Reference-guided Style-Consistent Content Transfer

Wei-Fan Chen[1], **Milad Alshomary**[2], **Maja Stahl**[2],
Khalid Al-Khatib[3], **Benno Stein**[4], **Henning Wachsmuth**[2]

[1]Conversational AI and Social Analytics (CAISA) Lab, University of Bonn, Germany,
[2]Institute of Artificial Intelligence, Leibniz University Hannover, Germany
[3]Center for Language and Cognition (CLCG), University of Groningen, Netherlands,
[4]Intelligent Information Systems (Webis), Bauhaus-Universität Weimar, Germany

wchen@uni-bonn.de, m.alshomary@ai.uni-hannover.de, m.stahl@ai.uni-hannover.de
khalid.alkhatib@rug.nl, benno.stein@uni-weimar.de, h.wachsmuth@ai.uni-hannover.de

Abstract

In this paper, we introduce the task of *style-consistent content transfer*, which concerns modifying a text's content based on a provided reference statement while preserving its original style. We approach the task by employing multi-task learning to ensure that the modified text meets three important conditions: reference faithfulness, style adherence, and coherence. In particular, we train three independent classifiers for each condition. During inference, these classifiers are used to determine the best modified text variant. Our evaluation, conducted on hotel reviews and news articles, compares our approach with sequence-to-sequence and error correction baselines. The results demonstrate that our approach reasonably generates text satisfying all three conditions. In subsequent analyses, we highlight the strengths and limitations of our approach, providing valuable insights for future research directions.

Keywords: Natural Language Generation, Text Analytics, Textual Entailment and Paraphrasing

1. Introduction

Text style transfer concerns changing the style of a given text while preserving its content. With the advances in neural network architectures, significant contributions have been made to text style transfer (Fu et al., 2018; Jin et al., 2022). However, there has been relatively little focus on the inverse direction: preserving the original style while modifying the content, a task referred to as *content transfer* (Qin et al., 2019; Prabhumoye et al., 2019).

Content transfer finds applications in different scenarios where text updates are necessary during the writing process. For instance, it allows for revising specific events in a narrative while maintaining the established writing style (West et al., 2022). In scientific writing, updating related work demands modifications while maintaining consistency in the article's overall style and tone. Also in collaborative writing environments with multiple authors, content transfer enables content updates without compromising a unified tone and style, creating a final document that appears as if drafted by a single author (Prabhumoye et al., 2019).

In this paper, we propose the following content transfer task: Given a reference statement, a desired style, and a preceding and subsequent context, generate a text to be inserted in between the context that meets three conditions: (1) *faithfulness* with respect to the reference statement, (2) *style adherence* to the discussed topic, and (3) *coherence* in relation to the context of the text. Table 1 exemplifies the task, where a new sentence is inserted into an existing text while meeting these three conditions.

We approach the proposed task with the hypothesis that all three outlined conditions need to be optimized jointly to ensure text quality. To operationalize this hypothesis, we adapt the idea of knowledge distillation (Hinton et al., 2015; Liu et al., 2021) to content transfer, where a generator is guided by pre-trained teacher models.

Our approach involves an initial training of classifiers, each dedicated to evaluating a specific condition, thereby determining the extent to which a given text complies with these conditions. The approach is generic in that the conditions could easily be replaced or extended by others, as detailed in Section 3. Given the classifiers, the knowledge encapsulated within these classifiers is then distilled into our model through a two-step process: In the first step, a multi-task learning model is trained, where one task focuses on generating the text, while the others revolve around validating the completion of the specified conditions. The classifiers, serving as a source of distilled knowledge, guide the model training process. Following this training phase, the model is used to generate multiple candidate texts. Subsequently, the classifiers assess the texts generated during the process, to select those that best align with the predefined conditions.

Our multi-task learning approach is applied to two content transfer domains: We conduct experiments on the *hotel review* corpus of Wachsmuth et al. (2014), where review statements conveying

(a) Hotel Review

Topic. Mattress

Reference. I really would expect a better mattress than a rock.

Text. Two complaints though: the bed, although not uncomfortable, was a little lumpy. **The mattresses felt like they were made of cement.** And the free WiFi never actually worked for me.

(b) News Article

Topic. Mike Huckabee

Reference. His book, "Quit Digging Your Grave with a Knife and Fork," became a bestseller and he made obesity a major policy initiative as governor.

Text. "And if someone can capture both the blue-collar, working-class Republicans, the conservatives, many of them even union members, as well as evangelicals, there's a real pathway to the nomination." **In 2009, he became one of the nation's most famous public officials for lifting the U.S. standard for personal appearance.** He was the longest-serving Arkansas governor, from 1996 to 2007.

Table 1: Examples of content transfer on a hotel review and a news article. The bold sentences are generated by the approach proposed in this work, given a *topic*, a *reference*, and the surrounding *text* as input. In (a), the generated sentence about mattresses can be entailed from the reference that the mattresses was like a rock. Similarly, in (b), the part on "lifting the U.S. standard for personal appearance" can be entailed from Mike Huckabee "made obesity a major policy initiative."

sentiments towards certain topics are the reference statements that we aim to integrate into the generated text. Also, we use the *news article* corpus of Fan et al. (2019), where sentences opinionated toward specific entities represent the reference statements. In Section 4, we detail how the reference statements are obtained in our experiments and how the relationships between these references and the generated texts are inferred.

