoBERTa-large)	**0.4985**	**0.4882**	**0.3563**	0.8494	0.8412	0.7167
Ours (D) (CNN/DM)	0.3824	0.3499	0.2528	0.6226	0.6000	0.4833
Ours (D) (XSUM)	0.3802	0.3502	0.2526	0.6126	0.5882	0.4667
Ours (R) (CNN/DM)	0.3733	0.3439	0.2495	0.5556	0.4735	0.4000
Ours (R) (XSUM)	0.3714	0.3445	0.2493	0.5447	0.4324	0.3667

Table D.2: Correlations of all considered metrics with Relevance aspect of CNN/DM, using annotations from (Fabbri et al., 2021). Reference-based metrics are calculated using 1 reference. Our (XSUM) metrics use $\mathbf{y} \rightarrow \mathbf{x}$ models based on XSUM alignment data, but still use $\mathbf{r} \rightarrow \mathbf{y}$ models based on CNN/DM alignment data. Accounting for both reference and input article on top of better alignment modeling, our metrics clearly outperform all other baselines.

Metric Name	Sample-Level Correlations			System-Level Correlations		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Reference-Based Metrics (11 References)						
ROUGE-1	0.3565	0.3160	0.2276	0.5486	0.7441	0.5833
ROUGE-2	0.2685	0.2564	0.1837	0.5659	0.6206	0.4333
ROUGE-L	0.3347	0.2990	0.2157	0.4364	0.3647	0.3667
BLEU	0.2750	0.2581	0.1851	0.6189	0.5471	0.3833
METEOR	0.3237	0.2936	0.2101	0.6217	0.7471	0.5500
ChrF	0.3561	0.3366	0.2418	0.7017	0.7441	0.5500
CIDEr	-0.0055	-0.0261	-0.0191	0.1654	0.1500	0.0833
S3-pyr	0.3469	0.3180	0.2272	0.3811	0.2765	0.2167
S3-rsp	0.3227	0.3101	0.2216	0.3975	0.2971	0.2333
SMS	0.2593	0.2467	0.1765	0.5156	0.4618	0.4000
BERTScore-f	0.3192	0.2961	0.2126	0.5991	0.5441	0.4000
MoverScore	0.3114	0.3108	0.2237	0.6419	0.5382	0.3500
Reference-Free Metrics						
SummaQA-prob	0.1370	0.1474	0.1039	0.6894	0.8235	0.6333
SummaQA-f	0.1665	0.1528	0.1071	0.5217	0.4412	0.3333
BLANC	0.2552	0.2355	0.1679	0.4690	0.3529	0.3167
FactCC-prob	0.2009	0.1576	0.1109	0.3487	0.3162	0.2353
SUPERT	0.3282	0.2848	0.2036	0.4569	0.3618	0.3667
Our Metrics (11 References)						
Ours (E) (BERT-base)	0.3906	0.3547	0.2544	0.5032	0.3765	0.3167
Ours (E) (RoBERTa-large)	**0.5198**	**0.4990**	**0.3671**	0.7539	0.7324	0.6167
Ours (D) (CNN/DM)	0.4423	0.3962	0.2862	0.5821	0.4794	0.3833
Ours (D) (XSUM)	0.4426	0.3991	0.2878	0.5691	0.4794	0.3833
Ours (R) (CNN/DM)	0.4115	0.3617	0.2644	0.4999	0.3824	0.3333
Ours (R) (XSUM)	0.4121	0.3680	0.2687	0.4906	0.3353	0.2833

Table D.3: Correlations of all considered metrics with Relevance aspect of CNN/DM, using annotations from (Fabbri et al., 2021). Reference-based metrics are calculated using 11 references. Our (XSUM) metrics use $\mathbf{y} \rightarrow \mathbf{x}$ models based on XSUM alignment data, but still use $\mathbf{r} \rightarrow \mathbf{y}$ models based on CNN/DM alignment data. Accounting for both reference and input article on top of better alignment modeling, our metrics clearly outperform all other baselines.

