MoFE: Mixture of Factual Experts for Controlling Hallucinations in Abstractive Summarization

Anonymous authors
Paper under double-blind review

Abstract

Neural abstractive summarization models are susceptible to generating factually inconsistent content, a phenomenon known as hallucination. This limits the usability and adoption of these systems in real-world applications. To reduce the presence of hallucination, we propose the Mixture of Factual Experts (MoFE) model, which combines multiple summarization experts that each target a specific type of error. We train our experts using reinforcement learning (RL) to minimize the error defined by two factual consistency metrics: entity overlap and dependency arc entailment. We construct MoFE by combining the experts using two ensembling strategies (weights and logits) and evaluate them on two summarization datasets (XSUM and CNN/DM). Our experiments on BART models show that the MoFE improves performance according to both entity overlap and dependency arc entailment, without a significant performance drop on standard ROUGE metrics. The performance improvement also transfers to unseen factual consistency metrics, such as question answer-based factuality evaluation metric and BERTScore precision with respect to the source document.

1 Introduction

Neural abstractive summarization systems trained by maximizing the likelihood of a reference summary (MLE) given its source document have been shown to generate plausible summaries with high lexical overlap with the references. However, human analyses (Fabbri et al., 2021; Pagnoni et al., 2021; Tejaswin et al., 2021) and automatic evaluations (Falke et al., 2019; Kryscinski et al., 2020; Maynez et al., 2020; Durmus et al., 2020) show that state-of-the-art neural models, trained on widely used XSUM (Narayan et al., 2018) and CNN/DM (Hermann et al., 2015) datasets, tend to hallucinate information with high frequency. For instance, by performing human evaluations on 2250 model generated summaries from CNN/DM and XSUM datasets, Pagnoni et al. (2021) found that 60% of the summaries contained at least one factual error. The hallucinations are broadly classified as *extrinsic*, when a model adds information that is not present in the source document, and *intrinsic*, when the model distorts information present in the source document into a factually incorrect representation. The type and degree of a model's hallucinations correlate with the quality of training data. As noted by Pagnoni et al. (2021), models trained on the XSum data, which include extrinsic hallucinations in reference summaries, tend to generate a higher proportion of extrinsic hallucination as compared to models trained on the cleaner CNN/DM dataset.

In this paper, we propose the Mixture of Factual Experts (MoFE), a simple framework that applies an ensemble of factual experts to control hallucinations in summarization systems. We define *factual expert* as a model that generates summaries with certain desirable factual qualities (e.g. fewer extrinsic hallucinations). Each constituent factual expert in MoFE is trained to target a unique type of factual quality. The training of the experts is motivated by two broad observations. First, the *data* on which the model is trained may influence the factual consistency of the model (Pagnoni et al., 2021). Therefore, we employ a data pre-processing step that filters training samples such that the references exhibit the desirable factual qualities. Second, the maximum-likelihood *loss function* may overlook factual consistency. Therefore we employ reinforcement learning (RL) to train a model using explicit signals of factual consistency. Further, we augment standard reward-based RL loss with a KL divergence loss between the expert and pre-trained model's distributions on *pivot*

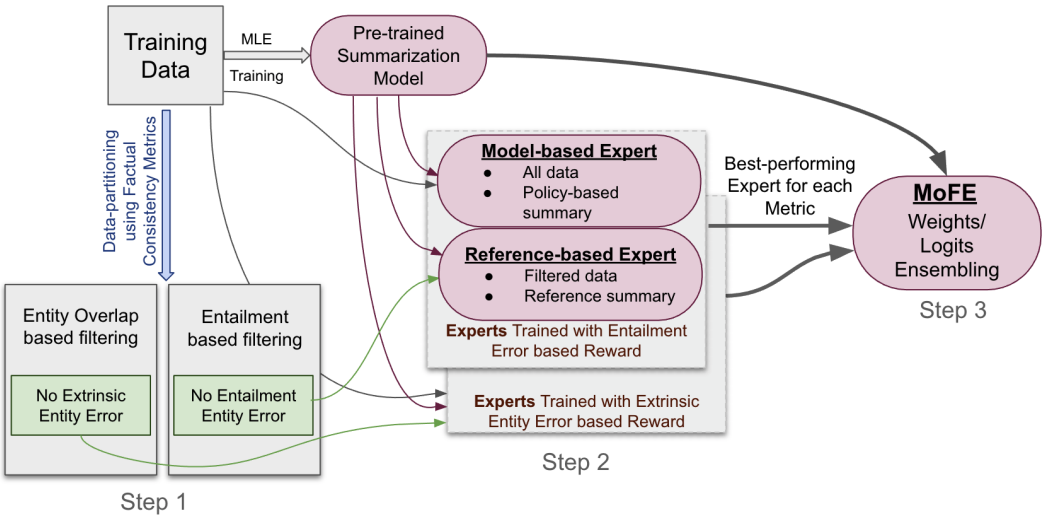


Figure 1: Schematic view of steps for building the MoFE model. In the first step, it uses automated factual consistency metrics to filter out training samples with the desirable factual quality. Then in the second step, it trains reference- and model-based expert models on the filtered and whole training set respectively. Finally, in the third step, it combines the best-performing experts through weights and logits ensembling.

summary[1] to prevent the former from deviating too far from the latter. We propose to choose the pivot summary depending on the number of factual errors in training samples. When all samples in the training data possess desirable factual consistency, we use reference as the pivot summary. On the contrary, if training data contains factual errors, we use the expert-sampled summary as the pivot. We show the schematic view of MoFE in Figure 1.

We use entity overlap and dependency arc entailment (DAE) accuracy (Goyal & Durrett, 2020) metrics as measures of extrinsic and intrinsic hallucinations, respectively, and accordingly use both metrics to define rewards for training experts targeting both types of hallucination. Entity overlap evaluates the number of entities in summary that are absent from the source document and is a direct measure of extrinsic hallucination. Intrinsic hallucination, on the other hand, is broader and includes errors such as incorrect predicates or their arguments, coreference errors, discourse link errors, etc. (Pagnoni et al., 2021). Since DAE accuracy measures the fine-grained entailment relations at the dependency arc level, we consider it a reasonable proxy for measuring intrinsic hallucinations (Goyal & Durrett, 2020; 2021). Additionally, given that experts trained on both entity overlap and DAE metrics try to improve precision and are prone to reducing factual recall, MoFE also includes an entity recall-based expert. Subsequently, we combine the above three experts through logits and weights ensembling.