We empirically evaluate our approach both automatically and manually in Section 5. Compared to a sequence-to-sequence training method from our previous paper (Chen et al., 2021b) and an error correction model (Thorne and Vlachos, 2021), our approach improves content transfer in terms of all three conditions (faithfulness, style adherence, and coherence) in a crowd-based user study, while being on par with the baselines in automatic metrics. A key observation from our evaluation is the effectiveness of retaining review sentiment and the complexity associated with acquiring the necessary knowledge for news articles.

Altogether, we provide three main contributions:

1. We propose a generic reference-based content transfer framework.
2. We instantiate the content transfer framework as a multi-task learning approach employing knowledge distillation.
3. We present empirical evidence demonstrating that our approach generates text that is more faithful to references compared to existing methods in two domains.

2. Related Work

Content transfer is closely related to text style transfer by concept. The notion of "style" is treated broadly in the literature, and sometimes the border between style and content is blurred. Style transfer tasks include changing positive sentiment to negative sentiment (He and McAuley, 2016; Shen et al., 2017), formal writing to informal writing (Rao and Tetreault, 2018), one topic to another (Huang et al., 2020), and one frame to another (Chen et al., 2021b). However, there are only a few works explicitly on content transfer so far (Prabhumoye et al., 2019; West et al., 2022). They study how to insert a faithful text in an existing document based on information from a reference document. They showed that the model using information from the existing document and the reference achieved the best performance. With this in mind, our approach uses multi-task learning considering the information from the context and the reference.

The faithfulness issue exists in many natural language generation tasks (Li et al., 2022), such as machine translation (Zhou et al., 2021), dialogue generation (Welleck et al., 2019), and, in particular, abstractive summarization (Cao et al., 2020; Dong et al., 2020; Zhu et al., 2021). The first step to tackle this issue is to define what is considered *faithful* and what *not faithful*, so that faithfulness can be detected (Maynez et al., 2020). Goodrich et al. (2019) model facts as subject-relation-object triplets to then compare facts extracted from generated texts to these triplets. However, the authors mention that their approach cannot extract negation relations. In contrast, the faithfulness in our approach is a free text and thus negation relations can be captured as well. We borrow the entailment/contradiction concept from research on natural language inference for this purpose (MacCartney, 2009). Recently, Laban et al. (2022) also used natural language inference to detect faithfulness inconsistency. Our approach also uses natural language inference to help improving faithfulness.

Several studies on improving faithfulness or factuality relate to abstractive summarization (Zhang et al., 2020b; Nan et al., 2021; Thorne and Vlachos, 2021). For example, Cao et al. (2020) train a

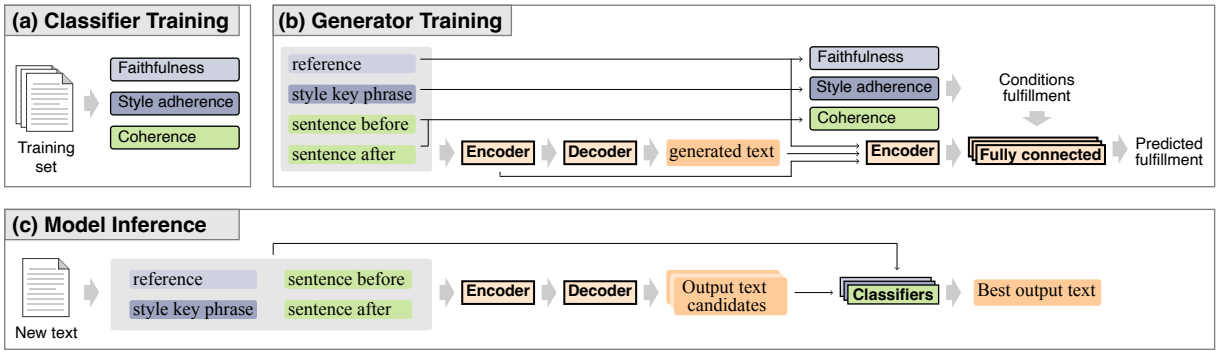

Figure 1: The three stages of our approach: (a) Training classifiers for each condition. (b) Training the multi-task learning model with the knowledge from the trained classifiers. Three fully-connected layers (one for each condition) are used to predict the condition fulfillment. The model learns to generates the text while satisfying the conditions. (c) Selecting the output text best fulfilling the conditions using the classifiers. Different conditions require different inputs.

sequence-to-sequence model where the inputs are the corrupted and a reference summary, and the output is the correct summary. Zhu et al. (2021) integrate knowledge into their model, where the knowledge is again defined as subject-relation-objects. Closer to our idea, previous works also propose selecting the best output among the candidates. For example, Chen et al. (2021a) use contrastive candidate generation and selection. Likewise, Falke et al. (2019) and Barrantes et al. (2020) use the pre-train natural language inference model to select the most factually correct output. Different from summarization, content transfer requires a coherent integration of given and new content.

The idea of knowledge distillation, used in the present work, was first formalized by Hinton et al. (2015) as a method of transferring knowledge from one teacher model to a student model. The knowledge learned may be feature-based, relational-based, or response-based (Gou et al., 2021). In the field of NLP, knowledge distillation has been used, for example, to train smaller versions of transformer-based models (Sanh et al., 2019). Our approach distills knowledge from pre-trained classifiers in order to better control text generation.

3. Approach

In this section, we define the reference-guided content transfer task we consider, and then we present our approach to the task using knowledge distillation. Figure 1 gives an overview of the approach.