Metric Name	CNN/DM Correlations			XSUM Correlations		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Reference-Based Metrics						
ROUGE-1	0.2874	-	-	0.1322	-	-
ROUGE-2	0.1772	-	-	0.0895	-	-
ROUGE-L	0.2409	-	-	0.0886	-	-
METEOR	0.2665	-	-	0.1003	-	-
BLEU-1	0.2968	-	-	0.1176	-	-
BLEU-2	0.2565	-	-	0.1168	-	-
BLEU-3	0.2396	-	-	0.0841	-	-
BLEU-4	0.2145	-	-	0.0564	-	-
BERTScore-f	0.2763	-	-	0.0251	-	-
Reference-Free Metrics						
FactCC-prob	0.4158	0.4837	0.3758	0.2968	0.2588	0.2118
QAGS	0.5453	-	-	0.1749	-	-
Our Metrics						
Ours (E) (BERT-base)	0.6083	0.5180	0.4074	0.1436	0.1437	0.1176
Ours (E) (RoBERTa-large)	0.6091	0.5229	0.4141	0.0548	0.0489	0.0400
Ours (D) (CNN/DM)	0.6188	**0.5640**	**0.4500**	0.3085	0.2952	0.2416
Ours (D) (XSUM)	0.6205	0.5362	0.4260	**0.3222**	**0.3149**	**0.2576**
Ours (R) (CNN/DM)	0.6468	0.5252	0.4180	0.2157	0.1949	0.1594
Ours (R) (XSUM)	**0.6612**	0.5445	0.4348	0.2718	0.2509	0.2053

Table D.4: Sample-Level correlations of all considered metrics with Consistency aspect of CNN/DM and XSUM, based on annotations from (Wang et al., 2020). Spearman and Kendall-Tau correlations for baseline metrics were not reported, except for `FactCC` which we computed on our own. System-level correlations are not reported due to dataset limits. On CNN/DM, all of our metrics outperform previous metrics. On XSUM, our `D`-based metric trained on the same domain also achieves the best performance. The `E`-based metrics see a catastrophic drop likely due to the higher abstractiveness of XSUM that renders embedding matching inadequate.

E All Style Transfer Results

Metric Name	Sample-Level Correlations			System-Level Correlations		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Lexical-Matching Based Metrics						
BLEU	0.4361	0.4303	0.3125	0.8476	0.8042	0.6364
ROUGE-1	0.4808	0.4981	0.3614	0.7749	0.6503	0.4545
ROUGE-2	0.4664	0.4589	0.3421	0.8444	0.7622	0.5455
ROUGE-L	0.4833	0.4975	0.3606	0.7814	0.7133	0.5152
METEOR	0.4764	0.4993	0.3592	0.8291	0.7622	0.5455
ChrF	0.5048	0.5028	0.3632	0.8026	0.7343	0.5758
Embedding Based Metrics						
EmbedAvg	0.3248	0.4127	0.2959	0.7272	0.5944	0.3939
GreedyMatch	0.4542	0.4760	0.3434	0.7431	0.6084	0.4242
VectorExtrema	0.4571	0.4589	0.3306	0.7702	0.6364	0.4242
WMD	0.4902	0.5047	0.3646	0.7776	0.6713	0.5455
Pre-Trained Model Based Metrics						
BERTScore	0.5185	0.5187	0.3751	0.8078	0.7133	0.5152
MoverScore	0.5209	0.5148	0.3734	0.8308	0.7622	0.5455
BLEURT	0.5043	0.4934	0.3566	0.8673	0.7902	0.6061
Our Metrics						
Ours (E) (BERT-base)	0.5147	0.5169	0.3740	0.8096	0.7133	0.5152
Ours (E) (RoBERTa-large)	0.5142	0.5150	0.3752	0.8618	0.7832	0.5758
Ours (E) (RoBERTa-large-MNLI-9)	**0.5216**	**0.5236**	**0.3805**	0.8081	0.7133	0.5152
Ours (D)	0.4974	0.4952	0.3579	0.8385	0.7483	0.5152
Ours (R)	0.5060	0.5059	0.3645	0.8226	0.6993	0.4848

Table E.1: Correlations of all considered metrics with Preservation aspect of Yelp, using annotations from (Mir et al., 2019). Explicitly accounting for two-way input-out alignments in an "F1"-style harmonic mean aggregation (Eq.4), our metrics (E) achieve competitive or better performance than previous metrics. Our D- and R-based metrics fall behind slightly, likely because one-to-one token matching is more suitable for two text pieces with similar information volume.