We evaluate our MoFE on the two benchmark abstractive summarization datasets, XSUM and CNN/DM. We use a diverse set of metrics, including entailment, entity overlap, and question answering (QA)-based metrics to measure factual errors. We find that MoFE models strongly outperform the state-of-the-art BART model (Lewis et al., 2020), obtaining up to $\sim 6\%$ absolute improvement on factual consistency metric (summary-level DAE accuracy) used to train experts, with marginal degradation (< 0.74) on ROUGE scores. Further, MoFE performs better than BART on two QA-based metrics (FEQA (Durmus et al., 2020) and QuestEval (Scialom et al., 2021)) on XSUM dataset. However, on CNN/DM, we find contrasting results, MoFE improves on QuestEval but not on FEQA, which may be attributed to the differences in questions generated by the two systems. Finally, we use the SummVis tool (Vig et al., 2021) to analyze MoFE and BART models' summaries on the XSUM dataset. We find that, relatively, MoFE reduces factual errors, some of which are clear cases of hallucinations while others may represent world knowledge (e.g. replacing the UK with the

[1]We name the summary used to prevent expert from diverging too far from the pre-trained model as *pivot*.

United Kingdom). In the appendix, we show some example cases where MoFE-generated summary removed, added, or ignored factual errors in the BART-generated summary.

2 Related Work

Factual consistency metrics and analysis Abstractive text summarization metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) evaluate lexical and semantic overlap respectively but fail to sufficiently evaluate factuality and faithfulness (Tejaswin et al., 2021). This has led to a line of research dedicated to evaluating factual consistency and hallucination in text summarization using new metrics and analyses. Some recent works have applied natural language inference (NLI) based models to test for factual consistency (Falke et al., 2019; Kryscinski et al., 2020; Maynez et al., 2020). Goyal & Durrett (2020) applied such entailment models at the dependency level and showed them to be more effective at localizing factual errors in generated summaries.

Another line of work uses question generation and question answering for evaluating text summarization. This includes the APES (Eyal et al., 2019), an RL-based metric SummaQA (Scialom et al., 2019), QAGS (Wang et al., 2020), FEQA (Durmus et al., 2020), and QuestEval (Scialom et al., 2021). All of these metrics can be used as proxies for entity-focused factual consistency evaluation and in particular QAGS, FEQA, and QuestEval have fact-based evaluation as their primary motivation. More recently, Nan et al. (2021) proposed entity-precision metric focusing on entity level factual consistency.

The slew of work on factual evaluation metrics has given rise to research focused on comparing, analyzing, and benchmarking these metrics on various text summarization datasets. Gabriel et al. (2021) show that although QA metrics are better than general metrics for evaluating factuality, they are extremely sensitive and there is no clear winner. Some of the analysis work has focused on collecting human annotations for factual consistency errors, categorizing the errors, and measuring their correlations with automated metrics (Fabbri et al., 2021; Pagnoni et al., 2021; Goyal & Durrett, 2021; Tejaswin et al., 2021). These evaluation studies have contradicting observations. For instance, Durmus et al. (2020) found that entailment-based automated metrics have lower correlation with faithfulness than the QA-based metrics. On the other hand, Pagnoni et al. (2021) concluded that entailment-based FactCC and semantic overlap-based BERTScore precision with respect to the source document exhibit the highest correlations with the human judgment of factuality, and the correlation between FEQA and factual consistency is insignificant. Given the variations in findings from different human analyses of popular factual consistency evaluation metrics, we select a few metrics from each of the entailment, entity overlap, and QA-based evaluations, as well as use ROUGE and BERTScore metrics for evaluating MoFE.

Methods for enforcing factual consistency Along with the growing body of work on analysis and evaluation of factual consistency, there has been some recent work on developing methods to enforce factual consistency in pre-trained language models. These include sampling techniques such as constrained decoding (Mao et al., 2020) and neurologic decoding (Lu et al., 2020). Another strategy is to control generation either by using language models to guide a base language model as in GeDi (Krause et al., 2020) and DExperts (Liu et al., 2021a) or via a hallucination knob (Filippova, 2020). Although these methods claim to be generic, they haven't been successfully applied to constrain summary generation on the source document.

Comparatively, there are fewer papers that propose methods for factual consistency in text summarization. Most of these focus on posthoc correction such as SpanFact (Dong et al., 2020), contrast entity generation and selection (Chen et al., 2021), and encoding SRL structure (Cao et al., 2020). Aralikatte et al. (2021) use focus attention and sampling to improve diversity and faithfulness of summaries while Liu et al. (2021b) use data augmentation with the contrastive loss for factual consistency of abstractive summarization applied to customer feedback.

3 MoFE Model

We propose Mixture of Factual Experts (MoFE) to improve the factual consistency of text summarization systems. As illustrated in Figure 1, MoFE consists of three main steps. First, we filter the training dataset to obtain samples that are factually consistent, using automated metrics between

source document and reference summary (§3.2). We discuss the automatic metrics used for building and evaluating MoFE in §3.1 Then, we use reinforcement learning to train two variants of expert models for each factual consistency metric, which both learn to minimize corresponding factual errors while also minimizing KL divergence between the expert's and pre-trained model's distributions on the so-named pivot summary. 1) *Reference-based expert* is trained on filtered samples that are factually consistent according to the given metric and use human-written summaries as the pivot (§3.3.1). 2) *Model-based expert* is trained on the entire training dataset and use summaries sampled from the then expert as the pivot (§3.3.2). Finally, we select one expert with the least factual error, between the reference and model-based experts, for each metric and combine them through weights (Izmailov et al., 2018) or logits ensembling to construct the final MoFE summarization system (§3.4).

3.1 Automated Metrics for Measuring Factual Consistency

There are three popular paradigms for evaluating the factual consistency of summaries generated by a model. 1) The simplest method includes measuring token-level overlap between the information of interest (e.g. named entities) in the summary and source document (Nan et al., 2021). This metric can be used as a proxy to measure simpler cases of hallucinations, such as extrinsic entity errors. We use *entity-overlap* to both train and evaluate factual experts. 2) The second type of evaluation builds on NLI and evaluates if the facts claimed in a summary is entailed by the source document (Kryscinski et al., 2020; Goyal & Durrett, 2020; Maynez et al., 2020). Two popular entailment-based metrics include FactCC (Kryscinski et al., 2020) which measures entailment at the summary-level and DAE (Goyal & Durrett, 2020) which measures fine-level entailment by breaking summary into smaller claims defined by dependency arcs[2]. Pagnoni et al. (2021) finds that DAE correlates with the human judgment of factuality, and has the highest correlation with complex discourse errors, such as entity coreference. Therefore, we use DAE to identify cases of intrinsic hallucinations, both during training and evaluation. 3) The most complex methods for evaluating factuality rely on question generation (QG) and question answering (QA) (Durmus et al., 2020; Scialom et al., 2021). They first use a QG module to generate questions based on summaries and then use another QA module to find answers in the source document. They are computationally expensive to use to train experts. Therefore, we use them exclusively to evaluate the generalizability of MoFE to new factual evaluation metrics.