3.1. Reference-Guided Content Transfer

As previous work (Prabhumoye et al., 2019; Chen et al., 2021b), we model reference-guided content transfer as a sentence generation task:

> *Given a reference statement, a desired style, and a surrounding context, generate a text that is consistent with the content of the reference, adheres to the style, and is coherent with the context.*

We assume both the reference and the output text to be a single sentence each, and the style to be specified as a topic (e.g., "room").[1] For context, we follow our previous work (Chen et al., 2021b) in providing the sentence before and the sentence after the text to be generated. Examples are shown in Figure 2.

3.2. Knowledge Distillation

We approach the three conditions of the given task (faithfulness, style adherence, and coherence) using a multi-task paradigm, where the primary task is text generation while the fulfilling of each condition serves as one auxiliary task. Most multi-task learning methods produce all output at once (Collobert and Weston, 2008), generating the text and predicting the conditions' fulfillment at the same time. However, we argue that the prediction should be made *after* generation, so the misfulfillment of the conditions can be backpropagated to guide the generation.

In line with Liu et al. (2021), we realize our approach in three stages, as illustrated in Figure 1:

(a) *Classifier training.* Given the three conditions, we train one classifier for each.

(b) *Generator training.* The classifiers are used as teacher models to guide the training of the generation model.

(c) *Model inference.* Given output text candidates, the classifiers find out the text best fulfills the conditions.

[1] It is arguable whether topic is content or style. We discuss the notion of style in Section § 2.

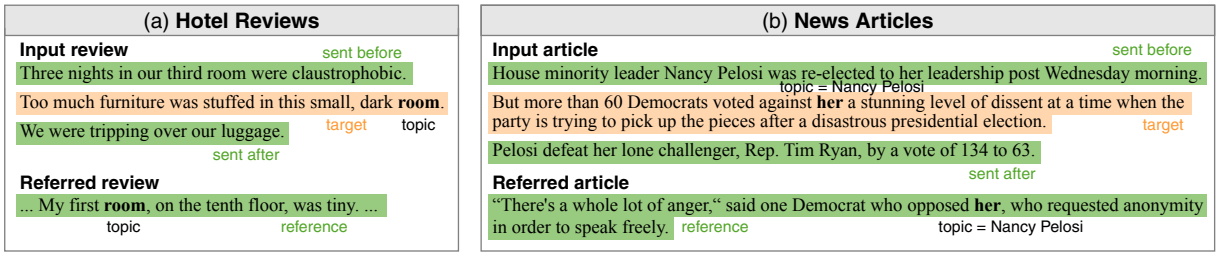

Figure 2: Examples from the corpora used: (a) A training instance from the hotel reviews. Both the *target* and the *reference* have negative sentiment toward the *topic*, "room". (b) A training instance from the news articles. Note that the original annotation does not include the location of the topic in this dataset. We label "her" here for better understanding. Similarly, both the *target* and the *reference* have negative sentiment toward the *topic*, "Nancy Pelosi". In this figure, the inputs ($sent_{before}$, $sent_{after}$, and *reference*) are in green and the output (*target*) is in orange.

Inspired by research on knowledge distillation (Hinton et al., 2015; Liu et al., 2021), we use the classifiers as the teacher models and distill their knowledge into our approach. This could also be extended by further conditions, simply by using further classifiers and concatenating the embeddings with different inputs, if necessary. The main difference between Liu et al. (2021) and ours is that we add (c) to further improve the performance.

Classifier Training For faithfulness, the classifier predicts the relation between the generated text, $\hat{y}$, and the reference, r. Borrowing ideas from natural language inference (MacCartney, 2009), we distinguish three cases: (1) *faithful*, i.e., one text entails the other text, (2) *non-faithful*, i.e., the two texts contradict each other, and (3) the texts are in a *"neutral" relationship*. For style adherence, a binary classifier predicts whether the generated text covers the desired topic t. Lastly, a binary classifier predicts whether the three sentences (sentence before $sent_{before}$, generated text, and sentence after $sent_{after}$) are coherent in sequence.

Generator Training For generation, we use an encoder-decoder architecture that learns to generate the target output text $\hat{y}$ given the inputs (reference, style, sentence before, and sentence after), as shown in Figure 1(b). After generation, we pass $\hat{y}$, r, t, $sent_{before}$, and $sent_{after}$ to the trained classifiers to get the predicted labels L, one label per classifier. We also feed $\hat{y}$ and all other inputs to the encoder again and concatenate their embeddings. Three fully-connected layers (one for each condition) are used to predict labels $\hat{L}$, where the target labels L are given by the classifiers. Note that, unlike most multi-task learning approaches, the labels of the conditions are given by the teacher models on-the-fly and are not known before training.

Multiple losses have to be optimized: $\mathcal{L}_g$, the generation loss from comparing y with $\hat{y}$, and one condition loss $\mathcal{L}_i$ each from comparing $l_i \in L_i$ with $\hat{l}_i \in \hat{L}_i$ for condition i. The overall loss is as follows, where n is the number of conditions:

$$\mathcal{L} = \alpha_g \cdot \mathcal{L}_g + \sum_{i=1}^{n} \alpha_i \cdot \mathcal{L}_i, \quad (1)$$

where $\alpha_g \geq 0$, $\alpha_i \geq 0$ and $\alpha_g + \sum \alpha_i = 1$. The optimization of the weights α_i can be found in the appendix.