F All Dialog Results

Metric Name	Sample-Level Correlations			System-Level Correlations		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Reference Based Metrics						
F-1	0.0473	0.0132	-	0.9956	1.0000	-
BLEU-1	-0.1081	-0.0922	-	0.2599	0.6000	-
BLEU-2	-0.1048	-0.1010	-	0.6816	0.4000	-
BLEU-3	-0.1247	-0.1151	-	0.6668	0.4000	-
BLEU-4	-0.1359	-0.1242	-	0.8413	0.8000	-
METEOR	-0.0458	0.0116	-	0.9065	0.8000	-
ROUGE-L	-0.1456	-0.1354	-	0.1710	0.0000	-
BERTScore (base)	0.0325	0.0491	-	0.5173	0.8000	-
BERTScore (large)	-0.0418	-0.0245	-	0.2410	0.0000	-
Reference-Free Metrics						
FED-Interesting	0.1818	0.0255	0.0182	0.9277	1.0000	1.0000
USR-MLM	-0.1045	-0.1007	-	-0.2842	-0.4000	-
USR-DR (x=c)	0.0606	0.2634	-	0.8202	1.0000	-
USR-DR (x=f)	-0.0022	-0.0039	-	-0.0178	-0.2108	-
USR	0.0315	0.0171	-	0.8084	1.0000	-
Word Length	0.3910	0.4220	0.3267	0.8965	0.8000	0.6000
Our Metrics						
Ours (E) (BERT-base)	0.5003	0.5490	0.4193	0.9061	1.0000	1.0000
Ours (E) (RoBERTa-large)	0.4081	0.4502	0.3375	0.9003	0.9000	0.8000
Ours (D) (PersonaChat)	0.5265	0.5793	**0.4412**	0.9425	1.0000	1.0000
Ours (D) (TopicalChat)	0.5317	**0.5818**	0.4409	0.9447	1.0000	1.0000
Ours (R) (PersonaChat)	**0.5320**	0.5692	0.4346	0.9433	1.0000	1.0000
Ours (R) (TopicalChat)	0.4933	0.5333	0.4043	0.9244	1.0000	1.0000

Table F.1: Correlations of all considered metrics with Engagingness aspect of PersonaChat, using annotations from (Mehri and Eskenazi, 2020b). Kendall-Tau correlations for baseline metrics were not reported. By measuring aligned information volume (sum) and with accurate estimation models, our metrics with different implementations all improve over previous methods by large margins.

Metric Name	Sample-Level Correlations			System-Level Correlations		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Reference Based Metrics						
F-1	0.0869	0.1056	-	0.9956	1.0000	-
BLEU-1	0.0737	0.0729	-	0.2599	0.6000	-
BLEU-2	0.1083	0.0722	-	0.6816	0.4000	-
BLEU-3	0.0999	0.0594	-	0.6668	0.4000	-
BLEU-4	0.0698	0.0528	-	0.8413	0.8000	-
METEOR	0.1678	0.1719	-	0.9065	0.8000	-
ROUGE-L	0.0710	0.0632	-	0.1710	0.0000	-
BERTScore (base)	0.0719	0.0465	-	0.5173	0.8000	-
BERTScore (large)	0.0271	0.0094	-	0.2410	0.0000	-
Reference-Free Metrics						
FED-Informative	0.0165	-0.0405	-0.0315	0.9015	0.9000	0.8000
USR-MLM	-0.0782	-0.0756	-	-0.2842	-0.4000	-
USR-DR (x=c)	0.4508	0.6309	-	0.8202	1.0000	-
USR-DR (x=f)	-0.0927	-0.0903	-	-0.0178	-0.2108	-
USR	0.4027	0.3177	-	0.8084	1.0000	-
Word Length	0.3171	0.3051	0.2467	0.7698	0.5000	0.4000
Our Metrics						
Ours (E) (BERT-base)	0.5761	0.5683	0.4492	0.8720	0.9000	0.8000
Ours (E) (RoBERTa-large)	0.3758	0.3652	0.2862	0.8140	0.7000	0.6000
Ours (D) (PersonaChat)	0.5683	0.5674	0.4505	0.8684	0.9000	0.8000
Ours (D) (TopicalChat)	0.4056	0.4172	0.3270	0.7536	0.5000	0.4000
Ours (R) (PersonaChat)	0.6597	0.6689	0.5338	0.9151	1.0000	1.0000
Ours (R) (TopicalChat)	**0.6819**	**0.7113**	**0.5636**	0.9420	1.0000	1.0000

Table F.2: Correlations of all considered metrics with Groundedness aspect of PersonaChat, using annotations from (Mehri and Eskenazi, 2020b). Kendall-Tau correlations for baseline metrics were not reported. Trained with aggregated alignment scores and benefiting from the expressivity of end-to-end models, our regression-based (R) metrics strongly outperform all other metrics.