3.2 Training Data Filtering

Recent studies show that reference summaries in common text summarization datasets often contain factual errors (Tejaswin et al., 2021; Nan et al., 2021), which accounts for one of the known sources of hallucination in summarization models. Therefore, in the first step, we apply automatic factual consistency evaluation metrics to filter factually consistent training samples. We apply metrics that target extrinsic and intrinsic hallucinations, and create a filtered training subset for each. To identify extrinsic hallucinations, we measure entity overlap between the source document and the reference summary, using SpaCy (Honnibal et al., 2020) to identify named entities. We filter training samples in which all the entity tokens in reference summary are also mentioned in the source document. To identify intrinsic hallucinations, we measure the dependency arc entailment (DAE) (Goyal & Durrett, 2021) between the source and reference summary. We filter all training samples where all of the dependency arcs in the summary are entailed by the source documents. Subsequently, we use the above two filtered subsets to train reference-based (§3.3.1) experts targeting extrinsic and intrinsic hallucinations as well as the factual recall.

3.3 Training Factual Expert Models

In addition to factual errors in training data, the MLE training objective is another known source of hallucination. A model trained by maximizing the log-likelihood of reference summaries can efficiently learn to generate summaries with high n-gram overlap but may fail to learn to enforce factual consistency. Therefore, we train our factual experts by directly optimizing for the factual

[2]Dependency arcs define grammatical structures in a sentence and often describe semantic connections between words, such as predicate-argument relations. It provides a fast mechanism to identify intrinsic errors involving relationships between entities.

consistency using the self-critic algorithm (Rennie et al., 2017), a frequently use reinforcement learning technique for training NLP models.

We consider parameters of an expert (θ) as the policy model and define action as predicting the next token in a summary sequence. Given a factual consistency metric M, we define the action reward $R_{(y,\hat{y})}$ as the score of the generated summary (y) according to M. Here, $\hat{y}$ is the source document for precision-based factual consistency metrics (e.g. DAE accuracy), and the reference summary for fact recall-based metrics (e.g. Entity recall). Further, in accordance with the self-critic training, we use the test-time greedy decoding strategy (i.e. $argmax$) to obtain a summary and calculate the baseline reward ${R^a}_{(y,\hat{y})}$. We subtract the baseline reward from the action-based reward ($R_{(y,\hat{y})}$) and use the resulting reward signal to train our experts. This minimizes the variance of the gradient estimate and importantly adjust the reward scale to provide both positive and negative values. Overall, we train our expert policy to minimize the negative of expected reward difference which, after Monte Carlo approximation (Williams, 1992), is defined as:

$$L_\theta^{fc} = -\mathbb{E}_x[(R_{(y,\hat{y})} - {R^a}_{(y,\hat{y})})\ log\ p_\theta(y|x)] \tag{1}$$

Following standard reinforcement learning-based sequence training formulations, we initialize the policy model with a text summarization model ϕ trained on human-annotated datasets. Further to prevent the policy from collapsing to single mode[3] or significantly deviating away from ϕ, we add an additional KL divergence loss (eq. 2) between the next token probabilities of the policy θ and baseline ϕ[4]. We train experts using the weighted sum of the two losses $\lambda L_\theta^{fc} + (1-\lambda)L_\theta^{kl}$.

$$L_\theta^{kl} = \mathbb{E}_x[p_\phi(y^*|x)\ log(p_\phi(y^*|x)\ /\ p_\theta(y^*|x))] \tag{2}$$

Equations 1 and 2 describe the general framework for training our experts. Note that we call the summary y^* in eq. 2 as the *pivot summary* to simplify description of our experts. Next, we explain the two variants of our expert, reference-based expert (§3.3.1) and model-based expert (§3.3.2) in the following two sections.

3.3.1 Reference-based Experts

We hypothesize that human-written reference summaries are generally more natural and preferable than the summaries generated by a summarization model. So, on training samples that do not contain factual errors, we propose to use reference as the pivot summary[5]. Since it uses reference as pivot summary, we call it *reference-based* expert. The reference-based expert is similar to prior work, such as Paulus et al. (2017), Li et al. (2018) and Pasunuru & Bansal (2018), that uses RL-based training to directly improve ROUGE, saliency, entailment, or other text quality metrics for a text summarization system. However, different from prior work, we optimize for factual consistency metrics that are more fine-grained and defined at word-level (entity overlap) or word pair-level (DAE). Secondly, noting that a significant percentage of summaries in commonly used text summarization datasets contain factual error, we propose to train reference-based experts on samples filtered according to the factual consistency metric that is being used to define the reward. For instance, a reference-based expert for DAE metric is trained using the samples filtered for intrinsic hallucinations.

3.3.2 Model-based Experts

When dataset contains frequent factual errors, minimizing KL divergence with respect to reference summary encourages the model to continue to uniformly increase probability mass on factually inconsistent references. This is problematic and may lower the gain from reward based loss. Under such scenarios, we propose to use summary sampled following probabilities from then expert (policy) model as the pivot summary. We call this expert *model-based* expert given it uses summary generated by the expert model as pivot. Intuitively, it prevents the expert from losing significant probability mass for sampled summary unless the sampled summary contains many factual error or the pre-trained model assigns very low probability to the summary (high perplexity).

[3]Policy learns to assign entire probability mass to a single token, setting both $R_{(y,\hat{y})}$ and $\hat{R}_{(y,\hat{y})}$ to zero and thereby reducing gradients to zero.

[4]Note that the KL divergence loss reduces the policy exploration. However, we believe this to be a reasonable trade-off for a high-entropy task, such as abstractive summarization, where factually consistent summaries are very few among all possible summary sequences. Further, as noted by Pang & He (2021), the benefit of exploration in training text generation systems is limited in the absence of perfect reward functions.

[5]Alternatively, we can replace the KL divergence loss in eq. 2 with the standard cross-entropy loss.

3.4 MIXING FACTUAL EXPERTS

Following the data filtering and RL training steps described in §3.2 and §3.3, we train both reference- and model-based experts for intrinsic and extrinsic hallucination using DAE accuracy and entity overlap precision metrics as rewards, respectively. Also, because experts for both intrinsic and extrinsic hallucinations are trained to improve precision with respect to the source document, they may negatively impact the content recall. So, we train entity-recall experts to maximize recall of salient entities between the generated summary and the reference summary. We train reference-based experts for DAE metric on data filtered for intrinsic hallucination, and for entity overlap precision and entity recall metric on data filtered for extrinsic hallucination. The model-based experts for both metrics are trained on the entire training dataset.