Dataset	Training	Validation	Test
Hotel Reviews	23,471	5,158	6,489
News Articles	4,064	533	1,457

Table 2: The number of instances in training set, validation set, and test set for each dataset.

Model Inference During inference, we first generate a set of candidate texts $\hat{y}_j$ and then use the classifiers to predict the probabilities of condition fulfillment, as shown in Figure 1(c). The text with the highest aggregated probability is selected as the output. In particular, given all predicted probabilities $p_{i,j}$ of condition i and candidate j, the best output text is computed as follows, with the same weights as in Equation 1:

$$\hat{y} = \arg\max_{\hat{y}_j} \sum_{i=1}^{n} \alpha_i \cdot p_{i,j}, \quad (2)$$

4. Experiments

In this section, we report on the experiments we conducted to investigate the extent to which the proposed approach can perform reference-guided content transfer. We present the setup, the dataset preparation, and the considered baselines.

4.1. Datasets

We consider two text registers in our experiments: hotel reviews and news articles. The reviews have

shorter sentences and are more homogeneous, since their topic-related scope is limited. By contrast, news articles are longer and more diverse, likely making them more challenging. The distribution of the two datasets can be seen in Table 2.

Hotel Reviews We use the corpus of Wachsmuth et al. (2014), which contains a balanced set of hotel reviews from TripAdvisor with 300 reviews each for seven hotel locations, 60 each per star rating from 1 to 5. We use the hotels located in Amsterdam, Seattle, Sydney, and Berlin as the training set, hotels located in San Francisco as the validation set, and hotels located in Barcelona and Paris as the test set. In each review, the sentiment of each statement is classified as positive, negative, or neutral. We focus on the sentences where a topic (so-called *product feature* in the paper), and positive or negative sentiments are annotated, for example *room* in Figure 2(a) with negative sentiment. Given the corpus, we prepare a sentence generation task similar to Chen et al. (2021b) and West et al. (2022), as exemplified in Figure 2(a).

News Articles The other type of texts we considered is represented by the BASIL news corpus (Fan et al., 2019), which consists of three news articles each for 100 events. We randomly selected 70 events as the training set, 10 events as the validation set, and 20 events as the test set. We only use the sentences which have bias annotated. An example of the dataset can be seen in Figure 2(b). A biased sentence is annotated with a topic (so-called *target* in the paper), such as a person or an event, and the sentiment toward the topic.

4.2. Classifiers and Model

Here we detail the three training of the three classifiers outside the generation model. As a base model, we use `facebook/bart-base` from the Huggingface library (Wolf et al., 2019)[2]. Besides, we have the following three teacher classifiers.

Faithfulness Classifier Given a reference statement, r, and generated text, $\hat{y}$, this classifier predicts the probability of faithfulness. It is based on the pre-trained natural language inference model from `microsoft/deberta-base-mnli`. For a given sentence having a sentiment toward a topic (hotel or event), we took all sentences from other texts with the same topic and sentiment as those as *faithful* samples. To finetune the classifier on the training data, we also needed samples with *non-faithfulness* and *neutral* labels. For non-faithfulness, we selected sentences from other reviews/news for the same hotel/event mentioning the same topic but opposite sentiment. For neutral, we randomly selected sentences from other reviews/news mentioning a different topic. Table 3 shows the distribution of the labels in all datasets. For the hotel reviews, the macro-average F_1-score of the classifier is 0.86, and the F_1-score for the faithfulness label is 0.85. For news articles, the performance is limited only, with a macro-average F_1-score 0.47 for the classification and an F_1-score of 0.61 for the faithfulness label.

Style Adherence Classifier Given a topic t, and a generated text $\hat{y}$, this classifier predicts whether $\hat{y}$ is relevant to t. For the negative instances, we randomly chose a sentence with a different annotated topic. We used the pre-trained `bart-base`(Lewis et al., 2020) and finetuned it on our datasets. The data distribution can be seen in Table 4. The accuracy of the model is 0.98 and 0.74 for the hotel reviews and the news articles, respectively.

Coherence Classifier Given $sent_{before}$, $\hat{y}$, and $sent_{after}$, this classifier predicts if $\hat{y}$ is coherent in between the others. We also used the pretrained `bart-base` and finetuned it on the given data. As negative instances, we randomly chose sentences to replace the sentence in the middle. The data distribution is the same in Table 4. The accuracy of the model is 0.83 and 0.84 for the hotel reviews and the news articles, respectively.

At inference time, we use the classifiers' output probabilities of *faithfulness*, *style adherence*, and *coherence* for the three conditions in Equation 2. To train the classifiers, the positive and negative training instances are generated from the training instances in the content transfer experiment, and so for the validation and the test instances. As a result, we make sure the classifiers do not learn any information from the validation and test sets.

4.3. Baselines

As baselines for our approach, we select the following two models that can be considered state-of-the-art to the best of our knowledge. We trained both baselines and optimized their hyperparameters on the validation set to create strong baselines.[3]

[2]Since our goal is to compare models with and without the multi-task learning architecture, it is important to note that the base model can be substituted with other advanced large language models.