Metric Name	Sample-Level Correlations			System-Level Correlations		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Reference Based Metrics						
F-1	0.2523	0.2565	-	0.5944	0.6000	-
BLEU-1	0.3144	0.3343	-	0.8197	0.7000	-
BLEU-2	0.3184	0.3323	-	0.8099	0.9000	-
BLEU-3	0.2782	0.3247	-	0.9047	0.9000	-
BLEU-4	0.2322	0.3156	-	0.8883	0.9000	-
METEOR	0.3668	0.4391	-	0.9398	0.9000	-
ROUGE-L	0.2946	0.2995	-	0.8084	0.9000	-
BERTScore (base)	0.3512	0.3725	-	0.9108	0.9000	-
BERTScore (large)	0.3167	0.3349	-	0.8480	0.9000	-
Reference-Free Metrics						
FED-Interesting	-0.0004	-0.0328	-0.0230	0.8881	0.8286	0.7333
USR-MLM	0.3189	0.3337	-	0.4663	0.9000	-
USR-DR (x=c)	0.3533	0.4877	-	0.9233	0.7000	-
USR-DR (x=f)	0.2006	0.4110	-	0.8685	0.9000	-
USR	0.4555	0.4645	-	0.9297	1.0000	-
Word Length	0.4079	0.4089	0.3053	0.9662	0.8286	0.7333
Our Metrics						
Ours (E) (BERT-base)	0.4937	0.5047	0.3710	0.9606	0.9429	0.8667
Ours (E) (RoBERTa-large)	0.4471	0.4479	0.3288	0.9617	0.8286	0.7333
Ours (D) (PersonaChat)	0.5124	0.5245	0.3878	0.9572	0.6571	0.6000
Ours (D) (TopicalChat)	**0.5163**	**0.5253**	**0.3873**	0.9657	0.8286	0.7333
Ours (R) (PersonaChat)	0.4542	0.4588	0.3357	0.9529	0.8286	0.7333
Ours (R) (TopicalChat)	0.4653	0.4643	0.3395	0.9492	0.8286	0.7333

Table F.3: Correlations of all considered metrics with Engagingness aspect of TopicalChat, using annotations from (Mehri and Eskenazi, 2020b). Kendall-Tau correlations for baseline metrics were not reported. By measuring aligned information volume (sum) and with accurate estimation models, our metrics with different implementations all compete with or improve over previous methods.

Metric Name	Sample-Level Correlations			System-Level Correlations		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Reference Based Metrics						
F-1	0.1495	0.1485	-	0.5970	0.6000	-
BLEU-1	0.2888	0.3033	-	0.8357	0.7000	-
BLEU-2	0.2819	0.3066	-	0.8309	0.9000	-
BLEU-3	0.2442	0.3106	-	0.9259	0.9000	-
BLEU-4	0.2126	0.3096	-	0.9084	0.9000	-
METEOR	0.3328	0.3909	-	0.9534	0.9000	-
ROUGE-L	0.3099	0.3273	-	0.8333	0.9000	-
BERTScore (base)	0.2847	0.2947	-	0.9308	0.9000	-
BERTScore (large)	0.2909	0.3167	-	0.8703	0.9000	-
Reference-Free Metrics						
FED-Informative	0.0311	0.0243	0.0209	0.9340	0.7143	0.6000
USR-MLM	0.2195	0.2261	-	0.5070	0.9000	-
USR-DR (x=c)	0.2285	0.4179	-	0.9155	0.7000	-
USR-DR (x=f)	0.2220	0.4468	-	0.8884	0.9000	-
USR	0.3175	0.3353	-	0.9469	1.0000	-
Word Length	0.2624	0.2681	0.2084	0.9859	0.8286	0.7333
Our Metrics						
Ours (E) (BERT-base)	0.4293	0.3949	0.3075	0.9784	0.9429	0.8667
Ours (E) (RoBERTa-large)	0.3122	0.3141	0.2421	0.9868	0.8286	0.7333
Ours (D) (PersonaChat)	0.3697	0.3691	0.2856	0.9784	0.9429	0.8667
Ours (D) (TopicalChat)	0.3099	0.3159	0.2421	0.9842	0.8286	0.7333
Ours (R) (PersonaChat)	0.4026	0.3788	0.3137	0.9742	0.7143	0.6
Ours (R) (TopicalChat)	**0.5235**	**0.4768**	**0.3838**	0.9674	0.8857	0.7333