Next, for each of the three kinds of experts, we select one of the reference- and model-based experts having least factual error (or maximum factual recall) on validation dataset and combine them through weights or logits ensembling. We use the element-wise weighted average of all the parameters of pre-trained summarization model (ϕ) and expert models (θ^i_{expert}) for weights ensembling (eq. 3). For logits ensembling, we use the weighted average of logits (z_t) from all the experts ($z_{t,\theta^i_{expert}}$) and the pre-trained model ($z_{t,\phi}$) during decoding, as described by the eq. 4.

$$\theta_{final} = \sum_i (\alpha^i \theta^i_{expert}) + (1 - \sum_i \alpha_i)\phi \tag{3}$$

$$p(y_t|x, y_{<t}) = softmax(\sum_i (\alpha^i z_{t,\theta^i_{expert}}) + (1 - \sum_i \alpha_i) z_{t,\phi}) \tag{4}$$

4 EXPERIMENTS

4.1 DATA AND EVALUATION METRIC

We evaluate MoFE on XSUM (Narayan et al., 2018) and CNN/DM (Hermann et al., 2015) datasets. The XSUM data is highly abstractive and noisy while CNN/DM is more extractive but contains fewer factual errors (Tejaswin et al., 2021). We use standard ROUGE-1/2/L (R1/R2/RL), DAE-arc accuracy (DAE-A), and DAE-summary accuracy[6] (DAE-S), entity precision with respect to source (NER-PS) and entity recall with respect to the reference (NER-RT) as primary evaluation metrics for individual experts and the MoFE model. Among these seven metrics, DAE-A/S and NER-PS evaluate the factual consistency of a summary with respect to the source document. Separately, we also evaluate the MoFE on BERTScore precision (BScore-P) and recall (BScore-R) with respect to source and two question answer-based evaluation metrics, FEQA and QuestEval (QEval).

4.2 MODEL

We use the *BART* (Lewis et al., 2020) released with Huggingface's transformer (Wolf et al., 2020) (*bart-cnn-large/ bart-xsum-large*) as the base summarization models. From the human-based analyses, Pagnoni et al. (2021) finds that BART generated summaries have the least number of factual errors. We adopt the standard hyperparameters for BART during the inference, e.g. beam size of 6, minimum and maximum sequence length of 11 and 62, etc. for the XSUM model, and beam size of 4, minimum and maximum sequence length of 56 and 142, etc. for the CNN/DM model.

Training Experts: We use Huggingface Transformers library (Wolf et al., 2020) (PyTorch (Paszke et al., 2017)) to implement our experts. We initialize each expert with the pre-trained BART model and fine-tune the decoder module on the weighted sum of RL and KL divergence losses (eq. 1 and 2). We keep encoder parameters fixed during the training. All experts are trained for 1 epoch with batch size of 32 using default training hyperpaperameters (optimizer: Adam, learning rate: 5e-5, adam β_1: 0.9, adam$_{\beta 2}$: 0.999, adam ϵ: 1e-8). We experiment with 3 values of λ: 0.9, 0.5, and 0.1.

We train three experts corresponding to three metrics: DAE accuracy (DAE), entity overlap precision with source (NER-P), and entity recall with reference (NER-R). We construct two variants of MoFE, MoFE$_{weights}$ and MoFE$_{logits}$ using weights and logits ensembling respectively. Note that

[6]We consider a summary accurate if all dependency arcs in summary are entailed by the source document.

Model	DAE-A	DAE-S	NER-PS	NER-RT	R1	R2	RL
BART	76.16	34.75	63.82	53.66	**45.34**	**22.21**	37.13
DAE	**83.83**	**46.83**	**69.09**	51.82	44.32	21.20	36.11
NER-P	76.81	36.02	67.37	53.69	44.51	21.58	36.48
NER-R	75.48	33.56	63.50	**55.04**	45.19	22.04	36.98
MoFE$_{weights}$	80.36	41.08	66.74	53.20	45.00	21.92	36.80
$\Delta_{Improve}$	+4.20	+6.33	+2.92	-0.46	-0.34	-0.29	-0.33
MoFE$_{logits}$	80.70	41.06	66.81	53.40	45.18	22.03	36.94
$\Delta_{Improve}$	+4.54	+6.31	+2.99	-0.26	-0.16	-0.18	-0.19

Table 1: DAE accuracy, entity precision, entity recall and ROUGE scores on XSUM test set.

Model	DAE-A	DAE-S	NER-PS	NER-RT	R1	R2	RL
BART	96.26	75.0	98.44	58.92	44.05	21.07	40.86
DAE	97.17	77.92	98.19	60.15	44.13	21.13	40.91
NER-P	95.38	68.18	98.31	61.11	**44.46**	**21.36**	**41.24**
NER-R	**98.36**	**88.76**	**99.21**	61.45	42.18	19.70	38.83
MoFE$_{weights}$	96.73	75.77	98.26	61.15	44.10	21.11	40.75
$\Delta_{Improve}$	+0.47	+0.77	-0.18	+2.23	+0.05	+0.04	-0.11
MoFE$_{logits}$	97.51	81.29	98.54	**62.89**	43.31	20.60	39.84
$\Delta_{Improve}$	+1.25	+6.29	+0.10	+3.97	-0.74	-0.47	-1.02

Table 2: DAE accuracy, entity precision, entity recall and ROUGE scores on CNN/DM test set.

we include an expert in MoFE only if it does not under-perform the BART model by more than 5% on any of the DAE-A/S, NER-PS/RT, and ROUGE metrics. We find experts'/BART's mixture weights using grid search, assigning a minimum value of 0.1 to each model and incrementing weights by the step size of 0.2 for XSUM data. On CNN data, we exclude NER-P expert from the MoFE given it failed to improve NER-PS accuracy on the validation set and degraded DAE-S accuracy by greater than 5%. Similar to XSUM, we use grid search to find mixture weights for CNN data, but we assigned a minimum weight of 0.2 to each expert and the BART model.

4.3 Results

Tables 1 and 2 summarize the results on XSUM and CNN/DM datasets respectively. First, all three experts outperform the BART model, on their respective factual consistency metric, for XSUM data. However, on CNN/DM, NER-P expert model performs slightly worse (-0.13) than the BART on the entity precision (NER-PS) metric. This is unsurprising given BART is consistent against extrinsic entity hallucination on CNN/DM (NER-PS of 98.44) and has a very small room for improvement. This aligns with the findings from the human evaluation that the BART model has very few extrinsic entity errors (Pagnoni et al., 2021). Secondly, DAE expert performs better than (or comparable to) NER-P expert on NER-PS metric on both datasets. Intuitively, dependency arc error subsumes extrinsic entity error since dependency arcs corresponding to extrinsic entities can not be entailed by the source document. On ROUGE and other factual consistency metrics that are not part of the expert training, we observe mixed performance. For instance, DAE expert improves entity recall (NER-RT) and ROUGE scores on CNN but not on XSUM. By combining multiple experts, we reduce the variations in performance across different metrics and resulting MoFE$_{weights}$ and MoFE$_{logits}$ outperforms BART across all factual consistency metrics, except MoFE$_{weights}$ that marginally lowers the entity precision (-0.18) on CNN/DM. Also, neither of the MoFE models lowers ROUGE scores substantially on XSUM or CNN/DM, the worst being 0.74 drops for MoFE$_{logits}$ on CNN/DM. Between logits vs weights ensembling, we find the former slightly more effective on factual consistency metrics. However, by calculating logits for all experts and the pre-trained model at each decoding step, logit ensembling increases the decoding time linearly in the number of experts. Weights ensembling, on the other hand, does not increase the inference time and provides a lightweight method for combining experts.