[3]As mentioned in Section §2, the commonly used subject-relation-object triplets fail to capture negation relations and are thus not covered by our baselines. Also, we do not compare to pre-trained masked language models. While we could mask the target sentence in the given case, these models normally deal with text where only about 15% is masked and are not suitable to

Dataset	Training			Validation			Test		
	Faithful.	Neutral	Non-faith.	Faithful.	Neutral	Non-faith.	Faithful.	Neutral	Non-faith.
Hotel Reviews	23,471	34,478	16,925	5,158	7,225	3,861	6,489	9,205	4,890
News Articles	4,064	3,182	3,466	533	528	333	1,457	1,039	917

Table 3: The number of instances in the training, validation, and test set for the faithfulness classifier training. We randomly selected negative samples to have equal number as positive ones as much as possible.

Dataset	Training		Validation		Test	
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
Hotel Reviews	6931	6931	1445	1445	1841	1841
News Articles	960	960	124	124	315	315

Table 4: The number of instances in the training, validation, and test set for the style adherence and the coherence classifiers. The positive label (Pos.) means style adherence or coherence, respectively; the negative label (Neg.) means no style adherence or incoherence.

Sequence-to-Sequence Model We compare to the closely related sequence-to-sequence training strategy from our previous work (Chen et al., 2021b). As a conditional text generator, the inputs are the sentence before, $sent_{before}$, the sentence after, $sent_{after}$, the topic, t, and the reference, r. The target output is the sentence in the middle. Similar to Chen et al. (2021b), the four inputs are concatenated together using special tokens as

$$\texttt{[SB]}\ sent_{before}\ \texttt{[/SB][SA]}\ sent_{after}\ \texttt{[/SA]}\ \texttt{[T]}\ t\ \texttt{[/T]}\ \texttt{[R]}\ r\ \texttt{[/R]},$$

where the bracketed symbols are special tokens. As for our approach, the base model is `bart-base`.[4]

Error Correction Model On the other hand, we consider the architecture proposed by Thorne and Vlachos (2021). Given an input claim, the model conditionally generates a corrected version of the claim based on a retrieved evidence from Wikipedia. In our case, instead of using the retrieval component, we directly provide the ground-truth evidence to the input. In particular, we train a sequence-to-sequence model whose input consists of the sentence before, $sent_{before}$, the sentence after, $sent_{after}$, as well as the topic, t, concatenated with special tokens. The target then is the sentence in the middle. We used the trained model to generate the first draft output. Then, we trained an error correction model with the draft output and the reference, r, as input, and the target, y, as output. In other words, we have an ideal error correction case here: we exactly know the best r to correct the faithfulness error in the draft output.

recover a whole sentence (Wettig et al., 2022).

[4]We use the base model outlined in Section 4.2. Our focus is not on employing the most advanced model, but on evaluating how the multi-task learning paradigm enhances overall effectiveness.

	Hotel Reviews			News Articles		
Approach	R.-1	R.-2	R.-L	R.-1	R.-2	R.-L
Chen et al. (2021b)	30.9	11.9	27.0	17.2	3.2	13.2
Error correction	**31.0**	**12.1**	**27.2**	16.5	2.7	12.6
Our approach	**31.0**	12.0	27.0	**17.4**	**3.2**	**13.4**

Table 5: Rouge-{1, 2, L} F_1-scores of the two baselines and our approach on the hotel reviews and news articles. The best score in each column is marked bold.

5. Results and Discussion

This section discusses the automatic and manual evaluation results of our approach and the baselines. We analyze selected examples qualitatively and discuss the hyperparameters of the model.

5.1. Automatic Evaluation

We first evaluate the generated texts and the fulfillment of the input conditions using ROGUE F_1-scores and available automatic metrics:

ROUGE F_1-Scores Table 5 shows that the *error correction* model does best on the hotel reviews, while *our approach* outperforms both baselines on the news articles. Yet, the differences between the scores of the three approaches are small.

Condition Fulfillment We consider the following automatic metrics to evaluate the condition fulfillment.[5] For faithfulness, we used the BERTScore (Zhang et al., 2020a) with its best model `deberta-xlarge-mnli` from Microsoft to predict the similarity between the generated text and the reference. For style adherence,

[5]We refrain from using the teacher classifiers as evaluators since they are integrated in our approach.

	Hotel Reviews			News Articles		
Approach	**Fai↑**	**Sty↑**	**Coh↓**	**Fai↑**	**Sty↑**	**Coh↓**
Chen et al. (2021b)	**.651**	**.968**	40.7	.590	.412	18.2
Error correction	.643	.956	40.2	**.615**	**.571**	18.7
Our approach	.647	.965	**38.4**	.591	.479	**17.9**
w/o selection	.647	.956	40.0	.586	.391	18.1

Table 6: Automatic evaluation: Proportion of texts fulfilling the faithfulness (*Fai*), style adherence (*Sty*) and coherence (*Coh*) conditions on each dataset. *w/o selection* denotes the proportion before candidate selection. The best score per column is marked bold.

we follow Yin et al. (2019) to use a vanilla `bart-large-mnli` model as the zero-shot topic classifier to predict the probability that the generated text has the desired topic. Finally for coherence, we concatenated $sent_{before}$, generated text, and $sent_{after}$ as a single string and then computed the perplexity based on GPT-2 (Radford et al., 2019).[6]

Table 6 shows that *our approach* has the lowest perplexity while it has the second-best performance in faithfulness and style adherence in both datasets. Both baselines are the best on one dataset and the worst on the other. Such unstable results illustrate the limitation of these two baselines. We also see that the candidate selection (see Figure 1c) improves fulfillment of the three conditions, especially style adherence. However, the two baselines and the two variations of our approach are all very close to each other using automatic evaluations.