Table F.4: Correlations of all considered metrics with Groundedness aspect of TopicalChat, using annotations from (Mehri and Eskenazi, 2020b). Kendall-Tau correlations for baseline metrics were not reported. Trained with aggregated alignment scores and benefiting from the expressivity of end-to-end models, our regression-based (R) metric trained on TopicalChat strongly outperforms all other metrics.

Metric	PersonaChat		TopicalChat	
	Swapped	Ours	Swapped	Ours
Engagingness (E)	**0.5082**	0.5003	**0.5180**	0.4973
Engagingness (D)	0.4725	**0.5265**	0.4708	**0.5163**
Engagingness (R)	0.4679	**0.5320**	**0.4902**	0.4653
Groundedness (E)	0.5323	**0.5761**	0.3870	**0.4293**
Groundedness (D)	0.4945	**0.5683**	**0.3752**	0.3099
Groundedness (R)	0.4798	**0.6597**	0.3237	**0.5235**

Table F.5: Ablation Studies: Pearson correlations with engagingness and groundedness for dialog tasks with swapped formulas vs our definition. By swapping, we use our engagingness metric to measure groundedness, and vice versa. PersonalChat swaps see across-the-board decreases in correlations, indicating the importance of using our designed formulas on this dataset. TopicalChat swaps see correlation increases more frequently, but the best methods still retain their edge.

Metric Name	PersonaChat				TopicalChat			
	U	N	MC	O	U	N	MC	O
Reference-Based Metrics								
F-1	-0.0340	0.0815	0.1073	0.1422	0.0425	0.0301	0.1290	0.1645
BLEU-4	0.0537	0.1081	0.1467	0.1353	0.2010	0.1799	0.1307	0.2160
METEOR	0.0820	0.0989	0.2500	0.2527	0.2452	0.2121	0.2495	0.3365
ROUGE-L	0.0346	0.0096	0.1135	0.0659	0.2069	0.1760	0.1928	0.2745
BERTScore (base)	0.0676	0.0606	0.1770	0.1690	0.2611	0.2164	0.2432	0.3229
Reference-Free Metrics								
FED	0.0314	0.0870	-0.0634	-0.0786	0.0469	0.0482	-0.1915	-0.1393
USR-MLM	0.1313	0.0999	0.1805	0.0795	0.3264	0.3370	0.3099	0.3086
USR-DR (x=c)	0.0728	0.1733	0.6021	**0.4814**	0.1500	0.1325	0.3391	0.3245
USR-DR (x=f)	-0.0390	-0.0033	-0.0198	-0.0495	0.0881	0.0313	0.0594	0.1419
USR	0.0997	0.1862	**0.6065**	0.4693	0.2932	0.2260	**0.4160**	**0.4192**
Response Length	-0.0525	-0.0342	0.0901	0.2526	0.0845	0.1253	0.2458	0.3343
Our Metrics								
Ours (E)	0.1185	0.1891	0.2786	0.3690	0.2168	0.1528	0.1669	0.2483
Ours (D)	**0.1421**	**0.3384**	0.3837	0.4500	**0.3433**	**0.3653**	0.2969	0.3620
Ours (R)	0.0639	0.1595	0.2518	0.2076	0.0105	-0.0524	-0.0248	-0.0338

Table F.6: Sample-level Pearson correlations for the remaining aspects in the annotations of (Mehri and Eskenazi, 2020b), including understandable (U), natural (N), maintains context (MC) and overall (O). Our metric here is the average alignment confidence from response $\mathbf{y}$ to dialogue history $\mathbf{x}$ and knowledge $\mathbf{c}$, i.e. $\text{mean}\,(\textit{align}(\mathbf{y} \rightarrow [\mathbf{x}, \mathbf{c}])$, which outperforms existing metrics on *understandability* and *naturalness*.