In table 3, we report results for BART and MoFE models on BERTScore and QA-based metrics. Recent work on benchmarking different evaluation metrics suggests that BERTScore precision with

Models	XSUM				CNN/DM			
	BScore-P	BScore-R	FEQA	QEval	BScore-P	BScore-R	FEQA	QEval
BART	88.93	79.86	25.77	36.54	93.26	82.62	**38.22**	59.24
MoFE$_{weights}$	89.21	79.89	**27.87**	37.32	93.26	82.95	35.72	59.77
MoFE$_{logits}$	**89.24**	**79.94**	27.74	**37.43**	**93.67**	**83.46**	33.13	**60.39**

Table 3: BERTScore- and QA metrics-based evaluations of MoFE models on XSUM and CNN/DM.

respect to the source document correlates with the human judgment of factuality (Pagnoni et al., 2021), though BERTScore precision is not exclusively a metric for evaluating factual consistency. We find that MoFE models improve BERTScore precision (BScore-P) and recall (BScore-R) on both XSUM and CNN/DM datasets. Similarly, MoFE models improve on the QA-based QuestEval metric on both XSUM and CNN/DM datasets. However, both MoFE$_{weights}$ and MoFE$_{logits}$ perform much worse than the BART model on the FEQA metric for CNN/DM data. The contrasting observations between FEQA and QuestEval may be explained by the variation in question-generation (QG) modules used in both metrics. We observe that the QG model used in FEQA tends to copy the entire summary into the questions *(e.g. "**when** is the sigma alpha epsilon fraternity fighting back against claims that racism is stitched into the fabric of the fraternity ? one of the university of oklahoma students who took part in the infamous racist chant wrote that ' the song was taught to us ' ")*. This behavior does not pose serious problems for shorter summaries, like those in the XSUM. However, for longer summaries, questions become abruptly complicated for the QA model to find the correct answer in the source document (e.g. QA model answers this question by selecting the bolded phrase *"...racism is stitched into the fabric of the fraternity - by **mandating that all members of the organization undergo diversity training**"*.). On the other hand, the QG model in the QuestEval generates straightforward questions *(e.g. "**When** did the executive director announce changes to the Sigma Alpha Epsilon fraternity?")*. It is also worth noting that Pagnoni et al. (2021) found FEQA negatively correlated with the human judgment of factuality on CNN/DM data. However, the correlation was found to be positive on XSUM, though statistically insignificant.

We evaluate MoFE models using a diverse set of factual consistency evaluation metrics and find them effective in reducing hallucinations. However, the automatic evaluations can only provide anecdotal evidence. Therefore, next, we analyze 30 samples from each of the MoFE$_{logits}$ and BART models on XSUM data. We use *SummVis* (Vig et al., 2021) for our analysis. We show 8 interesting samples from the analyzed 30 in Appendix (§A). Looking at the examples where MoFE and BART differ in factual consistency, we find cases where MoFE: *I)* removes some of the factual errors but the new summary remains factually inconsistent, Fig. 3 and 4; *II)* removes all factual errors, Fig. 5; *III)* replaces one factual error with another, Fig. 6; *IV)* adds factual error, Fig. 7; and *V)* adds or removes world knowledge, Fig. 8 and 9. Ignoring world knowledge hallucination, in total, we find 3, 4, 4, and 2 examples for cases I, II, III, and IV respectively. The remaining summaries were both factually consistent (12 examples)/ inconsistent (5 examples) for both BART and MoFE. It is also worth noting that in all 4 examples of case II, BART summaries have exactly one factual error. From our analyses, we conclude that generally MoFE helps reduce factual errors, but it is most effective in cases where BART summaries contain a few factual errors. In more complex cases of hallucinations, MoFE can only partially remove factual errors.

Next, we analyze how data quality affects reference and model-based experts (§4.3.1). Given that XSUM is much noisier (Tejaswin et al., 2021; Nan et al., 2021) and the corresponding BART model performs poorly on factual consistency metrics, it provides an ideal avenue for our analyses. Lastly, we discuss the extractiveness-faithfulness trade-off for the BART and MoFE models (§4.3.2).

4.3.1 Reference vs Model-based Experts

In Table 4, we report the validation performance of both reference and model-based variants of DAE and NER-P experts trained on filtered XSUM training subset and whole XSUM training data. We observe that both variants of experts improve performance on their respective factual consistency metrics when trained on the filtered subset. However, the margin of improvement is higher for reference-based experts, implying the advantage of using reference as the pivot summary when training samples are free from factual errors. On the whole training data that includes factually inconsistent samples, we find that reference-based experts degrade the performance on DAE-A/S

	DAE				NER-P	
	All		Filtered		All	Filtered
	DAE-A	DAE-S	DAE-A	DAE-S	NER-PS	NER-PS
BART	76.67	35.79	76.67	35.79	64.30	64.30
Reference	75.55	31.33	**82.53**	**44.09**	60.87	**69.06**
Model	**84.1**	**46.92**	80.27	41.70	**67.84**	66.88

Table 4: Validation performance of DAE and NER-P experts trained with reference and sampled summary-based KL loss on all training data and filtered subset of training data.

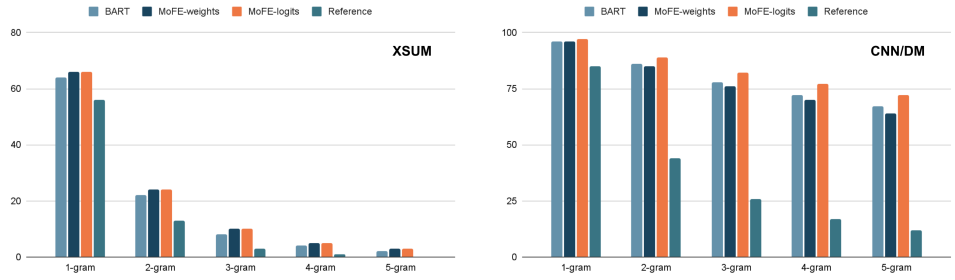


Figure 2: Percentage of overlapped n-grams in XSUM and CNN/DM summaries.

accuracy or entity precision (NER-PS) metrics. On contrary, we find model-based experts effective, with model-based DAE expert trained on the whole data even outperforming reference-based DAE expert trained on filtered subset by 1.57% and 2.83% on DAE-A and DAE-S metrics respectively. Overall, empirical results suggest that the factual quality of training data affects the performance of experts. On factually consistent samples, we can train either of the model or reference-based experts. However, when samples contain factual errors, reference-based experts may not be effective.

4.3.2 Faithfulness vs Abstractiveness

We compare the ratio of n-grams in summaries that appear in the source document. As shown in Figure 2, all BART and MoFE models are highly extractive on CNN/DM datasets. Also, the difference in n-gram overlap between reference and model-based summaries is much higher on the CNN/DM data. On the contrary, models have fewer n-gram overlaps on XSUM, but they still generate summaries with higher n-gram overlap than the reference. It is generally observed that neural models, including BART, tend to increase the extractiveness (Durmus et al., 2020).