5.2. Manual Evaluation

Since the automatic evaluation only approximates the actual quality, we carried out a manual study where humans judged the faithfulness, style adherence, and coherence of the generated sentences. For the study, we randomly selected 100 instances from each dataset (200 in total). We presented participants with the generated text as well as two sentences: one before and one after, along with the topic and the reference. On this basis, we posed three questions to them about the generated text.

Q1. What is the relationship between the sentence and the reference?
{ The sentence entails the reference (1)
| The reference entails the sentence (1)
| The sentence partly entails the reference(0.5)
| The reference partly entails the sentence(0.5)
| Neither entailment nor contradiction (0)
| They partly contradict each other (-0.5)
| They contradict each other (-1) }

Q2. Does the sentence cover the given topic?
{*yes (1)* | *partially (0.5)* | *no (0)* }

Q3. Is the sentence coherent with the others?
{*yes (1)* | *partially (0.5)* | *no (0)* }

[6]Although perplexity is not a direct measure of text coherence, it has been shown to reflect coherence to a certain degree.

	Hotel Reviews			News Articles		
Approach	**Fai↑**	**Sty↑**	**Coh↑**	**Fai↑**	**Sty↑**	**Coh↑**
Chen et al. (2021b)	0.49	0.88	0.96	0.70	0.73	0.98
Error correction	0.53	0.83	0.98	0.72	0.69	**0.99**
Our approach	**0.54**	**0.89**	**0.99**	**0.76**†	**0.76**†	**0.99**

Table 7: Manual evaluation (main results): Mean scores of faithfulness (Fai), style adherence (Sty), and coherence (Coh) on each dataset. The best score in each column is marked bold. The †symbols denoting a significance ($p < 0.05$) comparing to the second best.

The numbers in parentheses indicate the scores of each answer. For faithfulness (Q1), entailment from either direction has a positive score while contradiction has a negative score.

We used Amazon Mechanical Turk to acquire four annotations per instance, at a cost of $0.33 per instance. To get more reliable results, we required the workers to have at least 95% approval rate and at least 10,000 approved HITs. Furthermore, we applied the multi-annotator competence estimation (MACE) measure (Hovy et al., 2013) to estimate the final label. MACE achieved around 90% of accuracy in predicting the labels.

Table 7 presents the manual evaluation results. For *faithfulness*, our approach has the best score compared to the other two baselines. For hotel reviews, it outperforms the baselines with a small margin, while the difference is more significant on the news articles. In terms of *style adherence*, our approach is still the best while error correction is the worst. The reason for this could be the two-step process harms more to the style adherence. For the *coherence* condition, all approaches perform almost equally well on both datasets, only Chen et al. (2021b) is slightly worse.

Comparing the two baselines, we found that Chen et al. (2021b) is better in generating style-adherent texts, while the error correction model is better in generating faithful texts. Conceptually, Chen et al. (2021b) was developed to capture the frame (style) in the texts, while the error correction model was developed to fix the *error* (faithful mistakes) in the texts. The evaluation results also reflect such designed advantages of the two models.

Generally, we observe that all models perform

Topic. view
Reference. We arrived to a friendly welcome and a very comfortable King Premier room, which I had requested at the front of the hotel so that there was a *view of the city and waterfront*.
Chen et al. (2021b). Fantastic Hotel, Brilliant staff, let down by NOISE. **And the view was spectacular.** I would have given this hotel a 4 or 5 rating.
Error correction. Fantastic Hotel, Brilliant staff, let down by NOISE. **Great location, great view of the Eiffel Tower.** I would have given this hotel a 4 or 5 rating.
Our approach. Fantastic Hotel, Brilliant staff, let down by NOISE. **I have stayed at this hotel before and had a fantastic view of the city from my room.** I would have given this hotel a 4 or 5 rating.

Table 8: Sample from a hotel review (Wachsmuth et al., 2014). The bold sentences are generated by the baselines and our approach, given topic and reference.

significantly better on news articles in terms of faithfulness when compared to hotel reviews. This difference could be attributed to the nature of reviews, where opinions play a central role. Determining if one opinion implies another appears to be more challenging in this context.

In contrast, we assume that the higher style adherence on the hotel reviews is due to the different text layout. In news articles, the topics are usually more invisible. For example, the text in Figure 2b mentioned Nancy Pelosi with the pronoun "her". Such a writing style makes it more difficult to generate topic-relevant texts.

5.3. Qualitative Analysis

In addition to quantitative evaluations, we conducted a qualitative analysis to gain deeper insights into the approach's effectiveness. Particularly, we look at one sample of generated texts of the three approaches for a hotel review (Table 8) and one for a news article (Table 9).

For the former, the model was asked to generate a sentence regarding the topic "view", given a reference and the surrounding sentences. We see that the model of Chen et al. (2021b) generates rather generic text. While the error correction model output mentions the great view, the hotel is actually in Barcelona, not in Paris, so the statement regarding the Eiffel Tower cannot be true. In contrast, our model correctly states that the hotel has a city view which can be inferred from the reference.

In the more complex example in Table 9, the reference text is about the comment that Trump was sidestepping the congress to access certain federal funds. Here, the model of Chen et al. (2021b) generates an ironic sentence where Trump criticizes himself. The error correction model also talks about Trump, but "the repeal of the Affordable Care Act" has nothing to do with the reference. Our approach negatively comments that "It's a dangerous precedent" with the respect to both the reference and the subsequent sentence on the White House.