We find that both MoFE$_{weights}$ and MoFE$_{logits}$ increase n-gram overlaps on XSUM data. However, on CNN, MoFE$_{logits}$ increases the n-gram overlap while MoFE$_{weights}$ decreases the overlap. Since we train our experts using RL that maximizes or minimizes probability mass on summaries generated by them (not the reference summary as in MLE training), we expect them to increase extractiveness. Notably, MoFE models do not consistently increase n-gram overlap (e.g. CNN), and the margin of increase in extractiveness is much lower than the difference between the reference and BART. We consider the minor increase in overlapped n-grams tolerable for improved factual consistency. Our findings are similar to Aralikatte et al. (2021), suggesting a diversity-faithfulness trade-off, where increasing faithfulness decreases the novel n-grams.

5 Conclusion

We present MoFE to reduce content hallucinations in abstractive summarization models. We first train different experts to exclusively minimize extrinsic and intrinsic hallucinations that are defined using automated factual consistency evaluation metrics. Then, we combine them with the BART model through weights or logits ensembling to control the hallucinated content. We evaluate MoFE on XSUM and CNN/DM datasets using a diverse set of metrics, finding that MoFE effectively reduces hallucinations without a significant drop on ROUGE scores or increase in extractiveness.

6 REPRODUCIBILITY STATEMENT

We describe our models, hyperparameters, and training procedures in §4. We use publicly available datasets and libraries for all our experiments and analyses, and will release our code and trained models upon acceptance.

REFERENCES

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. Focus attention: Promoting faithfulness and diversity in summarization. arXiv preprint arXiv:2105.11921, 2021.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. Factual error correction for abstractive summarization models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6251–6258, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.506. URL https://aclanthology.org/2020.emnlp-main.506.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 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, pp. 5935–5941, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.475. URL https://aclanthology.org/2021.naacl-main.475.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. Multi-fact correction in abstractive text summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9320–9331, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.749. URL https://aclanthology.org/2020.emnlp-main.749.

Esin Durmus, He He, and Mona Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5055–5070, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.454. URL https://aclanthology.org/2020.acl-main.454.

Matan Eyal, Tal Baumel, and Michael Elhadad. 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), pp. 3938–3948, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1395. URL https://aclanthology.org/N19-1395.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9:391–409, 2021.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 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, pp. 2214–2220, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1213. URL https://aclanthology.org/P19-1213.

Katja Filippova. Controlled hallucinations: Learning to generate faithfully from noisy data. In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 864–870, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.76. URL https://aclanthology.org/2020.findings-emnlp.76.

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

Tanya Goyal and Greg Durrett. Evaluating factuality in generation with dependency-level entailment. In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3592–3603, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.322. URL https://aclanthology.org/2020.findings-emnlp.322.

Tanya Goyal and Greg Durrett. Annotating and modeling fine-grained factuality in summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1449–1462, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.114. URL https://aclanthology.org/2021.naacl-main.114.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. CoRR, abs/1506.03340, 2015. URL http://arxiv.org/abs/1506.03340.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL https://doi.org/10.5281/zenodo.1212303.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407, 2018.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. Gedi: Generative discriminator guided sequence generation. arXiv preprint arXiv:2009.06367, 2020.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In Proceedings of the 27th International Conference on Computational Linguistics, pp. 1430–1441, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1121.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume

1: Long Papers), pp. 6691–6706, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.522. URL https://aclanthology.org/2021.acl-long.522.

Yang Liu, Yifei Sun, and Vincent Gao. Improving factual consistency of abstractive summarization on customer feedback. arXiv preprint arXiv:2106.16188, 2021b.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Neurologic decoding:(un) supervised neural text generation with predicate logic constraints. arXiv preprint arXiv:2010.12884, 2020.

Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. Constrained abstractive summarization: Preserving factual consistency with constrained generation. arXiv preprint arXiv:2010.12723, 2020.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. Entity-level factual consistency of abstractive text summarization. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 2727–2733, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.235.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 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, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206.

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

Richard Yuanzhe Pang and He He. Text generation by learning from demonstrations, 2021.

Ramakanth Pasunuru and Mohit Bansal. Multi-reward reinforced summarization with saliency and entailment. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 646–653, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2102. URL https://aclanthology.org/N18-2102.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization, 2017.

S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1179–1195, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.131. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.131.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 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), pp. 3246–3256, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1320. URL https://aclanthology.org/D19-1320.

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

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. How well do you know your summarization datasets? In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 3436–3449, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.303. URL https://aclanthology.org/2021.findings-acl.303.

Jesse Vig, Wojciech Kryscinski, Karan Goel, and Nazneen Rajani. SummVis: Interactive visual analysis of models, data, and evaluation for text summarization. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pp. 150–158, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.18. URL https://aclanthology.org/2021.acl-demo.18.

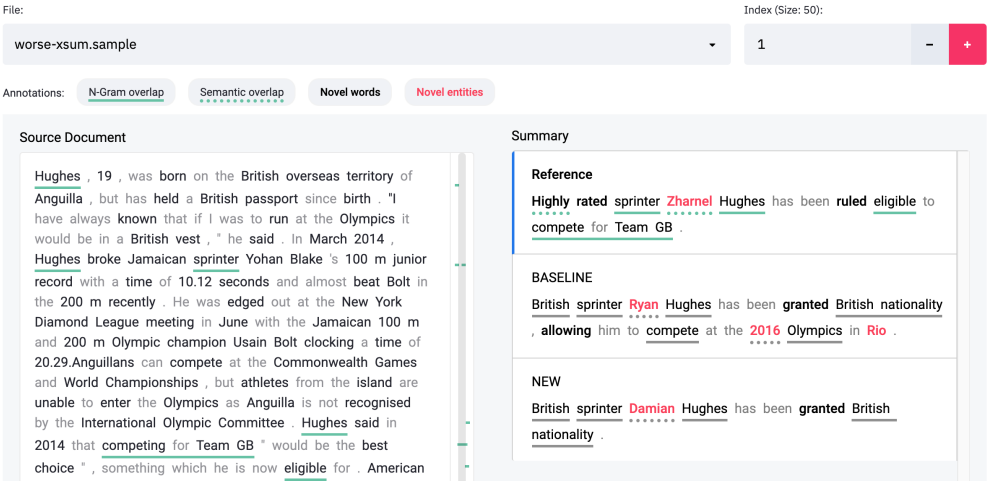
Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5008–5020, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.450. URL https://aclanthology.org/2020.acl-main.450.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3-4):229–256, 1992.

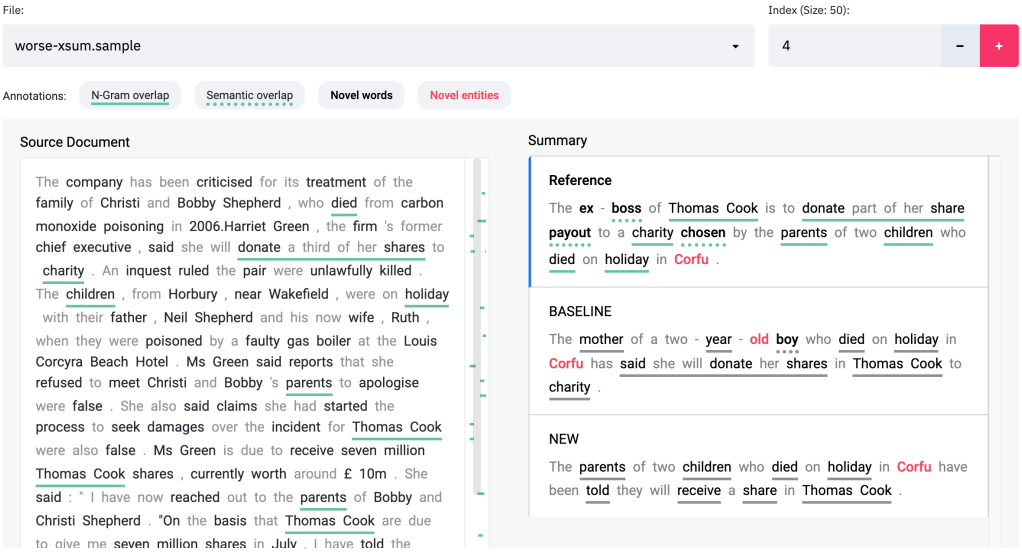
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest,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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations, 2019.

A SUMMVIS: ANALYSIS



(a) In this example, BART hallucinates 2016 Olympic and Rio which get corrected by MoFE. But both BART and MoFE incorrectly generate the first name *(Ryan vs Damian)*, as well as *"granted British nationality"*.



(b) In this example, BART hallucinates the age of children which gets corrected by MoFE. But both BART and MoFE hallucinate Corfu. In addition, both BART (parents will donate shares) and MoFE (parents will receive shares) summaries possess intrinsic hallucinations.

Figure 3: Examples where MoFE generates fewer novel entities (highlighted in **red**) that are absent from the source article.

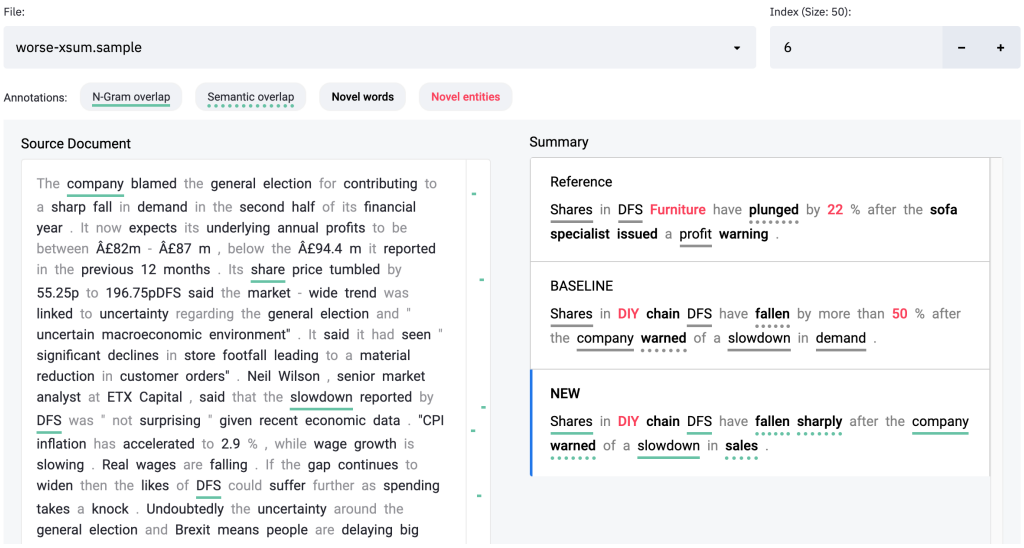


Figure 4: In this example, BART hallucinates percentage amount (50%). MoFE replaces percentage amount to a generic word *sharply*. Both BART and MoFE hallucinates DIY.

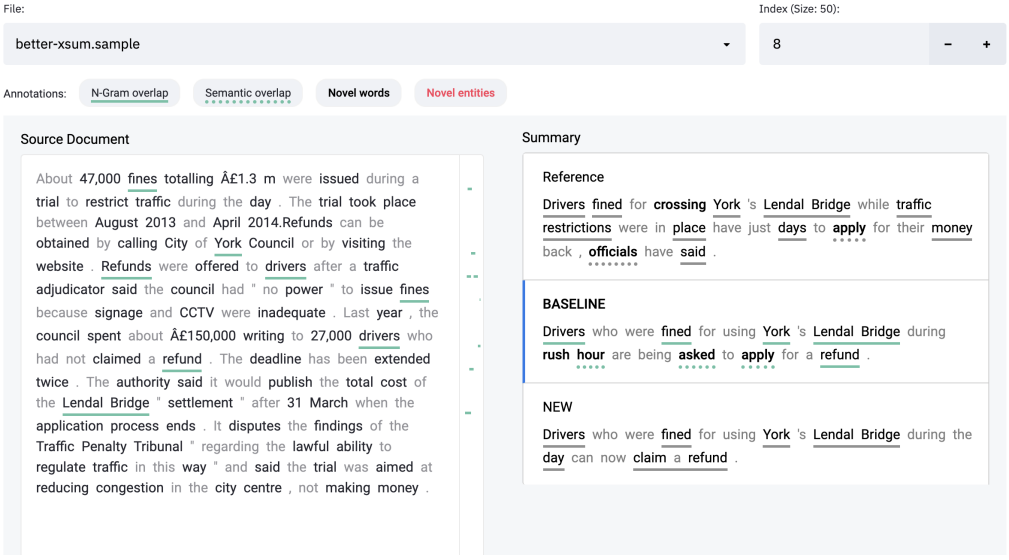


Figure 5: In this example, BART hallucinates *rush hour*. In contrast, MoFE generates factually correct summary.