Topic. Donald Trump
Reference. In this case, Trump is sidestepping Congress so that he can gain access to certain federal funds without congressional approval.
Chen et al. (2021b). Ms. Coulter fired back shortly afterward. **"This is a power grab by a president who has disregarded the will of the American people," Mr. Trump said.** White House officials rejected criticism from across the ideological spectrum that Mr. Trump was creating a precedent that future presidents could use to ignore the will of Congress.
Error correction. Ms. Coulter fired back shortly afterward. **Mr. Trump, who has repeatedly called for the repeal of the Affordable Care Act, called the bill "unconstitutional" and called it a "disaster" for the country.** White House officials rejected criticism from across the ideological spectrum that Mr. Trump was creating a precedent that future presidents could use to ignore the will of Congress.
Our approach. Ms. Coulter fired back shortly afterward. **"It's a dangerous precedent" Mr. White House officials said in a brief statement.** White House officials rejected criticism from across the ideological spectrum that Mr. Trump was creating a precedent that future presidents could use to ignore the will of Congress.

Table 9: Sample from a news article (Fan et al., 2019). The bold sentences are generated by the baselines and our approach, given topic and reference.

In summary, the two examples suggest that the model of Chen et al. (2021b) tends to generate text without much detail. The error correction model can provide more precise details, but part of the details tend to go wrong. Overall, our approach generates the most reasonable texts.

5.4. Hyperparameters

To optimize our approach and the baseline models, we tuned their hyperparameters to maximize performance on the validation set. In particular, we considered the number of training steps and weights in Equation 1. We logged the results in every half epoch, with maximum 5 epochs of training. For the weights, we set the minimum value to be 0.1 and the maximum to be 0.7 (since $0.1 + 0.1 + 0.1 + 0.7 = 1$). We validated the model in every combination of the weights with a grid size 0.1. This gives a total of 84 combinations.

In the end, all models were saturated between the second and the third epoch. For the weights,

we found the best combination for hotel reviews is 0.4 for generation loss and 0.2 for all conditions. For news articles, the best combination for hotel reviews is 0.7 for generation loss and 0.1 for all conditions. The difference between the best combination and the worst one was about five points in terms of ROUGE score. The higher generation loss for news articles suggests that it is harder to generate news texts, so the models have to learn more from the generation loss.

6. Conclusion

Recent generation models often produce faithfully incorrect text, which raises ethical concerns and limits their impact in many real-world applications. For content transfer tasks, this paper has therefore focused on generating faithful text, integrating generation with faithfulness, style adherence, and coherence constraints in a multi-task learning approach. Compared to the baselines in our manual evaluation, our approach succeeded in generating faithful, style-adherent, and coherent text.

A key strength of our approach is that we can use any reference as input for the faithfulness constraint. At the same time, using a variable set of classifiers provides a new, flexible mechanism to fulfill the desired conditions. While we thereby contribute towards faithfulness in context transfer, our results have also indicated much room for further improvement. In future work, we plan to study more conditions in text generation and consider other conditional text generation tasks.

7. Limitations

Our experiments involve only one type of conditional text generation task, namely, content transfer. Further research is needed to verify the transferability of our results to other text generation tasks, such as text summarization and question answering. Moreover, we have included three conditions only. We expect that more conditions will make the task more complex. Finally, the applicability of our approach is limited to settings where references are provided as natural language texts about a topic in focus or can be derived reliably in some way. In other scenarios, such kind of references may have to be generated or retrieved from other sources.

8. Ethical Concerns

The intended application of our approach is to improve faithfulness in text generation and thereby further reduce the amount of misinformation that text generation models produce. We, therefore, target positive ethical consequences with our work in general.

However, the actual faithfulness of an output text strongly depends on the quality of the data used as underlying information. In our experiments, we do not verify the faithfulness of the information coming from reference sentences. The references could contain different biases, for example, political bias in the case of the news article dataset and subjective or even invented information in the case of the hotel reviews. Thereby, our approach may reinforce present biases of misinformation by transferring them to other texts as well. Furthermore, we cannot prevent our approach from being misused to intentionally generate texts containing misinformation or bias using respective reference texts. Ultimately, this does not create new ethical concerns, though, but rather leads back to general correctness problems in text generation.

Even with faithfully-correct reference texts, the non-perfect effectiveness of our approach can lead to faithfully incorrect sentences being generated. Ideally, the approach should be used in a semi-automatic setting with human post-checking to prevent the spread of misinformation.

Finally, we note that no personal information was collected from any participant in our crowd-based user study. There is no way to match the created annotations to their identities. As stated in the paper, the participants were paid on a per-instance level. The reward ($0.33) was defined, so as to achieve a fair salary. We did not receive feedback from crowd-workers that indicates otherwise.

Acknowledgments

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Center “On-The-Fly Computing” (SFB 901/3) under the project number 160364472, and by the AISafety Project, funded by the BMBF under the grant proposal 05D23PD1. This research was also supported by the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

9. Bibliographical References

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128. Association for Computational Linguistics.

Ioannis Arapakis, Filipa Peleja, Barla Berkant, and

Joao Magalhaes. 2016. Linguistic benchmarks of online news article quality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1893–1902.

Mario Barrantes, Benedikt Herudek, and Richard Wang. 2020. Adversarial NLI for factual correctness in text summarisation models. *CoRR*, abs/2005.11739.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *CoNLL 2016*, pages 10–21.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6251–6258. Association for Computational Linguistics.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Kunal Chawla and Diyi Yang. 2020. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2340–2354. Association for Computational Linguistics.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021a. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2021b. Controlled neural sentence-level reframing of news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2683–2693. Association for Computational Linguistics.