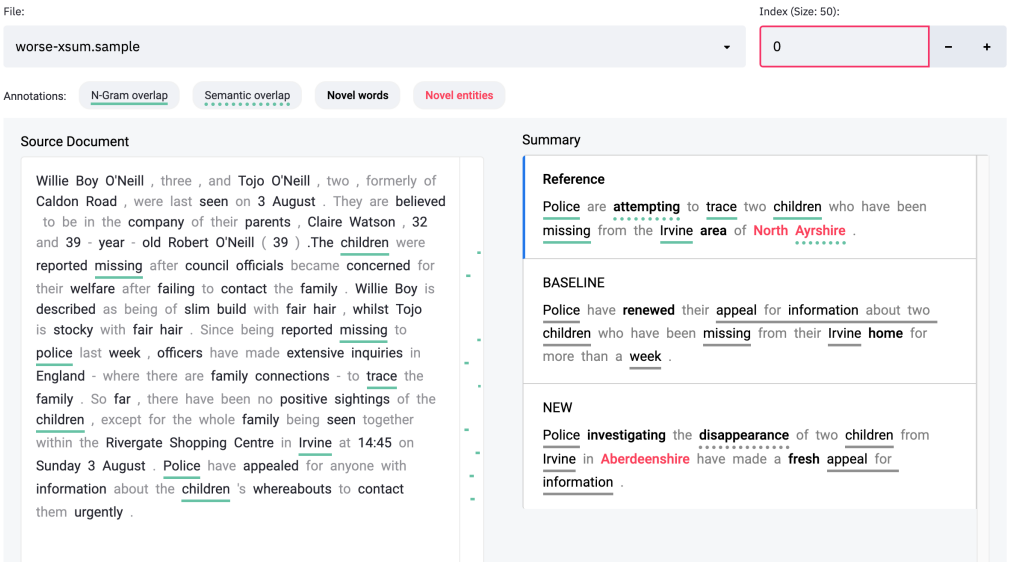


Figure 6: Both BART and MoFE generate different factual errors, BART hallucinates *more than a week* and MoFE hallucinates *Aberdeenshire*.

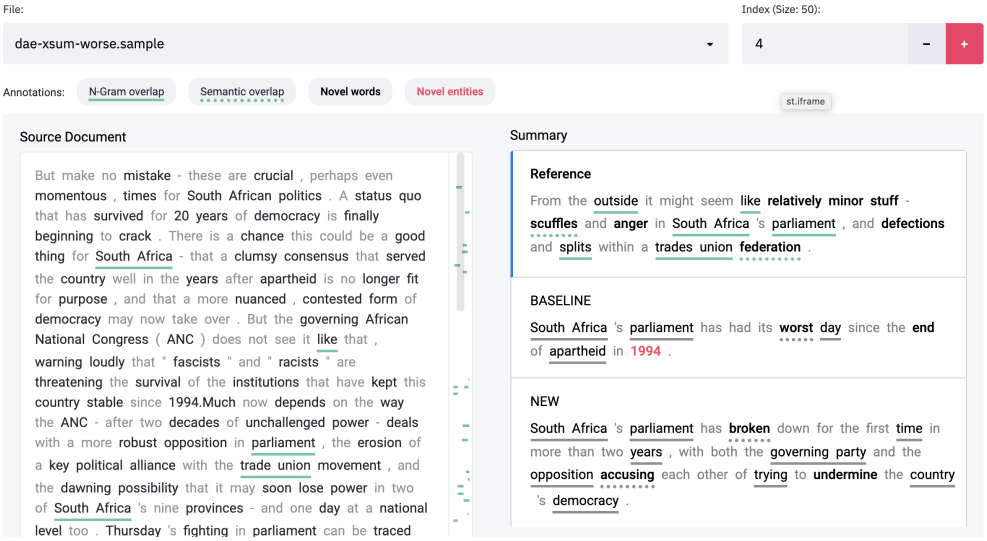


Figure 7: In this example, BART incorporates world knowledge *"end of apartheid"* and is factually consistent otherwise. MoFE adds factual error *"two years"*.

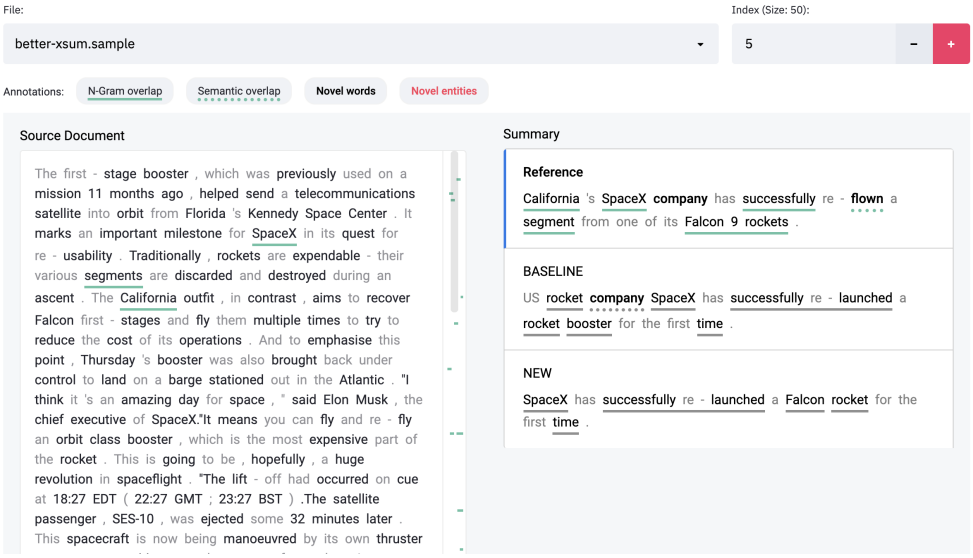


Figure 8: Both BART and MoFE are factually correct, though BART generates *US rocket company* which can not be inferred from the source document (hallucinations vs world knowledge).

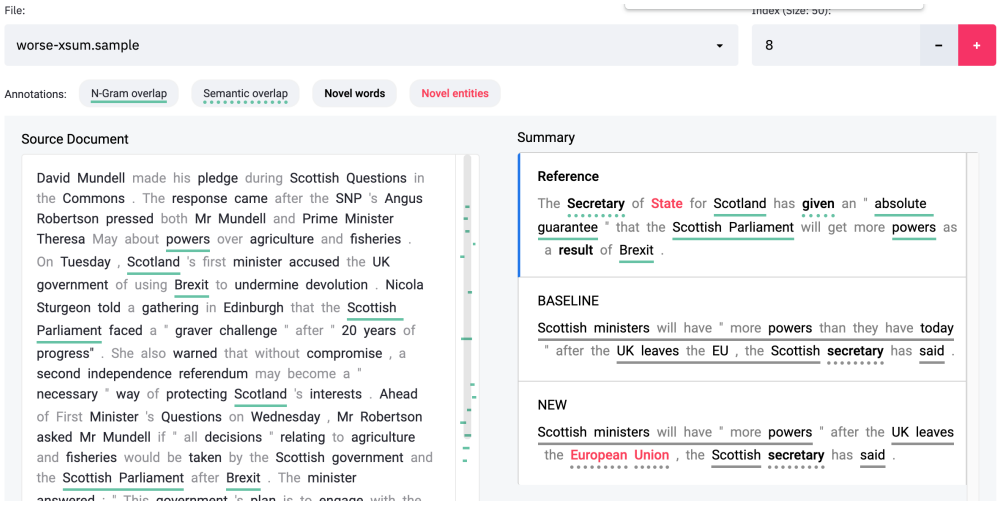


Figure 9: Both BART and MoFE are factually correct, though MoFE replaces *EU* with *European Union* (world knowledge).