Yu Cheng, Zhe Gan, Yizhe Zhang, Oussama Elachqar, Dianqi Li, and Jingjing Liu. 2020. Contextual text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2915–2924. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA. Association for Computing Machinery.

Ulla Connor. 2002. New directions in contrastive rhetoric. *TESOL quarterly*, 36(4):493–510.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 623–632.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9320–9331. Association for Computational Linguistics.

Amy Einsohn. 2011. *The copyeditor's handbook: A guide for book publishing and corporate communications*. University of California.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2214–2220. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6342–6348. Association for Computational Linguistics.

Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. 2012. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 63–72. ACM.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second Conference on Artificial Intelligence*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 634–642.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 166–175. ACM.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, et al. 2009. *International corpus of learner English*, volume 2. Presses universitaires de Louvain Louvain-la-Neuve.

Tim Groseclose and Jeffrey Milyo. 2005. A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2017. Generating sentences by editing prototypes. *arXiv preprint arXiv:1709.08878*.

Edward T Hall. 1976. Beyond culture. garden city. *NY: Anchor*.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 507–517, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jonathan Herzig, Michal Shmueli-Scheuer, Tommy Sandbank, and David Konopnicki. 2017. Neural response generation for customer service based on personality traits. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 252–256.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Benjamin D Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. *arXiv preprint arXiv:1803.10124*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning whom to trust with MACE. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 1120–1130. The Association for Computational Linguistics.

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

Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. 2020. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2213–2223, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Ann M Johns. 1986. The ESL student and the revision process: Some insights from schema theory. *Journal of Basic Writing*, 5(2):70–80.

Robert B Kaplan. 1966. Cultural thought patterns in inter-cultural education. *Language learning*, 16(1-2):1–20.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *The 2nd International Conference on Learning Representations*.

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 2020, Online, November 16-20, 2020*, pages 9332–9346. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 1489–1500.

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.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *North American Association for Computational Linguistics*.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *CoRR*, abs/2203.05227.

Yu-Ru Lin, James P. Bagrow, and David Lazer. 2011. More voices than ever? Quantifying media bias in networks. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*.

ZC Lipton, S Vikram, and J McAuley. 2016. Generative concatenative nets jointly learn to write and classify reviews. *CoRR*, 1511:1–11.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Donghong Liu and Fang Wan. 2020. What makes proficient writers' essays more persuasive? A Toulmin perspective. *International Journal of TESOL Studies*, 2(1):1–13.

Siyi Liu, Sihao Chen, Xander Uyttendaele, and Dan Roth. 2021. Multioped: A corpus of multi-perspective news editorials. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*,

pages 4345–4361. Association for Computational Linguistics.

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

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Jonas Mueller, David Gifford, and Tommi Jaakkola. 2017. Sequence to better sequence: Continuous revision of combinatorial structures. In *International Conference on Machine Learning*, pages 2536–2544.

Feng Nan, Cícero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen R. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6881–6894. Association for Computational Linguistics.

Jin Kyung Paek and Yusun Kang. 2017. Investigation of content features that determine korean EFL learners' argumentative writing qualities. *English teaching*, 72(2):101–122.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1534–1543, Baltimore, Maryland. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 543–552, Barcelona, Spain (Online). Association for Computational Linguistics.

Shrimai Prabhumoye, Chris Quirk, and Michel Galley. 2019. Towards content transfer through grounded text generation. 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 2622–2632, Minneapolis, Minnesota. Association for Computational Linguistics.

Anushka Prakash and Harish Tayyar Madabushi. 2020. Incorporating count-based features into pre-trained models for improved stance detection. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 22–32, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga. 2021. Annotating argumentative structure in English-as-a-foreign-language learner essays. *Natural Language Engineering*, pages 1–27.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill 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, Florence, Italy. Association for Computational Linguistics.

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

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. 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 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1278–1286.

Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 440–450.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66.

Stephanie Schoch, Wanyu Du, and Yangfeng Ji. 2021. Contextualizing variation in text style transfer datasets. In *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021, Aberdeen, Scotland, UK, 20-24 September, 2021*, pages 226–239. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083.

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*, pages 6833–6844.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor.*, 19(1):22–36.

Tony Silva. 1993. Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL quarterly*, 27(4):657–677.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3298–3309. Association for Computational Linguistics.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 1680–1691. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics*,

pages 115–127, Berlin Heidelberg New York. Springer.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting cognitive and emotional empathic writing of students. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4063–4077, Online. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*, pages 3731–3741. Association for Computational Linguistics.

Peter West, Chris Quirk, Michel Galley, and Yejin Choi. 2022. Probing factually grounded content transfer with factual ablation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3732–3746, Dublin, Ireland. Association for Computational Linguistics.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2022. Should you mask 15% in masked language modeling? *arXiv preprint arXiv:2202.08005*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3881–3890.

Tae Yano, Philip Resnik, and Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 152–158. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China. 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 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

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*.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis P. Langlotz. 2020b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5108–5120. Association for Computational Linguistics.

Chunting Zhou and Graham Neubig. 2017. Multi-space variational encoder-decoders for semi-supervised labeled sequence transduction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 310–320.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1393–1404. Association for Computational Linguistics.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 718–733. Association for Computational Linguistics.